Lecture Notes in Artificial Intelligence 4693

Edited by J. G. Carbonell and J. Siekmann

Subseries of Lecture Notes in Computer Science

Bruno Apolloni Robert J. Howlett
Lakhmi Jain (Eds.)

Knowledge-Based Intelligent Information and Engineering Systems: KES 2007 - WIRN 2007

11th International Conference, KES 2007
XVII Italian Workshop on Neural Networks
Vietri sul Mare, Italy, September 12-14, 2007
Proceedings, Part II

Springer

Series Editors

Jaime G. Carbonell, Carnegie Mellon University, Pittsburgh, PA, USA
Jörg Siekmann, University of Saarland, Saarbrücken, Germany

Volume Editors

Bruno Apolloni
Dipartimento di Scienze dell'Informazione
Università degli Studi di Milano
20135 Milano, Italy
E-mail: apolloni@dsi.unimi.it

Robert J. Howlett
University of Brighton
Centre for SMART Systems, School of Engineering
Brighton, BN2 4GJ, UK
E-mail: R.J.Howlett@bton.ac.uk

Lakhmi Jain
University of South Australia
Knowledge-Based Intelligent Engineering Systems Centre
SA 5095, Australia
E-mail: lakhmi.jain@unisa.edu.au

Library of Congress Control Number: 2007934283

CR Subject Classification (1998): I.2, H.4, H.3, J.1, H.5, K.6, K.4

LNCS Sublibrary: SL 7 – Artificial Intelligence

ISSN 0302-9743
ISBN-10 3-540-74826-1 Springer Berlin Heidelberg New York
ISBN-13 978-3-540-74826-7 Springer Berlin Heidelberg New York

This work is subject to copyright. All rights are reserved, whether the whole or part of the material is concerned, specifically the rights of translation, reprinting, re-use of illustrations, recitation, broadcasting, reproduction on microfilms or in any other way, and storage in data banks. Duplication of this publication or parts thereof is permitted only under the provisions of the German Copyright Law of September 9, 1965, in its current version, and permission for use must always be obtained from Springer. Violations are liable to prosecution under the German Copyright Law.

Springer is a part of Springer Science+Business Media

springer.com

© Springer-Verlag Berlin Heidelberg 2007
Printed in Germany

Typesetting: Camera-ready by author, data conversion by Scientific Publishing Services, Chennai, India
Printed on acid-free paper SPIN: 12120574 06/3180 5 4 3 2 1 0

Preface

These three volumes are a collection of the contributions presented to the joint conferences of KES 2007, the 11th International Conference on Knowledge-Based and Intelligent Information and Engineering Systems, and the WIRN 2007, the 17th Italian Workshop on Neural Networks, held in Vietri sul Mare, Italy, in September 2007.

The formula by which KES conferences gather over 500 people each year from the four corners of the globe to discuss the topic of knowledge-based and intelligent information and engineering systems is: an open mind with rigor. Within the vast universe of the conferences centered around the keywords "information" and "computational intelligence," we encourage in our meetings the offering of new ideas and techniques to give solutions to the never-ending series of problems and challenges that our own intelligence poses.

As a precious attribute of the human brain, we will never be disturbed by the novelty, and possibly the provocation, of new mental paradigms and hazardous conjectures, especially if they are raised by fresh research teams. At the same time, we have riddled each contribution using the sieve of scientific quality, checking the rigor with which the ideas are illustrated, their understandability and the support of the theory or of the experimental evidence.

The structure of the conference reflects this philosophy. In addition to regular tracks on the main fields of the discipline, we invited scientists to propose sessions focused on topics of high interest. Their response was generous and based on all sources, we received some 1203 submissions. From this number we collected 11 general track sessions and 47 invited sessions to make a total of 409 papers after a severe referee screening, an acceptance rate of 34%. Thus the reader may have from these volumes an almost exhaustive overview of researcher's and practitioner's current work in the field of information extraction and intelligent systems.

WIRN 2007 was the annual meeting of the Italian Society for Neural Networks (SIREN). Joining the two conferences provided the double benefit of giving the meeting a qualified and interested local committee on the one hand, and putting the local scientific community in touch with an international cluster of researchers in similar fields, on the other. The efficiency and effectiveness of the result suggest that it would be beneficial to replicate this formula in the future.

We wish to express our sincere thanks to many people who worked hard for the success of the conference, to the authors who sent their contributions, and to the whole scientific community who contributed to intelligence and knowledge values being preserved and enhanced for future generations.

September 2007

Bruno Apolloni
Robert J. Howlett
Lakhmi C. Jain

Organization

KES 2007 was organized by KES International – Innovation in Knowledge-Based and Intelligent Engineering Systems. WIRN 2007 was organized by IIASS – Istituto Italiano Alti Studi Scientifici.

KES 2007 and WIRN 2007 Conference Chairs

General Chair: B. Apolloni (University of Milan, Italy)
Executive Chair: R. J. Howlett (University of Brighton, UK)
Honorary Invited Session Committee Chair: L. C. Jain (University of South Australia)

KES Conference Series

KES 2007 is part of the KES Conference Series
Conference Series Chairs, L. C. Jain and R. J. Howlett
KES Executive Chair, R. J. Howlett (University of Brighton, UK)
KES Founder, L. C. Jain (University of South Australia)

WIRN Conference Series

WIRN 2007 is part of the WIRN Conference Series
Conference Chairs: M. Marinaro (IIASS, Italy) and B. Apolloni (University of Milan, Italy).

Local Organizing Committee

S. Bassis, S. Gaito, D. Malchiodi, G. L. Galliani, L. Valerio, A. Zippo (University of Milan, Italy)
M. Marinaro, A. Esposito: IIASS

International Program Committee and KES 2007 Board of Reviewers

A. Abe
J. Abe
M. Abulaish
Y. Adachi
A. Adli
S. Akama
G. Alfonso
E. Al-Hasel
U. Amato
P. Angelov
D. Anguita
C. Angulo-Bahon

M. Anisetti
A. Antonella
B. Apolloni
J. Arima
S. Arima
M. Aritsugi
A. Azzini
N. Baba
I. Bae
S. Bae Cho
J. Bajo
B. Balachandran
S. Balzarotti
S. Bandini
B. Baruque
R. Bashar
S. Bassis
K. Basterretxea
R. Batres
L. Bedini
K. Belda
V. Bellandi
B. Berendt
A. Bertoni
P. Beullens
M. Bianchini
F. Biassoni
M. Bielikov
Y. Bin Kwon
L. Bobrowski
G. Bogdan
J. Bohm
A. Bonaldi
N. Borghese
A. Bouchachia
P. Bouquet
A. Brega
D. Brown
L. Brun
I. Buciu
H. Byun
C. Caiafa
P. Campadelli
V. Capasso
F. Capkovic
C. Carpineto
M. Ceccarelli
P. Ceravolo
B. Chabr
C. Chan
H. Chan Cho
J. Chan Lee
C. Chang
D. Chen
K. Chen
M. Chen
W. Chen
Y. Chen
G. Chetty
L. Chilwoo
W. Chou
J. Chung
S. Chung
A. Ciaramella
M. R. Ciceri
A. Colla
S. Colucci
A. Columbari
D. Cook
M. Corazza
E. Corchado
J. Corchado
R. Corchuelo
P. Cosi
A. Cosio
R. Cox
P. Crippa
M. Cristani
A. Cuzzocrea
C. d'Amato
E. Damiani
A. DAnjou
L. D'Apuzzo
P. Davidsson
C. de Campos
S. De Capitani di Vimercati
M. Degemmis
D. Deng
E. Di Claudio
E. Di Iorio
T. Di Noia
E. di Sciascio
D. Difabio
X. Ding
M. do Carmo Nicoletti
Y. Dourisboure
L. Dumitriu
R. Duro
A. Edman
A. Eleuteri
A. Esposito
F. Esposito
L. Eun-Sur
J. Fahimi
P. Falcoz
N. Fanizzi
M. Fansi
G. Fassano
J. Feng
A. Fernandez-Caballero
S. Ferraresi
S. Fiori
A. Formisano
F. Frati
T. Fuchino
C. Fugazza
S. Fujii
T. Fujinami
M. Fukumi
T. Fukumoto
H. Funaoi
C. Furlanello
A. Gabillon
B. Gabrys
S. Gaito
L. Galliani
G. Gao
K. Gao
M. Garcia-sebastian
P. Gastaldo
T. Gavrilova
D. Gendarmi
H. Ghodosi
F. Gianfelici

G. Gianini
P. Giorgini
S. Giove
W. Goh
S. Gon Kong
L. Gonzalez
E. Gouardères
G. Gouardères
M. Grana
M. Graña
K. Grant
D. Gu
H. Guo
T. Guy
K. HaengKon
M. Hagenbuchner
M. Haindl
A. Håkansson
B. Hammer
A. Hara
K. Harada
F. Harris
R. Hartung
S. Hasegawa
Y. Hashimoto
A. Hassanien
Y. Hayashi
X. He
M. Hemmje
M. Hiot Lim
K. Hiraishi
T. Hochin
S. Ho-Jun
X. Hong
S. Hori
A. Hotho
R. Howlett
P. Hraber
E. Hsiao
X. Huang
Y. Huang
F. Hussain
S. Hyun Kim
T. Ichikawa
T. Ichimura
K. Iizuka
N. Inuzuka
Y. Iribe
H. Ishibuchi
Y. Ishida
N. Ishii
H. Ito
J. Itou
Y. Iwahori
S. Iwashita
L. Jain
R. Jain
M. Jason
D. Jeng
M. Jeng
I. Jeon
J. Jiang
H. Joo Lee
S. Joon Yoo
J. Jung
S. Jung
K. Juszczyszyn
J. Kacprzyk
H. Kanai
T. Kanda
Y. Kang
M. Karny
W. Karwowski
R. Katarzyniak
N. Kato
S. Kato
P. Kazienko
L. Kenneth
A. Keskar
D. Keysers
B. Kim
D. Kim
H. Kim
I. Kim
S. Kim
Y. Kim
S. King
M. Kinnaert
D. Kitakoshi
P. Klan
T. Kojiri
T. Kokogawa
S. Kollias
H. Kosaka
A. Koukam
D. Król
N. Kubota
K. Kubota
S. Kunifuji
H. Kunimune
C. Kuroda
Y. Kurosawa
P. Kyu Rhee
K. Lam
K. Le
C. Lee
Y. Lee
F. Leporati
P. Leray
L. Lhotska
J. Li
L. Lin
P. Linh
H. Liu
Y. Liu
B. López
P. Lops
S. Luan
W. Ma
M. Maggini
L. Magnani
M. Majid
S. Makrogiannis
D. Malchiodi
J. O. Maldonado
D. Malerba
L. Mal-Rey
M. Mancini
S. Marinai
M. Marinaro
S. Marrara
G. Martinelli
R. Martone
F. Mason
F. Masulli

J. Matas
N. Matsuda
N. Matsui
H. Matsumoto
N. Matsumura
M. Matsushita
G. Mauri
Q. Meng
F. Menolascina
K. Mera
Y. Mi Kwon
F. Michaud
S. Miguet
H. Minami
H. Mineno
K. Misue
H. Mitsuhara
Y. Mitsukura
H. Miura
M. Miura
T. Mizuno
M. Mizuta
D. Mladenic
H. Mochizuki
Y. Mogami
M. Mohammadian
D. Monekosso
A. Montuori
I. Morgan
A. Morici
P. Motto Ros
N. Mukai
C. Mumford
J. Munemori
M. Muselli
M. Nachtegael
I. Nagy
T. Nakada
K. Nakamatsu
S. Nakamura
T. Nakamura
R. Nakano
T. Nakano
J. Nam Jung
Y. Nara
J. Nascimento
O. Nasraoui
D. Nauck
D. Ndedi Monekosso
M. Negoita
N. Nguyen
G. Nicosia
C. Niederée
A. Nijholt
T. Nishida
K. Nishimoto
T. Nishiura
H. Nobuhara
A. Nowak
M. Nowostawski
A. Nuernberger
Y. Ochi
S. Oeda
R. Oehlmann
L. Oestreicher
N. Ogata
Y. Ohsa
Y. Ohsawa
M. Okada
T. Okamoto
M. Ozden
V. Palade
F. Palmieri
D. Pan
M. Paprzycki
R. Parisi
T. Parisini
G. Park
Y. Park
F. Parra
E. Pasero
G. Pasi
W. Pedrycz
E. Pessa
T. Pham
L. Phong
F. Picasso
A. Pieczynska
L. Prevost
A. Ragone
G. Raiconi
G. Raimondo
J. Ramon
R. Ranawana
R. Rascuna
K. Rattan
L. Razmerita-Hockerts
M. Refice
P. Remagnino
M. Resta
L. Reyneri
A. Rohani
M. Ryoke
G. Ryung Uh
K. Saito
L. Saitta
M. Sakalli
E. Salerno
M. G. Sami
R. Sassi
M. Sato
Y. Sato
M. Sato-Ilic
A. Scarrelli
F. Scarselli
Z. Schindler
M. Schlegel
F. Schwenker
F. Scotti
G. Semeraro
C. Seng Chan
G. Sergiadis
R. Serra
S. Sessa
D. Shen
Y. Shiau
M. Shikida
B. Shizuki
V. Shkodirev
A. Sidhu
J. Smith
J. Sobecki
P. Somol
D. Soo Kim
F. Sorbello

Z. Sosnowski
A. Sperduti
A. Staiano
G. Stamou
R. Stecher
H. Stoermer
Y. Su Choi
T. Sugihara
K. Sugiyama
M. Suka
Z. Sun
I. Sun Choi
W. Sunayama
I. Tabakow
R. Tagliaferri
E. Takahagi
M. Takahashi
O. Takahashi
O. Takata
F. Takeda
H. Taki
H. Tamura
J. Tan
Y. Tanahashi
J. Tanaka
M. Tanaka-Yamawaki
P. Tann
Y. Tateiwa
C. Teeling
L. Tesař
H. Thai
C. Thanh Hoang
N. Thanh Nguyen
P. Tichavský
I. Ting
P. Tino
A. Tonazzini
D. Toshinori
D. Tran
E. Trentin
F. Trinidad
F. Trojani
K. Tsuda
Y. Tsuge
S. Tsumoto
N. Tsuyoshi
G. Tummarello
C. Turchetti
J. Tweedale
K. Umeda
A. Uncini
T. Ushiama
G. Valentini
I. Villaverde
S. Vitabile
I. Vlachos
T. Wadayama
D. Wan Kim
A. Wang
D. Wang
J. Wang
P. Wang
J. Wata
J. Watada
T. Watanabe
Y. Watanabe
Y. Wen
Y. Weo Lee
N. Wessiani
G. Wren
B. Wu
X. Wu
L. Xi
Y. Xiong
F. Xu
X. Xu
Y. Yabuuchi
T. Yamakami
Y. Yamashita
C. Yang
T. Yoshino
M. Young Sung
D. Yu
Z. Yu
T. Yuizono
M. Zalili
A. M. Zanaboni
A. Zeng
X. Zeng
B. Zhang
Y. Zhang
X. Zhou
G. Zhu
Y. Zhu
A. Zippo
I. Zoppis
R. Zunino

General Track Chairs

Generic Intelligent Systems Topics

Artificial Neural Networks and Connectionists Systems
Ryohei Nakano (Nagoya Institute of Technology, Japan)

Granular Computing
Detlef Nauck (BT, UK), Zensho Nakao (University of Ryukyus, Japan)

Machine Learning and Classical AI
Floriana Esposito (University of Bari, Italy)

Agent Systems
Ngoc Thanh Nguyen (Wroclaw University of Technology, Poland)

Knowledge-Based and Expert Systems
Anne Hakansson (Uppsala University, Sweden)

Miscellaneous Intelligent Algorithms
Honghai Liu (University of Portsmouth, UK)

Applications of Intelligent Systems

Intelligent Vision and Image Processing
Tuan Pham (James Cook University, Australia)

Knowledge Management and Ontologies
Guy Gouarderes (University of Bayonne, France), Gloria Wren (Loyola College in Maryland, USA), Lakhmi Jain (University of South Australia, Australia)

Web Intelligence, Text and Multimedia Mining and Retrieval
Andreas Nuernberger (University of Magdeburg, Germany)

Intelligent Signal Processing, Control and Robotics
Miroslav Karny (Czech Republic Academy of Science, Czech Republic)

Other Intelligent Systems Applications
Viacheslaw Shkodirev St. Petersburg State Poly. University, Russia)

Invited Session Chairs

Ambient Intelligence, *Cecilio Angulo-Bahon (Universitat Politecnica de Catalunya, Spain) Honghai Liu (University of Portsmouth, UK)*

Artificial Intelligence Applications in Digital Content, *Mu-Yen Chen (National Changhua University of Education, Taiwan), Hsiao-Ya Chiu (Yu-Da College of Business)*

Artificial Intelligence Applications in Security, *Emilio Corchado (University of Burgos, Spain) Rodolfo Zunino (Genoa University, Italy)*

Artificial Intelligence Methods for Information Processing (AIMIP 2007), *Lifeng Xi, Jifang Li, Kun Gao (Zhejiang Wanli University, Ningbo, China)*

Communicative Intelligence 2007, *Toyoaki Nishida (University of Kyoto, Japan) Ngoc Thanh Nguyen (Wroclaw University of Technology, Poland)*

Computation Intelligence for Image Processing and Pattern Recognition, *Yen-Wei Chen (Ritsumeikan University, Nojihigashi, Japan)*

Human Computer Intelligent Systems, *Takumi Ichimura, Kazuya Mera (Hiroshima City University, Japan)*

Hybrid Artificial Intelligence Systems Workshop (HAIS 2007-KES2007), *Juan M. Corchado (University of Salamanca, Spain) Emilio Corchado(University of Burgos, Spain)*

Innovations in Intelligent Data Analysis, *Mika Sato (University of Tsukuba, Japan) Lakhmi Jain (University of South Australia, Australia)*

Intelligent Agents and Their Applications, *Dharmendra Sharma, Wanli Ma (University of Canberra, Australia), Haeng Kon Kim (Catholic University of Daegu, Korea)*

Intelligent and Adaptive Systems in Economics, Finance and Management, *Marco Corazza (University Ca' Foscari, Venice) Norio Baba (Osaka Kyoiku University, Japan)*

Intelligent Automation Systems, *MuDer Jeng (National Taiwan Ocean University)*

Intelligent Control Theory and Applications, *Kazumi Nakamatsu (University of Hyogo, Japan) Scheng-Luen Chung (National Taiwan University of Science and Technology)*

Intelligent Data Processing in Process Systems and Plants, *Tetsuo Fuchino (Tokyo Institute of Technology, Japan) Yoshiyuki Yamashita (Tohoku University, Japan)*

Intelligent Mechanism for Knowledge Innovation, *Toyohide Watanabe (Nagoya University) Teketoshi Ushiama (Kyushu University)*

Intelligent Multimedia Solution and Security in the Next-Generation Mobile Information Systems (IMSS), *Dong Chun Lee (Howon University, Korea) Hyuncheol Kim (Namseoul University, Korea)*

Intelligent Techniques for Biometric-Based Authentication, *Ernesto Damiani, Antonia Azzini, Stefania Marrara (University of Milan, Italy)*

Logic-Based Intelligent Information Systems, *Kazumi Nakamatsu (University of Hyogo, Japan)*

Chance Discovery, *Akinori Abe (ATR Knowledge Science Laboratories, Japan) Yukio Ohsawa (University of Tokyo, Japan)*

Knowledge-Based Interface Systems I, *Naohiro Ishii (Aichi Institute of Technology, Japan) Yuji Iwahori (Chubu University, Japan)*

Knowledge-Based Interface Systems II, *Yoshinori Adachi (Chubu University, Japan) Nobuhiro Inuzuka (Nagoya Institute of Technology, Japan)*

Knowledge and Information Management in a Social Community, *Toyohide Watanabe (Nagoya University, Japan) Naoto Mukai (Tokyo Science University, Japan) Jun Feng (Hohai University, China)*

Knowledge and Ontological Engineering for Intelligent Information System Development (KOS), *Tatiana Gavrilova (St. Petersburg State Polytechnic University, Russia) Vyacheslav Shkodyrev (Polytechnic of St. Petersburg, Russia)*

Knowledge Engineering in Multi-Robot Systems, *Manuel Graa, Richard Duro (Universidad del Pais Vasco, Spain*)

Knowledge-Based Creativity Support Systems, *Susumu Kunifuji, Motoki Miura (JAIST, Japan) Kazuo Misue (Tukuba University, Japan)*

Knowledge-Based Multi-Criteria Decision Support, *Hsuan-Shih Lee (National Taiwan Ocean University)*

Knowleged-Based Systems for e-Business, *Kazuhiko Tsuda (University of Tsukuba, Japan) Masakazu Takahashi (Shimane University, Japan)*

Computational Learning Methods for Unsupervised Segmentation (CLeMUS), *Emanuele Salerno (Consiglio Nazionale delle Ricerche, Italy) Simon Wilson (Trinity College, Ireland)*

Computational Methods for Intelligent Neuro-Fuzzy Applications, *Gwi-Tae Park, Dongwon Kim (Korea University)*

Learning Automata and Soft Computing Techniques and Their Applications, *Norio Baba (Osaka Kyoiku University, Japan) Ann Nowe, Katja Verbeeck (Vrije Universiteit, Belgium)*

Learning from Uncertain Data, *Dario Malchiodi (University of Milan, Italy)*

Neural Information Processing for Data Mining, *Ryohei Nakano, Kazumi Saito (Nagoya Institute of Technology, Japan)*

Neural Networks: Advanced Applications, *Eros Pasero (University of Turin, Italy)*

Soft Computing Approach to Management Engineering, *Junzo Watada (Waseda University, Japan) Huey-Ming Lee (Chinese Culture University, Taiwan) Taki Kanda (Bunri University of Hospitality, Japan)*

Soft Computing in Electromagnetic Applications, *Raffaele Martone (University of Naples, Italy)*

Advanced Cooperative Work, *Jun Munemori, Takashi Yoshino (Wakayama University, Japan) Takaya Yuizono (JAIST, Japan)*

Behavior Support in Advanced Learning Collaborations, *Toyohide Watanabe, Tomoko Kojiri (Nagoya University, Japan)*

Context-Aware Adaptable Systems and Their Applications, *Phill Kyu Rhee (Inha University, Korea) Rezaul Bashar (Islamic University, Bangladesh)*

Engineered Applications of Semantic Web - SWEA, *Tommaso Di Noia, Eugenio di Sciascio (Politechnic of Bari, Italy) Giovanni Semeraro (University of Bari, Italy)*

Environment Support in Advanced Learning Collaborations, *Toyohide Watanabe, Tomoko Kojiri (Nagoya University, Japan)*

Immunity-Based Systems, *Yoshiteru Ishida (Toyohashi University of Technology, Japan) Giuseppe Nicosia (University of Catania, Italy)*

Interactive Visualization and Clustering, *Roberto Tagliaferri (University of Salerno, Italy)*

Multi-Agent Systems Design, Implementation and Applications, *Dharmendra Sharma, Bala M. Balachandran (University of Canberra, Australia)*

Multimedia Systems and Their Applications Focusing on Reliable and Flexible Delivery for Integrated Multimedia (Media 2007), *Yun Ji Na (Convergence Information Technology Research Center, Korea) Il Seok Ko (Dongguk University, Korea)*

Recommender Agents, *Dariusz Król, Janusz Sobecki (Wroclaw University of Technology, Poland)*

Skill Acquisition and Ubiquitous Human Computer Interaction, *Hirokazu Taki (Wakayama University, Japan) Satoshi Hori (Institute of Technologists, Japan)*

XML Security, *Stefania Marrara, Ernesto Damiani (University of Milan, Italy) Majirus Fansi, Alban Gabillon (University of Pau, France)*

Keynote Speakers

Jean-François Cardoso, École Nationale Supérieure des Télécommunications, France: *Independent Component Analysis: Concepts and Applications*

Stephanie Forrest, University of New Mexico, USA: *Self-Healing Systems and Autonomic Network Security*

Walter J. Freeman, University of California, Berkeley, USA: *Thermodynamic Model of Knowledge Retrieval in Brain Dynamics for Information Processing*

Mario Gerla, University of California, Los Angeles, USA: *Probing and Mining the Urban Environment Using the Vehicular Sensor Network*

Hans-Andrea Loeliger, ETH, Zurich, Switzerland: *The Factor Graph Approach to Model-Based Signal Processing*

Yoshiteru Ishida, Toyohashi University, Japan: *The Immune System Offered a Glimpse: What Makes Biological Systems Distinct from Artificial Ones*

Sponsoring Institutions



Università degli Studi di Milano



Seconda Università di Napoli



Comune di Vietri sul Mare



Comune di Salerno



Regione Campania



Ministero per le Riforme e le Innovazioni nella P A



Centro Regionale Information Communication Technology

Table of Contents – Part II

I Intelligence Everywhere

Ambient Intelligence

An Efficient Indexing Technique for Location Prediction of Moving Objects 1
Dong-Oh Kim, Kang-Jun Lee, Dong-Suk Hong, and Ki-Joon Han

Combining Smart Tags and Body Fixed Sensors for Disabled People Assistance 10
Juan Antonio Álvarez, Carlos Pérez, Cecilio Angulo, and Juan Antonio Ortega

Geometrical Representation of Quantity Space and Its Application to Robot Motion Description 18
Honghai Liu, David J. Brown, and George M. Coghill

Modeling Smart Homes for Prediction Algorithms 26
A. Fernández-Montes, J.A. Álvarez, J.A. Ortega, M.D. Cruz, L. González, and F. Velasco

Qualitative Correspondence for Object Tracking Using Dynamic Panorama 34
Farshad Fahimi, Honghai Liu, and David J. Brown

Artificial Intelligence Applications in Digital Content

A Hybrid System: Neural Network with Data Mining in an e-Learning Environment 42
David Wen-Shung Tai, Hui-Ju Wu, and Pi-Hsiang Li

A Personalization Recommendation Framework of IT Certification e-Learning System 50
Heng-Li Yang, Chen-Shu Wang, and Mu-Yen Chen

An Architecture for e-Learning System with Computational Intelligence 58
Marc El Alami, Nicolas Casel, and Denis Zampunieris

APD-A Tool for Identifying Behavioural Patterns Automatically from Clickstream Data 66
I-Hsien Ting, Lillian Clark, Chris Kimble, Daniel Kudenko, and Peter Wright

Breedbot: An Edutainment Robotics System to Link Digital and Real World 74
Orazio Miglino, Onofrio Gigliotta, Michela Ponticorvo, and Stefano Nolfi

Develop e-Learning Platform for Reinforcement Learning on Temperature Sensor 82
Wen-Jye Shyr

Modeling e-Learning System Performance Evaluation with Agent-Based Approach 90
Hsiao-Ya Chiu, Sheng Chieh Chung, and An-Pin Chen

Salable Vector Graphics (OpenVG) for Creating Animation Image in Embedded Systems 99
Sang-Yun Lee, Sunghwan Kim, Jihoon Chung, and Byung-Uk Choi

Artificial Intelligence Applications in Security

A Data Mining Methodology for Anomaly Detection in Network Data 109
Costantina Caruso and Donato Malerba

A Dedicated Hardware for Fingerprint Authentication 117
G. Danese, M. Giachero, F. Leporati, G. Matrone, and N. Nazzicari

Belnap's Logic in Sharing Protected Web Resources 125
Sylvia Encheva and Sharil Tumin

IDS Based on Bio-inspired Models 133
Paolo Gastaldo, Francesco Picasso, Rodolfo Zunino, Álvaro Herrero, Emilio Corchado, and José Manuel Sáiz

Network-Based Anomaly Intrusion Detection Improvement by Bayesian Network and Indirect Relation 141
ByungRae Cha and DongSeob Lee

Security Policies Tuning Among IP Devices 149
Simone Ferraresi, Emanuele Francocci, Alessio Quaglini, and Francesco Picasso

Artificial Intelligence Methods for Information Processing (AIMIP 2007)

A Novel Image Edge Detection Using Fractal Compression 159
Liangbin Zhang and Lifeng Xi

A Rapid Algorithm for Computing ST Reliability of Radio-Communication Networks 167
Fei Gao, Xuejun Liu, and Hanzhong Liu

A Second-Order Upwind Difference Scheme for a Singularly Perturbed Problem with Integral Boundary Condition in Netural Network 175
Zhongdi Cen and Xin Cai

Application of Genetic Algorithm to Hausdorff Measure Estimation of Sierpinski Carpet 182
Qili Xiao and Lifeng Xi

Decomposability of Implication Operators 189
Zhudeng Wang

Fault Detecting Technology Based on BP Neural Network Algorithm 194
Ran Jin, Kun Gao, Zhigang Chen, Chen Dong, Yanghong Zhang, and Lifeng Xi

Image Information Hiding Encryption Using Chaotic Sequence 202
Zhen Liu and Lifeng Xi

Improvement of an Image Encryption Algorithm Based on Combined Multidimensional Chaotic Systems 209
Chen Dong, Jifang Li, Lifeng Xi, Jie Hao, and Ran Jin

Investigation on Fitting Graph Based on Fractal Dimension's Pretreatment 217
Min Jin, Qin Wang, and Lifeng Xi

Practice of Video Encryption Algorithms Based on Chaotic Sequence and Scan Pattern 225
Qian Li

Research on WebGIS Based on Multi-agent 233
Taowei Wang, Liangbin Zhang, and Aimin Yang

Transient Air-Fuel Ratio Estimation in Spark Ignition Engine Using Recurrent Neural Networks 240
Yanhong Zhang, Lifeng Xi, and James Liu

Communicative Intelligence 2007

A Pattern Recognition Algorithm Used in Knowledge-Based Adaptive Control System to Select Strategy 247
Krzysztof Brzostowski and Jerzy Świątek

Alignment-Based Preprocessing of Personal Ontologies on Semantic Social Network 255
Jason J. Jung, Hong-Gee Kim, and Geun-Sik Jo

Browser Latency Impact Factors 263
Andrzej Siemiński

Global Adaptive Request Distribution with Broker 271
Leszek Borzemski, Anna Zatwarnicka, and Krzysztof Zatwarnicki

Global and Local Approach to Complex Systems Modeling Using Dynamic Neural Networks–Analogy with Multiagent Systems 279
Jarosław Drapała and Jerzy Świątek

Computation Intelligence for Image Processing and Pattern Recognition

A Content Dependent Visualization System for Symbolic Representation of Piano Stream 287
Alexander Adli, Zensho Nakao, and Yasunori Nagata

Audio Watermarking Based on the Psychoacoustic Model and Modulated Complex Lapped Transform 295
Hanane H. Mirza, Alexander Adli, Hien D. Thai, Yasunori Nagata, and Zensho Nakao

A New Optimization Algorithm of Kinoforms Based on Simulated Annealing 303
Shinya Nozaki, Yen-Wei Chen, and Zensho Nakao

Curvelet-Domain Image Watermarking Based on Edge-Embedding 311
Thai Duy Hien, Ikeda Kei, Hanane Harak, Yen Wei Chen, Yasunori Nagata, and Zensho Nakao

Linking KANSAI and Image Features by Multi-layer Neural Networks 318
Xinyin Huang, Shouta Sobue, Tomoki Kanda, and Yen-Wei Chen

Pattern Analysis and Prediction of *O*-Linked Glycosylation Sites in Protein by Principal Component Subspace Analysis 326
Yen-Wei Chen, Xuemei Yang, Masahiro Ito, and Ikuko Nishikawa

Prediction of the *O*-Glycosylation with Secondary Structure Information by Support Vector Machines 335
Ikuko Nishikawa, Hirotaka Sakamoto, Ikue Nouno, Kazutoshi Sakakibara, and Masahiro Ito

Human Computer Intelligent Systems

A Simple Model for Human-Robot Emotional Interaction 344
Isabella Cattinelli and N. Alberto Borghese

Applying Trust, Reputation and Intuition Aspects to Support Virtual Communities of Practice 353
Juan Pablo Soto, Aurora Vizcaíno, Javier Portillo-Rodríguez, and Mario Piattini

Evaluation of Trivial Dialogue Phrase Databases Through Practical Application to User-Computer Conversation - Case Study: English - Spanish 361
Calkin S. Montero and Kenji Araki

Formalizing Recognition of Sketching Styles in Human Centered Systems 369
Danilo Avola, Fernando Ferri, and Patrizia Grifoni

Fostering Multi-Modal Summarization for Trend Information 377
Tsuneaki Kato, Mitsunori Matsushita, and Noriko Kando

Interactive Virtual Oriental Brush with Pen-Tablet System 387
Won-Du Chang and Jungpil Shin

Multidimensional Emotional Appraisal Semantic Space (MEAS): Evaluating HM Affective Interactions 395
Maria Rita Ciceri and Stefania Balzarotti

Multimodal Sentence Similarity in Human-Computer Interaction Systems 403
Fernando Ferri, Patrizia Grifoni, and Stefano Paolozzi

Reconsideration of the Effectiveness on Extracting Computer Diagnostic Rules by Automatically Defined Groups 411
Yoshiaki Kurosawa, Akira Hara, Kazuya Mera, and Takumi Ichimura

Hybrid Artificial Intelligence Systems Workshop (HAIS07-KES2007)

Adaptive Mechanisms for Classification Problems with Drifting Data . . . 419
Zoheir Sahel, Abdelhamid Bouchachia, Bogdan Gabrys, and Paul Rogers

Architecture of an Hybrid System for Experimentation on Web Information Retrieval Incorporating Clustering Techniques 427
Montserrat Mateos and Carlos G. Figuerola

Automated Ham Quality Classification Using Ensemble Unsupervised Mapping Models 435
Bruno Baruque, Emilio Corchado, Hujun Yin, Jordi Rovira, and Javier González

FreeEnCal: A Forward Reasoning Engine with General-Purpose 444
Jingde Cheng, Shinsuke Nara, and Yuichi Goto

Hybrid Agents Based Architecture on Automated Dynamic Environments 453
Dante I. Tapia, Javier Bajo, Juan M. Corchado, Sara Rodríguez, and Juan M. Manzano

Hybrid Architecture for a Reasoning Planner Agent 461
Javier Bajo, Dante I. Tapia, Ana de Luis, Sara Rodríguez, Juan F. de Paz, and Juan M. Corchado

Learning Topologic Maps with Growing Neural Gas 469
José García-Rodríguez, Francisco Flórez-Revuelta, and Juan Manuel García-Chamizo

Using a Neurofuzzy Approach in a Medical Application 477
Constantinos Koutsojannis and Ioannis Hatzilygeroudis

Innovations in Intelligent Data Analysis

A Likelihood Ratio Test for Differential Metabolic Profiles in Multiple Intensity Measurements 485
Frank Klawonn, Claudia Choi, Beatrice Benkert, Bernhard Thielen, Richard Münch, Max Schobert, Dietmar Schomburg, and Dieter Jahn

Combining Bagging, Boosting and Dagging for Classification Problems 493
S.B. Kotsianti and D. Kanellopoulos

Functional Clustering and Functional Principal Points 501
Nobuo Shimizu and Masahiro Mizuta

Prescreening of Candidate Rules Using Association Rule Mining and Pareto-optimality in Genetic Rule Selection 509
Hisao Ishibuchi, Isao Kuwajima, and Yusuke Nojima

Term Clustering in Texts Based on Fuzzy Neighborhoods and Kernel Functions 517
Sadaaki Miyamoto, Satoshi Hayakawa, and Yuichi Kawasaki

Variable Based Fuzzy Blocking Regression Model 525
Mika Sato-Ilic

Intelligent Agents and Their Applications

A Design and Testing Technology for Ubiquitous Robot Companion Software Component 533
Sun-Myung Hwang, Yun-Koo Chung, and Hang-Kon Kim

A Multiagent Based Vehicle Engine Fault Diagnosis 541
Xiaobing Wu, Xueshan Gao, and Dharmendra Sharma

A Multi-agent Security Framework for e-Health Services 547
Rossilawati Sulaiman, Dharmendra Sharma, Wanli Ma, and Dat Tran

A Proactive Routing Algorithm Based on Bitmap Tables for Wireless Sensor Networks 555
Sangjoon Jung and Younky Chung

A Study on Mobile Agent Based Resource Management in Grid 565
Chen Zhao, Jian Yu, and Bencheng Chai

An Agent-Based Evolutionary Robotic System for Its Reconfiguration 571
Xueshan Gao, Koki Kikuchi, and Xiaobing Wu

Design of the Intelligence User Identification Method for the Improved Password Input Method and the Personality Security 581
Eun-Ser Lee, Haeng-Kon Kim, and Sang Ho Lee

Using Tuple Space to Coordinate Multiagent Activities 589
Wanli Ma, Dat Tran, and Dharmendra Sharma

Intelligent and Adaptive Systems in Economics, Finance and Management

Adaptive Use of Technical Indicators for Predicting the Intra-Day Price Movements 597
Mieko Tanaka-Yamawaki and Seiji Tokuoka

Credal Networks for Operational Risk Measurement and Management 604
Alessandro Antonucci, Alberto Piatti, and Marco Zaffalon

Fair Consistency Evaluation in Fuzzy Preference Relations and in AHP 612
Matteo Brunelli and Michele Fedrizzi

Making Financial Trading by Recurrent Reinforcement Learning 619
Francesco Bertoluzzo and Marco Corazza

Multiple Ant Colony Optimization for a Rich Vehicle Routing Problem: A Case Study 627
Paola Pellegrini, Daniela Favaretto, and Elena Moretti

Portfolio Optimization Through Elastic Maps: Some Evidence from the Italian Stock Exchange 635
Marina Resta

Prediction of Golden Cross and Dead Cross by Neural Networks and Its Utilization 642
Norio Baba and Kou Nin

Intelligent Automation Systems

An Agent-Based Early Manufacturability Assessment for Collaborative Design in Coating Process 649
Chii-Ruey Lin, Pei-Shu Fan, Yea-Jou Shiau, and MuDer Jeng

Design of a Simulated Environment for Flexible Manufacturing Systems 656
Yi-Sheng Huang and Pin-June Su

MyHome: A Residential Server for Smart Homes 664
Sheng-Luen Chung and Wen-Yuan Chen

Semi-automatic Production Testing of Spark Plugs 671
S.D. Walters, P.A. Howson, and R.J. Howlett

Intelligent Control Theory and Applications

A Gait Based Approach to Detect Directional Bias of Four-Legged Robots' Direct Walking Utilizing Acceleration Sensors 681
Ding-Jie Huang and Wei-Chung Teng

An Intelligent Algorithm for Scheduling Jobs on a Single Machine with a Common Due Date 689
Zne-Jung Lee, Chen-Chia Chuang, and Kuo-Ching Ying

Application of Intelligent Neural Networks to Prediction of Micro-electroforming for U-Type Micro-cavity 696
Sheau-Wen Shiah, Pai-Yu Chang, Tzeng-Yuan Heh, Po-Hung Lin, and Fu-Cheng Yang

Design of a Supervisor for Traffic Light Systems 704
Ta-Hsiang Chung and Yi-Sheng Huang

Intelligent Data Processing in Process Systems and Plants

A Knowledge-Based Approach for the Analysis of Abnormal Situations 712
Tetsuo Fuchino, Rafael Batres, and Yukiyasu Shimada

A Multi-agent Approach to Process Design 720
Rafael Batres, Hikaru Takashima, and Tetsuo Fuchino

A Sliding-Mode Control Approach in Ultra-Precision Positioning System 728
Insung Choi, Seungok Choi, Gunhaeng Heo, Kiheon Park, and Kwanho You

Integration of Multi-agent Controller and Scheduler for Multi-purpose and Multi-batch Plant 736
Takashi Hamaguchi, Masaki Inoue, Takamasa Yoshida, Koji Kawano, Hossam A. Gabbar, Kazuhiro Takeda, Yukiyasu Shimada, Teiji Kitazima, and Yoshihiro Hashimoto

Optimization Method for Design of Supercritical Water Oxidation Process Using Genetic Algorithm 744
Hiroya Yamasaki, Kazuhiro Takeda, Yukiyasu Shimada, Takashi Hamaguchi, Teiji Kitajima, and Tetsuo Fuchino

Intelligent Mechanism for Knowledge Innovation

An Implementation Design of a Fine-Grained Database Access Control Policy Consistency Checking Mechanism 752
Bat-Odon Purevjii, Masayoshi Aritsugi, Sayaka Imai, and Yoshinari Kanamori

Applying Privacy Preserving Count Aggregate Queries to k-Classification 761
Hidehisa Takamizawa and Masayoshi Aritsugi

Event-Centralized Management of Geographic Information Collected from Blog 769
Shoko Yasumura, Masakazu Ikezaki, Toyohide Watanabe, and Taketoshi Ushiama

Message Flow: Design and Architecture of a Message Flow Management System 777
Hideaki Ito and Teruo Fukumura

Scalable 2-Pass Data Mining Technique for Large Scale Spatio-temporal Datasets 785
Tahar Kechadi and Michela Bertolotto

Synchronizing Music and Video of Query Results in Cross-Media Retrieval System 793
Teruhisa Hochin and Wen Xue

Intelligent Techniques for Biometric Based Authentication

A Fuzzy Approach to Multimodal Biometric Authentication 801
Antonia Azzini, Stefania Marrara, Roberto Sassi, and Fabio Scotti

An Open Source Java Framework for Biometric Web Authentication Based on BioAPI 809
Elisardo González Agulla, Enrique Otero Muras, José Luis Alba Castro, and Carmen García Mateo

Feature Distribution of the Fingerprint Template Generated by the Geometric Hashing-Based Fuzzy Vault 816
Sungju Lee, Daesung Moon, and Yongwha Chung

Hand Geometry Verification Using Time Series Representation 824
Vit Niennattrakul, Dachawut Wanichsan, and Chotirat Ann Ratanamahatana

Navigation Dynamics as a Biometrics for Authentication 832
Ernesto Damiani and Gabriele Gianini

Logic Based Intelligent Information Systems

A Cooperative Grid Computing Approach to Automated Theorem Finding and Automated Problem Proposing 840
Jingde Cheng, Yuichi Goto, Shinsuke Nara, and Takahiro Koh

A Logic Grammar for Circuit Analysis: Problems of Recursive Definition 852
Takushi Tanaka

A Natural Deduction System for Annotated Predicate Logic 861
Seiki Akama, Kazumi Nakamtsu, and Jair Minoro Abe

An Intelligent Coordinated Traffic Signal Control Based on EVALPSN 869
Kazumi Nakamatsu, Jair M. Abe, and Seiki Akama

Intelligent Design of Diagnosable Systems: A Case Study of Semiconductor Manufacturing Machines 877
YuanLin Wen, ShengLuen Chung, LiDer Jeng, and MuDer Jeng

Knowledge Assessment Based on Many-Valued Logic 885
Sylvia Encheva and Sharil Tumin

Monadic Curry Algebras Qτ 893
Jair Minoro Abe, Seiki Akama, and Kazumi Nakamatsu

II Intelligent Knowledge

Chance Discovery

A Method for Visualising Possible Contexts 901
Shigeki Amitani and Ernest Edmonds

An Interface for Medical Diagnosis Support 909
Akinori Abe, Norihiro Hagita, Michiko Furutani, Yoshiyuki Furutani, and Rumiko Matsuoka

Creating Chances Through Cognitive Niche Construction: The Role of Affordances 917
Lorenzo Magnani

Discovering Color Semantics as a Chance for Developing Cross-Cultural Design Frameworks 926
Gyoung Soon Choi, Ruediger Oehlmann, Hilary Dalke, and David Cottington

Evaluating a Constructive Meta-learning Algorithm for a Rule Evaluation Support Method Based on Objective Indices 934
Hidenao Abe, Shusaku Tsumoto, Miho Ohsaki, and Takahira Yamaguchi

Information Literacy and Everyday Life Risks 942
Yumiko Nara

Logic of Discovery in Uncertain Situations–Deciding Algorithms 950
V. Rybakov

RFID Tags Without Customers ID in Book Library for Detecting Latent Interest of User Group 959
Yukio Ohsawa, Takuma Hosoda, Takeshi Ui, Misato Ueda, and Hirofumi Tanaka

Trigger to Switch Individual's Interest Toward Unconscious Preference 970
Yoshiharu Maeno, Yukio Ohsawa, and Takaichi Ito

Visualization of Similarities and Dissimilarities Between Rules Using Multidimensional Scaling 978
Shusaku Tsumoto and Shoji Hirano

Knowledge - Based Interface Systems (I)

A Unified Approach to Web Usage Mining Based on Frequent Sequence Mining 987
Nobuhiro Inuzuka and Jun-ichi Hayakawa

DACS Web Service 995
Kazuya Odagiri, Rihito Yaegashi, Masaharu Tadauchi, and Naohiro Ishii

Estimation of Initial Contour Based on Edge Background Subtraction for Self-affine Mapping System 1005
Haruki Kawanaka, Hirofumi Kato, Fuminori Matsubara, Yuji Iwahori, and Robert J. Woodham

Improvement of Accuracy for Gaussian Curvature Using Modification Neural Network 1013
Yuji Iwahori, Takashi Nakagawa, Shinji Fukui, Haruki Kawanaka, Robert J. Woodham, and Yoshinori Adachi

Shadow Removal Method for Real-Time Extraction of Moving Objects 1021
Shinji Fukui, Nozomi Yamamoto, Yuji Iwahori, and Robert J. Woodham

Knowledge - Based Interface Systems (II)

Delayed Learning and the Organized States 1029
Toshinori Deguchi and Naohiro Ishii

Educational System Using Self-monitor Study and Streaming 1037
Masahiro Ozaki, Yoshinori Adachi, Saori Takeoka, Ai Sugimura, and Naohiro Ishii

Influence of Presence of Frame on Writer Recognition 1045
Yoshinori Adachi, Masahiro Ozaki, Yuji Iwahori, and Naohiro Ishii

Information Extraction by XLM 1051
Masashi Okada, Naohiro Ishii, and Nariaki Kato

The Characteristics Evaluation of Stacked-Type Electrostatic Actuators by the Neural Network 1059
Masashi Kawaguchi, Katsuhisa Yamano, Kazuo Okuda, Takashi Jimbo, and Naohiro Ishii

Knowledge and Information Management in Social Community

A Piecewise Linear Representation Method of Time Series Based on Feature Points 1066
Yuelong Zhu, De Wu, and Shijin Li

A Regional Safety Information Sharing System Based on CMS and Online Map System 1073
Akira Hattori, Masato Goto, Takami Yasuda, and Shigeki Yokoi

Avoidance of Traffic Delay for Panicking Crowds Subject to Information Propagation 1081
Naoto Mukai, Toyohide Watanabe, and Jun Feng

Building the Virtual Community to Support Interregional Exchange Between Rural and Urban 1089
Yoshida Chiho, Yasuda Takami, and Yokoi Shigeki

Indexing of Moving Objects on Road Network Using Composite Structure 1097
Jun Feng, Jiamin Lu, Yuelong Zhu, Naoto Mukai, and Toyohide Watanabe

Optimization of Vehicle Assignment for Car Sharing System 1105
Kentaro Uesugi, Naoto Mukai, and Toyohide Watanabe

Knowledge and Ontological Engineering for (KOS)

An Application of Decision Trees for Rule Extraction Towards Telecommunications Fraud Detection 1112
Constantinos S. Hilas and John N. Sahalos

Assisting Dialogical Agents Modeled from Novice User's Perceptions 1122
David Leray and Jean-Paul Sansonnet

Combining Empirical Studies of Audio-Lingual and Visual-Facial Modalities for Emotion Recognition 1130
M. Virvou, G.A. Tsihrintzis, E. Alepis, I.-O. Stathopoulou, and K. Kabassi

Learning Objects Repository for Training of Power Systems Operators 1138
G. Arroyo-Figueroa, L. Argotte-Ramos, and G. Rodríguez-Ortíz

Mining Meaningful Student Groups Based on Communication History Records 1146
Yang-Sae Moon, Hun-Young Choi, Hea-Suk Kim, and Jinho Kim

Ontological Engineering for Practical Knowledge Work 1154
Tatiana Gavrilova

Semantic Modeling of Product Manuals 1162
Rossitza Setchi, Nikolaos Lagos, and Ammar Huneiti

Knowledge Engineering in Multi Robot Systems

Appearance-Based Multi-robot Following Routes Using Incremental PCA 1170
Luis Payá, Oscar Reinoso, Arturo Gil, Jose M. Pedrero, and Mónica Ballesta

Contribution to Legged Robot Visual Servoing 1179
Zelmar Echegoyen, Alicia d'Anjou, and Manuel Graña

Incremental Evolution of Stigmergy-Based Multi Robot Controllers Through Utility Functions 1187
P. Caamaño, J.A. Becerra, R.J. Duro, and F. Bellas

Lattice Independence and Vision Based Mobile Robot Navigation 1196
I. Villaverde, M. Graña, and J.L. Jimenez

ROBMAT: Teleoperation of a Modular Robot for Collaborative Manipulation 1204
Juan A. Escalera, Manuel Ferre, Rafael Aracil, and José Baca

Knowledge-Based Creativity Support Systems

Data Mining of Time-Series Medical Data by Formal Concept Analysis 1214
Kenji Sato, Yoshiaki Okubo, Makoto Haraguchi, and Susumu Kunifuji

A Case Study on Project-Management Training-Support Tools for Japanese/Chinese/Indian Offshore Development Engineers 1222
Sadamitsu Toyoda, Motoki Miura, and Susumu Kunifuji

A Modelling Framework for Sharing Knowledge 1230
Reyes Grangel, Ricardo Chalmeta, and Cristina Campos

An Assistant Interface for Finding Query-Related Proper Nouns 1238
Tomoya Iwakura, Kanji Uchino, and Seishi Okamoto

Community Support System for Extensive Readers: Case Study in a Domain 'Knowledge Science' 1246
Kohei Ishikawa, Kanayo Ogura, and Kozo Sugiyama

Design and Evaluation a Knowledge Management System by Using Mathematical Model of Knowledge Transfer 1253
Kouji Aoyama, Takanori Ugai, and Jun Arima

Development of a Mimamori-Care System for Persons with Dementia Based on the Real World-Oriented Approach 1261
Kenichi Nakagawa, Taro Sugihara, Hitoshi Koshiba, Ryozo Takatsuka, Naotaka Kato, and Susumu Kunifuji

Overview of Network Information by Using Anchored Maps 1269
Kazuo Misue

Spreadsheet Interface Which Simplifies the Visualization of the Network Structures 1277
Satoshi Koike, Kazuo Misue, and Jiro Tanaka

Knowledge-Based Multi-criteria Decision Support

A Fuzzy Ranking Approach to Data Envelopment Analysis 1285
Hsuan-Shih Lee, Pei-Di Shen, and Wen-Li Chyr

A New Information Fusion Method for Fuzzy Information Retrieval 1293
Hsuan-Shih Lee, Ming-Tao Chou, Wei-Kuo Tseng, Hsin-Hsiung Fang, and Chen-Huei Yeh

An Analytic Approach for Synthesizing Comparison Matrices in AHP 1299
Hsuan-Shih Lee, Cheng-Chi Chung, Su-Man Wang, Ming-Tao Chou, and Chen-Huei Yeh

Estimating Missing Values in Incomplete Additive Fuzzy Preference Relations 1307
Hsuan-Shih Lee, Ming-Tao Chou, Hsin-Hsiung Fang, Wei-Kuo Tseng, and Chen-Huei Yeh

Fuzzy Logic Based Reputation System for Mobile Ad Hoc Networks 1315
Jin-Long Wang and Shih-Ping Huang

Synthesizing Comparison Matrices of AHP Under Group Decision 1323
Ming-Tao Chou, Hsuan-Shih Lee, Ching-Wu Chu, and Chao-Yuan Cheng

Knowleged Based Systems for e-Business

A Proposal of Adequate and Efficient Designing of UML Documents for Beginners 1331
Masakazu Takahashi, Satoru Takahashi, and Yoshikatsu Fujita

Analysis of the Relation Between Stock Price Returns and Headline News Using Text Categorization 1339
Satoru Takahashi, Masakazu Takahashi, Hiroshi Takahashi, and Kazuhiko Tsuda

Evaluation of a Hierarchical Shaper as a Policy Execution Point 1346
Takeshi Aimoto, Takeki Yazaki, Takashi Isobe, Yoshihiko Sakata, and Kenichi Yoshida

Query Message Delivery over Community-Based Overlay Network 1354
Yoshikatsu Fujita, Yasufumi Saruwatari, Jun Yoshida, and Kazuhiko Tsuda

Relation Analysis on Information System Life Cycle Processes by KeyGraph Algorithms 1362
Tadashi Tsukahara, Atsuo Hazeyama, and Kazuhiko Tsuda

Author Index 1369

An Efficient Indexing Technique for Location Prediction of Moving Objects

Dong-Oh Kim, Kang-Jun Lee, Dong-Suk Hong, and Ki-Joon Han

School of Computer Science & Engineering, Konkuk University,
1, Hwayang-Dong, Gwangjin-Gu, Seoul 143-701, Korea
{dokim, kjlee, dshong, kjhan}@db.konkuk.ac.kr

Abstract. The necessity of the future index is increasing to predict the future location of moving objects promptly for various location-based services. However, the prediction performance of most future indexes is lowered by the heavy load of extensive future trajectory search in long-range future queries, and their index maintenance cost is high due to the frequent update of future trajectories. Thus, this paper proposes the Probability Cell Trajectory-Tree (PCT-Tree), a cell-based future indexing technique for efficient long-range future location prediction. The PCT-Tree reduces the size of index by building the probability of extensive past trajectories in the unit of cell, and predicts reliable future trajectories using information on past trajectories. Therefore, the PCT-Tree can minimize the cost of communication in future trajectory prediction and the cost of index rebuilding for updating future trajectories. Through experiment, we proved the superiority of the PCT-Tree over existing indexing techniques in the performance of long-range future queries.

Keywords: Moving Objects, PCT-Tree, Future Index, Future Trajectory, Cell, Probability Matrix.

1 Introduction

These days the advance of wireless communication technology and location information technology has brought the development of various location-based services (LBSs) [4,7,8]. In LBS, each moving object, which is a client, transmits information about its location to the server, and the server collects and indexes location information for moving objects and searches necessary location information. Such location indexing of moving objects for LBS is getting more important and research is being made actively on the future index for the fast information service about the present and future location of vehicles in logistics, car tracking and emergent situations [1,2,6].

The future index is a spatio-temporal index for querying present and future locations of moving objects. That is, the future index is used to obtain the future location of moving objects by processing present and future time queries such as "Which taxi will pass the city hall after 10 minutes?" and "Where will a moving object be after 10 minutes?" [6]. The range of future query time to be indexed has a

B. Apolloni et al. (Eds.): KES 2007/ WIRN 2007, Part II, LNAI 4693, pp. 1–9, 2007.
© Springer-Verlag Berlin Heidelberg 2007

significant effect on the size of the index and its performance. In general, the longer the query time is, the larger the index is and the lower the accuracy of queries is.

Future indexes are divided into network-based ones and non-network-based ones according to whether or not a road network is used. In the road network environment, a network-based future index is more effective [4,5]. With regard to network-based future indexes, there have been researches on the segment tracking technique [2] that predicts the location of moving objects using information on the road segment, the road network modification technique [1] that modifies road network by extending road segments in order to reduce segment changes, PLM (Prediction Location Model) [6] that generates future trajectories through probability-based prediction of road segments, etc. However, segment tracking has low reliability in the prediction of the future location and increases the communication cost between the server and the client. The road network modification technique changes many road segments in actual situations. PLM has many probability tables and generates an extensive probability search tree in long-range future location prediction.

To solve these problems in future indexes, this paper proposes the Probability Cell Trajectory (PCT)-Tree, an efficient future trajectory indexing technique for predicting the location of moving objects. The PCT-Tree is a future index that can support future trajectory prediction to solve the problem of slow prediction of future location in long-range future queries. The PCT-Tree manages past information on moving objects' entrance into and exit from cells and, by doing so, makes more reliable prediction of the future location of the moving objects. In addition, the PCT-Tree can guarantee faster future trajectory prediction by traversing the probability variables in long-range future location prediction.

This paper is organized as follows. Chapter 2 analyzes existing network-based future index techniques. Chapter 3 explains the PCT-Tree in detail and Chapter 4 discusses the results of performance evaluation through experiment. Finally, Chapter 5 draws conclusions.

2 Network-Based Future Index

This chapter analyzes existing network-based future index techniques that can predict the future location of moving objects. The basic intention of segment tracking is utilizing information on the road network when a client is moving. The server knows what segment of the road network the client is on using information sent by the client and, for this, the server uses the map matching technology. In segment tracking, a client's future location can be predicted as a point moving at a constant speed along a segment represented as a polyline [2]. Here, the speed is what has been reported most recently. However, segment tracking has a problem in that it cannot predict future location at the end of the segment and has to be switched to point or vector tracking and the client has to report to the server whenever it moves from the present segment to another segment.

The road network modification technique changes the road network by extending road segments in order to reduce segment changes. The purpose of road network modification is to connect segments in a way of minimizing segment changes when

an object moves around the road network [1]. Compared to segment tracking, road network modification can predict future location more effectively. If there are fewer segment changes predicted in the future index, the cost of communication and server load can be reduced together. However, because this technique considers the general properties of the road network (i.e., street number, main roads, trunk roads) and the characteristic of ordinary moving objects (i.e., straightness), it is distant from the preference of individual moving objects and, as a consequence, can change many road segments in actual situations.

In PLM, the server manages a probability matrix using the frequency of road segments selected by moving objects at junction points based on past trajectory information [6]. After the probability matrix for the past trajectories of moving objects is stored in the server, moving objects issue a future query to the server. PLM creates a probability search tree predictable from the location of a moving object using exit points determined by sub-network, and using the tree it can predict a continuous future trajectory of road segments to be passed from the present location of the moving object. However, because PLM has to manage the probability matrix at every junction point for each moving object, the number of probability variables is large, and if the time range of prediction is long the probability search tree becomes extensive and search time for future trajectory prediction becomes longer.

3 PCT-Tree

This chapter explains the general structure of the PCT-Tree, cell probability matrix, and PCT-Tree algorithms.

3.1 General Structure of PCT-Tree

Given a road network, the PCT-Tree builds a spatial index by dividing the network into cells of non-uniform size (called, *terminal cells*). Figure 1(a) is an example of a road network divided into non-uniform cells. The PCT-Tree divides cells using the number of road segments within a cell and the number of entrance/exit boundary points of the cell. If a cell is too large, the cell has to manage too many road segments, which increases the cost of internal trajectory management. In a terminal cell of the PCT-Tree, a *boundary point* means an entrance/exit point of the cell where a road segment in the cell meets the cell boundary.

In the PCT-Tree, the trajectory of a moving object can be represented as a series of locations where the object exits from a cell and enters into another cell and this is called a *cell trajectory*. A problem in the cell trajectory is that information on road segments within each cell disappears. Therefore, the PCT-Tree creates a trajectory within the cell through the shortest path algorithm.

The PCT-Tree has Quad-Tree structure, and has cell probability variables that form a cell probability matrix (CPM) for each terminal cell of the PCT-Tree. For each moving object, the PCT-Tree stores the experience of selecting a road segment inside a cell and boundary points of the cell in the probability variables of the cell

probability matrix. The number of boundary points in a cell determines the number of probability variables. Figure 1 shows an example of PCT-Tree structure.

As in Figure 1(a), the past trajectories of object O1 and O2 pass a number of cells. A past trajectory is converted to a cell trajectory which is composed of boundary points, entrance time and exit time. The cell trajectory of O1 is 00(1,2,t0,t1) →01(1,3,t1,t2)→02(0,1,t2,t3)→3(0,1,t3,t4)→20(0,1,t4,t5)→21(0,1,t5,t6)→22(0,1,t7, t8), and that of O2 is 00(0,3,t2,t3)→01(0,2,t3,t4)→1(0,1,t4,t5). Using the information, the record of cell object (CO) is created.

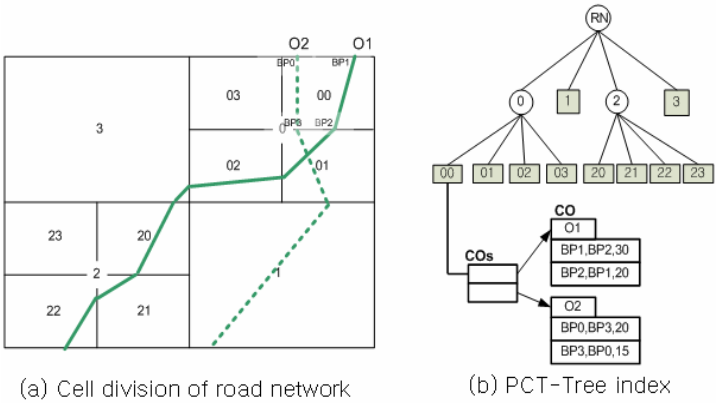


Fig. 1. Structure of PCT-Tree

Figure 1(b) shows a PCT-Tree index that stores the cell trajectories of the given road network. In Figure 1(a), there are two possible cases that O1 may enter and exit from cell 00, namely, BP1→BP2 and BP2→BP1. As in Figure 1(b), object O1 visited cell 00 30 times through BP1→BP2 and 20 times through BP2→BP1. Likewise, object O2 visited cell 00 20 times through BP0→BP3 and 15 times through BP3→BP0. In this way, if past trajectories of a moving object are entered more, the probability variables of the cell object store more experiences and are used as variables of the cell probability matrix when predicting future trajectories.

3.2 Cell Probability Matrix

The cell probability matrix has cell probability variables to generate the future trajectory of a moving object. Cell probability variables are generated only with boundary points that the moving object has passed, and search for future trajectory prediction is performed using the cell probability variables. Equation (1) shows the cell probability matrix for cell C with n the number of entrance/exit *boundary points*.

In Equation (1), time t_k is the present time, B_1, B_2,..., B_n are boundary points of the cell. $P(B_i|B_j)$ is the probability that a moving object enters through boundary point B_i of the cell and exits through B_j. In this way, entrance and exit data are used as the cell probability variables of the cell probability matrix. $P(B_i|B_1)$ (i=1,n) can be expressed as Equation (2) using frequencies.

$$\mathbf{CPM(C, t_k)} = \begin{pmatrix} P(B_1|B_1) & P(B_2|B_1) & \dots & P(B_n|B_1) \\ P(B_1|B_2) & P(B_2|B_2) & \dots & P(B_n|B_2) \\ \dots & \dots & \dots & \dots \\ P(B_1|B_n) & P(B_2|B_n) & \dots & P(B_n|B_n) \end{pmatrix} \quad (1)$$

$$\begin{aligned} P(B_1|B_1) &= P(N_{1,1}/N_1) \\ P(B_2|B_1) &= P(N_{2,1}/N_1) \\ &\dots \\ P(B_n|B_1) &= P(N_{n,1}/N_1) \end{aligned} \quad (2)$$

In Equation (2), N_i indicates the number of visits to B_i, and $N_{i,j}$ means the number of visits from B_j to B_i. Suppose that an object on boundary point B_1 of the cell can pass through B_2, B_3 or B_4. In this case, U-turn, namely, returning to its original location is excluded ($N_{i,i}$=0). The sum of the probabilities of all possible cases in N_1 is 1 (i.e., $P(N_1) = 1$). Therefore, $P(N_1) = P(N_{2,1})+P(N_{3,1})+P(N_{4,1}) = 1$ and resultantly $P(N_{2,1}) = P(N_{3,1}) = P(N_{4,1}) = 1/3$. Here, if N_1 is 3, $N_{2,1} = N_{3,1} = N_{4,1} = 1$. Therefore, probability $P(B_2|B_1) = P(N_{2,1}|N_1) = 1/3$. In the same way, $P(B_3|B_1) = P(B_4|B_1) = 1/3$. Therefore, the probability that a moving object exits from boundary point B_1, which has never been passed, through any of the other three boundary points is 33.33%.

Figure 2 shows PCT-Tree cell division using a road network. Here, C_i (i=1,2) indicate cells, B_j (j=1,4) boundary points, and E_k (k=1,7) edges (that is, road segments) linking the entrance boundary points and exit boundary points of the cells. Let us examine the cell probability matrix for each cell after 20 repetitions of past trajectory $E_1 \rightarrow E_3 \rightarrow E_5$. Equation (3) is the cell probability matrix of cell C_1 and C_2.

$$\begin{aligned} CPM(C_1,t_k) &= \begin{pmatrix} 0 & 1/23 & 21/23 & 1/23 \\ 1/3 & 0 & 1/3 & 1/3 \\ 1/3 & 1/3 & 0 & 1/3 \\ 1/3 & 1/3 & 1/3 & 0 \end{pmatrix} \\ CPM(C_2,t_k) &= \begin{pmatrix} 0 & 1/23 & 1/23 & 21/23 \\ 1/3 & 0 & 1/3 & 1/3 \\ 1/3 & 1/3 & 0 & 1/3 \\ 1/3 & 1/3 & 1/3 & 0 \end{pmatrix} \end{aligned} \quad (3)$$

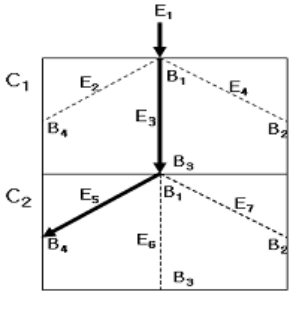


Fig. 2. Cell division

Let us find probability $P(E_1, E_3, E_5)$ for moving through $E_1 \rightarrow E_3 \rightarrow E_5$. Equation (4) is the Bayesian theory used to find the probability of cell trajectory.

$$P(b,a) = P(b|a) * P(a) \tag{4}$$

In Equation (4), $P(b,a)$ is the probability that b and a take place together, and $P(b|a)$ is the probability that b takes place after a has already done. If the present moving object exits from road segment E_1 and enters cell C_1, probability $P(E_1)$ is 1. Therefore, $P(E_3,E_1) = P(E_3|E_1)*P(E_1) = P(B_3|B_1)*1 = 21/23 = 0.913$. In the same way, $P(E_2,E_1) = P(E_4,E_1) = 1/23 = 0.0435$. Here, because E_1 and E_5 are independent in the probability space, $P(E_5|E_1,E_3) = P(E_5|E_3)$. Therefore, since $P(E_1,E_3,E_5) = P(E_1)* P(E_3|E_1)*P(E_5|E_1,E_3)$, we can obtain $P(E_1,E_3,E_5) = P(E_1)*P(E_3|E_1)*P(E_5|E_3) = 1*21/23 *21/23 = 0.8336$. Simiarly, we can get $P(E_1,E_3,E_6) = P(E_1,E_3,E_7) = 1*21/23*1/23 = 0.0397$.

The PCT-Tree uses the cell probability matrix without building a probability search tree for long-range future trajectory prediction, and traverses cell probability variables when predicting future trajectories. That is, traversing cell probability variables with the first-priority and second-priority probability and searching the future trajectory of the highest probability is the traversing rule of the PCT-Tree for the improvement in the speed of long-range future trajectory prediction at the cost of a slight prediction error.

3.3 PCT-Tree Algorithms

Figure 3 shows the search algorithm of the PCT-Tree. Line 1 adds the cell trajectory *CT* and the probability *Prob* of *CT* to *Cand_Prob_CT_List* if *Depth* is zero. *Depth* contains the limit value of the future trajectory prediction. Line 2 finds all cell object probabilities *cops* with the entrance boundary point *In_BP* in COPs of CO which is the cell object with *OID* in the cell *C*. Line 3 calculates total weight *totw* of *cops*. Line 4 finds the cell object probability *cop1* which has the highest weight in *cops*. Line 5 creates the cell trajectory segment CTS and adds it to *CT* to create the future cell trajectory *ct1*. Line 6 calls PCT_Search recursively to find the next future cell

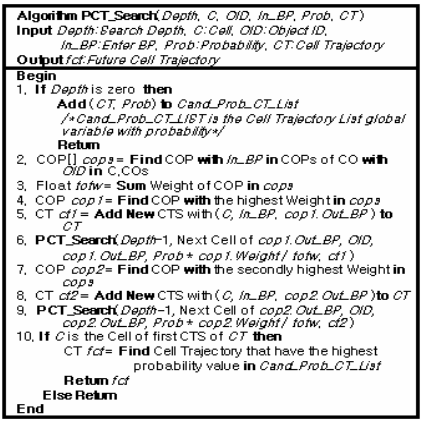


Fig. 3. Search algorithm

trajectory of *ct1*. Line 7~9 create the next future cell trajectory of the cell object probability *cop2* which has the secondary highest weight in *cops*. Line 10 returns the future cell trajectory *fct* which has the highest probability in *Cand_Prob_CT_List*, if C is the cell of the first CTS in *CT*.

Figure 4 shows the insert algorithm of the PCT-Tree. Line 1 converts the input trajectory *Traj* into cell trajectory *ct*. Line 2 takes the cell trajectory segment *cts* from *ct* and Line 3 finds the cell object *co* with *OID* in *cts.Cell*. Line 4 creates the cell object *co* newly if *co* with *OID* does not exist. Line 5 sets the access time entered into *co*. Line 6 finds the cell object probability *cop* with the entrance boundary point *cts.In_BP* and the exit boundary point *cts.Out_BP*. Line 7 creates the cell object probability *cop* with *cts.In_BP* and *cts.Out_BP* if *cop* does not exist. Finally, Line 8 increases the weight of the cell object probability *cop.Weight*.

Figure 5 shows the delete algorithm of the PCT-Tree. Line 1 converts the input trajectory *Traj* into the cell trajectory *ct*. Line 2 takes the cell trajectory segment *cts* from *ct* and Line 3 finds the cell object *co* with *OID* in *cts.Cell*. Line 4 goes to Line 2 if *co* with *OID* does not exist. Line 5 finds the cell object probability *cop* with the entrance boundary point *cts.In_BP* and the exit boundary point *cts.Out_BP*. Finally, Line 6 decreases the weight of the cell object probability *cop.Weight* if *cop* exists.

```
Algorithm PCT_Insert( OID, Traj )
Input OID:Object ID, Traj:Trajectory

Begin
1, CT ct= Convert Trajectory Traj to Cell Trajectory
2, For each Cell Trajectory Segment cts in ct do
3,    CO co = Find Cell Object with OID in cts.Cell
4,    If co not exists then co = New CO with OID
5,    Set access time into co
6,    COP cop = Find COP with cts.In_BP and cts.Out_BP in
         COPs of co
7,    If cop not exists then
         cop = New COP with ( cts.In_BP, cts.Out_BP)
8,    Increment cop.Weight by one
End
```

Fig. 4. Insert algorithm

```
Algorithm PCT_Delete( OID, Traj )
Input OID:Object ID, Traj:Trajectory

Begin
1, CT ct= Convert Trajectory Traj to Cell Trajectory
2, For each Cell Trajectory Segment cts in ct do
3,    CO co = Find Cell Object with OID in cts.Cell
4,    If co not exists in cts.Cell then goto 2
5,    COP cop = Find COP with cts.In_BP and cts.out_BP in
         COPs of co
6,    If cop exists then
         Decrement cop.Weight by one
End
```

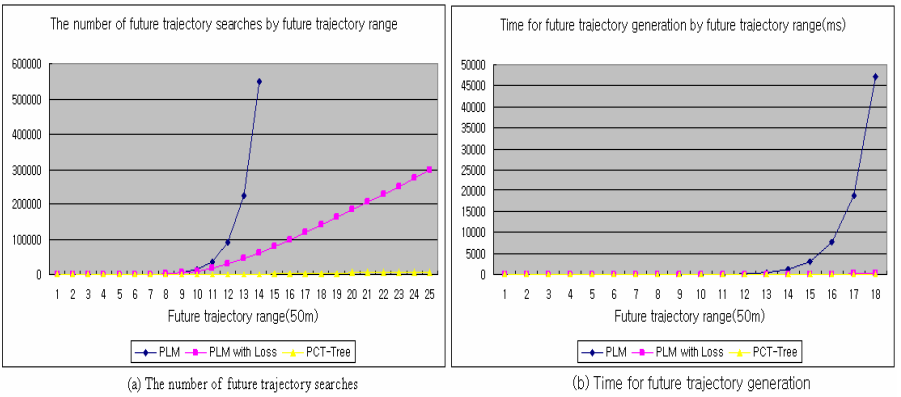
Fig. 5. Delete algorithm

4 Performance Evaluation

In PLM, all trajectories are checked 100% without loss in traverse for future trajectory prediction. Therefore, this paper considered PLM with loss that does not traverse trajectories with low probability. In the performance evaluation, we compared the number of future trajectory searches and time for future trajectory generation among PLM, PLM with Loss and the PCT-Tree. Figure 6 show graphs that show the performance comparison of future trajectory prediction among them.

As shown in Figure 6(a) and 6(b), the number of future trajectory searches and time for trajectory generation of PLM increase by the exponential time if the future trajectory range exceeds a certain limit. In PLM with loss, the number of future trajectory searches increases in proportion to the future trajectory range after passing a certain limit but time for future trajectory generation is similar to that of the PCT-Tree. However, in the PCT-Tree, the number of future trajectory searches and time for future trajectory generation for long-range future trajectory prediction are almost

constant, since they hardly change according to future trajectory range not only in short-range prediction but also in long-range prediction.



(a) The number of future trajectory searches (b) Time for future trajectory generation

Fig. 6. Performance comparison of future trajectory prediction

5 Conclusions

This paper proposed the PCT-Tree which is a cell-based future index for efficient long-range future location prediction, and proved its superiority to existing future indexing techniques through the performance evaluation. Since the PCT-Tree uses the cell probability matrix without building the probability search tree, this can reduce index size and enable prompt and efficient prediction of the future trajectories of moving objects. Therefore, we expect that the PCT-Tree can be used as an index for location prediction engines that can enhance the reliability of future location prediction in various location-based services and for traffic control systems that can predict future paths for guiding emergency or service vehicles to fast paths.

Acknowledgments. This research was supported by the Seoul Metropolitan Government, Korea, under the Seoul R&BD Program supervised by the Seoul Development Institute.

References

1. Civilis, A., Jensen, C.S., Pakalnis, S.: Techniques for Efficient Road-Network-Based Tracking of Moving Objects. IEEE Transactions on Knowledge and Data Engineering 17, 698–712 (2005)
2. Civilis, A., Jensen, C.S., Nenortaite, J., Pakalnis, S.: Efficient Tracking of Moving Objects with Precision Guarantees. In: Proc. of 1st. Int. Conf. on Mobile and Ubiquitous Systems: Networking and Services, pp. 164–173 (2004)
3. Ding, R., Meng, X., Bai, Y.: Efficient Index Maintenance for Moving Objects with Future Trajectories. In: Proc. of 8th Int. Conf. on Database Systems for Advanced Applications, pp. 183–192 (2003)

4. Ding, Z., Guting, R.H.: Managing Moving Objects on Dynamic Transportation Networks. In: Proc. of 16th Int. Conf. on Scientific and Statistical Database Management, pp. 287–296 (2004)
5. Frentzos, E.: Indexing Objects Moving on Fixed Networks. In: Proc. of 8th Int. Symp. on Spatial and Temporal Databases, pp. 289–305 (2003)
6. Karimi, H.A., Liu, X.: A Predictive Location Model for Location-Based Services. In: Proc. of 11th ACM Int. Symp. on Advances in Geographic Information Systems, pp. 126–133 (2003)
7. Lee, E.J., Rye, K.H., Nam, K.W.: Indexing for Efficient Managing Current and Past Trajectory of Moving Object. In: Proc. of 6th Asia-Pacific Web Conf. on Advanced Web Technologies and Applications, vol. 3007, pp. 782–787 (2004)
8. Saltenis, S., Jensen, C.S.: Indexing of Moving Objects for Location-Based Services. In: Proc. of 18th Int. Conf. on Data Engineering, pp. 463–472 (2002)

Combining Smart Tags and Body Fixed Sensors for Disabled People Assistance

Juan Antonio Álvarez[1], Carlos Pérez[2], Cecilio Angulo[2], and Juan Antonio Ortega[1]

[1] ETSII. University of Seville
Avenida Reina Mercedes, s/n. 41012 – Sevilla, Spain
jaalvarez@us.es, ortega@lsi.us.es
[2] CETpD – UPC. Technical University of Catalonia
Rambla de l'Exposició s/n, 08800 – Vilanova i la Geltrú, Spain
carlos.perez-lopez@upc.edu, cecilio.angulo@upc.edu

Abstract. The world population is rapidly aging in industrialized countries and at a growing rate. As well as the percentage of elderly people is constantly increasing, so related health costs are augmenting, demanding novel technological solutions that both, enhance elder daily life independence and reduce the social and economic overhead of their hospitalization. The general framework of the presented work is about technological platforms for helping elder and handicapped people in a non-intrusive manner. The research in human motion monitoring, having in mind the particular objective of fall prevention, and the use of RFID technology combined with text to speech techniques are merged on a general purpose assistive platform able to help all kind of people, including those who has not impairment but are seeking to improve their quality of live.

Keywords: Multimodal sensor fusion, assistive systems, body monitoring.

1 Introduction

Society demands adequate, accessible and affordable social services for disabled people and, in general, needing healthcare. Focusing in these needs as objective, this article presents on-going research that is creating novel knowledge on both, technology and accessible and adequate technological tools for people with dependencies needing medical monitoring or wanting to improve their quality of daily life. In particular, research is centred on human motion monitoring, having in mind the objective of fall prevention, and the use of RFID technology combined with text to speech techniques with the aim to make available an assistive platform able to help all kind of people.

Several works on human motion capture through artificial vision systems are in the literature concerning different application scenarios. These systems are usually based on the use on single or multiple cameras [1], producing in this form undesirable reactions from assisted people. A less intrusive system should be used in assistive environments, for example using fusion of data sensed with a device built in the form

B. Apolloni et al. (Eds.): KES 2007/ WIRN 2007, Part II, LNAI 4693, pp. 10–17, 2007.
© Springer-Verlag Berlin Heidelberg 2007

of wearable computing [2]. For particular diseases like Parkinson, there exist some interesting solutions [3, 4], but there are specifically focused in this disease, not for a general purpose. A more general system with a wider scope will be proposed in this work, considering as primary target users elder people with visual impairments and/or with motion difficulties.

There exist some projects with similar target users [4, 5], but they are mainly concerned on middleware or data mining procedures. In contrast, our work is offering complete electronics solutions based on body fixed sensors in the form of a wearable motion monitoring device, named KneeMeasurer, and approximate indoor positioning by using RFID tags technology, both integrated through a computing system running on a PDA unit. This technological platform signifies a non-intrusive communication link between the assisted users, their usual environments and their helpers. According to some inquiries, users prefer to be assisted by known people like their relatives, so a non-intrusive, secure and personal communication must be true.

In the next section an exemplified scenario is presented in order to illustrate the capabilities of the assistive platform. Next, the proposed architecture is presented. Some evidences of the completed work on the wearable motion monitoring device, the tag-based text-to-speech processing, and the integrated communication through a PDA unit are presented showing the viability of the proposed platform. Some conclusions and future work are presented to conclude.

2 Scenario

To show the possibilities of our system, a complete scenario is explained:

"John must take in some medicines at 4:00 PM, but he forgets them. Fortunately, a tag-based identification subsystem access to the stored information and detect that he hadn't took the object tagged with the medicine. The wearable PDA reads, using the Text to Speech technology, the reminder associate: "John, you must take the medicine sited on the cupboard of the kitchen, have you done? Touch your PDA-screen once if you do, twice if you don't". John touches it twice and goes to the kitchen; when he takes his medicine, he deactivates the reminder. Later, John has to do some exercise to keep fit according to doctor prescription, but he did not. At 7:00 PM, the reminder service explains John that he has not done too much exercise during the day, and it would be recommendable to do it before dinning. At 9:30 John is having a shower and unfortunately he fall on it and loss his consciousness. Although he is not carrying the wearable PDA and the motion monitoring device, the last place where the location service had information was the bathroom because he takes the tagged shampoo. The long time period with neither change in the location service nor tagged elements is detected as a risky situation so the communication service sends an alert text to the communication server that resends it as short messages (sms) to mobile phones of a relative or/and helpers.

3 Technology

RFID technology combined with text to speech techniques will be combined with a body fixed sensors motion monitoring system to create an assistive platform able to

help all kind of people in similar scenarios like that described above, including those who has not impairment but are seeking to improve their quality of live.

3.1 RFID

Radio frequency identification has been selected for the design of the positioning system [6, 7] because it is a growing technology being integrated on every element in our daily life, in the same form that bar codes were introduced some time ago. Nowadays, the most extended RFID technology is "High Frequency" (HF), working in the band of 13.56 MHz, being used as a basis for the communication system called NFC ("Near Field Communication"). The main disadvantage of this technology is that read distances range from 3 to 5 centimetres.

A second general RFID approach is "Ultra High Frequency" (UHF), working in the range of frequencies from 850 to 950 MHz. This technology is much less extended, but is the one that tends to more expand in the industrial identification fields due to its long reaches (until 8m in the latest versions of passive tags and 30 – 100m with active tags). In the next years this latter technology will prevail in the industrial area and a large number of products will come to our homes with the factory label, like now they are coming with bar codes. Unfortunately, nowadays the unique portable RFID reader Compact Flash for PDA devices is HF, so we are obligated to select this technology for our on-going applications.

3.2 KneeMeasurer

A novel methodology designed for measuring human joint angles, in a first version for knee angles, has been developed for the assistive platform in indoor and outdoor environments using a wearable gesture interface [8]. The knee-mounted KneeMeasurer device consists of two modules endowed with a dual axis accelerometer, a digital signal processing element and a wireless communication component. Accelerometers are commonly used to sense leg movements for monitoring the user gait in his/her daily life. Joint angles can be directly calculated in the processing units situated on the nodes of the KneeMeasurer and this information be sent by Bluetooth communication to a handled device. As a first application of this motion monitoring device, completely wireless on-line monitoring of the knee angle is performed.

The accelerometers used are Analog Devices ADXL203. They are precision dual-axis accelerometers with an scale of ±1.7 g. Adjusting the working bandwidth of them, it is possible to achieve a resolution of 13 mg, that is a typical measured error of 0.8 degrees of tilt. The accelerometers have two analogical outputs, one for each axis, that are directly connected to an analogical to digital converter. The sensor boards are both connected by cable to a process unit which captures all the sensor measures, calculates knee angles and sends it to a desktop, a laptop or a handled by Bluetooth communication. A deeper explanation of the system is showed on [8].

The power consumption is a critical factor in wireless devices, even more if the size of the system is also an important parameter. The sensors must been chosen with the power consumption as a premise. Each accelerometer consumes about 0.6 mA in

normal operation, and gyroscopes consume 4 mA approximately each one. For this reason, it has been tried to minimize the consumption of every part of the design by eliminating the use of the gyroscope.

A critical element to be also considered due to power cost is the microcontroller. The dsPIC30F processing unit consumes about 22 mA in normal operation, and together with the ZigBee module transmitting it consumes about 28 mA. The total power consumption of the system in its normal operation is of about 31 mA, but it can be reduced using communication interruptions and putting the micro controller in sleep mode during periods of inactivity. In the case of using continuous Bluetooth communication to a PDA, the total consumption increase until 65mA.

Another factor to take in consideration is the size of the prototype. One of the objectives to accomplish is to achieve a system the less annoying possible for the patient that has to wear it. So trying to miniaturize the components is also a priority. In the actual prototype, the build sensor board is sized about 30x30mm and the microcontroller board has approximately the same size.

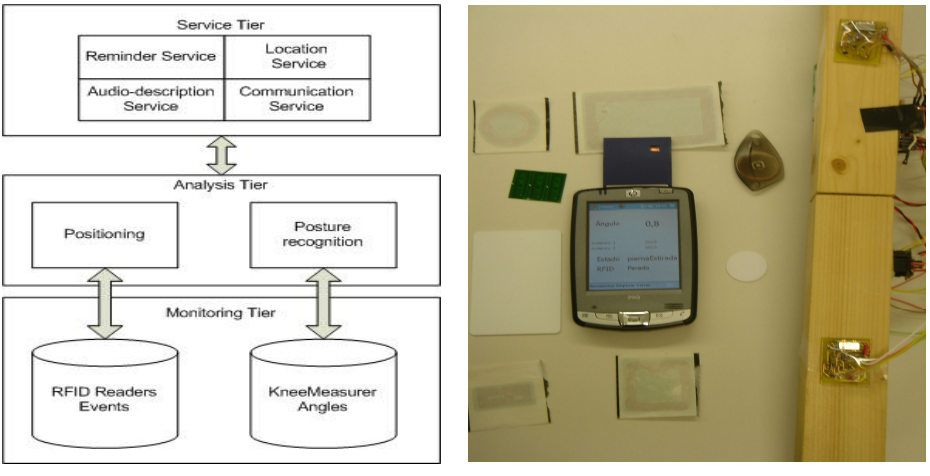


Fig. 1. Three-tier architecture and prototype image

4 Architecture

Three-tier architecture has been selected to deal with all the received and processed information. The first one is the Monitoring tier, in charge to store both, data collected from the body fixed sensors wearable motion monitoring device (the KneeMeasurer) and data coming from RFID readers. The second tier analyses the information stored on the first one and it interprets data for obtaining places associated to the tags identification and matching simple patterns like walking, sitting or lying down. Used technology has been described in the last Section and each service on the third tier will be now detailed.

4.1 Audio-Description Service

Audio description will be a main service for people with sight problems to fully inform about any kind of object tagged with a RFID tag. It is an additional narration that describes all significant information such as medicines composition or dosage or material, colour and description of clothes. The information for each tag will be stored in distributed databases, the system maintaining diverse databases at different levels. Figure 2 shows a hierarchical description of them.

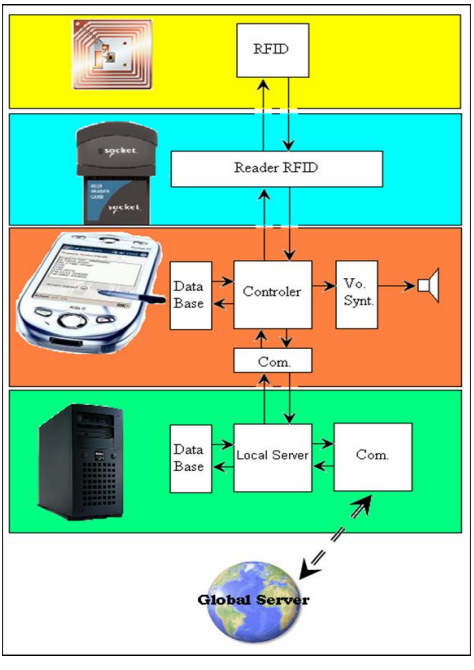


Fig. 2. Scheme of hierarchical databases

The first one is a small customized database in the device memory, keeping the personal user database. These data allow personal tags recorded by the user with his voice with no need to be linked to a remote server.

Whether the associated information to a tagged-product is not located on this PDA database, a second one is disposed on a local server, usually a PC-like machine at home. This machine will control the data "expiring" in the local database and will look for the data in Internet. Server will also incorporate data verification and renovation systems, maintaining reasonable size limits without loosing of speed. Fast processing is also a critical point for the system due to the time consumption on queries send through Internet.

Although a PDA can be connected by itself to Internet, it is being considered that this machine "bridge" becomes necessary because the power cost to connect a portable device, like a PDA, to Internet and to maintain a large size data base updated is excessive for the present batteries characteristics. On the other hand, it must be had in mind the present PDAs memory limitations, data base size and speed limits. In the future it cannot be discarded that the portable device can exert both functions.

Finally a Global server is considered, so that third, properly authorized, people can access to introduce global information on the product identifiers. This server will maintain the global data base to which the local servers will access, under automatic consultation, to update their databases.

This structure should be able to be integrated within standard ISO 18006, and more concretely, a desirable objective is to approach the system as much as possible to the EPCglobal recommendations, a standards precursory organization.

4.2 Location Service

Using the RFID infrastructure for the audio description service, a simple indoor location system was built. Although more accuracy would be desirable, the requirements for a reliable and continuous indoor position determination are not very wearable [9]. The constraints imposed over the service were hard, due to the kind of RFID readers (HF) and their range (5 cm). When the portable reader detect a position tag is possible to locate the person when he goes through the doors and when the reader detects an object is possible to associate the object with a place or an action.

Tags at shoulder-height were placed on both sides of doorframes on each door. When the user enters the room facing forward, the portable reader will reads the tag on the right of the doorframe so knowing the tag identification and the place located, deduction of position is done. Figure 3 shows how are placed the tags. A 5-seconds time window is waited to allow sequences of two tags read. A useful situation is when the user walks the corridor up or down.

For energy consumption optimality, tags reader is not always reading, otherwise batteries will be down in a short time. Reader tags will be active only when the KneeMeasurer is detecting activity movement of the user, i.e. whether the user is walking or he/she is actively changing the posture.

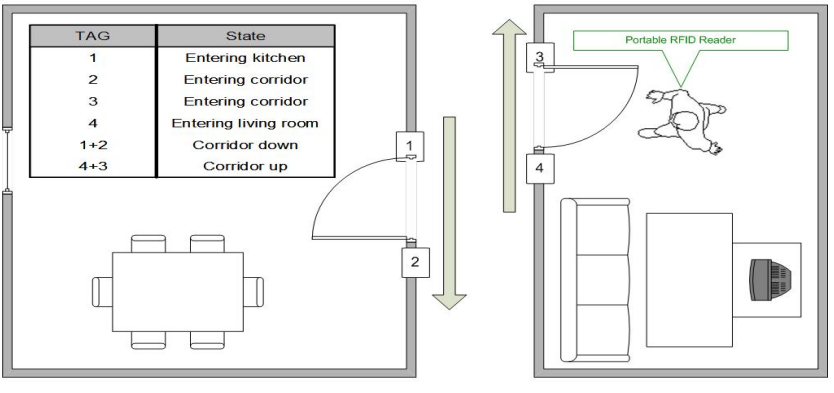


Fig. 3. Tags and static readers on strategic places

Typical approaches distribute different RFID readers on different places, such as the security readers at the doors of most major retail stores. A large infrastructure is needed compared to our approach where the RFID reader is located in a handled device carried by the user and tags are located in the frame of doors. This technique dramatically reduces costs when few people are tracked.

4.3 Reminder Service

This service is oriented to patients who must take some medicaments or/and do some physical exercise under medical supervision. To achieve this goal, a Web accessed database was implemented where all the data from different treatments and

medicaments were stored. Each user owns a prescription about exercises and medicines so reminders are programmed to be read during each day.

For the exercises, the PDA unit is reading via Bluetooth the KneeMeasurer motion monitoring system and periodically sends via WiFi the activity in work. Due to the variety of rehabilitation exercises, it can be distinguished a number of exercises in order to monitor them:

- Walk or run exercises: a fast frequency of changing angles is measured. When duration of this pattern matches with the proposed duration of the exercise the reminder will be deactivated.
- Exercises in a place. The location service is needed to know the change of room.
- Legs movement. For rehabilitation exercises, the movement of the knee wearing the sensors is monitored; the KneeMeasurer is useful so that the reminder would be deactivated.
- Legs static. In this case the reminder is always active.

The timetable for the rehabilitation exercises is softer than taking medicines and could be done during all the day. To avoid remind very late, the last hour to do it is 7:00 PM. If at this hour the recorded activity does not match with the duration and frequency of the exercise, the text that may be played is sent to the PDA and it reproduces using the text to speech technology. In order to make easy and simple the interfaces, speech recognition was avoided and the implemented one was tactile yes/no interface.

4.4 Communication Service

Each minute, when the information about read objects and, angles is sent to the server using Web Services, the server evaluates both, the information stored and retrieved. On the evaluation it takes account of the day and hour, the activity of the user, whether he/she is at home, the room where the objects are read by the RFID readers. If a risk pattern is detected, for example a large period of inactivity with a 'lied down' pattern and the room where the user is positioned has not any sofa or bed for sleeping, the server will send short messages to helper's mobiles.

5 Conclusions and Future Work

A technological platform for helping both, elder and handicapped people in a non-intrusive manner have been presented. The research in human motion monitoring and the use of RFID technology combined with text to speech techniques are merged on an assistive platform.

Main problems in the system are the position accuracy and the power consumption. The first one is being analysed in [8]. To overcome it, we are moving to use different biometric information. The second one is studied from the KneeMeasurer and the PDA energy consumption. The KneeMeasurer is currently handled using ZigBee communication and, furthermore, tests with the Wii-mote, having more sensors and actuators than our KneeMeasurer, show that its autonomy is about 50 hours.

From the PDA perspective, implementation under Windows Mobile OS, that does not stop until a hard reset is done, allows programming the reminder throw the system's clock such as the tasks reminder included on it. Hence, only the PDA clock is working and only programmed reminders activates all the processor.

A future work is to integrate sensors using compact flash or secure disk ports on the PDA discarding wireless payload.

Acknowledgments

This work has been partly granted by the projects FAMENET-InCare (TSI2006-13390-C02-02), EXODUS-ADA (DPI2006-15630-C02-01), and SAPIENS (IMSERSO 106/5 2005) from the Spanish Ministry of Education and Culture.

References

1. Chella, A., et al.: A System for Simultaneous People Tracking and Posture Recognition in the context of Human-Computer Interaction. In: International Conference Computer as a Tool, EUROCON 2005. vol. 2, pp. 991–994 (November 21-24, 2005)
2. Farella, E., et al.: A Wireless Body Area Sensor Network for Posture Detection. In: Proceedings. 11th IEEE Symposium on Computers and Communications, ISCC '06, June 26-29, 2006, pp. 454–459. IEEE Computer Society Press, Los Alamitos (2006)
3. Moore, S.T., et al.: Long-term monitoring of gait in Parkinson's disease. In: Science Direct, Gait & Posture, Elsevier, Amsterdam (2006)
4. Han, J., et al.: Gait detection from three dimensional acceleration signals of ankles for the patients with Parkinson's disease. In: The International Special Topic Conference on Information Technology in Biomedicine (ITAB 2006)
5. Yiping, T., et al.: Elder abnormal activity detection by data mining. In: SICE 2004 Annual Conference. vol. 1, August 4-6, 2004, vol. 1, pp. 837–840 (2004)
6. RFID Whitepaper: alliedworld.com, Allied Business Intelligence Inc. (2002)
7. Jackson, R.J.: Radio Frequency Identification (RFID), A White Paper available on: http://home.agh.edu.pl/~luc/my_doc/what_is_rfid.pdf
8. Raya, C., et al.: KneeMeasurer. A Wearable Interface for Joint Angle Measurements. In: drt4all Conference (April 2007)
9. Retscher, G.: Location Determination in Indoor Environments for Pedestrian Navigation. In: Position, Location, and Navigation Symposium, 2006 IEEE/ION, pp. 547–555 (April 25-27, 2006)

Geometrical Representation of Quantity Space and Its Application to Robot Motion Description

Honghai Liu[1], David J. Brown[1], and George M. Coghill[2]

[1] Institute of Industrial Research, The University of Portsmouth
Portsmouth PO1 3HE, England, UK
[2] Department of Computing Science, The University of Aberdeen
Aberdeen, AB24 3UE, Scotland, UK

Abstract. We are interested in the problem of intelligent connection of perception to action, i.e., the connection between numerical data and cognitive functions. In this paper we extend conventional quantity space into that in a geometric vector context and then propose quantity arithmetic for quantity vector computation in a normalized quantity space. An example of motion abstraction of a Puma robot is provided to demonstrate the effectiveness of the proposed method.

1 Introduction

Intelligent connection of perception to action [1] is an inherent problem to most research in the field of AI and robotics due to an intelligent robotic system comprises non-symbolic components and symbolic components. This problem involves many research topics, e.g., cognitive science, psychology and computing, it is a combination of several research problems, e.g., symbol grounding, pattern recognition and perceptual anchoring [2]. This problem is understood in our research context as the intelligent connection of non-symbolic level control & sensing task and symbolic-level cognitive functions. We are interested in the connection of objects and their behaviours; our aim is to generate a vector space of atomic symbols and functions to connect non-symbolic and symbolic modules. On the one hand, the space automatically communicates with numerical space without involvement of hand-coded techniques, e.g., joint space in robotics; on the other hand, the space provides atomic symbols and functions for inputs and outputs of symbolic systems. Domain knowledge and the proposed method have been separated by normalizing quantity spaces into a quantity vector space of normalized quantity spaces and a converting module vector. It should be noted that symbolic systems still are able to have their own functionalities and strategies for planning or reasoning tasks as long as their interaction modules to non-symbolic systems can be represented by the proposed elements.

The idea behind of the proposed vector-based approach is that encapsulating domain knowledge into a vector representation of quantity spaces, which allows us to focus on the mathematical nature of the connection problem. Research has proved that symbols and symbolic behaviours can of course be interpreted as meaning or standing for something, but the interpretation would clearly be derivative rather than intrinsic. For intelligent systems, a common solution is to introduce domain knowledge providing necessary context background.

B. Apolloni et al. (Eds.): KES 2007/ WIRN 2007, Part II, LNAI 4693, pp. 18–25, 2007.
© Springer-Verlag Berlin Heidelberg 2007

The concept of quantity space has sufficient ability to contribute to the problem solving [3]. A quantity-space is a set that has the property that a distance function is defined for any two elements in the set. The distance function is employed to represent a quantity value by ordinal relationships with specially chosen points in the space; intervals and their fuzzy versions; order of magnitude representations; finite algebras, among others. Possible values for a quantity can be defined in a quantity space, on the other hand, a specific value from its quantity space can be assigned to a quantity in model fragments and scenario's. Quantity space has been successful in the applications [4, 5]. For instance, Shen & Leitch [6] use a fuzzy quantity space, which allows for a more detailed description of the values of the variables. The extension principle and approximation principle of fuzzy sets support the calculations in terms of the fuzzy sets of the fuzzy quantity space. Obviously domain knowledge is able to categorize quantities into their corresponding quantity space. Besides, ordinal relationships of quantity elements in a quantity space allow easily to interact with other types data in both non-symbolic level and symbolic level. For instance, relative order of magnitude using fuzzy relations [7] has been applied to ambiguity problem solving in qualitative reasoning. Not only can it mechanize the commonsense reasoning of engineers simplifying complex equations and computing approximate solutions, but also it can be applied to provide a fuzzy finistic semantics to plausible reasoning with qualitative probabilities.

The rest of the paper is organized as follows. Section 2 proposes normalized quantity space which includes vector representation of quantity space, fuzzy number normalization and quantity arithmetic. Section 3 implements the proposed method into motion description of a PUMA robot. Section 4 concludes this paper with future work.

2 Normalized Quantity Space

Our understanding of a quantity space is a vector space $\mathbb{Q}^n$, a vector is given by n coordinates and can be specified as $(\mathbb{Q}_1, \mathbb{Q}_2, ..., \mathbb{Q}_n)$. For avoiding the confusion could be caused by Euclidean terms, a vector in quantity space is named as a quantity vector. Each coordinate is a conventional quantity space [3, 6]. An element of $\mathbb{Q}^n$ is written $\mathbb{Q} = (q_1,\ q_2,\ ..., q_n)$ which is a set of n elements of conventional quantity space, e.g., fuzzy numbers, intervals. The canonical basis for the n-dimensional quantity space is the basis obtained by taking the n basis vectors,

$$\begin{aligned} \mathbb{E}_1 &= \big(\mathbb{I}, \mathbb{O}, \cdots, \mathbb{O}\big) \\ \mathbb{E}_2 &= \big(\mathbb{O}, \mathbb{I}, \cdots, \mathbb{O}\big) \\ &\vdots \\ \mathbb{E}_n &= \big(\mathbb{O}, \mathbb{O}, \cdots, \mathbb{I}\big) \end{aligned} \tag{1}$$

where $\{\mathbb{E}_i : 1 \leq i \leq n\}$, $\mathbb{I}$ is a subset of $[a_1\ a_2\ \ldots\ a_n]$ for a_i=1; $\mathbb{O}$ is a subset of $[a_1\ a_2\ \ldots\ a_n]$ for b_i=0. Besides, the length, the number of entries of a $\mathbb{O}$, must be the same as that of the $\mathbb{I}$ in the same column in equation 1, this condition is for computation purpose.

Say $\mathbb{Q}$ is an arbitrary vector in $\mathbb{Q}^n$, which can be written in the form, $\mathbb{Q} = \sum_{i=1}^{n} q_i \mathbb{E}_i$. An element of a quantity space vector is a quantity value by ordinal relationships with

points in the vector space. It could be a set of intervals and fuzzy numbers, order of magnitude representations, finite algebras or among others. Let us denote a normalised coordinate vector as $\mathbb{P}$, equation 2 can be rewritten as,

$$\mathbb{Q} = \sum_{i=1}^{n} \lambda_i p_i \mathbb{E}_i = \Lambda \sum_{i=1}^{n} p_i \mathbb{E}_i = \mathbb{P}\Lambda \tag{2}$$

where p_i is the ith normalized element. λ_i is named as ith converting module, Λ is converting module vector. $\mathbb{P}$ is a normalized n-dimensional vector. Besides, we convert elements of a quantity space into those in terms of midpoint-radius representation. For literature on interval midpoint-radius (MR) representation, please see [8, 9]. Basically we map vectors in a n-dimensional space into those in its normalized quantity space in MR representation, which allows to handle different types of elements of quantity spaces in a standard platform. Due to each normalised quantity space provided with domain knowledge, we are able to map between numerical, semi-numerical data and their symbols. For instance, there are three ordinal symbols in a normalized quantity space used to describe speed context, it is easier to ground a interval value $[0.2,\ 0.1]$ to a suitable symbol.

Definition 1. Quantity vector $\mathbb{Q}$ is a quantity position vector if and only if the length of entries $\mathbb{I}$s in $\mathbb{E}_i$ for i=n are one. Otherwise, $\mathbb{Q}$ is a quantity patch vector.

It is obvious to understand a quantity position vector easier with Euclidean space background; a patch quantity vector is fundamentally important for supporting reasoning techniques. Patch vectors are able to take a set of adjacent entries of a quantity space into a reasoning system. The reason for introducing patch quantity vectors are two fold, it supports atomic elements-existing reasoning systems, e.g., [10]; on the other hand, it allows to carry out parallel propagation.

For a position quantity vector, its metric or distance function on $\mathbb{Q}^n$ can be defined using the norm as the following form,

$$d\left(\mathbb{Q}_i, \mathbb{Q}_j\right) = \left\|\mathbb{Q}_i - \mathbb{Q}_j\right\| = \sqrt{\sum_{i=1}^{n} \left(q_{ik} - q_{jk}\right)^2} \tag{3}$$

We have two metric measurements, i.e., patch metric and average metric, for a patch quantity vector, denote the length of a patch vector $\mathbb{Q}_i$ as m; $\mathbb{Q}_j$ as n. The patch metric is the set of the distance functions between entries in $\mathbb{Q}_i$ and $\mathbb{Q}_j$ using the metric in equation 3.

$$D\left(\mathbb{Q}_i, \mathbb{Q}_j\right) = \left\{d\left(\mathbb{Q}_i, \mathbb{Q}_j\right) | 1 \leq i \leq m, 1 \leq j \leq n\right\} \tag{4}$$

The average distance function is the average of its patch metric results, it is given below,

$$\bar{D}\left(\mathbb{Q}_i, \mathbb{Q}_j\right) = \frac{\sum_{j=1}^{n} \sum_{i=1}^{m} d\left(\mathbb{Q}_i, \mathbb{Q}_j\right)}{m+n} \tag{5}$$

Note that the MR representation separates metric measurement into two parts, midpoint metric and radius metric. The midpoint metric measures the propagation of the centre

of a quantity vector, the radius metric tells the propagation degree of a quantity vector drifting from its centre. The separation characteristic benefits not only propagation techniques but also aggregation techniques [11].

2.1 Fuzzy Numbers Normalization

There is a wide variety of fuzzy number types, for instance, standard fuzzy numbers, e.g., triangular, Gaussian fuzzy numbers, and non-standard fuzzy numbers. Normalizing fuzzy numbers has two stages, firstly, converting each fuzzy number into that in the middle-point radius (MR) representation terms; secondly, normalizing those in the MR representation terms from stage 1.

For the conversion of any fuzzy number to that in its MR representation, say A is a fuzzy set, $[A]^{\alpha}$ is a compact subset of R, and $M(A)$ is a closed interval bounded by the lower and upper possibilistic mean values $[M_*(A), M^*(A)]$ of A [12]. Then we can give the following definition,

Definition 2. *α-centre of a fuzzy set is its possibilistic mean value.*

$$C\left(A_{\alpha}\right) = \frac{M^{*}\left(A_{\alpha}\right) + M_{*}\left(A_{\alpha}\right)}{2} \tag{6}$$

where

$$M^{*}\left(A_{\alpha}\right) = \frac{\int_{0}^{1} Pos[A \geq b(\alpha)] \times \max[A]^{\alpha} d\alpha}{\int_{0}^{1} Pos[A \leq b(\alpha)] d\alpha}$$
$$M_{*}\left(A_{\alpha}\right) = \frac{\int_{0}^{1} Pos[A \leq a(\alpha)] \times \min[A]^{\alpha} d\alpha}{\int_{0}^{1} Pos[A \leq a(\alpha)] d\alpha}$$

and Pos denotes possibility.

Definition 3. *α power support of a fuzzy set is the proportion of its α-power to its α-height.*

$$PS\left(A_{\alpha}\right) = \begin{cases} \int_{R} \mu_{A}^{\alpha}(x)\, dx & \alpha = 1 \\ \frac{\int_{R} \mu_{A}^{\alpha}(x) dx}{1-\alpha} & \alpha \neq 1 \end{cases} \tag{7}$$

A fuzzy number A in a α-level can be described by its α-centre and α power support.

$$A_{\alpha} = \left[C\left(A_{\alpha}\right), PS\left(A_{\alpha}\right) \right] \tag{8}$$

Further by employing intervals normalization, we obtain,

$$\mathbb{P}_{i} = \left(p_{i1}, p_{i2}, \cdots p_{in} \right) \tag{9}$$

where its middle point p_{ij}^{m} and radius p_{ij}^{r} are,

$$p_{ij}^{m} = \frac{C_{ij}^{m}(A_{\alpha})}{C_{in}^{m}(A_{\alpha})}, q_{ij}^{r} = \frac{PS_{ij}^{r}(A_{\alpha})}{\sum_{k=1}^{n} PS_{ik}^{r}(A_{\alpha})}$$

and

$$\lambda_i = \left[C_{in}^m(A_\alpha), \sum_{k=1}^{n} PS_{ik}^r(A_\alpha) \right]$$

If the elements of a quantity space are four tuple fuzzy numbers, normalized quantity space in equation 9 is α-cut version of the normalized fuzzy quantity space given by Shen [6] in terms of the MR representation. We introduce quantity arithmetic to support vector propagation in a quantity space. The quantity arithmetic is based on the midpoint-radius representation. Denote two quantity vectors $\mathbb{P}_i(p_i^m,\, p_i^r)$ and $\mathbb{P}_j(p_j^m,\, p_j^r)$, the arithmetic basic operators are given in [13], The proposed quantity arithmetic can be used to generate quantity propagation. For simplicity and without loss of generality, let us consider vector propagation in a 3-dimensional quantity space. Point and patch vector propagation is two fold, on the one hand, a point vector can be propagated or transformed by either point vectors or patch vectors, on the other hand, a patch vector can be propagated by both two as well. We use point vector propagation to describe rigid transformation of quantity point vectors. Please refer to robotics textbook for transformation terminology in Euclidean space. Understanding geometrical motion of quantity vectors will benefit mathematical description of variable propagation for handling and transferring different types of parameters in reasoning systems. Understanding mathematical nature of reasoning methods, e.g., qualitative reasoning, fuzzy reasoning and probabilistic reasoning will speed up the development of automatic specification and verification tools. It will also help to overcome limitations like state explosion, undecidability, etc.

Say, a point quantity vector $\mathbb{P}(p_1,\, p_2,\, p_3)$ in equation 2, its transformation consists of translation transformations $\mathbb{P}_T$ and orientation transformations $\mathbb{P}_R$. The transformations can be defined by either numerical vectors or quantity vectors or both of them. There is no propagation but transforming a quantity vector to another numerical location if simply applying real numbers to the transformation matrices as we do in Euclidean space. Herein we will consider a translation matrix denoted by quantity vectors, an orientation matrix by either real numbers or quantity vectors. For instance, let us consider the following transformation,

$$\mathbb{P} = \mathbb{P}_R \mathbb{P}_T \mathbb{P}_0 \tag{10}$$

Set quantity space of the range $[-1,\ 1]$ as twenty-five elements, twenty-four intervals plus a zero element. Due to no orientation vectors defined yet in the framework of quantity space, first we give a numerical orientation angle $\pi/4$ around y axis. Its simulation results are shown in Fig. 1, the small rectangle is the initial position $\mathbb{P}_0$, another one is propagated position. Obviously, the propagated position can be mapped back to the quantity space, and ground them back to symbols if symbolic meaning has been provided. Then we replace the numerical angle with qualitative angles from fuzzy qualitative trigonometry [14], and set a full orientation circle as 16 qualitative states. Fig.2 shows the result with the second qualitative state as a rotation angle. The initial position remain the same, the clustering circles are the centres of the propagated quantity positions in midpoint-radius representation. It shows fuzzy qualitative trigonometry makes significant contribution to the propagation of the quantity position. It indicates that the propagation will grow if the employed numerical data and quantity vectors are replaced by patch vectors. The propagated results include all transformed possibilities. Using

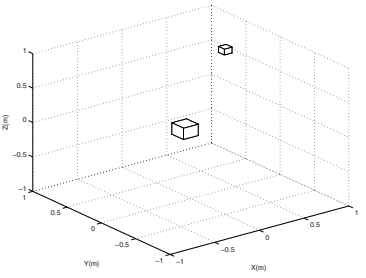


Fig. 1. Point vector propagation with an numerical angle and quantity translation vector

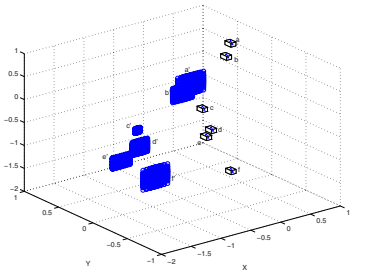


Fig. 2. An example for quantity non-point vector propagation

quantity point vector propagation, quantity vector propagation can be extended to linear and non-linear quantity trajectory vector generation which can be used to adapt if-then statements in mathematically qualitative description. Any data type of quantitative or qualitative functions can be described in terms of quantity vector propagation. For instance, the propagated intervals of a sequence of intervals a, b, c, d, e are shown in Fig.2 as a', b', c', d', e'.

3 An Example of Robot Motion Description

In order to demonstrate robot motion description, we have implemented the proposed method into a PUMA 560 robot. It is common in behaviour-related research to assume that robots have rigid bodies and to consider the joint angles as variables only. Hence we denote the joint angles as fuzzy qualitative variables and the other parameters for the PUMA 560 as numeric, i.e. a_2=432mm, a_3=20mm, d_3=124mm and d_4=432mm. The orientation parameter was set as p as 60, the translation parameter q as 40 and the robotic configuration was set as $QS_a(\theta) = [[0.007, 0.015]\,[0.391, 0.017]\,[0.024, 0.016]]$, $QS_a\left(\dot{\theta}\right) = [[2, 0.12]\,[5, 0.14]\,[3, 0.12]]$. The center of the PUMA robot's tool frame is calculated as,

$$QS_d(\mathbf{p}_{tcp}) = \begin{bmatrix} C\theta_1\left[a_2C\theta_2 + a_3C\theta_{2+3} - d_4S\theta_{2+3}\right] - d_3S\theta_1 \\ S\theta_1\left[a_2C\theta_2 + a_3C\theta_{2+3} - d_4S\theta_{2+3}\right] + d_3C\theta_1 \\ -a_3S\theta_{2+3} - a_2S\theta_2 - d_4C\theta_{2+3} \\ I \end{bmatrix} \tag{11}$$

Table 1. A trajectory of qualitative states for a PUMA 560 robot

$QS\,No$	1	2	3	4	5	6
$QS(\theta_1)$	1	4	2	8	15	18
$QS(\theta_2)$	24	30	38	42	36	50
$QS(\theta_3)$	2	7	12	12	18	25

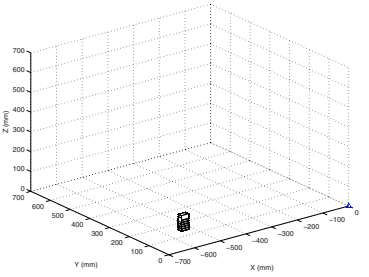


Fig. 3. A fuzzy qualitative position for a PUMA robot

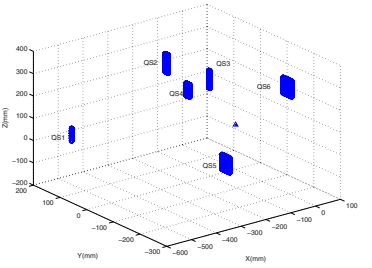


Fig. 4. Fuzzy qualitative positions for a PUMA robot

where θ_{2+3} stands for $QS_a(\theta_2) + QS_a(\theta_3)$.

The simulation results of fuzzy qualitative positions are shown in Figs 4-7. The position of the initial robotic configuration is given in Fig. 4 in the interval terms. The trajectory of qualitative states in Table 1 is given in Fig. 5 using middle-point representation. It shows each qualitative position is composed of the centres of overlapped intervals which are labelled using circles. The figure also shows that the qualitative positions do not have the same interval regions, indicating that each qualitative position has different position behaviours. In order to extract a suitable behaviour from the qualitative trajectory, we employ max, min and median aggregation operators. Positions labelled as '□' are the trajectory points generated by the maximum operator, positions labelled as '△' are generated by the minimum operator and positions labelled as '⋆' are generated by the median operator. Fig. 6 shows the behaviours of the line-up centre points of aggregated qualitative values based on all qualitative states; Fig. 7 shows the behaviours of the line-up centre points of aggregated qualitative values based on boundary qualitative states. Each group of positions are lined up to show the trajectory sequence. Suitable interpolation techniques can be applied to generate smooth trajectories for non-symbolic modules such as motion control modules.

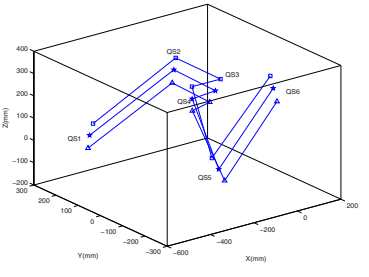


Fig. 5. Motion behaviour description using all qualitative states

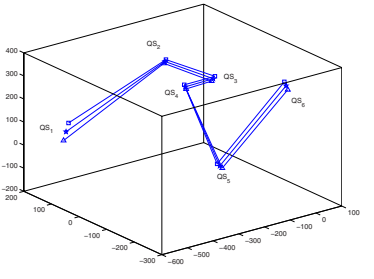


Fig. 6. Motion behaviour description using boundary qualitative states

4 Concluding Remarks

We have proposed a vector-based approach for geometric representation of quantity vector and have applied it to robot motion description. The proposed method has been implemented into our MATLAB toolbox XTRIG. An example based on a PUMA robot has been provided to demonstrate robot motion abstraction. Future work will be focused on improvement of effectiveness in describing behaviours of intelligent robotic systems. Apart from the proposed interval computation and aggregation operators, other relevant methods should be investigated to see the possibility of replacing the role of the quantity computation and aggregation operators.

References

1. Brady, M.: Artificial intelligence and robotics. Artificial intelligence 26, 79–121 (1985)
2. Coradeschi, S., Saffiotti, A.: Journal of robotics and autonomous systems, special issue on perceptual anchoring (2003)
3. Forbus, K.: Qualitative process theory. Artificial Intelligence 24, 85–168 (1984)
4. Kuipers, B.: Qualitative reasoning. MIT Press, Cambridge (1994)
5. Coghill, G.M.: Mycroft: A Framework for Constraint-Based Fuzzy Qualitative Reasoning. PhD thesis, Heriot-Watt University (1996)
6. Shen, Q., Leitch, R.: Fuzzy qualitative simulation. IEEE Transactions on Systems, Man, and Cybernetics 23(4), 1038–1061 (1993)
7. Ali, A., Dubois, D., Prade, H.: Qualitative reasoning based on fuzzy relative orders of magnitude. IEEE transactions on Fuzzy Systems 11(1), 9–23 (2003)
8. Neumaier, A.: Interval methods for systems of equations. In: Encyclopedia of mathematics and its applications, Cambridge University Press, Cambridge (1990)
9. Rump, S.: Fast and parallel interval arithmetic. BIT 39(3), 534–554 (1999)
10. Ohlbach, H., Koehler, J.: Modal logics, description logics adn arithmetic. Artificial Intelligence 109(1-2), 1–31 (1999)
11. Yager, R., Rybalob, A.: Full reinforcement operators in aggregation techniques. IEEE Transactions on Systems, Man and Cybernetics 28, 757–769 (1998)
12. Carlsson, C., Fuller, R.: On possibilistic mean value and variance of fuzzy numbers. Fuzzy Sets and Systems 122, 315–326 (2001)
13. Liu, H., Brown, D.: An extension to fuzzy qualitative trigonometry and its application to robot kinematics. In: Proc. IEEE International Conference on Fuzzy Systems, Canada, pp. 1111–1118 (2006)
14. Liu, H., Coghill, G.: Fuzzy qualitative trigonometry. In: Proc. IEEE International Conference on Systems, Man and Cybernetics, Hawaii, USA (2005)

Modeling Smart Homes for Prediction Algorithms

A. Fernández-Montes, J.A. Álvarez, J.A. Ortega, M.D. Cruz, L. González, and F. Velasco

Departamento de lenguajes y sistemas informáticos. ETSI Informática Avda. Reina Mercedes s/n, Sevilla
alejandro.fdez@gmail.com, jaalvarez@us.es, ortega@lsi.us.es, manuel3086@gmail.com, velasco@us.es, luisgon@us.es

Abstract. This paper reviews the goals of the Domoweb project and the solutions adopted to achieve them. As a result we enjoy a great support to develop smart home techniques and solutions. As a consequence of the acquired experiences a Smart home model is proposed as a division of four main categories. In relation with the smart home model, we show the essential features a smart environment prediction algorithm should satisfy and a procedure to select relevant information from the model to achieve artificial intelligence based solutions.

1 Introduction

Smart home technologies are often included as a part of ubiquitous computing. Mark Weiser [1] outlined some principles to describe Ubiquitous Computing (ubicomp) from which we emphasize that the purpose of a computer is to help you do something else.

Home technologies have tried to help home inhabitants since its creation. Nowadays, due to the popularization of computational devices, ubiquitous computing is called to be the revolution to develop smart systems with artificial intelligence techniques.

Domoweb [2] is a research project which was originally developed as a residential gateway implementation over the OSGi (Open Services Gateway Initiative) service platform. Domoweb implements the standard services any residential gateway must have like http server, web interfaces, device manager, user manager and other basic services.

Nowadays Domoweb conform a great platform where researchers from different disciplines converges and where we can deploy, develop and test smart home related solutions, due to the component based model, and the service oriented architecture that Domoweb and OSGi supports.

This article focuses on modeling smart homes and the features the prediction algorithms should implement. A good smart environment model must represent the

B. Apolloni et al. (Eds.): KES 2007/ WIRN 2007, Part II, LNAI 4693, pp. 26 – 33, 2007.
© Springer-Verlag Berlin Heidelberg 2007

properties, states, attributes and any other characteristic that could be useful in building smart environments solutions as is proposed in section 2.

Artificial intelligent methods can be supported by this model like prediction algorithms where is centered this article. The features these algorithms must be aware of are defined in section 3 and finally a procedure to discriminate significant information is released in section 4.

2 Smart Home Model

Artificial intelligence algorithms need a solid base of knowledge to work. This fact demands us a great effort for building a model of the smart home and its environment. Other projects have helped us to compose the model [3-5] which has been arranged in four main categories explained in next sections and expounded in table 1.

2.1 Device Related

This category is the most obvious one so is related with the main elements in a smart home environment. Ambient intelligence algorithms should be aware of next main fields:

- **Status.** Algorithms must know the current states of devices installed on the smart home. Obviously this is essential for these algorithms, and one of the most important domains to build future predictions.
- **Location.** Devices usually occupy a location for a long time and this location may be useful for ambient intelligence algorithms. The model must be able to consider non-still devices as well, like motorized cleaner robots and others.

2.2 Inhabitants Related

Smart home algorithms must be aware of the inhabitants' status to offer appropriate predictions for any user or for the whole group of inhabitants. On this line we discuss some of the necessary fields to infer inhabitants-aware predictions:

- **Personal data.** This field includes all the data concerning to a particular person like name, age, sex and so on.
- **Location.** Inhabitants can move over the home rooms, so smart home systems have to know where each inhabitant is, and should be able to identify them.
- **Physical state.** This field is related with the illness or injuries that an inhabitant can suffer during his life. Smart home technologies must adapt to this situations and offer appropriate replies.
- **Mental state.** The state of mind of a person can be defined as a temporary psychological state. A depressed inhabitant behavior usually differs from a euphoric one, so smart home must be consistent with these circumstances.

Table 1. Smart Home Model

CATEGORIES	FIELDS	DESCRIPTION	EXAMPLE
DEVICE RELATED	Status	Current state the devices are.	Sensor temperature is at 25ºC
	Location	Where the devices are.	Cleaner robot is in Living room.
INHABITANTS RELATED	Personal data	Name, Age, sex.	Diane is 45.
	Location	Where the inhabitants are.	Mark is at bedroom.
	Physical state	Illness, injuries, and others states inhabitants are.	Roy has a cold.
	Mental state	Psychological state the inhabitants are.	David is depressed.
ENVIRONMENT RELATED	Date, time	Temporary information	Current time is 13:36.
	Environmental conditions	the phenomena that currently occur in the atmosphere	It's rainy
HOME BACKGROUND	Inert entities location	Where these entities are.	Sofa is at living room.
	Home limits properties	The properties of the home structure and limits.	Bedroom window opacity is 70%

2.3 Environment Related

This category probably is the most diffused due to it covers heterogeneous and difficult-to-limit fields as we discuss in the following list:

- **Date, time, season.** Obviously smart home behavior should be different in different temporal conditions and it may depend in these temporal factors as the air condition policies will differ between summer and winter.
- **Environmental conditions.** This field comprises current environmental conditions (sunny, cloudy, rainy and others). Smart home should make a request for a weather forecast as well, which could be significant to assess future decisions.

2.4 Home Background

This category must comprise all the relevant things concerned to inert entities and its properties or qualities. This could be the less relevant category discussed, but anyway could be significant in some concrete applications. We propose a couple of fields related in next listing:

- **Furniture location and position.** Furniture occupies space at home and can be moved. Location (room where the furniture is) and position (place where the furniture is placed in a room) should be known by the smart home systems due to it could be useful by concrete applications like robot movement related algorithms, or presence detection related algorithms.
- **Home limits properties.** The texture of a floor, the color of a wall, or the opacity of the windows could be significant in several cases such as temperature adjustment applications.

3 Features of Prediction Algorithms

In this section we present the features that a smart home system must implement, specially related with the prediction algorithms the systems may have. The article doesn't focuses on artificial intelligence techniques like the studies of [5-10] do, but on what are the most important and indispensable features that must be considered to develop prediction algorithms. Much of the ideas presented below could be useful to implement others smart home algorithms.

3.1 Prediction Supported by Last Events and States

Prediction algorithms should consider as input data two main aspects. First it should analyze the last events occurred in the home's performance field which have changed the home status in any manner.

Second it should analyze current state and previous states as well. This way the algorithms should determine last changes in the home status, and which events have been involved in these changes.

In section 5 we discuss further about these aspects and propose a way to consider states and events.

3.2 Predictable by the Inhabitants

Smart homes should learn inhabitant behaviors and habits, and build predictions, but it has to be predictable by the inhabitants to achieve no unexpected behaviors.

3.3 Understandable Decisions

Prediction algorithms have to offer an explanation of their predictions and/or actions. This way the inhabitants will be more trusted with the decisions the system took. This will improve the user acceptance of the smart home predictions.

3.4 Wrong Decisions Detection and Related Improvement

If the smart home executes a wrong decision it should be aware of this failure and be able to learn from its errors. A possible scenario should be when someone arrives home at night and the system orders the hall lights to switch on, but immediately the user performs the opposite action (which should be switch off the lights). The system has to notice this failure and extract some knowledge from this experience to face future similar situations with guarantees of success.

3.5 Anomalies Detection

In some scenarios, the smart home should consider that a wrong decision executed as a consequence of a wrong prediction could be produced due to an anomaly. We can consider this scenario, an inhabitant wake up all working days at 7.00 a.m., so smart home switch on the coffee-maker some minutes before. But when the wake-up alarm goes off, the smart home detects no movement so it could be desirable to request an emergency service with a standard phone call or other mechanism.

3.6 Quick Response When Required

Some situations require the smart home prediction algorithms to return a response with time limits. These algorithms are responsible for detection of this situations and they must be able to adapt to these circumstances in order to provide a quick response.

3.7 General Policies and User Adaptation

Smart homes technologies should implement the mechanisms to support some home policies like security, energy or comfort. These policies can be collected in different levels. The inhabitants could define some general policies which could be customizable by concrete inhabitant preferences. We can discuss the following scenario; inhabitant A gives preference to energy consumption over comfort. To satisfy these preferences smart home system should try to minimize the energy

consumption produced as a result of its predictions. On the other hand, inhabitant B gives preference to comfort over energy consumption, so algorithm has to adapt to this user preferences policy.

This way if the prediction algorithm determines that inhabitant A arrives home, it will switch on air conditioner system only at the arrival of the inhabitant, however if the inhabitant B is going to arrive home, the smart home should switch on the air conditioner system sometime before the arrival of the inhabitant.

4 Window of States

Smart home prediction algorithms usually make use of last states and events occurred at home environment. A state can be defined as the whole set of pairs field-value according to the model presented in section 3. We could also distinguish subsets of these states for each category so we shape the related sub-states. This way we can consider sub-states to include devices related, inhabitants related, environment related and background related information.

Events can be defined as something that happen at a given place and time. In smart home contexts, we can add to this definition that the event must cause a state change of the smart home. Events that do not cause a state change shouldn't be considered by smart home algorithms as something significant.

When building a smart home solution we can discuss what states and events should be considered. To resolve this situation we propose the use of a window which envelops the states (and events) that are going to take part in the giving response as shown in figure 2.

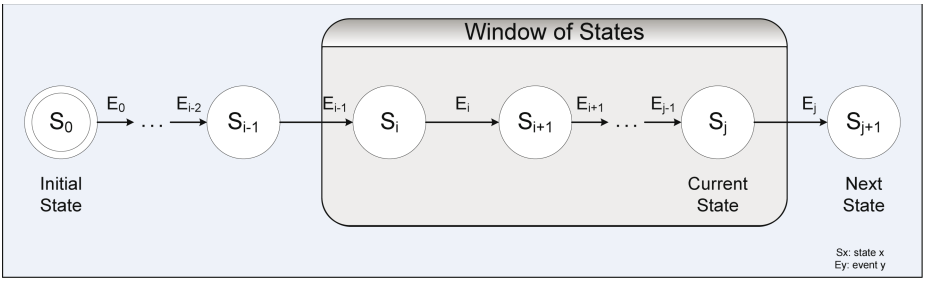


Fig. 1. Window of states

States are represented as circumferences labeled from S_0 to S_{j+1}, transitions between states as arrows labeled with the event which caused the state change.

Reader should notice that states from S_0 (initial state) to S_{j-1} are past time states, S_j is the current state and S_{j+1} is the next state which is currently unknown.

The size of the window $\boldsymbol{n}$ equals to the number of states and events to be processed. This size is deduced from the equation $\boldsymbol{n = j-i+1}$.

4.1 Dynamic vs. Static Window

Two approaches can be considered related with the size of the window. Algorithms designer could prefer the use of a static and predefined size of this window to build smart predictions. This approach can be useful in non-complex solutions, so it's easy to implement. The size could be estimated by means of empirical techniques and the own designer experience.

However dynamic window size is a more powerful approach for determining the size in real time. Algorithms could extend and reduce the size of the window depending on the requisites of current application like time of processing, hardware capabilities, solution accuracy and others.

5 Conclusions and Future Work

The previous sections have discussed some basis related with smart environments methods. A first model approach have been proposed as a start point to develop successful solutions, and prediction algorithms features form a guideline to implement these algorithms.

Prediction algorithms promise to be applicable to many areas within the smart home and smart environments so future work should be centered in the study of the methods and artificial intelligence techniques for prediction as in the improvement of the model definition.

Acknowledgements. This research is supported by the MEC I+D project InCare. Ref: TSI2006-13390-C02-02.

References

[1] Weiser, M.: The computer for the 21st century. Scientific American 265(3), 66–75 (1991)
[2] Álvarez, J.A., Cruz, M.D., Fernández, A., Ortega, J.A., Torres, J.: Experiencias en entornos de computación ubica mediante arquitecturas orientadas a servicios. In: JSWEB 2005 (W3C) (2005)
[3] Cook, D.J., Youngblood, M., Das, S.K.: A Multi-agent Approach to Controlling a Smart Environment. Designing Smart Homes, 165–182 (2006)
[4] Li, J., Bu, Y., Chen, S., Tao, X., Lu, J.: FollowMe: On Research of Pluggable Infrastructure for Context-Awareness. In: Proceedings of the 20th International Conference on Advanced Information Networking and Applications, IEEE, Los Alamitos (2006)
[5] Das, S.K., Cook, D.J.: Designing Smart Environments: A Paradigm Based on Learning and Prediction. In: Pal, S.K., Bandyopadhyay, S., Biswas, S. (eds.) PReMI 2005. LNCS, vol. 3776, Springer, Heidelberg (2005)
[6] Leake, D., Maguitman, A., Reichherzer, T.: Cases, Context, and Comfort: Opportunities for Case-Based Reasoning in Smart Homes. Designing Smart Homes, 109–131 (2006)
[7] Hagras, H., Callaghan, V., Colley, M., Clarke, G., Pounds-Cornish, A., Duman, H.: Creating an ambient-intelligence environment using embedded agents. IEEE Intelligent Systems 19 (2004)

[8] Choi, J., Shin, D., Shin, D.: Research on Design and Implementation of the Artificial Intelligence Agent for Smart Home Based on Support Vector Machine. In: Wang, L., Chen, K., Ong, Y.S. (eds.) ICNC 2005. LNCS, vol. 3610, Springer, Heidelberg (2005)
[9] Roy, N., Roy, A., Das, S.K.: Context-Aware Resource Management in Multi-Inhabitant Smart Homes:A Nash H-Learning based Approach. In: IEEE PERCOM'06 (2006)
[10] Yamazaki, T.: Beyond the Smart Home. In: ICHIT'06 (2006)
[11] Jiang, L., Llu, D.-Y., Yang, B.: Smart home research. In: Proceedings of the Third International Conference on Machine Learning and Cybernetics, Shanghai, pp. 26–29 (2004)

Qualitative Correspondence for Object Tracking Using Dynamic Panorama

Farshad Fahimi, Honghai Liu, and David J. Brown

Institute of Industrial Research, The University of Portsmouth
Portsmouth PO1 3HE, England, UK

Abstract. We propose a novel method to identify the correspondence of objects in different spaces using panorama generation and qualitative reasoning, the spaces are namely image space, camera fuzzy qualitative space and real world space. The correspondence is carried out in a three-layered image understanding framework. The first layer consists of single cameras which is to extract quantitative meausres using off-the-shelf image algorithms with a target of providing local feature information; The second layer targets at fusing qualitative information of single cameras at the level of cameras network; The third layer is intended to generate semantic description of object behaviours using nature language generation. This paper is focused on qualitative correspondence of objects in the first layer, which is realized by a two-stage tracking cycle consisting of panorama generation and object tracking. A case study is given to demonstrate the effectiveness of the method.

1 Introduction

Until very recently the attention of relevent research communities and the public have been drawn to surveillance studies and its industrial applications specially due to the recent tragic attacks such as 911 event [1, 2]. The current issues in security imaging can be summarized as follows: firstly, cooperation between multiple sensors is not in efficient manner; secondly, is the reliance on intense human monitoring for anomaly detection; finally is that image understanding is considerably far from the advanced stage to be accomplished.

To grasp the surveillance system foundation it is a good practice to break it into building blocks. One common classification is single or multiple camera systems. Single camera surveillance is the fundamental challenge in automated monitoring systems [2, 3, 1]. However, no single camera is omnipotent and systems often employ a network of cameras [4, 5]. Each camera is designed to overcome inadequacies of others. Complementary information from these elements can be used to determine image sequence charactristics which otherwise were inadequate, inaccurate or not easily extractable for image understanding application. On the other hand, multiple camera networks can be created out of common, inexpensive elements. Information and image fusion is the critical point multisensory systems. The fusion algorithms selected based on many parameters as well as charactristics of the fundamental elements in the network. Selection of the algorithm is connected to the application needs and the scenario.

B. Apolloni et al. (Eds.): KES 2007/ WIRN 2007, Part II, LNAI 4693, pp. 34–41, 2007.
© Springer-Verlag Berlin Heidelberg 2007

In this paper our focus is on object correspondence in human surveillance of an indoor environment where design plays a key role. The reason being is two-fold. Firstly fusion algorithm heavily depends on the architecture and secondly several algorithm's inefficiencies can be tackled with the right design. Two common network architectures are competitive(Non-overlaping) and cooperative(Overlaping) modes. Our research take advantage of a network of PTZ(Pan, Tilt and Zoom) installed in cooperative mode. At last research presented in this paper is our attempts to solve the qualitative correspondence of object tracking in different spaces. It includes dynamic panorama creation and conventional motion tracking in multiple camera environment with a focus on single camera layer of our proposed framework.

The remainder of this paper is organized as follows: Section 2 presents the three-layered image understanding framework with the highlights of the motivation of this work and differences from other representations. Section 3 a case study on single camera scenarios is proivided. Section 4 demonstrates our experimental results on a single camera. Finally, we conclude this paper with remarks and future work.

2 Three Layered Framework

The proposed framework consists of three layers shown in Fig. 1, namely Single camera layer, Camera network layer and semantic generation Layer. Single camera layer targets at generating feature in formation of each camera by employing off-the-shelf image/video processing algorithms, it includes background substraction, motion feacture extraction, etc. The second layer reads the updated information from the specified memory of the corresponding layer, and fuses them for subsequent inference at the level of camera network. Finally third layer provides the ability of generating semantic explanation using natural language processing. Novelty from other system architectures is the manner in which layers are defined. Most of the components specifically in the first layer are conventional parts of every surveillance system.

2.1 Layers Definitions

Single Camera Layer. Fig. 1 demonstrates a component view of the proposed framework. The Single Camera Layer (Processing and fusion in image Level) is the lowest layer of the framework. The input for this layer is the video stream from each camera fed into the system. Internal components are based on image and video processing functions. Output of this layer is Object Parameters (OP) and Motion Parameters (MP).

OPs are the parameters related to object in a single view. Features like edges and lines, points on boundaries, invariant features or parameters that can uniquely be used in the identification tasks. These as well can be the parameters related to the human body, human body parts or vehicles in the scene. Motion Parameters are parameters related to the motion of objects. These parameters are connected to position feature of the target.

These Parameters act as output for higher layers. The best representation for parameters in physical world is packets communicated between layers. These parameters should then be carefully chosen according to their effect on the system functionality in

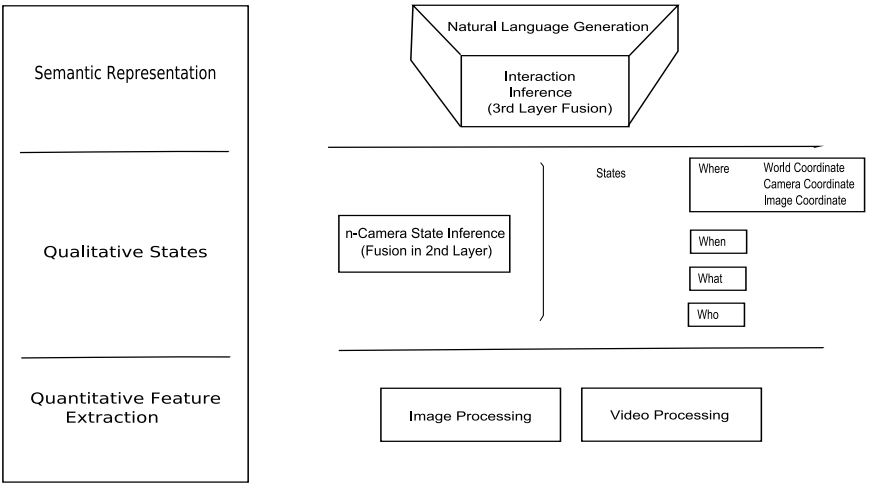


Fig. 1. The Three Layered Framework

real-time. For this purpose, as mentioned earlier, avoiding communication of unnecessary information is of utmost importance because firstly it keeps the system from having a bottleneck between layers, secondly the internal communication between layers is not affected and hirdly because of feedback and forward links in the system we will avoid any kind of delay or ignore losing of data in short periods.

Multiple Camera Layer. As stated the input of this layer is OPs and MPs from many single cameras. To explain, the components are seen in relation to one of the classes in Haritaoglu [1] in a bounding box, please see Fig. 1. For a high-level understanding of the scene there are several questions to be answered. High-level refers to anything that can be explained qualitatively. In this framework the combination of methods with some improved features that are suitable for the application is considered. Some Firstly the communications and representation of the medium layer is in fuzzy qualitative state measures [6]. Secondly in the future work a qualitative approach is proposed for 3D Geolocation, Homography and Calibration and as a result, communication between layers will be in this format.

The result of basic object classification is then passed to the higher and lower layers. In the higher layer it is used as a symbolic representation. Parameters passed to the lower layer works as a feedback which makes tracking tasks easier and body part classification, head tracking and other algorithms more consistent.

Semantic Layer. After qualitative representation of data is completed, states are passed to the higher layer. This layer interprets state sequences and state relations, finds the fittest corresponding scenario and generates a semantic representation for them. Qualitative state definition and state sequence understanding plays an important role in the process. 2.a).

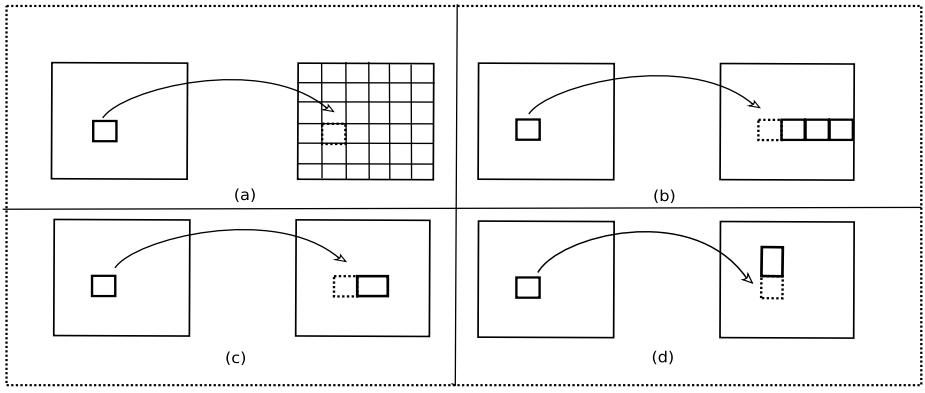


Fig. 2. SAD algorithm with and without filter. (a) Basic SAD, (b) SAD with ordinal qualitative values, (c,d) SAD with qualitative states.

3 Case Study

Human surveillance can be seen as a feature extraction, representation and classification problem. Since one of the most important type of these features are position related parameters, that is the focus of this section. Position related parameters are defined as any parameters that can be calculated having the position in a time period. Velocity, acceleration and length are some examples of these parameters. After extracting the positions they should be represented mathematically. The original features and inherited information are then used for further inference.

Since this paper is focused on human surveillance the first step is to reduce the search space through the image sequence and extract the human features. First was simple background subtraction to detect motion in the video. Skin colour detection is used for further reduction and filtering. The principles of background subtraction is to take the static background and subtract it from image sequence at time samples to find the motion in the image. This method could be considered inaccurate and therefore a number of alternative methods have been proposed. One problem for these algorithms is that to use them with an active camera an extension is required. This is due to the fact that the background is static in the basic definition of the background subtraction algorithms. One proposed solution is to generate panorama and extract the right background at each time instant. Subsequent to extraction of the area of interest, the features can be extracted. Position features can be features relating to the target. Centroid, bounding box and blob features are examples of the features. Calibration is the traditional method to find the position of a target in the image. Our method presented in the subsequent sections is based approximate position of cameras and a prior knowledge of the environment.

3.1 Qualitative States and Their Relations

Three spaces exist in common computer vision shown in Fig. 3 are image space, fuzzy qualitative space and real world space. the states in these three spaces are names as

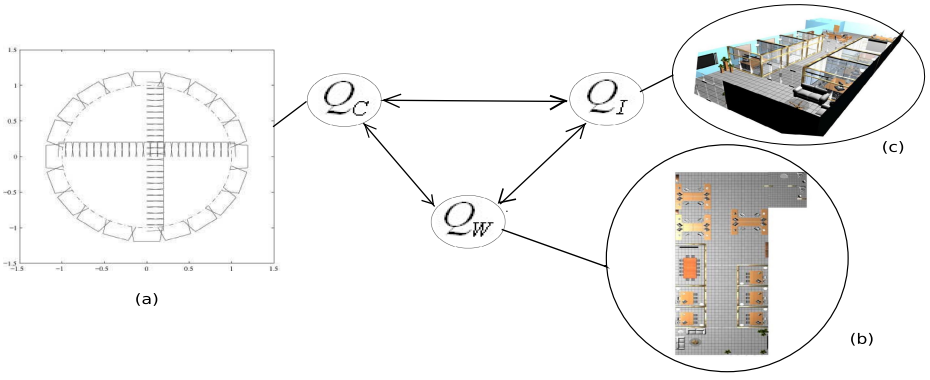


Fig. 3. 3 State Spaces (a) Fuzzy camera states demonstrated in a unit circle (b) World states presented in a map of the environment (c) Image state

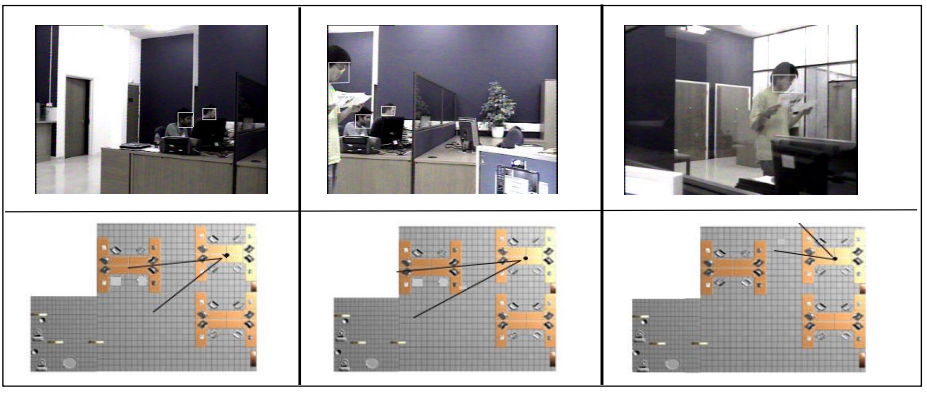


Fig. 4. (Top): Features are extracted and tracked, (Bottom): The correspondent world states are shown on the map

image states, camera states and world states. These states are bound together, and from two of the three it is possible to calculate the other one. World, Camera and Image States can be defined based on application needs. There are numerous advantage to this type of representation which will be discussed through out the text. where Q_c is the qualitative state of the camera. Where ϕ_m and θ_m stands for maximum rotation degree. The following unit circle (Fig. 3) is a demonstration of these states for camera angles. In our first step these steps are only conidered as regions which can be generalized to the fuzzy format at later stage. The regions are again constant and equal and can be selected based on required resolutions. Q_I is the qualitative image regions. The definition is similar to that of Q_c. Lastly Q_w or world states shows the positions in qualitative measures.



Fig. 5. Sample of generated panorama

Algorithm 1. General View of Panorama Creation and Extraction

Initialize *states* Q_c, Q_i and Q_w and Probability Correspondence Matrix
Record the background
repeat
  if $Q_c^{t+1} \neq Q_c^t$ **then**
    Do(Panorama Extraction) : Extract the right window
    Do(Panorama Generation) : Generate and/or update the panorama
  else
    Do SAD *(Only inside a landmark)* to calculate small shifts
    if $\mathrm{Image}^{t+1} \cong \mathrm{Image}^{t}$ **then**
      Update background for that state
    else
      Update panorama cosidering the small shift
    end if
  end if
until (STOP SIMULATION)

Algorithm 2. Panorama Generation

if $Q_c\,(t) \rightarrow Q_c\,(t+1)$ *: Camera state transition due to movement* **then**
  ($Q_i\,(t) \rightarrow Q_i\,(t+1)$ *: Image state transition*)
end if
$\Delta Q = Q_i\,(t+1) - Q_i\,(t)$: *Reduces the search space for SAD*
Shift is calculated and panorama is generated

In this simple scenario a map of the environment is used for visualization purposes. Camera states are collected from the shaft encoder which are coupled with the motor. Image states are infered from the extracted features in the image sequence. As mentioned having these two states an approximation of world states can be calculated. The positions estimated from a single camera are not identical. The previously defined states aligned on the map shows that the estimated world states can be narrowed down. Algorithms developed has been given in algorithms 1-4, an example of generated panorama is given in Fig. 5.

Algorithm 3. Panorama Extraction

1: Request current Q_c
2: Find correspondence Q_i in image database
3: Apply SAD for last stage accuracy (In window extraction)

Algorithm 4. People Tracking

1: Request current Q_c
2: Find correspondence Q_i in image database
3: Apply SAD for last stage accuracy (In window extraction)
4: Do(Background Subtraction)
5: DO(Skin Colour Detection)
6: Record the Q_c and Q_i states for the target
7: Comparing to probability matrix of the environment find the most possible position
8: Record position

4 Results

In this section firstly the primary steps taken to extract the features are explained and secondly the correspondences in conjunction to maps are described. Panorama creation is the first step to background subtraction in dynamic camera systems. The reason we used panorama is the dynamic background in active camera scenarios. Simple panorama creation algorithms join the images based on the calculated shift between the image sequence. One popular method to calculate the shift is to find the correspondences in the images and subtract the positions to find the estimated shift. The disadvantages to this method is that firstly one correspondence in the images does not provide us with a good approximation of the shift, which affects the panorama creation and consequently feature extraction. Secondly to searching for correspondences is a slow process. For more accurate estimation multiple search for correspondences is a solution which does not allow the algorithms to work as fast (Fig.

The proposed method to generate panorama (Fig. 5) has two advantages. Firstly the camera state constraint reduces the search space or thus acting like a filter. Secondly this helps generating panorama in online mode because the direction of movement is hidden in the states of the camera. (Fig. 2) The next stage is human extraction from image. First the right window in the image is selected. The camera states are defined and coupled with image states which simplifies the process. Second a colour detection algorithm is used to extract human skin colour (Fig. 4). Finally the image states and camera states are used to calculate the possible positions. Due to multiple choices in a single camera a matrix of world states are presented as a probability matrix. This matrix with the available previous states define the current state. A demonstration can be seen in Fig. 6.

5 Conclusion

A novel three-layered image understanding framework has been proposed, it includes single camera layer, camera network and semantic generation. This paper is focused on

qualitative correspondence of objects in the first layer, which is realized by a two-stage tracking cycle consisting of panorama generation and object tracking. A case stduy is given to demonstrate the effectiveness of the qualitative representation of human position using a single camera. Symbolic representation [7] is one of several fields that can improve anomaly detection and human-readable explanation [8] in surveillance systems. Future work consists of generalising this method in the multiple camera case for cooperative cameras to improve the intelligent functioning of the system. Also trajectory analysis and qualitative fusion from multiple cameras is one of our targets.

References

1. Haritaoglu, I., Harwood, D., Davis, L.: W4: A real time system for detecting and tracking people, 962–962 (1998)
2. Wren, C.R., Azarbayejani, A., Darrell, T., Pentland, A.: Pfinder: Real-time tracking of the human body. IEEE Transactions on Pattern Analysis and Machine Intelligence 19(7), 780–785 (1997)
3. Haritaoglu, I., Harwood, D., Davis, L.: W4s: A real-time system for detecting and tracking people in 2 1/2-d, I, 877 (1998)
4. Wu, G., Wu, Y., Jiao, L., Wang, Y., Chang, E.: Multicamera spatio-temporal fusion and biased sequence-data learning for security surveillance (2003)
5. Ukita, N., Matsuyama, T.: Real-time multi-target tracking by cooperative distributed active vision agents. In: AAMAS '02: Proceedings of the first international joint conference on Autonomous agents and multiagent systems, pp. 829–838. ACM Press, New York (2002)
6. Liu, H., Coghill, G.: Fuzzy qualitative trigonometry. In: Proc. IEEE International Conference on Systems, Man and Cybernetics, IEEE Computer Society Press, Los Alamitos (2005)
7. Lin, J., Keogh, E., Lonardi, S., Chiu, B.: A symbolic representation of time series, with implications for streaming algorithms. In: DMKD '03: Proceedings of the 8th ACM SIGMOD workshop on Research issues in data mining and knowledge discovery, pp. 2–11. ACM Press, New York (2003)
8. Chan, C.S., Liu, H., Brown, D.J.: Human arm-motion classification using qualitative normalised templates. In: Gabrys, B., Howlett, R.J., Jain, L.C. (eds.) KES 2006. LNCS (LNAI), vol. 4251, pp. 639–646. Springer, Heidelberg (2006)

A Hybrid System: Neural Network with Data Mining in an e-Learning Environment

David Wen-Shung Tai, Hui-Ju Wu, and Pi-Hsiang Li

Department of Industrial Education and Technology,
National Changhua University of Education
Changhua, 500, Taiwan, ROC
ietaiws@cc.ncue.edu.tw, d94311001@mail.ncue.edu.tw,
d94311007@mail.ncue.edu.tw

Abstract. This paper proposed a hybrid system combining the self-organizing map (SOM) of a neural network with the data mining (DM) method, for course recommendations in the e-learning system. SOM systems have been successfully used in several domains of artificial intelligence. Although many researches focused on e-learning system implementation and personal curriculum design, they do not give e-learners useful suggestions for selecting potential courses according to their interests or background. In order to enhance the efficiency and capability of e-learning systems, we combined the SOM method to deal with the cluster problems of the DM systems, SOM/DM for short. The experiment was carried out in a business college of a university in Taiwan, by applying the SOM/DM method to recommend courses to e-learners. The results indicated that the SOM/DM method has excellent performance.

1 Introduction

e-Learning is defined as the instructional content or learning experiences delivered or enabled by electronic technology [2], particularly computer networks and stand-alone computers. It is one of the main innovations that is increasingly diffusing in corporate settings. Web-based Intelligent e-Learning Environments (WILE) attract considerable attention worldwide, since they have the potential to improve the quality of e-learning applications and services [12]. WILE can overcome the main shortcoming of e-learning technologies of all learners receiving the same learning material and activities, the "one-size-fits-all" approach. It does so by facilitating personalized learning experiences that are adapted to the particular characteristics of each learner.

According to the International Data Corporation (IDC), the size of the e-learning market in Western Europe was said to be $358 million in 2003 (or some £200 million) [9]. This was predicted to grow to $994 million in 2007. According to this same IDC report, the UK e-learning market was forecast to be some $188 million in 2004 and $257 million in 2005. It is said to be growing at a compound annual rate of 28 percent a year. The corporate e-learning market in Asia/Pacific countries is expected to be worth almost $233 million by 2005, growing at a rate of 25 percent a year. These growth figures reveal that the number of e-learning initiatives in corporate training settings is increasing at a steady and rapid pace.

B. Apolloni et al. (Eds.): KES 2007/ WIRN 2007, Part II, LNAI 4693, pp. 42–49, 2007.
© Springer-Verlag Berlin Heidelberg 2007

In recent years, numerous Web 2.0 applications have been developed, such as portal websites (Google; Yahoo; MSN; AOL), multimedia community websites (Flickr; Riya; YouTube; Podbop; Wretch), various commercial websites (Amazon; eBay), and so on, demonstrating the increasing maturity of the internet. Consequently, the rapid growth of information on the Web has created a problem of information overload [3, 4], such that internet users are unable to find the information they require [1]. To help internet users search more efficiently, many powerful search tools have been proposed, such as the Google search engine (Google), or the Citeseer website (NEC Research Institute ResearchIndex). Most of these search tools provide personalized mechanisms to enable users to filter out uninteresting or irrelevant search results. In other words, personalized service has received considerable attention [6, 7, 13] recently because of the different information needs among users.

This paper constructs a hybrid system with artificial neural network (ANN) and data mining (DM) techniques. First, we use ANN to cluster the e-learner types. Based on these e-learner groups, a user can obtain course recommendations from the group's opinion. When groups of related interests have been established, DM will be used to find out the rules of "course vs. course". This approach allows for an inclusive curriculum to be suggested to a group of learners with the same interest. This system is ideal to stimulate the learners' motivation and interests. Moreover, it can serve as a reference when learners are choosing between classes.

The remainder of this paper is organized as follows. Reviews of related literature are described in Section 2. Section 3 demonstrates the design and the methodology of the SOM/DM system. The system performance is presented in Section 4. Finally, we draw our conclusions and discuss future research in Section 5.

2 Preliminaries

2.1 The Self-Organizing Map (SOM)

ANN is composed of richly interconnected non-linear nodes that communicate in parallel. The connection weights are modifiable, allowing ANN to learn directly from examples without requiring or providing an analytical solution to the problem. The most popular forms of learning are:

- Supervised learning. Patterns for which both inputs and outputs are known are presented to the ANN. The task of the supervised learner is to predict the value of the function for any valid input object after having seen a number of training examples. ANN employing supervised learning has been widely utilized for the solution of function approximation and classification problems.
- Unsupervised learning. Patterns are presented to the ANN in the form of feature values. It is distinguished from supervised learning by the fact that there is no *a* priori output. ANN employing unsupervised learning has been successfully employed for data mining and classification tasks. The SOM and Adaptive Resonance Theory (ART) constitute the most popular exemplars of this class.

The SOM is an unsupervised neural network model devised by Kohonen (1982) [11]. Basically it entails a non-linear projection of the probability density function of the input data space into a lower dimensional (in the case of this study a

2-dimensional) net of units (output space), each represented by a discrete reference weight vector or representative. In this paper, the SOM is used to group similar cases from the past case-base. After the grouping process is implemented by the SOM, it can reduce the time required to retrieve data from the past case base for the new case.

The notations used in the SOM for e-learner clustering are:

j^* denotes the winner, the neighborhood center, in output layer after competition, indicating the most representative factor for e-learner data.

(C_x, C_y) is the topological coordinates of the neighborhood center.

(X_j, Y_y) is the topological coordinates for processing element j in the output layer of the e-learner data.

R^t is the neighborhood radius.

$$R^t = 1\text{-}\frac{t}{T} \tag{1}$$

T is the total number of learning and training, and t is current number of learning and training.

r_j is the neighborhood distance.

$$r_j = \sqrt{(X_j - C_x)^2 + (Y_j - C_y)^2} \tag{2}$$

α_t is the learning rate.

α_rate is the reduced factor of learning rate.

R_factor is the neighborhood coefficient.

$$R_factor_j = f(R, r_j) \tag{3}$$

X_t is the factor of e-learners data.

2.2 Data Mining

Data mining (DM), known as “knowledge discovery in databases”, is the process of discovering meaningful patterns in huge databases. In addition it is an application area that can provide significant competitive advantage for making the right decision [8]. DM is an explorative and complicated process involving multiple iterative steps. Figure 1 shows an overview of the data mining process. In the first step, the main datasets to be retrieved from operational databases or business transactions warehousing are selected. In the second step, the selected datasets then go through data cleaning and preprocessing for removing discrepancies and inconsistencies among data types to improve quality. In the third step, the dataset is examined to identify patterns that express relationships among data by applying some algorithms, such as decision trees, clustering, Bayesian network and others.. Then the patterns are

validated with new datasets or a testing sample. It should be possible to transform these patterns into decision-making rules that are likely to help the administrators in accomplishing their objectives. Finally, the steps in the mining process are reiterated until meaningful knowledge is extracted [5].

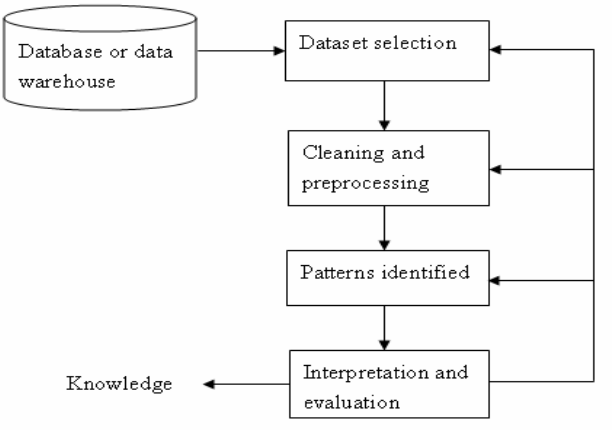


Fig. 1. Data mining phases

DM has been successfully applied to many domains. Some recent examples are as follows. Tseng and Lin (2006) propose a data mining method named SMAP-Mine that can efficiently discover mobile users' sequential movement patterns based on the requested services [14]. SMAP-Mine is shown to deliver excellent performance in terms of accuracy, execution efficiency and scalability. Recently, Kirkos et al. (2007) investigated the usefulness of Decision Trees, Neural Networks and Bayesian Belief Networks in the identification of fraudulent financial statements [10]. Huang et al. (2007) presented a model of a chronic disease prognosis and diagnosis system integrating DM and case-based reasoning (CBR) [8]. Their findings are considered helpful references for doctors and patients in the treatment of chronic diseases.

It is evident from the above literature that many research contributions applied data mining techniques to their applications. However, there are few researches that focused on hybrid data mining approaches and empirical investigation of e-learning related topics. Therefore, we will use association rules to enhance the accuracy of course recommendation in the e-learning system. It is our hope that the results of this study will provide useful information for educators to group e-learners more accurately.

3 SOM/DM System Design and Methodology

The proposed recommendation methodology in this paper is shown in Figure 2. The rationale of this proposed approach is that if e-learners have a similar background or behavior, then they may have similar interests for some courses. In other words, factors such as e-learners' gender, year, and major should be included in the variables to divide the e-learners. Thus, our approach follows two steps to identify e-learners' behavioral

pattern. First, gender, study year and major are used to divide them into groups with similar behavioral patterns. Then the SOM method is employed to facilitate this clustering according to the three above mentioned factors. Second, the association rule mining approach is applied to extract the recommended rules from each cluster. The extracted behavioral patterns of the e-learners are used to forecast their behaviors such as courses that are potentially interesting or useful to them. Therefore, the approach presented in this paper recommends potential courses for e-learners, based on their behavioral patterns by means of the association rule mining technique.

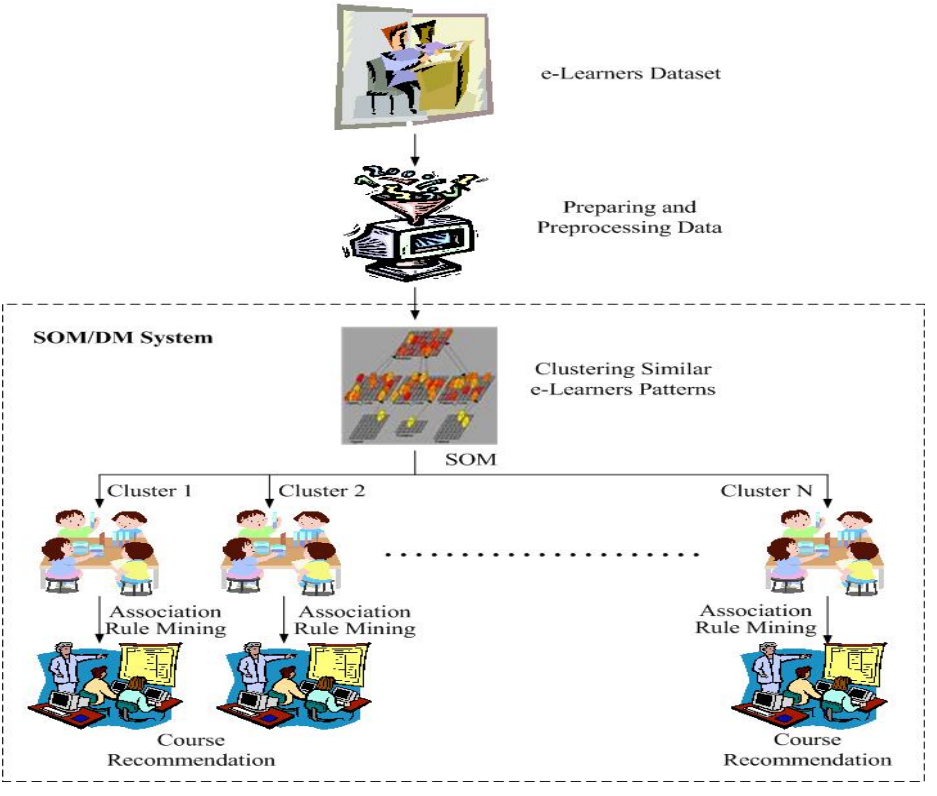


Fig. 2. Research Methodology

4 Performance of SOM/DM and Discussion

4.1 Data Description

A case study is used to illustrate the methodology. The subject is a business college of one of the universities in Taiwan. It offers degrees in business administration, international trade, accounting, information management and finance, with over 50 elective courses. Data on the background of 850 e-learners in the university had been collected over the past year. Some historical data were processed in advance prior to extracting the course records of these e-learners.

4.2 Data Analysis and Normalization

Based on the opinions of educational experts, three factors that could influence course recommendations were selected and were listed complete with definition in Table 1. In this research, the collected data could in the form of categorical data (gender, year, major) This necessitates transforming the collected data into a numerical form since the data to be used must be in numerical form for most clustering methods.

Table 1. Definition of Selected Factors

Factors	Definitions
Gender	Based on the e-learners' gender, the indexes were divided into two categories: male and female, indicated by values 1, and 2 respectively.
Year	Based on the study year, the indexes were divided into four categories: freshman, sophomore, junior, and senior, indicated by values 1, 2, 3, and 4 respectively.
Major	All types were divided into five categories based on the e-learners' major: business administration, international trade, accounting, information management and finance, indicated by value 1, 2, 3, 4, and 5 respectively.

The steps of transforming categorical data into numerical data are as follows.

(a) Calculate the average selected elective courses for all variables in each factor.

(b) Calculate the related distances for the above variables based on the above average selected elective courses.

(c) Use the following formula to normalize the above related distances, so that the value will be between 0 and 1.

$$F = \frac{X_i - \min(x_1, x_2, x_3, ...x_n)}{\max(x_1, x_2, x_3, ...x_n) - \min(x_1, x_2, x_3, ...x_n)} \tag{4}$$

X_i is average sales volume for each variable.

F_i is normalized value for each variable.

After above steps to normalize collected data, five factors are now values between 0 and 1.

4.3 Experimental Results

This paper used the values of root mean square (RMS) to determine if the clustering process was finished or not during the calculation of the SOM of ANN. When the interactions reached 20,000 times, the value of RMS was 0.006824 and remained robust. The SOM model then calculated one output value for each training sample. These output values were then used as the criterion for e-learners clustering. All output values were divided into five groups in a best clustering condition as shown in Table 2.

To further investigate whether the e-learners in each cluster would choose potential courses, association rule mining was then used to extract the prediction rule set RSi

from the background of the e-learners associated with each cluster. The cluster Ci to which a e-learner, e, belonged was first identified. Then, RSi, the recommendation rule set extracted from Ci was used to predict the candidate future potensial elective courses for e-learner e. All candidate potential elective courses were classified and ranked according to the associated confidence of the recommendation rules. Then the highest ranked candidate potential elective courses were selected as the recommended rule. The minimum confidence level was set to 0.8 and the minimum support to 0.1. The following rule is an example of a quantitative association rule, where X is a variable representating an e-learner. In this example, the e-learner belongs to Cluster 1 and has already taken the **"**Data Structure**"** course. Therefore the mining result will suggest for this e-learner to select **"**Visual Basic Programming**"** as his potential elective course.

Gender(X, **"**Male**"**) $\wedge$ year(X, **"**Junior**"**) $\wedge$ Major(X, **"**Business Administration**"**) $\wedge$ Course(X, **"Data Structure"**)

$\Rightarrow$ Courses Recommendation (X, **"Visual Basic Programming"**)

Table 2. Clustering Condition for e-Learners

Cluster 1	**Gender**	**Year**	**Major**
	Male	Freshman	Business Administration
	Female	Sophomore	International Trade
		Junior	Accounting
		Senior	Information Management
			Finance
Cluster 2	**Gender**	**Year**	**Major**
	Male	Freshman	Business Administration
	Female	Sophomore	International Trade
		Junior	Accounting
		Senior	Information Management
			Finance
Cluster 3	**Gender**	**Year**	**Major**
	Male	Freshman	Business Administration
	Female	Sophomore	International Trade
		Junior	Accounting
		Senior	Information Management
			Finance
Cluster 4	**Gender**	**Year**	**Major**
	Male	Freshman	Business Administration
	Female	Sophomore	International Trade
		Junior	Accounting
		Senior	Information Management
			Finance
Cluster 5	**Gender**	**Year**	**Major**
	Male	Freshman	Business Administration
	Female	Sophomore	International Trade
		Junior	Accounting
		Senior	Information Management
			Finance

5 Conclusion

In this paper we presented an e-learning recommendation system called SOM/DM. This paper was the first to utilize ANN to determine the clusters of e-learners. Based on these e-learner's groups, a user can obtain course recommendations from the group's opinion. When groups of related interests have been established, the a priori algorithm will be used to find out the rules of recommendation courses. This paper makes three important contributions: (1) it is an empirical study to apply demographic variables to separate e-learners; (2) the ANN and DM techniques were used to enhance the accuracy of the course recommendations; (3) the results of this study provides useful information for educators to classify their e-learners or students more accurately and to plan their teaching strategies. The experiments prove that our proposed method is ideal for stimulating a learner's motivation and interest. Finally, the demonstration confirms that our proposed method is both feasible and practicable.

References

[1] Arasu, A., Cho, J., Garcia-Molina, H., Paepcke, A., Raghavan, S.: Searching the Web. ACM Transactions on Internet Technology 1(1), 97–101 (2001)
[2] Aydin, C.H., Tasci, D.: Measuring Readiness for e-Learning: Reflections from an Emerging Country. Educational Technology & Society 8(4), 244–257 (2005)
[3] Berghel, H.: Cyberspace 2000: Dealing with information overload. Communications of the ACM 40(2), 19–24 (1997)
[4] Borchers, A., Herlocker, J., Konstanand, J., Riedl, J.: Ganging up on information overload. Computer 31(4), 106–108 (1998)
[5] Bose, I., Mahapatra, R.: Business data mining - a machine learning perspective. Information & Management 39, 211–225 (2001)
[6] Chen, C.M., Lee, H.M., Chen, Y.H.: Personalized e-learning system using Item Response Theory. Computers and Education 44, 237–255 (2005)
[7] Herlocker, J., Konstan, J.: Content-independent, task-focused recommendations. IEEE Internet Computing 5, 40–47 (2001)
[8] Huang, M.J., Chen, M.Y., Lee, S.C.: Integrating Data Mining with Case-based Reasoning for Chronic Diseases Prognosis and Diagnosis. accepted to appear in Expert Systems with Applications 32(3), 856–867 (2007)
[9] International Data Corporation Website, http://www.idc.com/
[10] Kirkos, E., Spathis, C., Manolopoulos, Y.: Data Mining techniques for the detection of fraudulent financial statements. Expert Systems with Applications 23(4), 995–1003 (2007)
[11] Kohonen, T.: Self-organized formation of topologically correct feature maps. Biological Cybernetics 43(1), 59–69 (1982)
[12] Ma, Z.: Web-Based Intelligent e-Learning Systems: Technologies and Applications, Information Science Publishing, pages 388 (2006)
[13] Rashid, A.M., Albert, I., Cosley, D., Lam, S.K., McNee, S., Konstan, J.A.: Getting to know you: Learning new user preferences in recommender systems. In: Proceedings of the 2002 international conference on intelligent user interfaces, pp. 127–134 (2002)
[14] Tseng, V.S., Lin, K.W.: Efficient mining and prediction of user behavior patterns in mobile web systems. Information and Software Technology 48(6), 357–369 (2006)

A Personalization Recommendation Framework of IT Certification e-Learning System

Heng-Li Yang[1], Chen-Shu Wang[1], and Mu-Yen Chen[2]

[1] Department of MIS, National Cheng-Chi University,
64, Sec. 2, Chihnan Rd., Mucha Dist, Taipei 116, Taiwan
{yanh, 93356506}@nccu.edu.tw
[2] Department of Accounting, National Changhua University of Education
Changhua, 500, Taiwan, ROC
chenmy@cc.ncue.edu.tw

Abstract. Obtaining IT certifications is a common and permanent goal of almost all IT students and IT workers. However, approximately 200–400 computer-related certifications exist and constantly increase or change. Examinee indeed requires the personalization recommendation about these certifications and so on. This study proposed a personalization recommendation framework contains two reasoning mechanisms: experience-based and capability-based reasoning. The experienced-based reasoning provides personalized suggestions based on the sophistic examinee's experience, while the capability-based reasoning considers the learner's background only. The proposed framework can provide the beneficial certification path to examine and a practice scenario is presented to illustrate the feasibility of proposed framework.

Keywords: IT certification, examinee personalization recommendation, certification examination preparation.

1 Introduction

IT certification is an increasingly important part of recruitment, often even becoming a vital screening mechanism. Venator (2006) contended that IT certification, like educational background, can be considered as a standard for determining the suitability of IT workers [9]. Additionally, Jo (2005) concluded that IT certification increases women's career opportunities, particularly in jobs related to information security [6]. IT certification seems to be the panacea for working opportunity and career promotion.

However, IT certification examination is a bottomless pit [2]. Approximately 200–400 computer-related certifications exist. The only way to obtain these certifications is to pass the related exams. Unfortunately, these certifications exams constantly increase and change by about 10% to 15% every year. Certification exams are thus a cause for concern for many IT certification examinees. Even exam veterans are concerned about particular issues when preparing and attending these IT certification

B. Apolloni et al. (Eds.): KES 2007/ WIRN 2007, Part II, LNAI 4693, pp. 50–57, 2007.
© Springer-Verlag Berlin Heidelberg 2007

exams: which certification exams to attend, the current certification; restrictions on applying the exam, and how to prepare for the certification. However, these questions are hard to answer consistently, due to the diverse professional backgrounds of examinee. Examinees may range from new-entry level to veteran and sophisticated. A Hence, a personalization recommendation mechanism is required to deal with these problems according to examinee personal conditions [3].

Professional certifications are becoming increasingly important. Most trades and professions are now developing their own professional certifications. Whether IT certification enables IT workers to be regarded as professionals is still a controversial issue [8]. However, almost all researches positively support IT certification [7,9,10]. The information requirement for IT certification exams can be expected to rise constantly. This study presents a framework for personalization recommendation of IT certification e-learning systems. The proposed framework handles the following problems:

(1) With reference to veteran and sophisticated examinees' experience, the proposed framework would provide personalization recommendations (suggestion for IT certification exam application) based on examinee conditions.
(2) Moreover, an examinee may be interested in a particular certification. Hence, the proposed framework can estimate the domain knowledge gap between a certification and the examinee's background.
(3) Finally, the proposed framework recommends a series of certifications and learning paths from lifelong learning perspective.

2 Proposed Model

Revised from Dolog and Sintek (2004), Figure 1 illustrates the ontology and metadata describing an IT certification examinee [3]. Every examinee is considered as a learner within the IT certification e-learning system, since some learners may only expect knowledge learning despite IT certification. The e-learning system maintains a learner's basic demographic information and profile. The profile includes learner's working experience (industry, title and duration of work) and experienced training classes. Additionally, a learner's profile records the certifications that he has previously obtained, and the learner's preference. The preference presents the important weights of each compared item and is applied to learner's personalization recommendations.

Additionally, learner may obtain the same domain knowledge from different sources, such as school, working experience and self-learning [5]. This study disregards knowledge sources, since learners may obtain different levels of knowledge even when they experience the same training classes. This study verifies learners' domain knowledge by exam instead of knowledge source. Restated, if a learner claims to have particular domain knowledge, then the relevant exams within the e-learning system are evoked. If a learner passes the exam, then the e-learning system updates his exam log, which presents his domain knowledge.

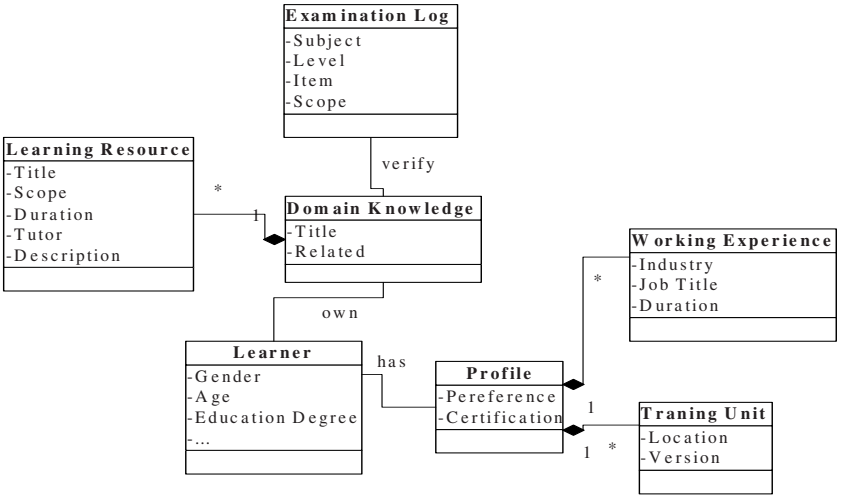


Fig. 1. Ontology of Learner

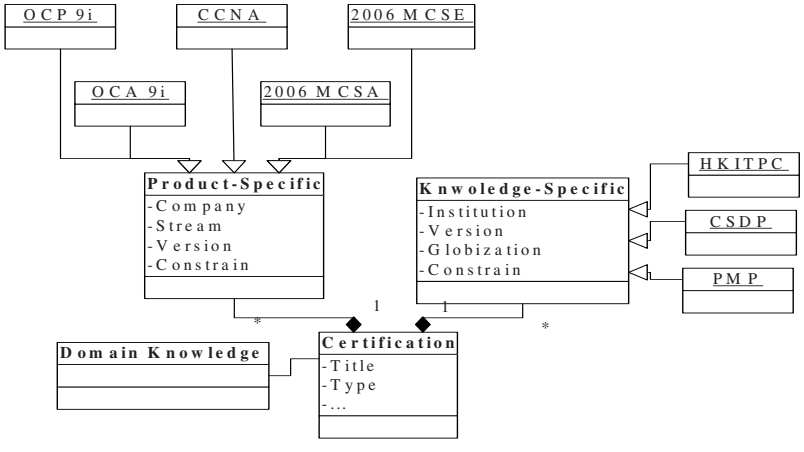


Fig. 2. Ontology of Certification

Figure 2 presents the ontology describing IT certification. IT certifications are generally classified into two types, namely product-producer specific IT certification, and the domain-knowledge specific. Both these two types of certification contain version and application constrains. For product specific certification, the product producer and certification stream are added. For instance, the Cisco Certified Network Associate (CCNA) certification belongs to the networking management stream of Cisco Career Certification. For knowledge-Specific certification, the approval institution and globalization are extended. For instance, the project management professional (PMP) is approved by the Project Management Institute globally, whereas the Hong Kong IT Professional Certification (HKITPC) is approved locally

by the Hong Kong Institute. The proposed framework contains two perspectives reasoning mechanisms, experience-based and capability-based. Figure 3 shows the proposed framework, in which learners interact with the e-learning system through an interactive interface, and may update their profiles and submit queries to achieve recommendation. If the system invokes the exam mechanism of the e-learning system to verify learners' domain knowledge, then update their exam logs.

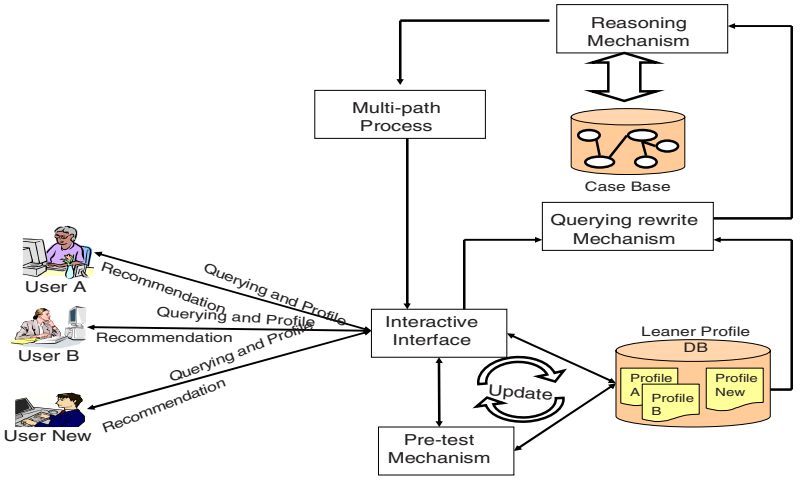


Fig. 3. Proposed recommendation framework of IT Certification e-learning system

2.1 Experience-Based Reasoning

The learner's query is translated by a query rewrite mechanism before experience-based reasoning. The query rewrite mechanism extends an original query by adding learner profiles, e.g. working experience and exam log of learner, and the revised query can be considered as compared items. Additionally, the preference records the learner's importance weight for each compared item. Thus, four items are compared, namely working experience, exam log, certification (obtained by learner) and basic demographic information of learner. The learner assigns to each item an importance weight, which is considered as his (her) preference. The experience-based reasoning algorithm estimates the similarity between learner and case according to the priority within preference. The comparison algorithm then estimates the similarity of each case by Eq. (1), in which the weight of each compared item represents its importance.

$$Sim(L, c_i) = \frac{1}{\sum_{n=k,p,q,m} w_n} \left(\frac{w_k}{k} \sum_k \Psi(L_k, C_k^i) + \frac{w_p}{p} \sum_p \Psi(L_p, C_p^i) + \frac{w_q}{q} \sum_q \Psi(L_q, C_q^{i)} + \frac{w_m}{m} \sum_m \Psi(L_m, C_m^i) \right) \quad (1)$$

Equation (1) estimates the similarity between learner and $case_i$, where Learner is abbreviated to L, and $case_i$ is abbreviated to C_i. Four items are compared, namely k, p, q and m, which denote the working experience, exam log, basic demographic data and certification, respectively. The weights prior to the item comparisons indicate their priorities within the preferences. If a learner discards preferences, then these four

compared item are treated equally. The similarity is between 0 (totally different) and 1 (exactly like). The match function, denotes as Ψ, compares attributes within a compared item, e.g. the industry type of working experience: if attribute value of leaner is consistent with $case_i$, then return 1, else return 0.

Experience-based reasoning implies that this reasoning mechanism provides recommendation by reference to the experience of veteran and sophisticated learners. As in case-based reasoning [1], each case estimates the similarity by Eq. (1). Then, the most similar case is retrieved and recommended to the learner for IT certification exam application reference.

2.2 Capability-Based Reasoning

The second reasoning perspective is capability-based reasoning. As described above, a learner may expect the recommendation based on his (her) background, and not refer to other veteran and sophisticated learners' experience.

Capability-based reasoning mechanism can infer the most beneficial IT certification for a leaner. Additionally, the relationship between IT certifications can also be inferred from domain knowledge coverage as describe in Figure 2. The core concept of capability-base reasoning is domain knowledge coverage, which is similar to the intersection operation in set theory. Capability-based reasoning considers a learner's domain knowledge as set A, treat domain knowledge of each IT certification as set B, and takes the intersection of sets A and B ($A \cap B$). The most similar set B is that of the exam for which a learner needs to the least preparation and which would be the most beneficial to the learner. Additionally, it can also infer the relationship between IT certifications. Then, the most similar certification stream can also be recommended to learner.

3 Example Scenario

The example scenario can reveal the proposed recommendation framework for the IT certification e-learning system. Tony has been a member of the IT certification e-learning system since 2005. He obtained his information management master degree in 2003, and then served as a database management administer (DBA) in ABC software company. He once attended the CISCO project, a networking management curriculum, in school, where he obtained the CCNA certification. Tony later obtained the Oracle OCA 9i certification. In the e-learning system, Tony attended some classes and passed the exams from 2005 to 2006. His exam log recorded that he had verified that his knowledge contained oracle database performance tuning, project management, installation and configuration of Microsoft Windows 2000 system.

ABC Software Company is currently striving to obtain CMMI certification. Hence, the administrated policies are intended to encourage its employees to obtained related IT certification. Tony would like to take advantage of the opportunity, since he wishes to obtain further IT certification in order to gain further promotion. However, Tony has no idea about which exam he should attend and the current IT certification version, or how he should prepare for the certification exam. Usually, examinee, like

Tony, would attend the advance examination in the same certification series, such as OCA, OCP and OCM in the Oracle DBA certification series.

Therefore, Tony submits a query for recommendation. He expects to refer to other veteran and sophisticated examinees' exam application information. His order of preference is the working experience (the weight is 4), exam log(the weight is 3), certification (once obtained) (the weight is 2) and personal demographic data (the weight is 1). He asks the e-learning system to return the most similar case.

3.1 Experience-Based Reasoning Scenario

This scenario is very familiar to any IT certification exam examinee. Because Tony expects refer other examinees' experience, experience-based reasoning is adopted. Tony's original query is extended, to include profile and preference, by querying the rewrite mechanism. The revised query is regarded as the Learner query, and is compared to all cases within the case base by Eq. (1). Three cases with similarity above 0.5 (Cases No. 5, 13 and 22) are retrieved for Tony's situation, and are listed below.

(1) Case No.5: His working experiences are identical to Tony's. He passed all examinations that Tony did except for the windows 2000 system administration. He has one more certification, the OCP 9i certification. Additionally, he is younger (age 28), other demographic attributes are same as Tony.
(2) Case No.13: He has the job title, MIS, with three years of working experiences. He has one more certification, the OCA 10g certification. All of his examination logs and demographic attributes are the same as Tony's.
(3) Case No.22: He has "DBA" job title and 2 years of working experiences in 3C industry. Except for project management, he passed all examinations that Tony did. He has one more certification, the 2003 MCSA certification. Additionally, he is male with age of 25 and received bachelor degree rather than master degree.

$$Sim_a(Tony,\ Case5)=\frac{1}{10}\left(4\times\frac{3}{3}+3\times\frac{3}{4}+1\times\frac{2}{3}+2\times\frac{2}{2}\right)=0.891$$

According to current preference (named situation A), the similarity of each case is 0.891, 0.864 and 0.590 respectively and the most similar case No. 5 is suggested to Tony, who would therefore prepare for the Oracle OCP 9i certification exam. However, suppose that Tony would change his order of preference due to consideration of exam preparation. The exam log (weight is 4) is before working experience (weight is 3) in the new preference order (labeled situation B). The similarity of each case would naturally be re-evaluated. Assume that the same three cases are retrieved in situation B. However the similarity of each case becomes 0.866, 0.898 and 0.632 respectively. Hence, the most similar case is now case No. 13, which is therefore recommended to Tony. According to case No. 13, Tony would be recommended to prepare for the Oracle OCA 10g certification exam.

3.2 Capability-Based Reasoning

However, Tony discovers that the regulations concerning rewards are yet to be introduced. Therefore, Tony submits another new query for requiring the least exam

preparation. He expects that the e-learning system would suggest IT certification requiring no more or the least exam preparation according to his current domain knowledge. Capability-based reasoning is adapted to deal with this problem, by taking the intersection of domain knowledge between Tony (exam log) and each IT certification. The most beneficial IT certification (meaning the most similar domain knowledge coverage) is recommended to Tony.

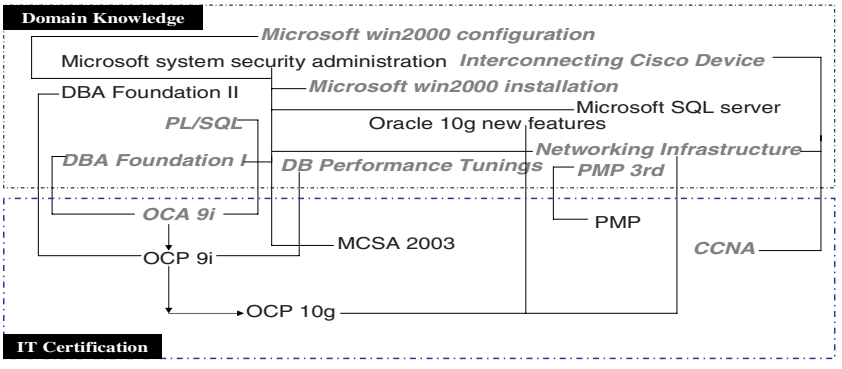


Fig. 4. Illustration of multi-path mechanism

Figure 4 shows multiple certification paths recommended to Tony. The PMP certification is first recommended to Tony, because it requires no further preparation. The domain knowledge coverage is thus 100%. Then, OCA 10g (another 9i upgrade: 10g new features domain knowledge required, the domain knowledge coverage is 80%) and the OCP 9i (another advance DBA foundation domain knowledge required, the domain coverage is 75%) are recommended. The multi-path mechanism arranges the priorities of certification paths based on their domain knowledge coverage.

Additionally, from the lifelong learning perspective, the proposed frame recommends a series of related IT certification. As Figure 4, the bold and italic text denotes the domain knowledge and certifications that Tony had verified and obtained. Hence, from the domain knowledge coverage concept, the same series such as OCM, would be recommended, except that OCP 9i and OCA 10g are removed. Additionally, because Tony has always installed Oracle database on windows 2000 (or above) systems, he is also familiar with some domain knowledge of Microsoft MCSA Certification (including MS-70210). Therefore, Microsoft MCSA or MCSE can be considered as candidates for exam application.

Finally, Tony decides to apply for PMP certification application. Then, after the application constraint is checked, he would be warned that he is not qualified for PMP. According to the regulations of the project management institution, all examinee should attend training classes (by an approval institution) for at least 30 hours. Tony now needs to decide whether to fulfill the PMP 30-hours training constraint or apply for other certifications.

4 Conclusions and Future Work

The age of professional certification has arrived. 78% of managers believe that certified employees are more productive than others [8]. The educational background is becoming an essential but not decisive factor for employee competition. IT certification exam never disappears but blossom.

Most examinees are currently following blindly the hottest "certification trend" such as the top 10 wanted IT certifications. Almost all examinees thus focus on Microsoft or Java series certifications, and neglect their interests and specialties. It is feasible but a little unfortunate. From the lifelong learning perspective, certification should be an outpost, not a lifelong career guarantee. Examinees should plan certification paths according to their interests, capabilities and career plan. This study presents a framework as a recommendation mechanism for an IT e-learning system. Examinees can set their preferences for each compared item. Two reasoning algorithm perspectives can be adopted. The experienced-based reasoning provides personalized recommendations based on the sophistic examinee's experience, while the capability-based reasoning considers the learner's background only.

Further work will be to implement the physical systems. Further constraints, such as certification fee [4] and the difficult level of examination, would be added to recommendation model. Additionally, the passing rate of certification exam and stratification of examinees could be adopted to verify the proposed framework.

References

1. Aamodt, A., Plaza, E.: Case-Based Reasoning: Foundational Issues, Methodological Variations, and System Approaches. In: AI Communications, vol. 7(1), pp. 39–59. IOS Press, Amsterdam (1994)
2. Brookshire, R.G.: Information Technology Certification: Is This Your Mission? Information Technology, Learning, and Performance Journal 18(2), 1–2 (2002)
3. Dolog, P., Sintek, M.: Personalization in Distributed e-Learning Environments. In: Proceedings of the 2004 WWW conference, pp. 170–179 (2004)
4. Foster, D.: The Price of IT Certification Testing. Certification Magazine, 14 (April 2005)
5. Harp, C., Sztzinger, J., Taylor, S.: Many Paths to Learning Software. Training & Development 51(5), 81–84 (1997)
6. Jo, S.R.: IT certification: Increasing Women's Career Opportunities. Certification Magazine (July 12, 2005)
7. Mulkey, J.: IT Certification on Campus: Credential in a League of Their Own. Certification Magazine, 40–43 (September 2003)
8. Pierson, L., Frolick, M., Chen, L.: Emerging Issues in IT Certification. Journal of Computer Information Systems. 42(1), 17–20 (2001)
9. Venator, J.: IT Certification: Still Valuable After All These Years. Education & Careers, 28–31
10. Zeng, F.: A New Approach to Integrate Computer Technology Certifications into Computer Information System Programs. In: Proceedings of the 2004 American Society for Engineering Education Annual Conference & Exposition (2004)

An Architecture for e-Learning System with Computational Intelligence

Marc El Alami, Nicolas Casel, and Denis Zampunieris

CICeL, Faculty of Sciences, Technology & Communication, University of Luxembourg, 6 rue Richard Coudenhove-Kalergi, L-1359 Luxembourg-Kirchberg, Grand-Duchy of Luxembourg
{marc.elalami, nicolas.casel, denis.zampunieris}@uni.lu
http://cicel.uni.lu

Abstract. In this paper, we introduce a new kind of software tools intended to offer a virtual educational and/or training environment online: proactive e-Learning Management Systems.

These computational intelligence based systems are designed to improve the users' online (inter)actions by providing programmable, automatic and continuous intelligent analyses of the users' behaviors, augmented with appropriate actions initiated by the LMS itself.

We show how we implemented the proactive part of our LMS on the basis of a dynamic rules-based expert system. We also sketch how it looks like from a user's point of view. Finally, We give some examples of intelligent analysis of users' behaviors coded into proactive rules.

Keywords: Learning Management System, Proactive Computing, Rule-Based Expert System.

1 Introduction

Learning Management Systems (LMS) or e-learning platforms are dedicated software tools intended to offer a virtual educational and/or online training environment. Unfortunately, current LMS are fundamentally limited tools. Indeed, they are only reactive software developed like classical, user-action oriented software. These tools wait for an instruction and then react to the user request.

Students using these online systems could imagine and hope for more help and assistance tools, based on an intelligent analysis of their (lack of) actions. LMS should tend to offer some personal, immediate and appropriate support like teachers do in classrooms.

In this paper, we introduce a new kind of LMS: a proactive e-Learning Management System, designed to improve the users' online (inter)actions by providing programmable, automatic and continuous intelligent analyses of the users' behaviors, augmented with appropriate actions initiated by the LMS itself.

B. Apolloni et al. (Eds.): KES 2007/ WIRN 2007, Part II, LNAI 4693, pp. 58–65, 2007.
© Springer-Verlag Berlin Heidelberg 2007

Our proactive LMS can, for example, automatically and continuously help and take care of e-learners with respect to previously defined procedures rules, and even flag other users, like e-tutors, if something wrong is detected in their behaviors; it can also automatically verify that awaited behaviors of e-users have been carried out, and it can react if these actions did not happen.

In the following text, we show how we implemented the proactive part of our LMS on the basis of a dynamic rule-based expert system, and we sketch how it looks like from a user's point of view. Finally, we give some examples of computational intelligence coded into proactive rules.

2 User Interface

Recent works (see [1] for example) also propose how to improve current web-based educational systems by adding intelligence in these systems, but these intelligent add-ons modules are as static as the initial LMS was. Indeed, they still need a click or an action from the user to activate it. Our goal was to design and develop a LMS, that is able to analyze a situation and to act spontaneously wrt. this situation without queries from its environment.

In our system, the user can receive information, help or hints sent by the proactive system at any time and with no actions needed from him. These messages should not disturb him in his work, that's why the interface has been thought in such a way that the information will be viewable at any time and in any context in the LMS.

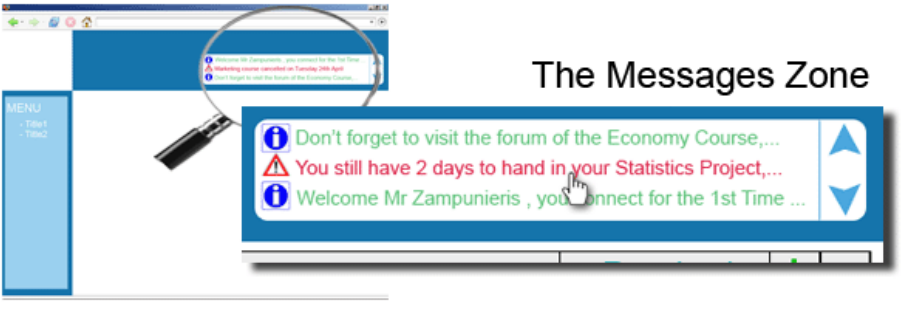


Fig. 1. The Message Zone

A messages zone has been dedicated in the header. This alert zone is a Flash application which is able to display server messages in real-time. Messages are following each other vertically and are colored differently according to their importance.

By clicking on a message, the user opens the Messages Manager and s/he can read more details on her/his alerts and save them.



Fig. 2. The Message Manager

3 Proactive Computing

Proactive systems (see [2], [3]) act on their own initiative on behalf of the user and without waiting for users explicit command. In other words, the user is "above the loop", that means that the system can manage himself with no human interactions. The user is here to supervise the system and take last decisions. These systems are more reactive to extern events and can communicate automatically with users using the usual ways of messaging like popup, emails, alerts zone or logs for administrators. The continuous computing of users (inter)actions is possible thanks to a set of programmed behaviours augmented with appropriated actions.

The aim of proactive LMS is to help users to better interact online in a learning environment. Guiding new user through the LMS is a common problem: dynamic help will be more efficient than an elementary static interface. Moreover, proactivity is useful in case of urgent situation: it enables focusing and dealing immediately with relevant information. For instance, a learner may have difficulties with an online exercise. Our system catches, analyses and transmits pertinent data to the tutor who can quickly respond. With standard LMS, the same piece of information is merely stored in a database.

How does it work? The proactive part of the LMS is based on few principles:

- Every proactive behaviours is coded in the Rules Running System (RRS)
- Rules are pushed in a FIFO queue
- The RRS is activated for a time frequency F and performs the N first rules of the FIFO queue
- Once executed, a rule is deleted from the system
- To regenerate, a rule should clone itself to be pushed again in the FIFO queue and be reactivated later

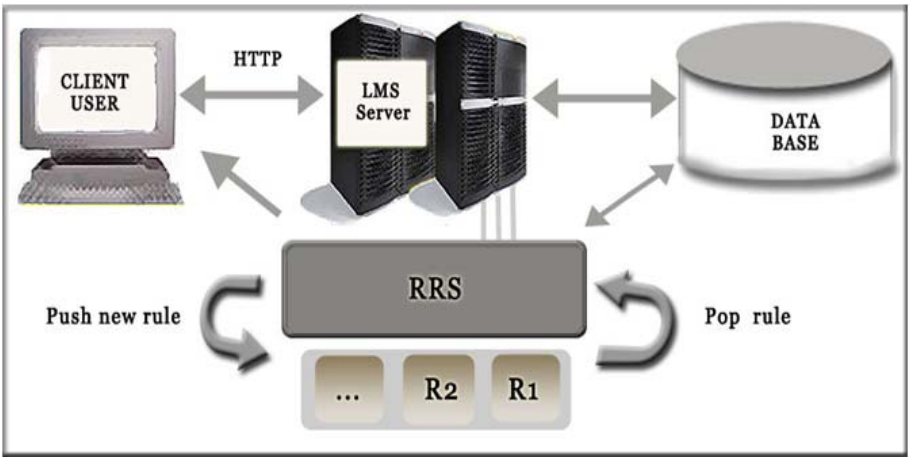


Fig. 3. The Rules Running System (RRS)

Figure 3 describes the different actors interfering in the system. The RRS is the extra element that wouldnt appear in a traditional LMS architecture. It is independent from the initial LMS server but works in complementarity with it by sharing the same database. The RRS also directly communicates with the Client and the LMS server. For more details, please see [4].

A rule is made of five parts: data acquisition, activation guards, conditions, actions and rules generation.

```
data acquisition :
        es = get_user(S)
        ec = get_course(C)
activation guards :
        es.isConnectedToCourse(C) == true
conditions :
actions :
        showMessageBox(es.session, "Welcome to the course " + ec.name)
        showMessageBox(es.session, "Do not forget to take a look at the forum "
                + "dedicated to the course " + ec.name + " called " + ec.forum)
rules generation :
        if (activated == false)
        then cloneRule(self)
```

Fig. 4. The Algorithm to Run the 'Welcome' Rule

The **data acquisition** allows a rule to get information from the LMS in order to use these data in its others parts. The second part (**activation guards**) is made of a set of tests that determine wether or not the rest of the rule will be processed. If the activation guards are evaluated positively, the conditions and actions parts are performed. Otherwise, these two parts are ignored and

the rules generation is executed. The third part (**conditions**) is a set of AND-connected tests on local variables that, once evaluated, determine if the action part will be executed afterwards. Syntax and semantics of conditions tests are equivalent to activation guards tests. The fourth part (**actions**) is made of a list of instructions that will be performed in sequence. The fifth and last part (**rules generation**) is realised at the end. It allows the rule to regenerate itself or others rules that will be pushed in the FIFO queue and executed later.

4 RRS Architecture

Before starting to develop the system, some technological choices were made:

- The system should be simple, effective and fast: therefore we choose the C language that has these advantages.
- The LMS is able to send information to the client at anytime without waiting for a client query. Indeed, using sockets was clearly necessary in order to keep a permanent connexion where both parts can be servers.
- The client is able to treat and post text in a browser automatically without refreshing: Flash is the right client application to meet these requirements. Widespread (97% of the browsers) and crossplatform, it can be coupled with several servers technologies (Java, JSP, PHP, ASP, etc) and the most important property is that it can handle a socket connexion.

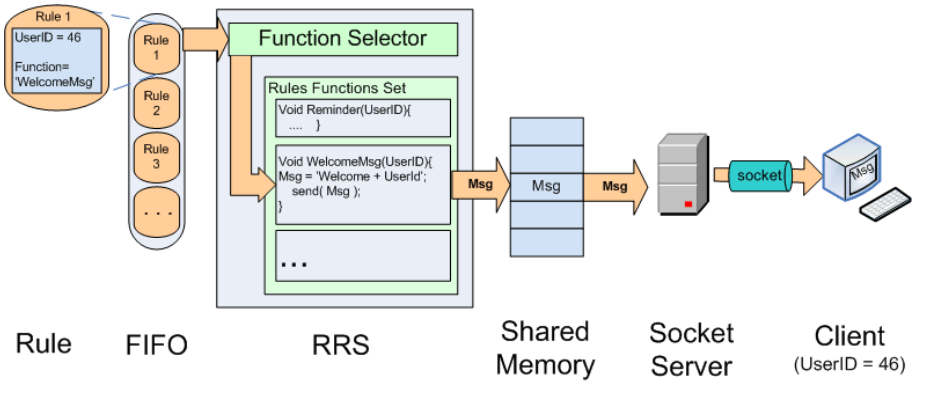


Fig. 5. The RRS Architecture

Figure 5 shows an overview of the Rules Running System. The RRS is able to pop a rule from the FIFO queue, check the type of this rule and parameters included. The reference to the function (or proactive behaviour) which performs the rule is given to the RRS. Actually, the RRS accesses a library that defines every rule and instructions to execute it. For instance, these actions can be

sending a message to the client through the socket connexion or by email. In this example, the rule is to send a welcome message to the user n°46. The message is build in the RRS and put in the shared memory. The server checks the shared memory and post the welcome message to the right online user. The use of the shared memory allows the RRS and the server to work asynchronously. The RRS can continue to handle new rules while the server is in charge of sending the message to the client.

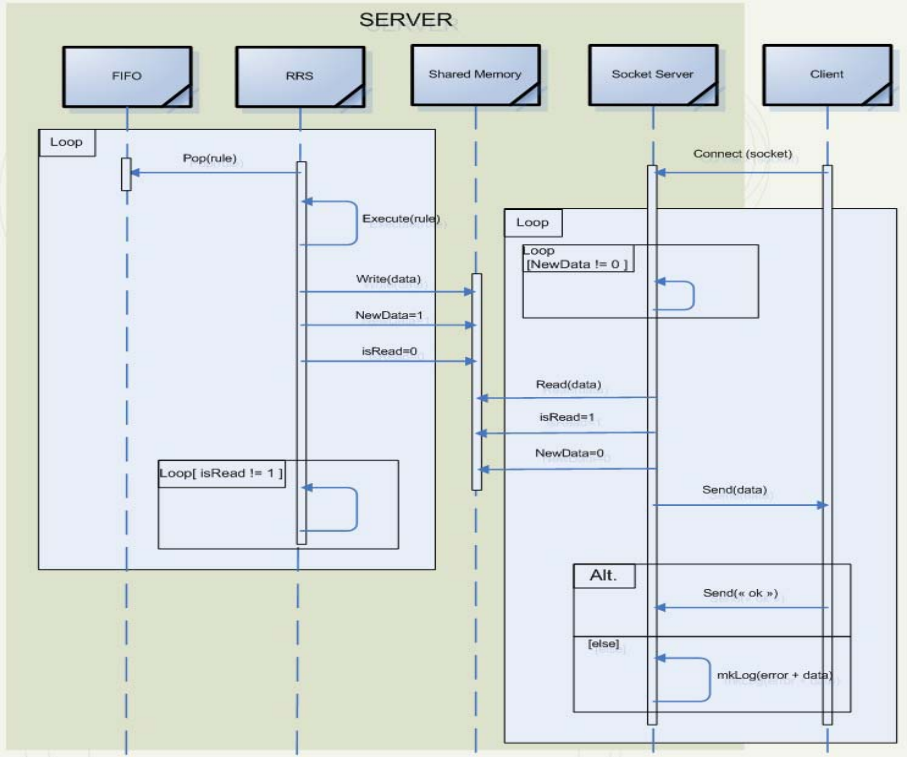


Fig. 6. The Message Sending Sequence Diagram

Figure 6 shows the path followed by a message from the state of a rule on server-side to the print on the client screen. First, the RRS pops the rule from the queue, tests the activation guards (ex: test if the user is connected) and executes the rule. The message is build and put in a shared memory where the socket server can easily access and read it. Before reading the shared memory, the server checks if there is a new message to send. Meanwhile, a client is connected and if the message concerns him, the server will use the open socket to address him the data read from the shared memory. Finally the client will send back a positive acknowledgment. If there is no response from the client after a given time, the server creates a log to inform the administrator. The RRS

waits for the socket server to read the message in the shared memory before continuing.

With the continuous evaluation of large sets of rules, an efficiency problem might appear with the system saturation. Indeed, as every rule makes an average of two to three requests to the LMS database, the mean reaction time of the LMS to a user request is increased, sometimes in a severe way. This problem can be solved with the introduction of "lazy evaluation" in the main algorithm. It is an optimization technique that attempts to delay computation of expressions until the results of the computation are known to be needed. For the interested reader, more information about this solution can be found in [5].

5 Rules for Intelligent Analysis

These proactive behaviours coded into rules are the bases for an automatic and intelligent analysis for the current state of the whole system and of its changes implied by user actions and interactions.

```
data acquisition :
        es = get_user(S)
        et = get_user(T)
        ec = get_course(C)
        date = get_date()
activation guards :
        date > ( ec.startDate + 7 days)
conditions :
        es.numberOfConnections(ec.forum) == 0
actions :
        sendLMSeMail(to = et.name, subject = "Warning", data = "e-student "
                + es.name + " did not use the forum " + ec.forum.name
                + " after one week... please check with her/him")
rules generation :
        if (activated == false)
        then cloneRule(self)
```

Fig. 7. Algorithm to Run the 'Forum-Warning' Rule

Figure 7 represents the structure of a specific rule that is intended to check that the e-student S used at least once the LMS forum communication tool dedicated to the e-course C one week after his registration to that e-course and if not, to notify it by an LMS email to the e-tutor T so that he can check with the e-student what is the problem. In the data acquisition, the system gets all the information needed (information about the student, the tutor, the course and the current date), then the activation guards are performed to determine if the rule should be activated or not; if yes, the rule enters the conditions part that determine if the rule should be processed. In that case the rule enter the actions part and accomplishes a list of instructions (in our case, the instruction is to send an email to the tutor with all the information). The fifth and last part (rules generation) allows to regenerate rules that have not been activated.

6 Conclusion

Current Learning Management Systems are fundamentally limited software tools: they are only reactive, user-action oriented software. These tools wait for an instruction and then react to the user request. They do not offer some personal, immediate and appropriate support like teachers do in classrooms.

We introduced a new kind of Learning Management System: proactive LMS. These e-learning platforms with computational intelligence, are designed to improve their users' online interactions by providing programmable, automatic and continuous analyses of users actions augmented with appropriate actions initiated by the LMS itself.

We showed how to implement such a proactive LMS on the basis of a dynamic rules-based expert system. We also gave some examples of proactive rules declarations. Finally, we sketched how it looks like from the users' point of view.

Future work includes the design and the implementation of sets of rules (packages) dedicated to common users needs, as well as to elaborated intelligent analysis coded into proactive rules.

These sets of rules will be of two kinds: standard packages that one will be able to use as is, as well as abstract packages (templates) that one will have to tailor to its specific needs by using appropriate tools.

References

1. Brusilovsky, P.: A component-based distributed architecture for adaptive Web-based education. In: Hoppe, U., Vardejo, F., Kay, J. (eds.) Artificial Intelligence in Education: Shaping the Future of Learning through Intelligent Technologies. Proceedings of AI-ED'2004, Sydney, Australia, July 20-24, 2004, pp. 386–388. OIS Press, Amsterdam (2003)
2. Tennenhouse, D.: Proactive computing. Communications of the ACM 43(5), 43–50 (2000)
3. Salovaara, A., Oulasvirta, A.: Six modes of proactive resource management: A user-centric typology for proactive behaviors. In: Proceedings of the Nordic conference on human-computer interaction, ACM international conference proceedings series, vol. 82, pp. 57–60 (2004)
4. Zampunieris, D.: Implementation of a proactive learning management system. In: Reeves, Th.C., Yamashita, Sh. F. (eds.) Proceedings of E-Learn – World Conference on E-Learning in Corporate, Government, Healthcare, & Higher Education, Hawaii, USA (October 2006) ISBN 1-880094-60-6
5. Zampunieris, D.: Implementation of efficient proactive computing using lazy evaluation in a learning management system. In: m-ICTE - International Conference on Multimedia and Information and Communication Technologies in Education, Seville, Spain (2006)

APD-A Tool for Identifying Behavioural Patterns Automatically from Clickstream Data

I-Hsien Ting, Lillian Clark, Chris Kimble, Daniel Kudenko, and Peter Wright

Department of Computer Science, The University of York
Heslington, York YO105DD, United Kingdom
{derrick, Lillian.clark, kudenko, kimble, pcw}@cs.york.ac.uk

Abstract. Clickstream can be a rich source of data for analysing user behaviour, but the volume of these logs makes it difficult to identify and categorise behavioural patterns. In this paper, we introduce the Automatic Pattern Discovery (APD) method, a technique for automated processing of Clickstream data to identify a user's browsing patterns. The paper also includes case study that is used to illustrate the use of the APD and to evaluate its performance.

Keywords: Web Usage Mining, Clickstream Data, Browsing Behaviour, Traversal Pattern.

1 Introduction

A thorough understanding of how users navigate or browse through a site is important in the development of web-based applications, especially when including features such as recommendations, advertising or personalisation. This understanding is also critical for performing usability testing [2], performance evaluation and website redesign. Analysis of Clickstream data can aid our understanding of user browsing behaviour by providing detailed information on the patterns generated by users as they navigate through a website hosted on that server. However, the size and nature of Clickstream logs can make pattern detection and classification difficult and time-consuming.

This paper is a paper that looks at the application of KDD to website design. In our previous work [9], we have introduced the "Footstep" graph as a means of visualising user's browsing patterns from Clickstream data. In addition, we have also shown how certain types of patterns such as "Stairs", "Mountain", and "Fingers" could be indicative of navigational problems in a site, e.g. in [3]. Our next challenge was to automate pattern generation and classification in order to process large Clickstream data files efficiently. In this paper, we propose a methodology known as Automatic Pattern Discovery (APD) method. APD is based on the concept of sequential mining to analyse browsing behaviour by detecting basic browsing elements and defining patterns for each user.

This paper is organised into six sections. Section 1 introduces the Footstep graph; user browsing patterns are described in section 2 and, in section 3, the three step APD process is introduced. In section 4, the basic browsing elements and the algorithms of

B. Apolloni et al. (Eds.): KES 2007/ WIRN 2007, Part II, LNAI 4693, pp. 66–73, 2007.
© Springer-Verlag Berlin Heidelberg 2007

the pattern detection steps are described and a case study to discover browsing pattern is presented in section 5. The experiment results and the evaluation of the APD algorithm are also shown in this section. The conclusion is presented in section 6.

2 Footstep Graph and Users' Browsing Patterns

In [4] we introduced the "Footstep" graph, a visualisation tool for identifying user's browsing patterns. The Footstep graph is based on a simple x-y plot where the x-axis represents time (in seconds) and the distance between points indicates the time between two nodes. The y-axis represents the nodes on the user's browsing route and the changes in the vertical axis indicate a transition from one node to another. The Footstep graph (figure 1) not only indicates the time-trends of a user's browsing history but also illustrates the relationship between each browsing node, transforming complex and unorganised Clickstream data.

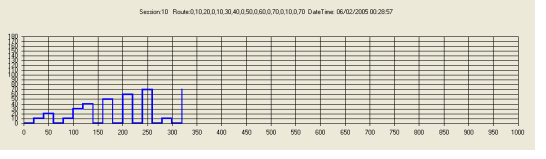
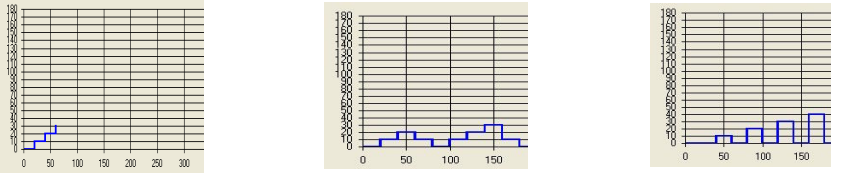


Fig. 1. A Sample Footstep Graph

2.1 Identifying User's Browsing Patterns

In our previous research, Footstep graphs were produced to demonstrate how this visualisation technique could help analyse user's browsing behaviour. From this research, we found certain frequently occurring patterns that were illustrative of user browsing behaviour. These patterns were named "Stairs" (including "Upstairs" and "Downstairs"), "Mountain" and "Fingers".

The Upstairs pattern is created when the user moves forward through previously unvisited pages in the website. An example of a Stairs pattern is shown in figure 2(a). These patterns are similar to Canter's "Path" pattern, which indicates that the user is exploring the website [1].



(a) (b) (c)

Fig. 2. (a) An example of a *Stairs* pattern (b) An example of a *Mountain* pattern **(c)** An example of a *Fingers* pattern

The Mountain pattern is a Downstairs pattern is immediately followed by an Upstairs one. An example of a Mountain pattern is shown in figure 2(b). This pattern is equivalent to Canter's "Loop" pattern, which indicates that the user is searching the site for a specific target. The Fingers pattern is found when the user moves directly from one page within the site to another and then directly returns to the original page see figure 2(c). This pattern is equivalent to Canter's "Spike" pattern and indicates that the user may have fallen into a browsing loop.

3 Automatic Pattern Discovery (APD)

The complete APD process contains three main steps: Clickstream Data Pre-processing, Browsing Route Transformation and Automatic Pattern Discovery. In turn, the automatic pattern discovery step contains three further sub-steps: browsing route segmentation, pattern definition and pattern detection. Figure 3 shows the whole APD process; the individual steps are discussed in detail below.

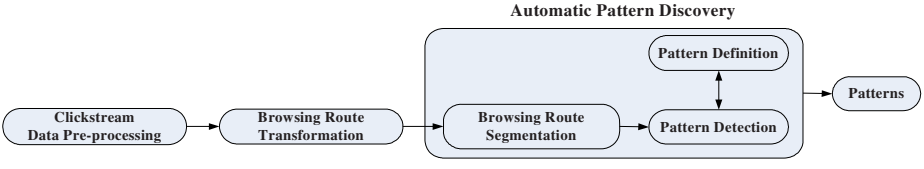


Fig. 3. The APD process

3.1 Clickstream Data Pre-processing

Clickstream data pre-processing is a necessary step not only for APD but also for any web usage mining technology. A standard data pre-processing process has been well-developed in current web usage mining research [4]. In general, the process should include the following steps: data filtering, data cleaning, user identification, session identification, 'bot' detection and data formatting and path restoration [9].

3.2 Users' Browsing Route Transformation

To discover a user's browsing pattern the browsing route must first be transformed to a number-based sequence. For example, a user's browsing route is shown in table 1. After transformation, a number-based sequence emerges as shown in table 2. In this case, the sequence starts from 0 and the increasing order of the sequence is 10. The transformation algorithm will search each node in the user's browsing route and assign each one a sequence number. In addition, the algorithm will also check whether each node has occurred before in the user's browsing route. If so, the algorithm will assign the same sequence number to the same nodes (e.g., the node No.3 /support/index.php has already occurred in the user's browsing route and has a sequence number '0', so the sequence number for the node No.3 is also '0').

Table 1. A sample user browsing route

No.	Date and Time	Accessed URL
1	15/06/2005,02:24:16	/support/index.php
2	15/06/2005,02:25:18	/support/contact.php
3	15/06/2005,02:26:20	/support/index.php

Table 2. A number-based sequence and time duration after the browsing route transformation

Number-Based Sequence	Time Duration	Accessed URL
0	0	/support/index.php
10	62	/support/contact.php
0	62	/support/index.php

3.3 Automatic Pattern Discovery

In sequential mining, the focus is always on the segmentation or transformation method of the sequence. A good segmentation method can help the sequential mining to generate better results. In APD, the segmentation method is easier than other sequential mining methods. As the APD methodology compares the sequence order of each node in entire route and the relationship between every two nodes, the segmentation method is to divide each node of a route into individual element. For example, a route A={0→10→20→30} (the arrow symbol in A means from one node to another node) will be divided to A'={0,10,20,30}, and a segmented route will never look like A2'={0,10,20→30}, because the continuous nodes are not allowed in the route after segmentation. The segmented route can then be used for the pattern detection step.

4 Basic Users' Browsing Elements, Pattern Detection and Pattern Definition

4.1 Basic Browsing Elements

(1) Level-1 Elements: Same, Up and Down

The segmented browsing route is transformed to basic level-1 elements by comparing the relation between every two nodes. These elements are known as "Same", "Up" and "Down". A Same element occurs when user continuously browses the same web pages by either refreshing a page, opening the same page or other activity. If there are two nodes in the route {ai, ai+1} and ai=ai+1 (e.g. {2, 2}) then the relationship between these two pages will be assigned the level-1 element Same (Figure 4). An Up element occurs when the user has moved through the website by using forward browsing behaviour, i.e. moving from one web page to another web page they have not yet visited. (Figure 5). A Down element occurs when the user has moved through a website by using backward browsing to one they have visited before, causing the order of the sequence to be lower. If there are two nodes {ai, ai+1} in a route and ai>ai+1, then the relationship between these two pages will be assigned the level-1 element Down (Figure 6).

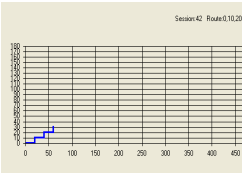


Fig. 4. The *Up* element

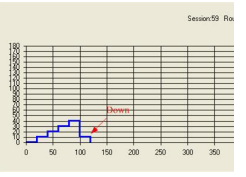


Fig. 5. The Down element

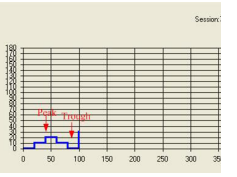


Fig. 6. *Peak* and *Trough* elements

(2) Level-2 Elements: Peak and Trough

The next level of browsing elements, Level-2 elements, are based on measuring the relationship between each contiguous element of the browsing route to discover changes in browsing directions or turning points. These Level-2 elements, Peak and Trough, are be used to define these turning points.

A Peak occurs when the browsing direction shifts from a forward to a backwards direction. For example if there are two nodes {bi, bi+1} in a route, and bi= Up and bi+1=Down (e.g., {Up, Down}) then the relationship between these two nodes will be assigned the level-2 element Peak. A Trough occurs when the browsing direction shifts from backwards to forwards. (Figure 7).

Table 3. Level-1 route transformation

```
For each Raw Users' Browsing Node Ai
  If Ai ≠ last node of route then
    If Ai<Ai+1 then
      Level-1 elements Bj='Up' (j=0…n)
    Else if Ai>Ai+1 then
      Level-1 elements Bj='Down'
    Else if Ai=Ai+1 then
      Level-1 elements Bj='Same'
```

Table 4. Level-2 route transformation

```
For each Level-1 Node Cp
  If Cp ≠ last node of route then
    If Cp='Up' and Cp+1='Down' then
      Level-2 elements Dq='Peak' (q=0…m)
    Else if Cp='Down' and Cp+1='Up' then
      Level-2 elements Dq='Trough'
```

4.2 Pattern Detection and Pattern Definition

(1) Level-1 and level-2 Element Detection

The pattern detection step of the APD sequentially processes the route based on the concept of level-1 and level-2 elements. The algorithms for level-1 and level-2 browsing route transformations are shown in table 3 and 4.

(2) Pattern Definition and Detection

Before pattern detection, it is essential for the website designer to define what patterns are of interest. However, the way in which the patterns are defined may vary. For example, some stakeholders may define a Mountain pattern as being {Up, Peak, Down, Trough, Up} rather than simply {Up, Peak, Down}, or maintain that a Fingers pattern should contain at least two fingers rather than one finger only {Trough, Peak, Trough}. The pattern definition must therefore be flexible as there is no standard definition for any one pattern. Once elements have been detected and desirable

patterns defined, the pattern detection step can detect these patterns automatically by matching the routes and the pattern rules according to the pattern detection algorithm.

5 Case Study and Experiment Result

In this section, we describe a case study in which the APD was used to discover browsing patterns automatically. Some results of the case study are shown in this section and the performance of the APD algorithm is evaluated [3].

5.1 Two Examples of Pattern Definition

For this case study, the researcher was studying student's use of a website related to one of the courses they were studying. The rules for Upstairs, Downstairs, Mountain, and Fingers patterns, based on the level-2 based routes, were defined as follows:

a. The *Mountain* and *Fingers* pattern
If there are continuous elements {Up, Peak, Trough}, {Up, Peak, Down}, {Trough, Peak, Down}, {Up, Peak}, and {Peak, Down} in the level-2 based route, then they should be recognised as a Mountain pattern. If there are continuous elements {Peak, Trough}, in the level-2 based route, then they should be recognised as a Fingers pattern. All remaining {Peak} elements in the level-2 based route should be recognised as a Fingers pattern.

c. The Downstairs and Upstairs pattern
All remaining {Down} elements in the level-2 based route should be recognised as a Downstairs pattern. All remaining {Up} elements in the level-2 based route should be recognised as an Upstairs pattern.

5.2 Experiment Result

After the pattern rules were defined, the APD method was used to analyse the Clickstream data. We used the standard data pre-processing approach; including a pattern restore method [9], to pre-process the raw Clickstream data. Then the pre-processed Clickstream data was transformed to number-based sequence and each route in the Clickstream data was segmented (table 5).

The segmented routes were then transformed to level-2 routes by using level-1 and level-2 based route transformation (Table 6). Finally, the patterns were automatically identified and categories (table 7).

Table 5. Segmented browsing routes after transformation

Session Number	Number-based sequence
1	0,1,2
2	0,1,2,1,2
10	0,1,2,0,1,3,4,0,5,0,6,0,7,0,1,0,7

Table 6. Level-1 and level-2 based user's browsing routes

Session Number	Level-1 based browsing routes	Level-2 based browsing routes
1	up, up	up, up
2	up, up, down, up	up, peak, trough, up
10	up, up, down, up, up, up, down, up, down, up, down, up, down, up, down, up	up, peak, trough, up, up, peak, trough, peak, trough, peak, trough, peak, trough, peak, trough, up

Table 7. The final users' browsing patterns

Session Number	Patterns
1	Upstairs
2	Mountain, Upstairs
10	Mountain, Mountain, Finger, Finger, Finger, Finger, Upstairs

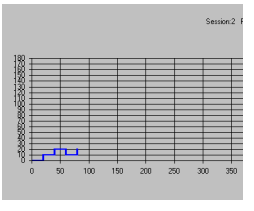


Fig. 7. Footstep graph of session 2

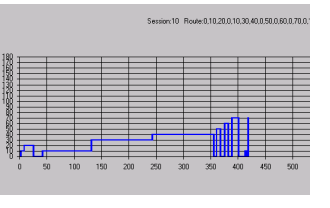


Fig. 8. Footstep graph of session 10

5.3 Performance Evaluation

A series of experiments was held to evaluate the performance of the APD algorithm. Tests were run using cleansed Clickstream ranging from 10 user sessions up to 1000 user sessions. The size of the cleansed Clickstream data files averaged 348 Kb for 500 sessions and 802Kb for 1000 sessions. The performance of the APD algorithm is shown in figure 9, where the x-axis represents the number of sessions per file and the y-axis represents the time in seconds.

According to figure 9, the execution time of the APD increases with number of sessions, and presents stable equimultiple increasing. Different average numbers of

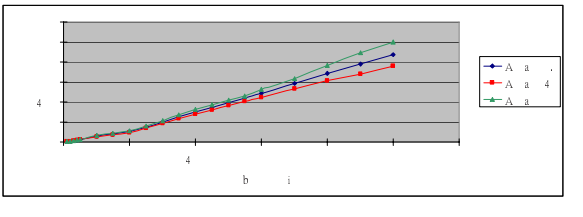


Fig. 9. Performance evaluation of the APD algorithm

browsing nodes do not affect performance. This shows the performance of the APD is under control and can be scaled-up to handle large volumes of data.

6 Conclusion

In this paper, described a novel sequence mining approach called APD, which was used to identify pre-identified patterns automatically. The paper also introduced the concepts of level-1 and level-2 elements of browsing behaviour, which constitute a vital part of the APD. This paper has shown that, by following the APD method, browsing patterns can be automatically identified efficiently, even when large amounts of Clickstream data are involved. We believe that such an approach can be an important tool for both web usage mining and for HCI research into browsing behaviour.

References

1. Canter, D., Rivers, R., Storrs, G.: Characterising users navigation through complex data structures. Behaviour and Information Technology 4(2), 93–102 (1985)
2. Chi, E.H.: Improving web usability through visualisation IEEE Internet Computing, March-April 2002, pp. 64–71 (2002)
3. Clark, L., Ting, I., Kimble, C., Wright, P., Kudenko, D.: Combining Ethnographic and Clickstream Data to Identify Browsing Strategies Information Research 11(2), paper 249 (2006), Available at http://InformationR.net/ir/11-2/paper249.html
4. Cooley, R., Mobasher, B., Srivastava, J.: Data Preparation for Mining World Wide Web Browsing Patterns. Journal of Knowledge and Information System 1(1), 5–32 (1999)
5. Cooley, R., Tan, P.N., Srivastava, J.: Discovery of Interesting Usage Patterns from Web Data, LNCS Vol. In: Masand, B., Spiliopoulou, M. (eds.) Web Usage Analysis and User Profiling. LNCS (LNAI), vol. 1836, pp. 163–182. Springer, Heidelberg (2000)
6. Eick, S.G.: Visual Analysis of Website Browsing Patterns. In: Borner, K., Chen, C. (eds.) Visual Interface to Digital Libraries, pp. 65–77 (2002)
7. Ezeife, C.I., Lu, Y.: Mining Web Log Sequential Patterns with Position Coded Pre-Order Linked WAP-Tree, Data Mining and Knowledge Discovery, 10, 5–38 (2005)
8. Ting, I.H., Kimble, C., Kudenko, D.: Visualising and Classifying the Pattern of User's Browsing Behaviour for Website Design Recommendation. In: Paper presented at the International Workshop on Knowledge Discovery in Data Stream, Pisa, Italy, vol. 24, pp. 101–102 (September 2004)
9. Ting, I.H., Kimble, C., Kudenko, D.: A Pattern Restore Method for Restoring Missing Patterns in Server Side Clickstream Data. In: Zhang, Y., Tanaka, K., Yu, J.X., Wang, S., Li, M. (eds.) APWeb 2005. LNCS, vol. 3399, pp. 501–512. Springer, Heidelberg (2005)
10. Xing, D., Shen, J.: Efficient Data Mining for Web Navigation Patterns. Information and Software Technology 46(1), 55–63 (2004)
11. Yen, S.J., Lee, Y.S.: An Efficient Data Mining Algorithm for Discovering Web Access Patterns, In Zhou, X. In: Zhou, X., Zhang, Y., Orlowska, M.E. (eds.) APWeb 2003. LNCS, vol. 2642, pp. 187–192. Springer, Heidelberg (2003)

Breedbot: An Edutainment Robotics System to Link Digital and Real World

Orazio Miglino[1,2], Onofrio Gigliotta[2,3], Michela Ponticorvo[1], and Stefano Nolfi[2]

[1] Department of Relational Sciences "G.Iacono", University of Naples "Federico II", Naples, Italy
[2] Laboratory of Autonomous Robotics and Artificial Life, Institute of Cognitive Sciences and Technologies, National Research Council, Rome, Italy
[3] Department of Psychology, University of Palermo, Palermo, Italy

Abstract. The paper describes Breedbot an edutainment software and hardware system that could be used to evolve autonomous agents in digital (software) world and to transfer the evolved minds in physical agents (robots). The system is based on a wide variety of Artificial Life techniques (Artificial Neural Networks, Genetic Algorithms, User Guided Evolutionary Design and Evolutionary Robotics). An user without any computer programming skill can determine the robot behaviour. Breedbot was used as a didactic tool in teaching Evolutionary Biology and as a futuristic toy by several Science Centers. The digital side of Breedbot is downloadable from www.isl.unina.it/breedbot.

Keywords: Edutainment, User Guided Evolutionary Design, Evolutionary Robotics.

1 Introduction

Nowadays, new-generation robots are widely used in Edutainment. There is a niche market that is fed by big and small industries which produce and sell didactic robotics kits. Between the many products available, a good example of this category are the Lego MindStorm kit (http://mindstorms.lego.com/) that is conceived mainly for schools and the small robot Surveyor (http://www.surveyor.com/) that is oriented to university and research center. Together with these kits, there are some robots that can be imagined as artificial pet companions: famous examples of this product line are Furby robots produced by Tiger Electronics (http://www.hasbro.com/furby/) and Paro (http://paro.jp/english/), a robot that is used also for cognitive rehabilitation. The main part of these robots presents an hardware structure that is sometimes very complex (formed by sensory apparatus with cameras, infrared sensors and ultra-sounds etc; by actuators as wheels, tracks, mechanical limbs and by powerful on-board computers), but this complexity is not counterbalanced on the behavioral side, as the behaviours they show, on the contrary, are quite rudimentary. Substantially we can observe in present edutainment robots a remarkable distance between their "bodies"

B. Apolloni et al. (Eds.): KES 2007/ WIRN 2007, Part II, LNAI 4693, pp. 74–81, 2007.
© Springer-Verlag Berlin Heidelberg 2007

sophistication (hardware) and poverty of their "minds" (Artificial Intelligence software underlying behaviors production).

The growth of this niche sector depends just on the individuation of methodologies and approaches that can reduce the gap between bodies and minds. In this work we describe a prototype made by us by which we try to apply Artificial Life techniques to robotics for Edutainment. In particular, it seems to us very promising to use the physical robot as a bridge between digital worlds created by Artificial Life and physical environments in which we and, obviously, the robots act. As it is known Artificial Life [3] is a discipline that tries to simulate in a digital world (software) the functional and structural features of biological systems. This discipline uses various techniques, models and methodologies such as Genetic Algorithms, Evolutionary Computation, Artificial Neural Networks, etc. that, in this framework, that is to say the A-life approach to technology, are used to build up digital objects by breeding, training or evolving them. Many applications are proposed in the context of User Guided Evolutionary Design [1] where a breeding digital environment allows users to "evolve" novel artificial objects. The mechanism is simple: on a computer screen a set of variants of the object users are seeking to evolve is proposed between which users can choose a subset of variants. The system then produces a new set of variants. The selection/production process proceeds until users have obtained the desired objects. There are applications in the field of evolution of artistic images [12], training human face profiles for crime investigations, objects design for furniture firms, etc. Most of those systems produce software images for static objects (cfr. for artistic images Artificial Painter [4], [8]). Moreover, within the broader field of Artificial Life, researchers are developing models and techniques to evolve actual robots [7].Our system, Breedbot, is built up by integrating the above cited A-life-like methodologies in order to permit children to evolve interesting mobile robots behaviors in an digital environment that can then be downloaded and tested on real robots. We attempted to try to avoid any form of "information processing methodology" (i.e.: programming) for children using hi-tech (i.e.: robots). This way the A-life techniques that are used are, to a certain extent, invisible for children who use it. Breedbot simulates the evolution of populations of digital agents (simulated robots) with elementary navigational capabilities. Based on Artificial Life and Evolutionary Robotics techniques, the digital environments allows the user to select agents with efficient sensory-motor behavior. This is done through a user- guided genetic algorithm (UGGA) that permits to select robots that are controlled by a neural network (i.e.: the model of the brain of an artificial pet). For the user- guided genetic algorithm the basic idea consists in showing the ongoing process (robots behaviors) on a computer screen (digital environment) and allowing the user to judge the robots actions, as a teacher or a breeder does in everyday life. In the following sections, 2 and 3, we describe the Breedbot system in detail. In section 4 we present some preliminary applications while section 5 is devoted to identify Breedbot limits and future directions for its improvement.

2 Breedbot: An Environment for Breeding Robots

Breedbot is an integrated hardware/software system that allows users with no technical or computer experience to breed a small population of robots. We will now describe Breedbot hardware and software in detail.

2.1 Breedbot's Hardware

Breedbot is shown in Fig.1a. It is built using motors, infra-red sensors, bricks and an on-board computer from the Lego Mindstorms kit. It is rectangular in shape. Its base measures 16*15 cm and it is 10 cm high. To move, it uses two large drive wheels, each controlled by its own small electric motor. Two wheels provide stability. All the wheels are fixed so there is no steering mechanism. The on-board computer (a Lego Mindstorms RCX) and the electric power supply are located on top of the motors. The sensor system - three Mindsensor infrared sensors - is placed above the on-board computer. The first sensor is mounted half way along the robot's short side and points in its direction of motion. The other two are fixed half way along the long sides. Each sensor produces a signal with a strength inversely proportional to its distance from an obstacle. The sensors can detect obstacles up to a maximum range of 15 cm.

2.2 Breedbot's Software

On software side, Breedbot uses a digital environment to simulate a process of artificial evolution that allows users with no technical or computer experience to breed a small population of robots. At the beginning of each simulation, the computer screen shows a first generation of robots in action. After a certain time, some of the robots are selected to produce offspring. Users can let the system select the "best robots" or make the decision themselves. If the system makes the decision, it rates the robots by their ability to explore the environment, and selects those with the highest scores. Human users, on the other hand, simply choose the robots they think have performed best. Once the selection procedure is over, the system creates clones of the selected robots. During this process it introduces random mutations into their control systems. The robots created in this way constitute a new generation. This selection/cloning/mutation cycle can be iterated until the "breeder" finds a particularly capable robot. At this point the brain of the simulated robot can be uploaded to a real robot and the user can see how it performs in the real world. Figure 1a shows a robot which has just received a "brain" developed with Breedbot. Figure 2 show's Breedbot's graphical interface.

On the left side there are the 9 robots while they explore a rectangular arena with walls. On the right side there is a display that indicates the current generation, radio buttons to choose between human and machine selection, STOP and RESET buttons and a graph, updated after each generation, that shows the mean fitness obtained in the exploration task along generations. The left hand side of the screen displays the behavior of the nine simulated robots in the arena,

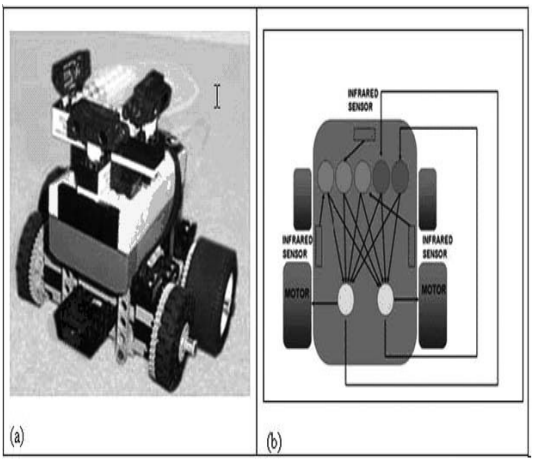


Fig. 1. 1a. A picture of the robot built with the Lego Mindstorms kit. 1b. A schematic representation of the robot and of its control system.

which is surrounded by walls. The right hand side provides information about the state of the system (the number of the current generation, a graph showing changes in the mean fitness of the population) along with a number of commands allowing the user to stop the system and to choose between human and artificial selection. The pull-down menu in the top left corner contains system utilities (to change the geometry of the environment, save configurations etc.).

Breedbot is designed to be easy to use for breeders of small robots. Breeders can use the system's graphical interface, to organize their own experiments in artificial evolution and if they want, they can select the individuals which will be allowed to reproduce. They can stop the program at any time, choose what they consider to be a well-adapted robot, and use the infrared port to upload its control system (its Artificial Neural Network) to a real Lego MindStorms robot (see Figure 3).

3 Artificial Life Techniques in Breedbot

Up to now, we have described the system without considering what is going on "inside", an aspect that is not directly perceivable by users. In this section we will open this black box to describe the Artificial Life techniques that are employed in the system.

The on-board computer on the hardware side and the Breedbot simulator on the software side, implement an Artificial Neural Network that is the system that controls the robot. The network, a simple perceptron, receives sensory stimuli from the infra-red sensors, processes the data and activates the robot's motors. The system's neural architecture consists of a layer of input neurons and a layer of output neurons (see Figure 1b). The input neurons receive stimuli from the

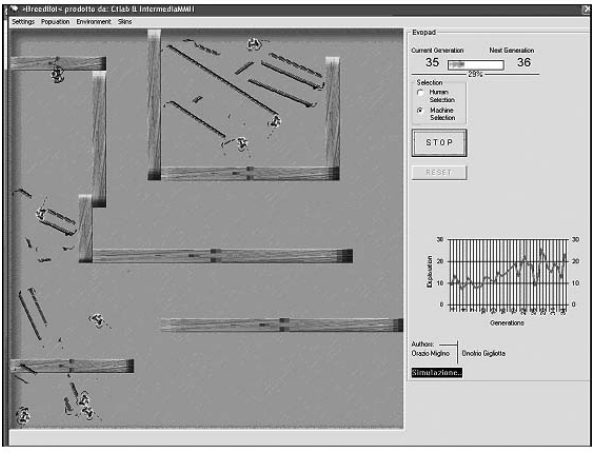


Fig. 2. A snapshot representing the Graphical Interface of Breedbot software

sensors and transmit these signals, through one-directional links ("connections") to the output neurons. Each connection is associated with a transfer value (its "weight"). This way, the signal arriving to the output neurons is filtered by the weights of the connections from neurons in the input layer. The input layer is made up of three sensor neurons, two proprioceptor neurons and two bias neurons. Each infrared sensor is associated with a single sensor neuron which receives its signal and activates the rest of the network. The two proprioceptor neurons have recurrent connections from the motor neurons (see Figure 1b). Thus the state of these neurons at time t+1 reflects the state of the motor neurons at time t. Finally, the bias neurons are always "on" (they always have an activation of 1). These neurons, which do not receive any kind of signal from the external environment, play an essential role, ensuring that the robot is always able to move, even when receives no input from the sensors. The output layer consists of two motor neurons: these neurons determine the robot's behavior at any given moment. Each motor neuron controls an electric motor. Its output is regulated by a threshold activation function. If the sum of the inputs to the neuron is equal to or higher than the threshold the neuron produces an output of "1". For values below the threshold, the output is 0. When a motor receives a "1" it turns clockwise for 2 seconds. When it receives a "0" it does nothing. In this way, the robot has three possible behaviors: it can move forward for 3 cm (when both motors are on); it can turn 10 degrees to the left (when the right motor is on and the left motor is off); it can turn 10 degrees to the right (when the right motor is off and the left motor is on). The software simulator replicates the physical characteristics of the robot and the training arena. Using the simulator we can conduct artificial evolution experiments with a population of 9 simulated robots. This aspect is based on Genetic Algorithms and Evolutionary Robotics. In terms of size, sensors, motors, and neural architecture, each individual in the population is identical to all the others. Only the weights of

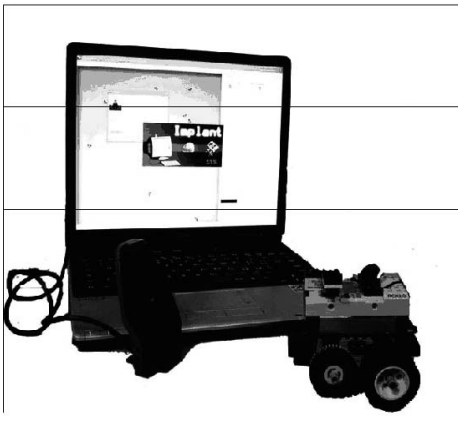


Fig. 3. The transfer of the control system (an Artificial Neural Network) from the Digital Environment to the real Lego MindStorms robot, through the infrared port

the connections in their control systems distinguish them. These are stored in their "genotypes". When breeding begins (the first generation of robots), the weights of the connections are extracted randomly from a uniform distribution in the range -1 - 1. For a certain time, the robots are allowed to move freely. Then the "breeder" (either the system or a human being) selects three robots for reproduction. Each robot's genotype (the values of its connections) is cloned three times, producing three offspring. But the clones are not perfect. During the copying process 3 per cent of the weights "mutate". The choice of which weights modify and the new value of the modified weight are random.

4 Preliminary Breedbot Application

The preliminary application of Breedbot has regarded two main spheres, both concerning Edutainment. First of all Breedbot has been used in various spreading exhibition in Italy, such as the Festival of Science in Genoa and in Rome or Futuro Remoto in Naples. In these demonstrations, Breedbot has been used by many children that could use both the software and the hardware of Breedbot prototype. Just relying on qualitative observations, we noticed that the joint use of the digital content and the physical robot led a better understanding of the themes that were meant to introduce.

For this reason we have run a pilot experimentation [6] in a school to verify if using software and hardware together could produce better results in learning. It is, in fact, well-known that computer simulation can be a powerful teaching tool, because they allow the learner to manipulate the most important variables involved in a process and observe what are the results [5], [10], [13]. It is not very much investigated if the use of simulation (digital content) and robots can enhance learning. So we compared two groups of students: in the first one the

students interact with a traditional multimedia hypertext (the control group), while in the other one the students could also manipulate the Breedbot prototype. These groups were formed by 22 Italian high school students (average age = 14.5 years) both attending a Biology Course. The questionnaire that was used to evaluate the students was developed together with biology teachers in order to measure the students competence in basic concepts of evolutionary biology. The questionnaire has 14 items with 4 different answers, of which only one is correct.

The following graph shows the preliminary results regarding Breedbot application in this context. The results of this pilot study indicated that the use of

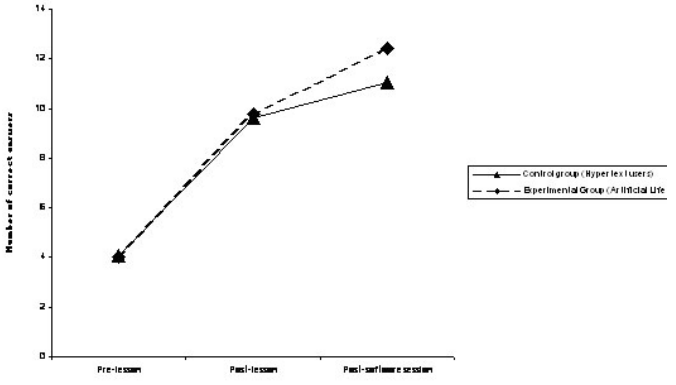


Fig. 4. The graphs show the average performance of the control group (solid line) and the experimental group (dotted line) across time. The subjects were tested on a questionnaire regarding evolutionary theory before a standard lesson, after the lesson, and once again after using either a hypertext tool (control group) or the artificial life tool Breedbot (experimental group).

Breedbot system, both software and hardware, improves learning (in the present case about a curricular subject such as Biology).

5 Conclusions and Future Directions

The preliminary results presented above suggest that the use of an integrated software-hardware system as Breedbot can be very useful in Edutainment. What we would like to stress is that its use allows to focus on the gap-link between digital and real world. As the process runs in simulation and can be transferred in reality, it makes it possible for users to ask about the limits and potentials both in simulation and in reality.

This is, of course, just a suggestion: the issue will be faced up in more detail in the next investigating efforts. Another issue we aim at tackling regards the extension of the Breedbot system. We now intend to improve the system

to handle with different robots. In particular it will be particularly interesting to try the system with the Surveyor robot (www.surveyor.com) that is quite different in size and shape and is provided with different actuators. The robot built with Lego Mindstorms Kit has two wheels, while Surveyor is a crawler, that means that it has tracks to move. This modifies completely the robot interaction with the external environment, thus underlying, once again, the importance of transferring digital agents in real world.

References

1. Bentley, P.: Evolutionary Design by Computers. Morgan Kaufman Academic Press, San Francisco (1999)
2. Jonassen, D.H.: Modeling with Technology: Mindtools for Conceptual Change, Merrill/Prentice-Hall, OH (2006)
3. Langton, C.G.: Artificial Life: An Overview. The M.I.T. Press/A Bradford Book, Cambridge, MA (1995)
4. Lund, H.H., Pagliarini, L., Miglino, O.: The Artificial Painter. In: Proceedings of ICANN 1995 International Conference on Artificial Neural Network, 5-13/6/95, Granada, Spain (1995)
5. Milheim, W.D.: Computer-Based Simulations in Education and Training: A Selected Bibliography. Educational Technology Publications Selected Bibliography Series, 8 (1992)
6. Miglino, O., Rubinacci, F., Pagliarini, L., Lund, H.H.: Using artificial life to teach evolutionary biology. Cognitive Processing: International Quarterly of Cognitive Science 5, 123–129 (2004)
7. Nolfi, S., Floreano, D.: Evolutionary Robotics: The Biology, Intelligence and Technology of Self-Organizing Machines. MIT Press/Bradford, Cambridge, MA (2000)
8. Pagliarini, L., Lund, H.H.: Art Robots and Evolution as a Tool For Creativity. In: Bentley, P., Corne, D. (eds.) Creative Evolutionary Systems, Morgan Kaufman, Seattle (2000)
9. Pan, Z., Aylett, R., Diener, H., Jin, X., Göbel, S., Li, L.: Technologies for E-Learning and Digital Entertainment. LNCS, vol. 3942. Springer, Heidelberg (2006)
10. Pappo, H.A.: Simulations for Skills Training. Educational Technology Publications (1998)
11. Singh, H.: Building effective blended learning programs. Educational Technology 43(6), 42–51 (2003)
12. Sims, K.: Artificial Evolution for Computer Graphics. Computer Graphics 25(4), 319–328 (1991)
13. Towne, D.M.: Learning and Instruction in Simulation Environments. Educational Technology Publications (1995)

Develop e-Learning Platform for Reinforcement Learning on Temperature Sensor

Wen-Jye Shyr

Department of Industrial Education and Technology,
National Changhua University of Education,
No. 1, Jin-De Road, Changhua 500, Taiwan, R.O.C.
shyrwj@cc.ncue.edu.tw

Abstract. This paper presents the development of e-learning platform for reinforcement learning on temperature sensor experimental module based on the graphical monitoring and control system. The e-learning platform in its current configuration aims at enabling experience of students in temperature sensor. It is integration graphical monitoring and control with four different temperature sensor devices. In this platform, the students can realize not only the characteristics, function and application of different temperature sensors but also learn to use the graphical monitoring and control system. In advance, the students can design e-learning platform to monitor and control the sensor module. The students learn, verify and realize lecture concepts by performing different sensor experiments in the laboratory via e-learning platform. They work in teams, design and implement their sensor systems and demonstrate their experimental results in orally present.

1 Introduction

Students have different learning styles and therefore can be classified as active, visual, inductive, sensor, and sequential [1-2]. Many studies have documented that traditional classroom teaching is not the best approach to teach college students. Traditional methods to the students on learning sensor are principle exploratory. Therefore, a new innovative teaching and reinforcement learning pedagogy is needed. As a result, educational institutions have started new approaches to enhance student learning [3-6]. The e-learning platform, the topic of this paper, integrates text, sensor devices, and graphical monitoring and control system. The main advantages of e-learning platform were reported to be the added value in organizational matters, such as easier access to information and the flexibility regarding time and place.

Sensor course aims to train the students to use for industrial process control, measurement, and automation, often involving light, temperature, pressure, flow, and level measurement. A survey about the research work in the sensor's laboratory at the TU Budapest was reported [7-8]. The individual areas are under research and development, and the results demonstrated here are not always complete. Ahmad [9] reported a laboratory experiment to teach some concepts on sensor-based robot assembly systems. These sensor-based motion strategies improve the reliability of a robot

B. Apolloni et al. (Eds.): KES 2007/ WIRN 2007, Part II, LNAI 4693, pp. 82–89, 2007.
© Springer-Verlag Berlin Heidelberg 2007

assembly system. The sensor-based assembly experiment, the laboratory setup, and the sensor-based motion strategies are discussed in detail. Peyton et al. [10] takes the view that the principles of sensor operation and instrumentation are important subjects to an electronics/mechatronics undergraduate course and are best presented as early as possible within the overall degree program. More generally, evaluation provides feedback about the effectiveness of e-learning platform [11].

This paper addresses one set of sensor experimental module based on the graphical monitoring and control system initiated in the department of industrial education and technology at National Changhua University of Education. First of all, this system gleans and sorts the data of several different sensors, then integrated one set of sensor experimental module via the make projects of students. The second, this system edit the e-learning platform of graphical monitoring and control system. Finally, it connects the software of graphical monitoring and control system with the hardware of different temperature sensor elements to achieve one set of experimental module. The experiment is developed to offer students with a team project experience of practice-oriented sensor and graphical monitoring and control system technology.

2 Description of Temperature Sensors

E-learning environments offer the potential to improve learning. Numerous e-learning applications have been developed to demonstrate laboratory experiments [12-13]. The development of a temperature sensor and graphical monitoring and control system offers new opportunities to enhance the quality of existing sensor teaching course.

A complete module in temperature sensor and graphical monitoring and control system fundamentals is through to an introduction to control. With this platform, the students can explore the principles and applications of different temperature sensor devices, all commonly found in industry.

Temperature sensors are often used within electronic systems to monitor temperature and provide protection from excessive temperature excursions. The most common temperature sensors for use within system are listed as follows: thermocouple, RTD, thermistor and AD590.

2.1 Thermocouple

Thermocouple uses the seeback effect, that is where two dissimilar metal alloys produce an Electro-Motive Force (EMF is a fancy way of saying "voltage") that is nearly proportional to temperature. Thermocouples are made by joining two wires of dissimilar metals. Thermocouple types differ by metallic composition and are designated by a single letter such as J, T, or K. With a known thermocouple type and the temperature can be calculated. The point of contact between the wires generates a voltage that is approximately proportional to temperature. It is not a linear relationship, however may be approximated as one. This voltage is amplified by signal conditioning equipment and then input to the computer through the DAQ card, thus turning a real life temperature measurement into a digital value on a computer that can be analyzed and presented. Fig. 1 shows the basic thermocouple circuit.

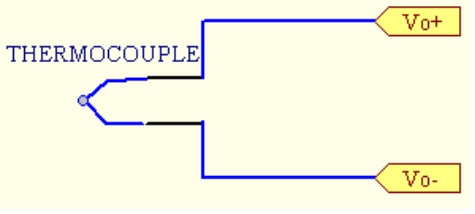


Fig. 1. The basic thermocouple circuit

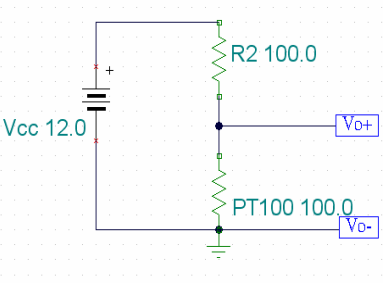


Fig. 2. The basic RTD (PT100) circuit

2.2 RTD (Resistive Temperature Detector)

Resistive temperature detector is made of coils or films of metals (usually platinum). When heated, the resistance of the metal increases; when cooled, the resistance decreases. By measuring this voltage, you determine its resistance, and thus determine its temperature. Resistance temperature detector, or RTD, is highly accurate temperature sensors. They are also known for their excellent stability characteristics. Due to their low resistance values, you must be careful with the RTD lead resistances. The basic RTD circuit is shown in Fig. 2.

2.3 Thermistor

Thermistor is temperature-dependent resistors, usually molded from conductive materials. The most common thermistor has a negative temperature coefficient (NTC) of resistance. Characteristics include moderate temperature range (up to 150°C), low-to-moderate cost (depending on accuracy), poor but predictable linearity, and the need for some signal conditioning. Thermistor is available in probes, in surface-mount packages, with bare leads, and in a variety of specialized packages. Maxim manufactures ICs that convert thermistor resistance to digital form. Thermistor has either a negative temperature coefficient (NTC) or a positive temperature coefficient (PTC). The first has a resistance which decreases with increasing temperature and the latter exhibits increased resistance with increasing temperature. A typical thermistor circuit is shown in Fig. 3.

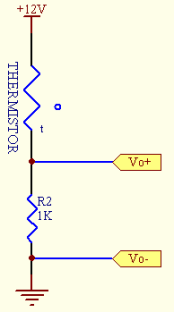


Fig. 3. A typical thermistor circuit

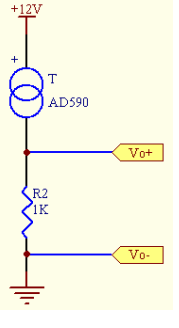


Fig. 4. The AD590 circuit

2.4 AD590

The AD590 is an integrated-circuit temperature transducer which produces an output current proportional to absolute temperature. The device acts as a high impedance constant current regulator, passing 1μA/°K for supply voltages between +4V and +30V. The inherent low cost of a monolithic integrated circuit combined with the elimination of support circuitry makes the AD590 an attractive alternative for many temperature measurement situations. Linearization circuitry, precision voltage amplifiers, resistance measuring circuitry and cold junction compensation are not needed in applying the AD590. In the simplest application, a resistor, a power source and any voltmeter can be used to measure temperature. In addition to temperature measurement, applications include temperature compensation or correction of discrete components, and biasing proportional to absolute temperature.

The AD590 is particularly useful in remote sensing applications. The device is insensitive to voltage drops over long lines due to its high-impedance current output. Any well insulated twisted pair is sufficient for operation hundreds of feet from the receiving circuitry. The output characteristics also make the AD590 easy to multiplex: the current can be switched by a CMOS multiplexer or the supply voltage can be switched by a logic gate output. The AD590 circuit is shown in Fig. 4.

3 e-Learning Platform System

E-learning platform, at its best, is learning that complements traditional methods and gives a more effective experience to the students on temperature sensor. Simply, e-learning platform is the use of technology to support the learning process. This section describes requirements of e-learning platform for the graphical monitoring and control system on temperature sensor experimental module.

3.1 General Description

(1) Work under this section includes providing graphical monitoring and control system equipment to provide the following functionality:
   (A) Monitoring of the sensor status,
   (B) Gathering and storage of sensor information,
   (C) Detailed analysis and trouble-shooting of the sensor system.
(2) An existing personal computer workstation provides for storage of collected data, system operation and running the analytical software.
(3) Sensor information shall be transmitted to and integrated into the building sensor experimental module based on the graphical monitoring and control system.

3.2 System Description

(1) The existing system includes sensor experimental module for graphical monitoring and control, device interface hardware, inter-communication wiring, personal computer workstation, software, interfacing, printer, and ancillary equipment.
(2) The existing system utilizes the high-speed backbone network.

(3) A personal computer workstation connected to the network has equal access to information provided by the graphical monitoring devices for centralizing data display, data logging, alarming, event recording, and other monitoring operations.
(4) This system allows direct access to data provided by the graphical monitoring devices for implementing different sensor automatic control.

4 Experiment Description

The major part of the system is dedicated to implement the temperature sensor experimental module based on the graphical monitoring and control system in laboratory sessions. Fig. 5 illustrates this experiment system. It is believed that the students take their time over "active learning" by making experiments is as important as take their time over "passive learning" through attending lectures. The same features come from some experiment or practice-oriented systems [14-15].

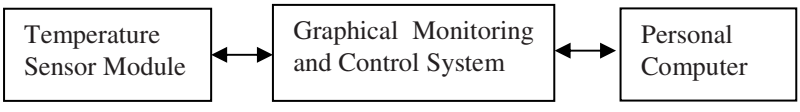


Fig. 5. Temperature sensor experimental module based on the graphical monitoring and control system

4.1 Thermocouple Sensor Module

The thermocouple sensor module is shown in Fig. 6. Without getting into a lot of boring theory, the problem with thermocouple is that once the students connect a pair of copper wires to them and run it to the model, the students created three dissimilar metal junctions, which introduce errors to the measurement. Also, a second junction at a known temperature is required for useful thermocouple temperature sensor measurements. In the past, that meant having an icebath out in the laboratory for the cold

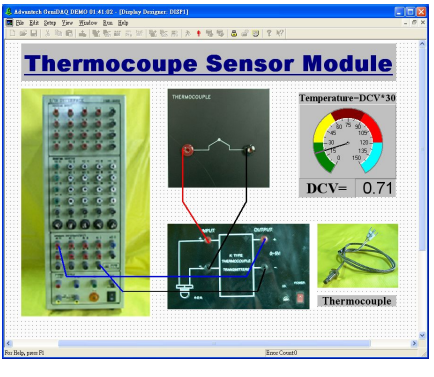


Fig. 6. Thermocouple sensor module

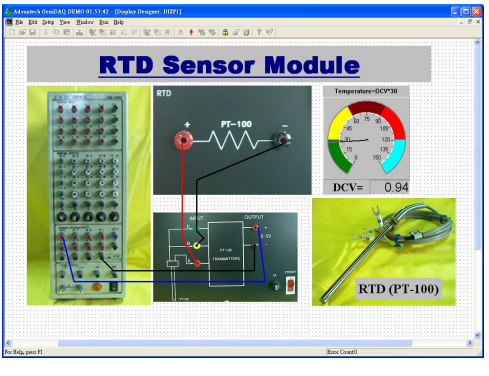


Fig. 7. RTD sensor module

junction. Thankfully, many DCV have a built-in cold junction for thermocouple temperature measurements; all you do is hook up the proper thermocouple and set the DCV to "temperature".

4.2 RTD Sensor Module

The RTD sensor module is shown in Fig. 7. The RTD (PT 100) is made from thin films of elemental metals, such as platinum. Because platinum has very low bulk resistivity, you need to create a thin film with a lot of squares to obtain a useful total resistance.

4.3 Thermistor Sensor Module

The thermistor sensor module is shown in Fig. 8. Thermistor is resistors (or perhaps semiconductors) that have predictable negative temperature coefficients. The negative temperature coefficients are abbreviated NTC. Thermistor is typically made of some type of ceramic. The resistance/temperature function of thermistor is not as linear as an RTD, but they are a whole lot cheaper. Note that when you choose a thermistor or RTD, you need to take into account the self-heating of the device due to current that you pass through it to obtain a voltage reading.

4.4 AD590 Sensor Module

The AD590 sensor module is shown in Fig. 9. This part provides a current in μA that is nearly exactly equal to its temperature in degrees Kelvin. This type of sensor is far less prone to errors due to lead resistance that a voltage sensor, thermistor or RTD might be.

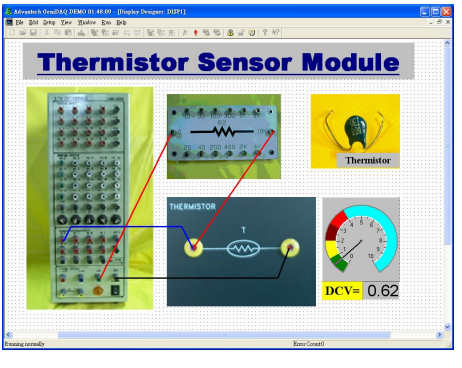


Fig. 8. Thermistor sensor module

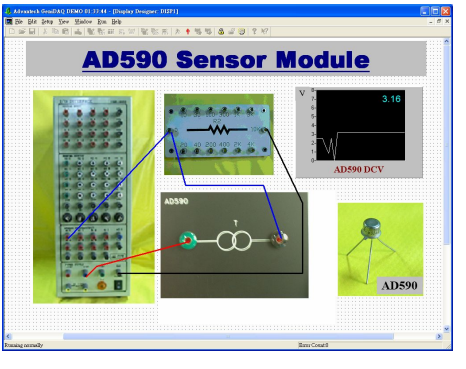


Fig. 9. AD590 sensor module

5 Evaluation

This innovative experiment and teaching methods have been inaugurated in the Department of Industrial Education and Technology at National Changhua University of

Education. The sensor course was attended by 38 students. Students were asked to complete a survey questionnaire in order to evaluate the e-learning platform. Feedback from the students was generally supportive of the sensor course contents and especially supportive of the e-learning introduced in the course. As the e-learning platform was designed to improve the practical sensor skills of students this is a positive result. Students indicated they would need further assistance to implement the methods from the course, in order to priority, from: case studies 70% more practical examples 54% course notes 38% and training assistance 22%. Students feel that they are able to learn more effectively in a practice-oriented sensor and graphical monitoring and control system technology. The design project, even though it is time consuming, is very well received by the students. Students completing this system have a real feeling of satisfaction and accomplishment since they realize that they have actually designed and constructed a sensor experimental module based on the graphical monitoring and control system via e-learning platform. Additionally, the result shows that how this intervention improves learning.

6 Conclusions

In the present study, an innovative concept of the e-learning system was been established and proved to be reinforcement to achieve an optimal approach in the understanding of temperature sensors. The proposed e-learning platform for temperature sensor experimental module, with its graphical user interface, created an experimental environment, which on its whole (integrating the various temperature sensors and visualization panels) provided adequate reinforcement learning elements. Both theory and practice are emphasized in this system. This e-learning platform sequence covers general sensor physics and technology as well as new and novel state of the art sensor design. It is anticipated that this system will bridge the gap between university education and industrial applications. Students work on a design project to implement and integrate an actual sensor device. The goal of the system sequence is to use our research and facilities in smart sensor development to give students a comprehensive background in the art and science of sensor and device technology, design, and integration.

Acknowledgement

This work based upon work supported by National Science Council, Taiwan, Republic of China under grants NSC 94-2516-S-018-010 and NSC 95-2516-S-018-005.

References

1. Felder, R.M.: Reaching the Second Tier-Learning and Teaching in College Science Education. Journal College Science Teaching 23(5), 286–290 (1993)
2. Felder, R.M., Silverman, L.K.: Learning and Teaching Styles in Engineering Education. Journal of Engineering Education, 574–581 (1988)

3. Rigden, J.S., Holcomb, D.F., Stefano, R.D.: The Introductory Physics Project. Physics Today 46, 32–37 (1993)
4. Barr, R., Pandy, M., Petrosino, A., Svihla, V.: Challenge-Based Instruction in an Engineering Technical Elective Course. In: Proceedings of the 2005 ASEE Gulf-Southwest Annual Conference (2005)
5. Burks, O.: Implementation of a Virtual Classroom in an Introductory Circuit Analysis Course. In: Proceedings of ASEE Annual Conference, vol. 1, pp. 187–191 (1995)
6. Christina, K., Lars, C.: Students' Perceptions of E-Learning in University Education. Journal of Educational Media 27 (2002)
7. Ramon, P.A., John, G.W.: Sensors and Signal Conditioning, 2nd edn. John Wiley & Sons, INC., Chichester (2001)
8. Harsanyi, G., Lepsenyi, I.: Sensors' Education and Related Student Research Projects. In: Proceedings of Electronic Components and Technology Conference, pp. 1034–1041 (2000)
9. Ahmad, S.: A Laboratory Experiment to Teach Some Concepts on Sensor-Based Robot Assembly Systems. IEEE Transactions on Education 31(2), 74–84 (1988)
10. Peyton, A.J., Burch, J.R., Bradshaw, A., Dorey, A.P.: The Teaching of Sensors and Instrumentation to Second Year Electronics/Mechatronics Students. IEE Colloquium on Sensors and Instrumentation Systems, Digest No: 1996-127, 11/1-11/3 (1996)
11. Unrau, Y.A., Gabor, P.A., Grinnell, Jr., R.M.: Evaluation in Social Work: The Art and Science of Practice. Oxford University Press, New York (2006)
12. Troy, J.L., Czapar, G.F.: Evaluating conservation practices using GIS and virtual reality modeling language. Journal of Natural Resources Life Science Education 31, 38–43 (2002)
13. Ramasundaram, V., Grunwald, S., Mangeot, A., Comerford, N.B.: Development of an environmental virtual field laboratory. Computers & Education 45(1), 21–34 (2005)
14. Halang, W.A.: A Curriculum for Real-Time Computer and Control Systems Engineering. IEEE Transactions on Education 33(2), 171–178 (1990)
15. Sloane, T.H.: Laboratories for an Undergraduate Course in Power Electronics. IEEE Transactions on Education 38(4), 365–369 (1995)

Modeling e-Learning System Performance Evaluation with Agent-Based Approach

Hsiao-Ya Chiu[1], Sheng Chieh Chung[2], and An-Pin Chen[3]

[1,2] The Department of Management Information, Yu Da College of Business
No.168, Hsueh-fu Rd, Chaochiao Township, Miaoli County, 361 Taiwan, ROC
kates.chiu@msa.hinet.net, sam-cc@umail.hinet.net
[3] Institute of Information Management, National Chiao Tung University,
1001 Ta Hsueh Road, Hsinchu, Taiwan 300, ROC
apc@iim.nctu.edu.tw

Abstract. Rapidly evolving information technology has dramatically changed the knowledge dissemination process. A proper e-learning environment is one of the most important knowledge tools in modern organizations. However, many of them lack a generic evaluation process to verify performance. In an attempt to solve this problem, this study propose an agent-based model which compose learning model, balanced scorecard and the option pricing approach to provide an dynamic, flexible framework for e-learning project's performance evaluations.

Keywords: E-learning, performance evaluation, option-pricing approach, Black-Scholes model, balanced scorecard approach, information agent systems.

1 Introduction

Most e-learning projects require huge amounts of organizational resource such as money, time, and human training. It is necessary to monitor the continuous performance change of e-learning project to help organizational managers making immediately decisions. However, it is difficult to evaluate e-learning performance because too many qualitative/quantitative factors exist. E-learning environment changes rapidly that decision makers have to dynamically adjust their development strategies. Thus, it is ideal to introduce concepts of risk management in financial engineering to help decision maker dealing with uncertainties with quickly response. This study propose an agent-based framework that combines balanced scorecard and real option analysis approaches to help organizations dynamically, automatically analyze e-learning project's performance and make fundamental decisions.

2 Preliminaries

The balanced scorecard (BSC) was first developed by in the early 1990s and became widely adapted for determining business strategies (Kaplan and Norton, [1][2][3]).

B. Apolloni et al. (Eds.): KES 2007/ WIRN 2007, Part II, LNAI 4693, pp. 90–98, 2007.
© Springer-Verlag Berlin Heidelberg 2007

The most widely adapted function of the BSC is the management performance system to align its vision and mission by demonstrating four different perspectives: financial, customer, internal process, and capacity. Some research indicates the possibility to apply BSC concept to plan and evaluate the learning project [4][12]. Meanwhile, this study determines the e-learning BSC perspectives based on Kirkpatrick's learning evaluation theory [5][6] which is based on four levels(reaction, learning, behaviour, and results) which is easy to mapping the general four visions of BSC.

However, BSC model cannot reveal the most important information: Will this project success? What is the key success/failure objective? Thus, this study applies ROA approach to answer these questions. The term "option" is a privilege sold by one party to another that offers the buyer the right, but not the obligation, to buy (call) or sell (put) the underlying asset at an agreed-upon price during a certain period of time or on a specific date. It has been explained by Cox and Ross (1976) that the option price is the "expected value" of payoff discounted at the risk-free interest rate over the risk-neutral distribution of the underlying asset [7]. Thus, the initial ROA framework focuses on the increased value of abandoning a project and liquidating the assets [8] that is based on the effectiveness of resource rearrangement to maximize the entire system performance [9]. This study finds out that an organization can use the traditional BSC model to perform top-down strategy planning and applies the ROA approach to evaluate the e-learning project's performance.

Agent-based systems have been well constructed for information gathering followed by performance monitoring in recent years [11] [13] [14][15], but yet been focused on learning performance. This paper proposes a framework of performance agents of e-learning systems with the main characteristics of agent-based systems such as autonomous, collaboration, communication, and dynamic alerts.

3 Constructing Agent-Based e-Learning Evaluation Framework

Supposing an organization is running an e-learning project that should be periodically examined (from hourly, daily, to monthly) for its performance monitoring for the decision makers who want to analyze the project's current status and ensure the success at deadline. The deadline can be an annual checkpoint for a long term project or be a checkpoint between development phases. Moreover, the decision makers also want to identify the key success and fail factors of the project. In this case, it is ideal for an organization to apply automatic agents to identify the e-learning project's perspectives and objectives in order to achieve the project's goals. Meanwhile, the 'satisfactory' level of each objective may change due to internal or external situations. Thus, this study applies ROA approach to analyze the BSC perspectives and detail objectives to help decision makers evaluate the e-learning project's performance under uncertainties.

3.1 The BSC Perspectives and Objectives

When using a BSC framework, a strategy maker has to firstly identify the organization's e-learning goals then properly define the major perspectives of the e-learning project. Each perspective has its objectives and targets that must be achieved. The four BSC perspectives identified by this study are Reaction, Learning, Behaviour, and Value.

Reaction evaluation is finding out how the learners feel about their learning experience. Learning evaluation is the measurement of the increase in knowledge, skill, and attitudes changed from before to after the learning experience. Behaviour evaluation is the extent to verify the behaviour change after the learners applied the knowledge they learned. Value perspective is typically the organizational key performance indicators such as volumes, values, percentages, timescales, return on investment, and other quantifiable aspects. But in most situations on e-learning planning, the "results" are generally intangible. This study follows the 'result' concept of Kirkpatrick model but replaces it with the term 'Value' to combine the general qualitative and quantitative factors into e-learning project evaluations.

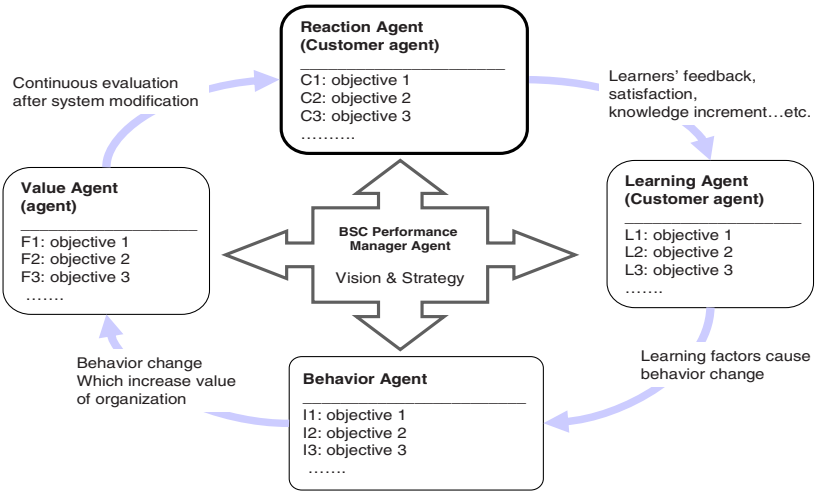


Fig. 1. Framework of e-learning performance agents

3.2 Applying the ROA Approach

This study uses ROA approach to obtain managerial information from BSC with Black & Scholes model (B & S model) [10] which have been widely adopted in real world applications to obtain option prices: (brief example also illustrated in Fig. 3).

Table 1. The notation for the B&S/BSC

Notation	Option pricing applications	BSC applications
C	The theoretical call price	The expected return of an objective that exceeds the target value
P	The theoretical put price	The expected return of an objective that fails to exceed the target value
S	Current value of the underlying asset	Current objective value
K	Exercise price of the option contract	The target value that an objective must exceed
σ	Volatility of the underlying asset	Standard deviation of an objective performance
T	Time to maturity	Time to the next checkpoint
r	Risk-less interest rate	The anticipated growth rate of an objective

Note that N(.) denotes the cumulated normal distribution, C is the call option price, P is the put option price, S is the current value of the underlying asset, K is the exercise price, T is the remaining time to maturity (in years), σ is the volatility of this underlying asset, and r is the risk-less interest rate.

The individual objective is measured periodically to investigate its score of satisfactory and then transferred into a measure to indicate its relative current value as the objective index. The standard deviation of a certain objective measure is called volatility. Each measure of a certain objective has its target value to be achieved after T years indicating the feasible performance (satisfactory degree) requirement. Different weighting values are assigned to indicate the importance of each objective while keeping the sum of all weighting values equal to 1. If we add or remove any objective we must readjust the sum of the existing weighting values to 1.

$$\sum_{i=1}^{n1} W_{Ci} = \sum_{i=1}^{n2} W_{Li} = \sum_{i=1}^{n3} W_{Ii} = \sum_{i=1}^{n4} W_{Fi} = 1 \text{ and } PW_c + PW_L + PW_I + PW_F = 1$$

Each objective is measured in relative value compared to the last measuring point. For example, the proper measurement will be "I am more satisfied with our e-learning system compared to last month: score 0 to 4" rather than "I am satisfied with our e-learning system: score 0 to 4".

The organization has to identify the investigating method and configure the weighting factors for indicating the importance of each perspective P_Γ and objectives W_Γ. The growth rate r_Γ can be greater (increase) or less (decrease) or equal (neutral) to zero. It is better to apply neutral perspective (r_Γ=0) to the growth rate terms unless the objective score can be properly forecasted. The perspective weights can be determined by the consumed resource (or cost) of each perspective. The analyst can initially determine the weight of each objective by its percentage of consumed resource to the perspective.

All questions must be designed so that the higher score indicates the higher performance. Then, collect data to evaluate the tangible objectives. All quantitative values are then compare to the last checkpoint and transform the performance into 0 to 4 scores for each objective so as to indicate the performance from strongly decreased (score = 0) to strongly increased (score = 4). The current measure is calculated with the following formula:

$$S_\Gamma = P_\Gamma(0.95 + X \times 0.025)$$

where $\Gamma \in \{C_1,..C_{n1}, L_1,..L_{n2}, I_1,..I_{n3}, F_1,..F_{n4}\}$

and

$X \in \{C_1,..C_{n1},L_1,..L_{n2},I_1,..I_{n3},F_1,..F_{n4}\}$

This formula indicates that the current measure will be 95% of the previous measure if the score is zero and will be 105% of the previous measure if the score is 4 (max variation is ±5%). For the first time issuing of the BSC, this study subjectively set P_Γ = 10.0 as its initial value. The 5% variance and initial value 10.0 is subjectively set by this study, an analyzer may flexibly adjust these values without influencing the final results. Meanwhile, the current measure is derived from the previous measure that assures the objective measure varies in a form of Xi+1/Xi to emulate the option's underlying asset's price variation in the real world. This study assumes that ln(Xi+1/Xi) follows a normal distribution and thus X follows a lognormal distribution as B&S model suggests.

The volatility can be derived as a standard deviation in nature log *ln(.)* from its previous measures. Assumes that the future volatility behaves like its previous n

measures, and then the periodically standard deviation can be calculated as $\sigma_\Gamma' = Std(P_{\Gamma,t}, P_{\Gamma,t-1}, ...P_{\Gamma,t-n})$, where *Std(.)* denotes the standard deviation function. Suppose that the BSC is periodically measured every *t* days, the annual volatility will be σ_Γ. For a quantitative objective, the objective score can be measured as (with continuous compound interest rate method):

$$m_1 = e^{\frac{\ln(\frac{v_1}{v_0})}{cp}}, m_2 = e^{\frac{\ln(\frac{v_2}{v_0})}{cp}} \qquad \sigma_\Gamma = \sqrt{\frac{365 \times (\sigma_\Gamma')^2}{t}}$$

$$score = 2 + \frac{cv - m_1}{(m_2 - m_1)/2}$$

where *cv* indicates the current incremental rate, *cp* indicates the checkpoint counts. For a qualitative objective, the objective score can be measured as:

$$m_1 = \frac{3 \times ct + 4 \times cs + 2(cp - ct - cs)}{cp}, m_2 = 4$$

$$score = 2 + \frac{cv - m_1}{(m_2 - m_1)/2}$$

where the objective's success point is measured as "at least *ct* satisfy counts and *cs* strongly satisfy counts cumulated" before deadline. Because the performance is evaluated by its slope (incremental rate), the target value K_Γ equals to the objective's initial value. As a result, both quantitative and qualitative objectives are normalized into the "satisfactory level" and thus can be summarized in the same scale.

An objective index is defined as a value that indicates the expected gain for each objective exceeds its target value after T years. An objective fail index is defined as the expected value that an objective fails to exceed its success condition.

$$\begin{cases} G_\Gamma = [S_\Gamma N(d1) - K_\Gamma e^{-r_\Gamma T} N(d2)] \times W_\Gamma \\ Z_\Gamma = [K_\Gamma e^{-r_\Gamma T} N(-d2) - S_\Gamma N(-d1)] \times W_\Gamma \end{cases} \quad \text{where} \quad d1 = \frac{\ln(\frac{S_\Gamma}{K_\Gamma}) + (r_\Gamma + \frac{\sigma_\Gamma^2}{2})T}{\sigma_\Gamma \sqrt{T}} \qquad d2 = d1 - \sigma_\Gamma \sqrt{T}$$

Comparing the expected success and fail values can generate the objective's success index. In order to provide a meaningful value, this study uses logarithm value V_Γ that positive means success (higher is better) and negative means fail:

$$V_\Gamma = \ln(\frac{G_\Gamma}{Z_\Gamma})$$

The perspective index value indicates the total performance summarized from its objective indexes:

$$PI_F = PW_F \times \sum_{i=1}^{n4} G_{Fi} \qquad PI_I = PW_I \times \sum_{i=1}^{n3} G_{Ii} \qquad PI_L = PW_L \times \sum_{i=1}^{n2} G_{Li} \qquad PI_C = PW_C \times \sum_{i=1}^{n1} G_{Ci}$$

The perspective fail index indicates the expected fail value that the entire perspective fails, which can be summarized from its objective fail indexes:

$$PZ_C = PW_C \times \sum_{i=1}^{n1} Z_{Ci}, \quad PZ_L = PW_L \times \sum_{i=1}^{n2} Z_{Li}, \quad PZ_I = PW_I \times \sum_{i=1}^{n3} Z_{Ii}, \quad PZ_F = PW_F \times \sum_{i=1}^{n4} Z_{Fi}$$

Similarly, the perspective's success index can be calculated as:

$$PV_C = \ln(\frac{PI_C}{PZ_C}), \quad PV_L = \ln(\frac{PI_L}{PZ_L}), \quad PV_I = \ln(\frac{PI_I}{PZ_I}), \quad PV_F = \ln(\frac{PI_F}{PZ_F})$$

The BSC index value can thus be summarized from four perspective indexes in (s1), also, the BSC fail index can be summarized from four perspective's fail indexes in (s2). Finally, the BSC success value can be calculated as (s3).

$$BSC\ index = PIC + PIL + PII + PIF \quad (s1)$$
$$BSC\ Fail\ Index = PZC + PZL + PZI + PZF \quad (s2)$$
$$BSC\ Success\ Value = ln(BSC\ index/BSC\ Fail\ Index) \quad (s3)$$

3.3 Design of Performance Agents

According the core concept of performance evaluation/monitor mentioned in 3.1 and 3.2, this study packages it into an agent-based system. Based on the definition of an information agent [11] "an autonomous, computational software entity that has access to one or multiple, heterogeneous and geographically distributed information sources…", the concept is suitable to develop a performance agent for continuous e-learning projects. A performance agent for e-learning systems is assumed to satisfy one or multiple of the following requirements: (1) autonomous information acquisition from on-line or off-line questionnaires for learner satisfaction (2) Information synthesis including quantitative and qualitative measurement (3) Intelligent user assistance for performance monitoring.

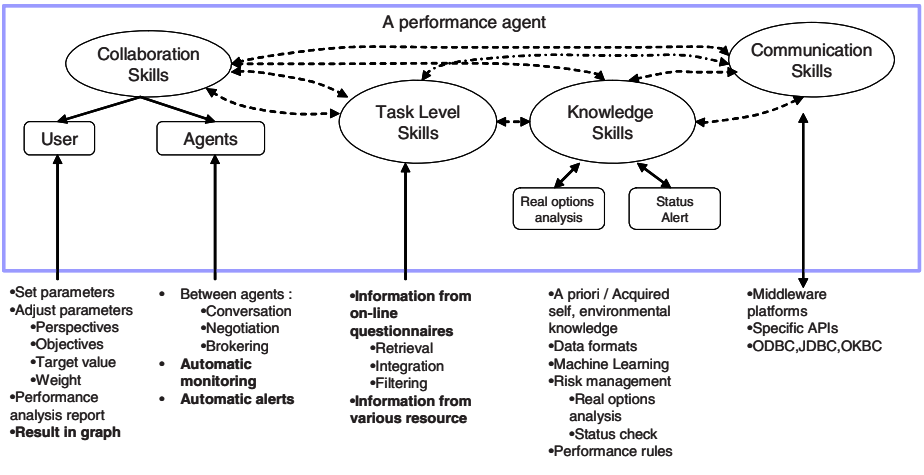


Fig. 2. Inside of a performance agent for e-learning systems

There are four functions inside of a performance agent. Each function is also a component agent which can active and proactive independently. Firstly, Collaboration agent takes care of the higher level interaction between users and other agents including conversation, matchmaking, and negotiation. It also maintains the collaboration between multiple agents according to BSC main perspectives which can be extended to detail objectives for management purpose. Communication functions communicate other agents, users, middleware and specific API such as ODBC, JDBC. Knowledge function copes with dynamically acquired data and the pre-given parameters, meanwhile, we pack the real options analysis and status check inside this function. Finally, the Task functions put the mainly focus on retrieving and monitoring information from various heterogeneous sources.

4 Brief Example

This session give a brief example. Suppose the entire BSC perspectives and objectives have been determined by the decision maker. C1 represents the first objective "Increase the enjoyment-of-use of e-learning environment" of Reaction perspective. The success point of C1 is "at least 3 cumulated satisfied score must be achieved". The decision maker also determined the Reaction perspective weight = 0.25 and objective weight of C1 is 0.6 according to the consumed resources. The previous measure of C1 at the first time and second time is 10.08 and 10.12 accordingly.

The questionnaire score of C1 equals to 2.26 in the investigation performed at the third checkpoint. Based on the success point "at least 3 cumulated satisfied score must be achieved", the objective score of C1 can be calculated and thus the objective score of C1 is 1.94. For example, the current measure of C1 is S_{C1}= 10.12* (0.95+1.94*0.025) = 10.11. The volatility of C1 can thus be calculated as σ_{C1}'=Std(10.08, 10.12, 10.11)= 0.0208. Because we collect the questionnaires every specific period, for example, 30 days (t= 30), thus σ_{C1}=0.0727. The decision maker want to explore if the performance can exceed the organization's goals the final deadline, the time to maturity can be calculated as T = (checkpoint 2 – checkpoint 1)/365 = 0.75 years (for example). Finally, the objective indexes of C1 can be calculated with parameters (S_{C1}, K_{C1}, σ_{C1}, T, r_{C1})= (10.11, 10.0, 0.0727, 0.75, 0) to obtain the objective index (G_{C1}=0.1852) and fail index (Z_{C1}=0.1220). Thus obtain the success value of C1 (V_{C1}=0.4169).

After applying (s1) to (s3), the BSC analyzing sheet can be completed:

Objective Score (X)	Expected Growth Rate (r_P)	Volatility (σ_P)	Previous Measure (P_P)	Current Measure (S_P)	Target Value (K_P)	Objective Weight (W_P)	Objective Index (G_P)	Fail Index (Z_P)	Success Value (V_P)
Reaction Perspective weight: *0.25* Perspective Index: *0.4245* Perspective Fail Index: *0.3569* Perspective Success Value: *0.1733*									
C1:1.68	0	0.0727	10.12	10.11	10	0.6	0.1852	0.1220	0.4169
C2:1.30	0	0.1716	10.11	10.01	10	0.4	0.2393	0.2349	0.0187
Learning Perspective weight: *0.27* Perspective Index: *0.7494* Perspective Fail Index: *0.6369* Perspective Success Value: *0.1627*									
L1:1.64	0	0.1042	10.08	10.06	10	0.5	0.1958	0.1659	0.1658
L2:1.90	0	0.2944	10.13	10.17	10	0.5	0.5537	0.4711	0.1615
Behavior Perspective weight: *0.24* Perspective Index: *1.1254* Perspective Fail Index: *0.9298* Perspective Success Value: *0.1909*									
I1:1.43	0	0.3437	10.29	10.28	10	0.3	0.4041	0.3190	0.2366
I2:1.20	0	0.2733	10.22	10.16	10	0.7	0.7213	0.6108	0.1662
Value Perspective weight: *0.24* Perspective Index: *1.7968* Perspective Fail Index: *1.3668* Perspective Success Value: *0.2736*									
F1:2.27	0	0.6595	10.39	10.57	10	0.4	1.0429	0.8152	0.2463
F2:1.58	0	0.3082	10.31	10.34	10	0.6	0.7539	0.5516	0.3125
BSC Index: *1.0098* BSC Fail Index: *0.8124* BSC Success Value: *0.2175*									

Reaction Agent: *objective 1: C1* *objective 2: C2* ……

Learning Agent *objective 1: C1* *objective 2: C2* ……..

Behavior Agent *objective 1: C1* *objective 2: C2* ……..

Value Agent *objective 1: C1* *objective 2: C2* ……

Fig. 3. A BSC analyzing sheet example

The system is originally developed via Excel and under construction to agent-based environment to fulfill the continuous, dynamical, extremely change environment. The decision maker can quickly figure out (1) the entire e-learning project is possible to success. (2) the best performing perspective is Value (3) the best performing objective is C1, and (4) objective F1 must be carefully monitored because it has largest objective index and fail index.

5 Conclusion

This study proposed a framework integrating the BSC and the B & S model to evaluate an organization's e-learning project performance. The applications of this framework not limit to e-learning but also other kinds of digital content systems. It can be flexibility adopted to evaluate any BSC-based investigations if the objectives are designed in relative measurement methods. The proposed framework provides meaningful information to help decision makers address the most success/weakest factors of the e-learning project under uncertainties. Finally, a framework of performance agent-based for e-learning evaluation proposes in this paper for further development.

References

1. Kaplan, R., Norton, D.: The balanced scorecard: measures that drive performance. Harvard Business Review 70(1), 71–80 (1992)
2. Kaplan, R., Norton, D.: Putting the balanced scorecard to work. Harvard Business Review 71(5), 134–147 (1993)
3. Kaplan, R., Norton, D.: Using the balanced scorecard as a strategic management system. Harvard Business Review 74(1), 75–85 (1996)
4. Forbes, L., Hamilton, J.: Building an international student market: educational-balanced scorecard solutions for regional Australian cities. International Education Journal 5(4), 502–520 (2003)
5. Kirkpatrick, D.L.: Techniques for evaluating training programs. Journal of the American Society of Training Directors 13, 3–26 (1959)
6. Kirkpatrick, D.L., Kirkpatrick, J.D.: Evaluating Training Programs: The Four Levels (3rd), Berrett-Koehler Publishers (1994)
7. Cox, J.C., Ross, S.A.: The valuation of options for alternative stochastic processes. Journal of Financial Economics 3, 145–166 (1996)
8. Myers, S.C., Majd, S.: Abandonment Value and Project Lift, Advances in Futures and Option Research 4, 1 (1990)
9. Brach, M.A.: Real Options in Practice, pp. 218–223. John Wiley & Sons, Inc., Chichester (2003)
10. Black, F., Scholes, M.: The pricing of options and corporate liabilities. Journal of Political Economy 81(3), 637–659 (1973)
11. Klusch, M., et al.: Information agent technology for the Internet: A survey. Data and Knowledge Engineering 36, 337–372 (2001)

12. Chiu, H.Y., Sheng, C.C., Chen, A.P.: Designing a Dynamic E-learning Project Performance Evaluation Framework. In: The 7th IEEE International Conference on Advanced Learning Technologies (accepted)
13. Jonker, C.M., Lam, R.A., Treur, J.: A multi-agent architecture for an intelligent website for insurance. In: Proceedings of cooperative agents III. LNCS (LNAI), pp. 86–100. Springer, Heidelberg (1999)
14. Preece, A.D., Hui, K.-Y., Gray, W.A., Marti, P., Bench-Capon, T.J.M., Jones, D.M., Cu, Z.: The KRAFT architecture for knowledge fusion and transformation. In: 19th SGES International conference on knowledge-based systems and applied artificial intelligence (ES'99), Springer, Heidelberg (1999)
15. Beer, M., Huang, W., Sixsmith, A.: Using agents to build a practical implementation of the INCA system. In: Intelligent agents and their applications, Springer, Heidelberg (2002)

Salable Vector Graphics (OpenVG) for Creating Animation Image in Embedded Systems

Sang-Yun Lee[1], Sunghwan Kim[2], Jihoon Chung[2], and Byung-Uk Choi[3]

[1] Dept. of Electronical and Telecommunication Engineering, Hanyang Univerity, Haengdang-dong Sungdong-gu, Seoul 133-791, Korea
syllee@etri.re.kr
[2] Pineone Communications Co., Ltd., R&D Division, Gumchuck Building 7F, 889-60, Daechi-dong, Gangnam-gu, Seoul, 135-280, Korea
{babynewton, jihoonc}@pineone.com
[3] Division of Information and Communications, Hanyang University, Haengdang-dong Sungdong-gu, Seoul 133-791, Korea
buchoi@hanyang.ac.kr

Abstract. We propose the reference implementation with software rendering of OpenVG for the scalable vector graphic hardware acceleration, which the Khronos group standardizes. We present the design scheme that enables EGL and OpenVG to be ported easily in an embedded environment. Moreover, we describe the background of selection of an algorithm, and the mathematical function adopted for the performance improvement, and we propose the optimum rendering method. We present displaying of vector image on a screen through the OpenVG implemented using software rendering method. And, we present the test result of the CTS which is compatibility test tool. And we show the performance comparison against the Hybrid corp.'s reference implementation.

Keywords: OpenVG, EGL, SVG, Vector Graphic, Software Rendering.

1 Introduction

Recently, the demand for the applications using the vector graphics technology has increased [1]. Particularly, in areas such as SVG viewer, hand-held guidance service, E-Book reader, game, scalable user interface, and etc, the vector graphics technology is widely applied. OpenVG™ is a royalty-free, cross-platform API that provides a low-level hardware acceleration interface for vector graphics libraries such as Flash and SVG [2]. OpenVG is the standard constituted by the Khronos group. And the version 1.0 was released at July 2005 for the first time [3].

When the standard needs to be verified, or when the OpenVG application needs to be operated through an emulator in advance, or when there is a no hardware supporting OpenVG, it is necessary to have the reference implementation (RI) operating in the software rendering mode. Additionally, it takes much time until the special-purpose hardware supporting OpenVG is produced. The RI also can reduce the cost.

B. Apolloni et al. (Eds.): KES 2007/ WIRN 2007, Part II, LNAI 4693, pp. 99–108, 2007.
© Springer-Verlag Berlin Heidelberg 2007

Besides, as embedded devices and CPU's performance is improved, the possibility of being replaced with the software rendering is high.

In this paper, we propose the OpenVG Reference Implementation which can be easily ported to the various embedded devices by using the software rendering method. And, we show that our RI is more excellent than the existing RI in the performance aspect through an experiment.

2 Design of OpenVG and EGL Engine

2.1 System Architecture

The system architecture of the OpenVG RI proposed in this paper is shown in Fig. 1.

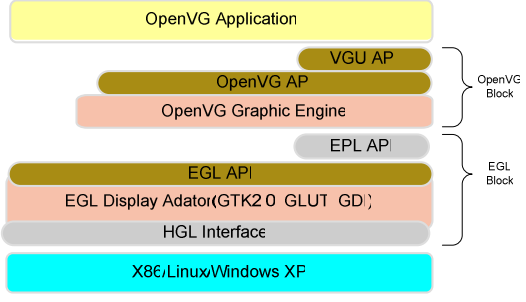


Fig. 1. The system architecture

The OpenVG RI is composed of Embedded Graphics Library (EGL) block and OpenVG block. EGL is an interface between rendering APIs such as OpenGL|ES or OpenVG (referred to collectively as client APIs) and an underlying native platform window system. EGL provides mechanisms for creating rendering surfaces onto which client APIs can draw, creating graphics contexts for client APIs, and synchronizing drawing by client APIs as well as native platform rendering APIs [4]. We designed so that, through the EGL Display Adapter, the client API could access the windowing system of the native platform.

In the EGL standard, the Embedded Platform Library (EPL) API is not included. However, it is necessary in order to implement client API, and must be ported according to a system. The API includes, for example, the functions of returning the frame buffer from Surface, the memory allocation, memory releasing, and etc. Hardware Graphic Library (HGL) interfaces performs the function of connecting EGL to the native graphics system.

EGL is itself the standard which is made to abstract the hardware system. However, it is necessary to have the separate porting layer like HGL so that EGL can be ported to the different native platform window systems through the minimum overhead. The OpenVG API is 2D vector graphics library and the VGU API is 2D vector graphics utility API of the high level.

2.2 Structure of the EGL and OpenVG Engine

The block structure of the EGL engine is shown in Fig. 2.

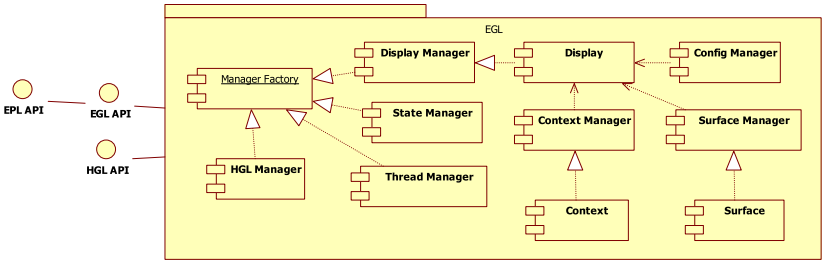


Fig. 2. The Structure of the EGL Engine

The Display Manager creates and manages the display object which takes charge of displaying graphics. The State Manager stores the error value generated when the functions executed. The Thread Manager provides the functions for avoiding the race condition in which several processes or the threads try to access EGL at the same time. The block structure of the OpenVG engine is shown in Fig. 3.

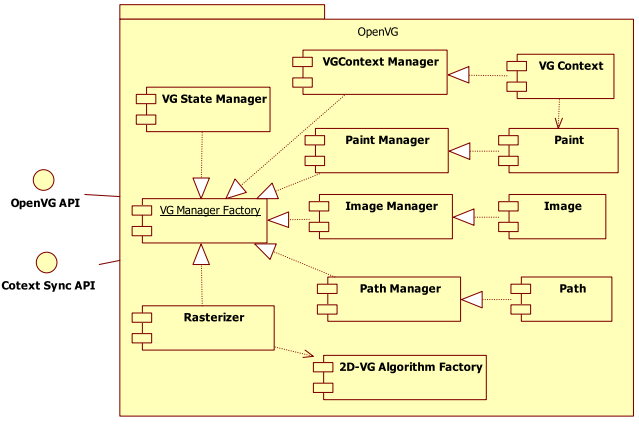


Fig. 3. The Structure of the OpenVG Engine

The Context Sync API provides synchronization between the VGContext generated in the OpenVG internally and the Context generated in EGL. The VG State Manager stores and manages the error value generated during the execution of OpenVG. The VG Context Manager, the Paint Manager, the Image Manager, and the Path Manager create and manage the VG Context object, the Paint object, the Image object, and the Path object respectively. The Rasterizer performs the function of drawing on the frame buffer provided by the surface of EGL with data combined of Path, Image, and Paint information.

2.3 Requirement for Designing the OpenVG/EGL Engine

OpenVG can operate not only on a desktop personal computer but also on a server. But it was developed to be mainly used in the embedded devices. So, there are always porting issues, because of wide variety of not only the hardware, but also the platforms and software. Therefore we must consider porting issues at as early as the architecture design phase, if we want to easily port once developed OpenVG to the various embedded devices. That is, the porting layer must exist so that the part to be modified according to the environmental change can be minimized. In addition, EGL and OpenVG must be loosely coupled.

Generally, in an embedded environment, the performance of a CPU is lower and the size of the memory is restrictive. Therefore, algorithms must be selected in such a way that the selected algorithms produce optimum performance and at the same time use as low memory and power as possible.

3 Novel Features of OpenVG Reference Implementation

3.1 Mathematical Function

In the operation process of drawing each graphic object of OpenVG, the use of the mathematical function is frequent. The method of calculating the mathematical function is classified into two ways. Firstly, it is the method of referring to the table value having the value calculated in advance. The second is the method of calculating the Taylor series of the finite order [5]. The Hybrid RI [6] adopted the later case. And as a result, it induced the performance degradation of the mathematics functional operation [7]. However, we adopted the table-look-up method and sought the performance improvement.

3.2 Sort Algorithm

In the OpenVG, sorting is used in the tessellation based rendering algorithm. That is, the vertex passing the scan line is arranged to the abscissa order. Or there is case where it arranges several scissoring rectangles. In the Hybrid RI, the bubble sort algorithm was adopted. But we adopted the merge sort algorithm in this paper.

3.3 Improved Raster Rendering Algorithm

A rendering refers to the operation of drawing the vector graphics object in the display [8]. There are the vector rendering mode in which the vector graphics object is drawn every time, and the raster rendering mode in which objects are drawn by calculating the color of each pixels of an image. Raster rendering algorithm normally calculates all the parts which are not in fact displayed in a screen, it spends the time owing to the unnecessary calculation.

In this paper, we propose the improved rendering algorithm in order to solve this problem. Fig. 4 shows the improved raster rendering algorithm proposed in this paper.

The procedure of the proposed rendering algorithm is as followings: (1) If the area overlapping with the path bound is discovered for each scissoring rectangle, the pixel

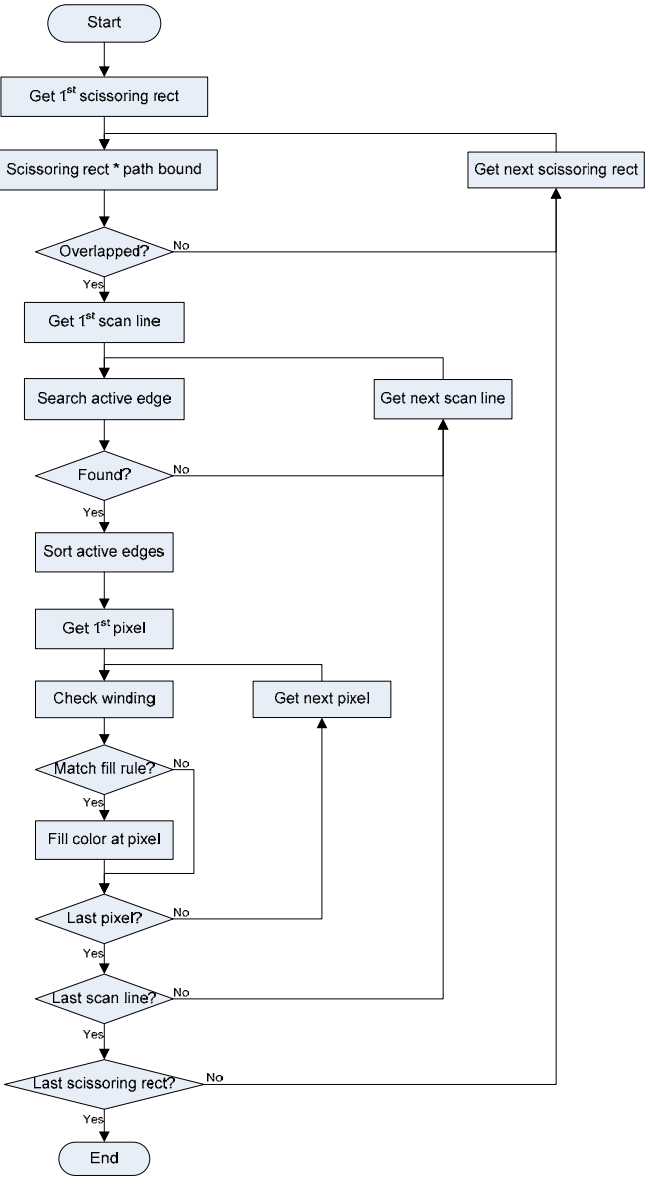


Fig. 4. The proposed Rendering Algorithm

color is calculated for the corresponding part; (2) If the area overlapping is not discovered, the same operation is performed for the next scissoring rectangle; (3) If the vertex intersecting with the corresponding scan line is not discovered, the scan line is moved to the next line.

The proposed method can display the vector graphics faster than the existing method by remarkably reducing the area visited in order to calculate the pixel value.

4 The Design Point for the Embedded Environment

The embedded environments have the following characteristics: most of the devices have limited memory capacity; the power usage is an important, because most of the devices operate on batteries; the devices have limited CPU power.

4.1 The Coherence of OpenVG and EGL

The Hybrid RI shares data structure between the objects which EGL and OpenVG create. Therefore, it has an effect on the other block if one block is modified among EGL or the OpenVG block. For example, the waste of resources occurs because the object for OpenVG is also generated within EGL when the data structure of EGL is expanded to support OpenGL|ES and is used, although we develop only the OpenGL|ES application.

In this paper, we separated data structure of OpenVG and EGL, in order to resolve these problems. Moreover, we designed so that OpenVG could use the function of EGL engine through the EGL API or the EPL API call.

4.2 The Language Dependency

The Hybrid RI was implemented with C++ language. C++ language is very powerful in the desktop environment, but it has many problems to be used in the embedded environment. Firstly, the speed of executing the inheritance or the virtual function is slow. Secondly, some compilers do not completely support C++ language. Therefore, we adopted C language that can be easily ported to an embedded device and the execution speed is fast.

4.3 The Singleton Pattern Design

There is an object in which several copy creations are not allowed among an object. Each Factory in the EGL block, and the context Manager, state Manager, image Manager, and algorithm Factory in the OpenVG block are those objects.

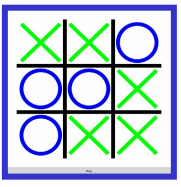
We introduced the singleton pattern for the guaranteed uniqueness of an object. If we design without the singleton pattern, each module has to recognize this object and avoid creating them. Or, these objects have to be registered in the global variables storage and used. However, existing code has to be modified to be ported to other platforms, because it is different in the structure of the global variables storage according to platforms.

5 Implementation and Experimental Results

5.1 Implementation

We verified whether our design method could be easily adapted to the embedded environment by implementing EGL and OpenVG and porting them to the various environments. We implemented EGL and OpenVG based on the Windows XP at first. And we modified the porting layer and could easily port it to the Linux and the WIPI

platform [9]. Fig. 5 shows vector graphics displayed on a screen through the proposed RI. Fig. 5(a) shows the image seen in the Tic-Tac-Toe game. Fig. 5(b) shows the tiger image that is the representative image of the vector graphics field. Fig. 5(c) shows the tiger image rendered on a cellular phone.



(a) Tic-Tac-Toe on PC



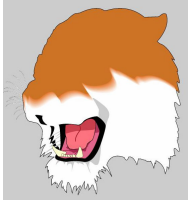
(b) tiger on PC



(c) tiger on handset

Fig. 5. Example of vector graphic displayed using the proposed RI

Fig. 6 shows the process of the tiger image being displayed with vector graphics. It shows when of 305 path, 25%, 50%, 75%, and 100% of the path are drawn.



(a) 25%



(b) 50%



(c) 75%



(d) 100%

Fig. 6. The process of drawing tiger

5.2 CTS Test Result

We performed test through the Conformance Test Suites (CTS) 1.0.0, the OpenVG compatibility test tool that the Khronos group distributes [10], in order to verify how our implementing RI adhered to the standard specification. Consequently, the success rate was 73% approximately. Table 1 and 2 shows the CTS success rate of the proposed RI and the Hybrid RI respectively according to the test item in detail.

Table 1. CTS test result of the proposed RI

Item	No.	Success	Item	No.	Success
Parameter	10	8	Path	48	37
Matrix	11	11	Image	10	6
Clearing	3	3	Paint	10	5
Scissoring	5	4	Image Filter	3	2
Masking	2	2	VGU	12	5

Table 2. CTS test result of the Hybrid RI

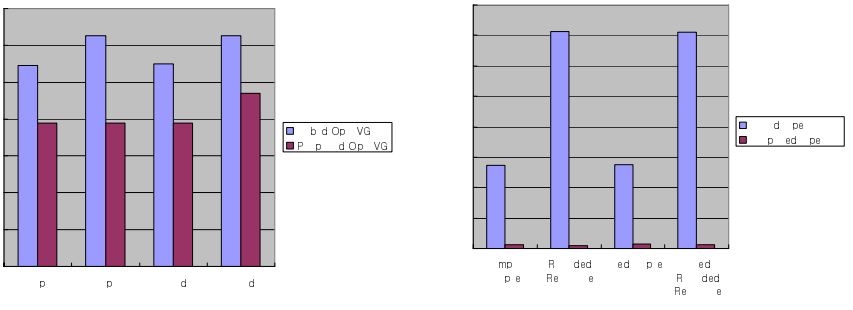
Item	No.	Success	Item	No.	Success
Parameter	10	8	Path	48	47
Matrix	11	11	Image	10	9
Clearing	3	3	Paint	10	6
Scissoring	5	4	Image Filter	3	2
Masking	2	1	VGU	12	9

Among the items that CTS reported as failure, there were items that correctly rendered resulting image cannot be differentiated from our reference-generated image visually. This was the case where the pixel value was off by one pixel. Moreover, there was a case where the already passed test image failed when a code was modified in order to pass another failing case. This is grasped that the test image of CTS is not yet stabilized.

5.3 Performance Evaluation

We developed performance measure program called Vgperf for 2D vector graphics. It was executed in the Windows XP for the comparison with the Hybrid RI. The test was progressed in the intel Core Duo 1.83GHz CPU, 2GB RAM, 2 MB (L2) Cash Memory, ATI Radeon X1400 Graphic card, and 128MB Video Ram environment.

The Fig. 7 shows the performance measurement result of Path drawing.



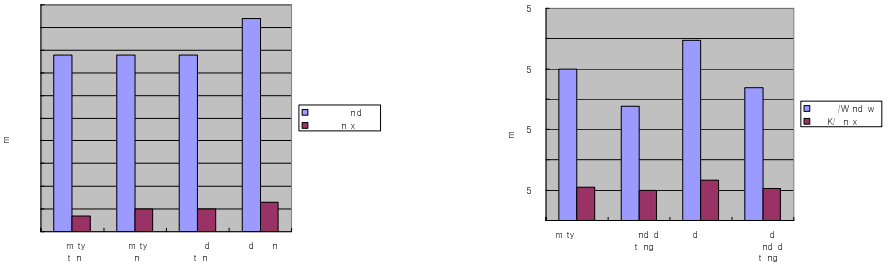
(a) Basic path performance (b) Complex path performance

Fig. 7. Performance test result of the proposed RI and the Hybrid RI

The Fig. 7(a) shows the measurement result of drawing basic diagram such as, a triangle or a square. The Fig. 7(b) shows the measurement result of drawing little more complicated figures like an ellipse or a rounded square.

As shown in the Fig. 7, the OpenVG RI proposed in this paper is faster than the Hybrid RI 1.3-1.6 times in case of the basic path, 4.6-76 times in case of the complicated path. As to the tiger image, the Hybrid RI took 7.3 seconds and our RI did 3.6 seconds, when the size of the drawing surface was 1,024 x 768.

The Fig. 8 shows the performance measurement result of Path drawing on the GLUT/Windows and the GDK/Linux.



(a) Basic path performance (b) Complex path performance

Fig. 8. Performance test result of the proposed RI on the Windows and the Linux

As shown in the Fig. 8, the proposed OpenVG RI on the GDK/Linux is faster than the RI on the GLUT/Windows 7.2-11.1 times in case of the basic path, 3.7-4.5 times in case of the complicated path. We can know that the GLUT had on bad effect on the rendering speed.

6 Conclusion

The Flash™ of the Macromedia Inc. already occupies the market more than 80%, in the vector graphics field. The OpenVG was initiated latter than the Flash, but has been settled as the industry standard. We expect that the OpenVG will expand market occupancy sooner or later, because most of the graphic card companies participate in standardization.

In this paper, we proposed the reference implementation of the OpenVG using software rendering. We showed the possibility of success of the software rendering mode, by showing the proposed RI outperforms the Hybrid RI. Moreover, we could port the proposed OpenVG RI easily to the various platforms due to systematically designing it considering an embedded environment.

The academic research or development case of the OpenVG has been scarcely reported because it is introduced recently. We expect that this paper will become the turning point that it activates the OpenVG research and development.

In the future work, we will enhance the success rate of the CTS and continue to research about the performance improvement by using software rendering mode. Moreover, we have a plan to research of supporting the SVG based on the OpenVG.

References

1. Pulli, K.: New APIs for Mobile Graphics. In: Proceedings of SPIE - The International Society for Optical Engineering, vol. 6074, art. no. 607401 (2006)
2. Pan, Z., Aylett, R., Diener, H., Jin, X., Göbel, S., Li, L. (eds.): Technologies for E-Learning and Digital Entertainment. LNCS, vol. 3942. Springer, Heidelberg (2006)

3. Khronos Group Std.: OpenVG, Kronos Grouop Standard for Vector Graphics Accelerations (2005), www.khronos.org
4. Khronos Group Std. EGL, Kronos Grouop Standard for Native Platform Graphics Interfaces (2005), www.khronos.org
5. Watt, A.: 3D Computer Graphics, 3rd edn. Addison-Wesley, Reading (2000)
6. Hybrid Graphics Forum, OpenVG Reference Implementation (2005), http://forum.hybrid.fi
7. Gonzalez, R.C., Woods, R.E.: Digital Image Processing, 2nd edn. Addison-Wesley, Reading (1992)
8. Harrington, S.: Computer Graphics A Programming Approach, 2nd edn. McGraw Hill, New York (2006)
9. KWISF: Wireless Internet Platform for Interoperability (2006), www.wipi.org.kr
10. Huone: Confermance Test Suite for OpenVG (2006), www.khronos.org

A Data Mining Methodology for Anomaly Detection in Network Data

Costantina Caruso and Donato Malerba

Dipartimento di Informatica, Università degli Studi di Bari
Via E. Orabona 4 - 70126 Bari - Italy
{caruso, malerba}@di.uniba.it

Abstract. Anomaly detection is based on profiles that represent normal behavior of users, hosts or networks and detects attacks as significant deviations from these profiles. Our methodology is based on the application of several data mining methods and returns an adaptive normal daily model of the network traffic as a result of four main steps, which are illustrated in the paper. The original observation units (the network connections) are transformed in symbolic objects and the normal model itself is given by a particular set of symbolic objects. A new symbolic object is considered an anomaly if it is dissimilar from those belonging to the model and it can be added to the model if it is ranked as a changing point, i.e. a new but legal behavior of the network traffic, otherwise it is an outlier, i.e. a new but illegal aspect of the network traffic. The obtained model of network connections can be used by a network administrator to identify deviations in network traffic patterns that may demand for her attention. The methodology is applied to the firewall logs of our Department network.

Keywords: Anomaly detection, ranking anomalies, decision boundary, adaptive model, data mining, machine learning, event monitoring, network traffic analysis.

1 Introduction and Related Work

Anomaly detection is one of the two main detection strategies (the second is misuse detection) to build an intrusion detection system (IDS) which is typically network-based if source data are network packets, or host-based if source data are systems calls and/or log data.While the latter is based on extensive knowledge, provided by human experts, of patterns associated with known attacks, the former is based on profiles that represent normal behaviour of users, hosts or networks and detects attacks as significant deviations from these profiles. Anomaly detection is potentially able to recognize new types of attacks, but it requires to define first what can be considered a normal behaviour for the observed entity (which can be a computer system, a particular user, etc.), and thereafter to decide how to label a specific activity as abnormal.

There are different aspects of anomaly detection that are important to underline.

B. Apolloni et al. (Eds.): KES 2007/ WIRN 2007, Part II, LNAI 4693, pp. 109–116, 2007.
© Springer-Verlag Berlin Heidelberg 2007

Classical supervised learning algorithms are hardly useful for anomaly detection tasks. They are designed for classification tasks and need labeled examples belonging to all the classes the analyzed system can present. In anomaly detection, we are typically given only one class of examples. In some cases, we can have normal (clean) data, i.e. data we know to be generated from the system in a normal operating condition, and we learn the model from this single class. However, it is difficult to have clean data because they can contain attacks if they are real, while they represent only a partial view of the system if they are simulated. The learning process must generate a model that has to be able to differentiate exactly between positive and negative cases but we often are not able to formalize this difference clearly. As a direct consequence, anomaly detectors present a serious drawback: false positive phenomena. Typically, false positive rate increases along the time and it is proportional to the system's variability rate.

Anomaly detectors should be able to recognize anomalies and to differentiate them in order to allow human operators to analyze them easily. Only recently the distinction between outliers and change points has emerged. Roughly, outlier detection is mainly interested in looking for isolated and exceptional points while change point detection looks for changing patterns. In most research works, outlier detection and change point detection are related explicitly and the adaptive properties, when considered, are built-in in the model. The difference between outliers and change points is well captured both in [1], which considers two different behaviors when analyzing static or dynamic datasets, and in [2], which deal with the issue of detecting outliers and change points from time series. Similarly, also the adaptive property appears in recent works (see [3]). However, most of works face the problem of creating the normal model by applying classical learning algorithms and then detecting anomalies in a static contest.

Works on anomaly detection reported in literature can be classified primarily by their type of training data (i.e., supervised, normal (one class) or unsupervised), and, secondarily, by their modeling technique. *Distance-based methods* [4, 5, 6] are widely used as unsupervised approach: they can work properly without previous knowledge. The idea in distance based methods is to represent every observation by means of a feature vector and to apply a distance function to measure how much the two observations are far/close. *Statistical approaches* [7,8] are used in both supervised or normal and unsupervised cases. Data points are modelled using a stochastic distribution; the drawback is that with high dimensions, it is difficult to estimate distributions. *Profiling techniques* [9, 10, 11] are applied when observation units concerning either human or machine or system behavior are both supervised and arranged in sequences or time series.

In this work we refine a methodology for the construction of an anomaly detector based on real observations (i.e. network connections), whose preprocessing phase starts with the application of a clustering (unsupervised) method to feature vectors representing daily network connections. Clusters are then described as sets of rules. The last steps of the preprocessing phase consist of the transformation of rules into symbolic objects and in the subsequent computation of similarities between symbolic objects in order to make the model adaptive. An anomaly is defined to be a SO which deviates too much from the model; a ranking mechanism is used to differentiate anomalies in order to identify more precisely change points, which express an

evolution of the system, and true outliers which can be caused by either malfunctioning or intrusions. The result is a longitudinal adaptive normal model of network connections that can be used by a network administrator to identify deviations in network traffic patterns.

2 The methodology

2.1 Generating the Normal Static Model

The first preprocessing step of the methodology has to be based on an Exploratory Data Analysis technique, since we know nothing about our system and we have no labeled examples. This justifies the use of clustering as our starting point. Clusters are groups of similar objects and are an extensional form of knowledge representation. Using all the observations to generate a model is the true solution in detecting anomalies but this is not feasible in our context where the amount of data to analyze daily is huge and clustering techniques have the disadvantage that they do not provide intensional descriptions of the extracted knowledge.

Intensional descriptions are essential in our context since they provide human-interpretable patterns of network traffic, and they are a lighter form of knowledge representation which is a good quality for a light computationally tool to use in real-time. Therefore, we adopt a two-stepped approach. First, we apply a clustering algorithm in order to decrease the size of the problem. Then we generate a set of rules whose consequents represent the cluster membership. The rules provide an intensional description of clusters.

The rules set ***R(t)*** generated for the time unit ***t*** is a static representation of the network traffic observed in ***t***. It should be noted that in this way we can drastically reduce data to treat and to store: the daily network traffic is given by few hundreds of rules vs. thousands of network connections.

A rule ***R*** corresponds to homogeneous groups of connections, that is, to second-order objects, or symbolic objects (SO), according to the terminology used in symbolic data analysis [14]. Symbolic objects simplify the change mining process on our data streams because we can easily compare them by means of dissimilarity measures.

Therefore, to provide the network administrator with a dynamic representation of network traffic, we propose to transform rules into symbolic objects and then to compute the dissimilarities between different SOs of different days. The transfomation of a rule ***R*** into the corresponding symbolic object is illustrated in [15] and represents the last step of the pre-processing phase.

The symbolic objects set ***So(t)*** generated for the time unit ***t*** is the *static normal model* of network traffic observed at time ***t.***

2.2 From Stationary to Adaptive

To build an adaptive model of network traffic ***M(t)*** by means of the daily normal static models ***So(t),*** given by sets of SOs, we need to compare them by computing similarities between SOs belonging to the subsequent daily normal static models.

Let ***M(t)*** be the adaptive normal model at time t^1 whose cardinality is |***M(t)***| and let ***So(t)*** be the static normal model at time ***t*** whose cardinality is |***So(t)***|. Given a threshold ***T*** and a symbolic object ***So*** in ***M(t)***, all the SOs in ***So(t)*** whose dissimilarity from ***So*** is less then ***T*** are considered similar to ***So***, otherwise they are tagged as anomalies.

Let $So_jM(t)$ be the j-th symbolic object in ***M(t)*** and $So_k(t)$ the *k*-th symbolic object in ***So(t)*** and let ***D*** be a dissimilarity measure between symbolic objects.

Definition 2.1. The symbolic object $So_k(t)$ in ***So(t)*** is an *anomaly* if

$$\mathbf{D_{jk}} = \mathbf{D}(So_jM(t), So_k(t)) > T \qquad \text{for each j=1,.., } |M(t)|$$

Therefore, if $\mathbf{D_{jk}} <= T$ for some *j*, the symbolic object $So_k(t)$ is considered a manifestation of a "known" behavior (it is already modeled by ***M(t)***), otherwise it is considered an "unknown" behavior, i.e. an anomaly.

2.3 Ranking Anomalies: Mean Dissimilarity and Minimum Dissimilarity

We need to differentiate anomalies on the fly in order to adapt the model only by means of novel events of the system and not by real outliers. In most research works, outlier detection and change point detection have not been related explicitly and the adaptive properties, when considered, are built-in in the model. We consider this approach unsound, and we propose an explicit ranking mechanism between anomalies. We define two further quantities to differentiate among the various type of anomalies.

When analyzing SOs at time t, it is significant to know *how much* a SO is similar to all the SOs belonging to the adaptive model ***M(t)*** in order to rank its level of dissimilarity. This information about a symbolic object $So_k(t)$ can be obtained by computing its dissimilarity mean value defined as follows:

$$D_{mean}(So_k(t)) = \frac{\sum_{j=1}^{|Mj|} D(So_k(t), So_jM(t))}{|M_j|}$$

Another interesting parameter is the minimum value of dissimilarity of a SO; indeed, a SO with high mean dissimilarity could be similar to few others but very dissimilar from the remaining ones and the mean value is not able to capture this situation. Therefore we compute the minimum value of dissimilarity between a fixed SO and all the SOs belonging to ***M(t)***:

$$D_{\min}(So_k(t)) = \min_{j=1}^{|M_j|} D(So_k(t), So_jM(t))$$

2.4 The Normal Adaptive Model *M(t)*

The notions of mean and minimum dissimilarity are used to give the complete definition of normal model. Let T_{mean} and T_{min} be two system-defined threshold values.

[1] The time unit we choose in our experiments is the entire day but the approach is general and the most suitable time unit can be used: seconds, minutes, hours or years and so on.

Definition 2.2. A symbolic object So_k in $So(t)$ is an anomaly of *prevalent type* PA_k if and only if the following condition holds:

$$(D_{mean}(So_k(t))) <= T_{mean}) \ \wedge \ (D_{min}(So_k(t))) <= T_{min})$$

Definition 2.3. A symbolic object So_k in $So(t)$ is a anomaly of *secondary type* SA_k if and only if the following condition holds:

$$(D_{mean}(So_k(t)) > T_{mean}) \ \wedge \ (D_{min}(So_k(t)) <= T_{min})$$

Definition 2.4. A symbolic object So_k in $So(t)$ is an *outlier* O_k if and only if the following condition holds:

$$(D_{mean}(So_k(t)) > T_{mean}) \ \wedge \ (D_{min}(So_k(t)) > T_{min})$$

Definition 2.5. Let $ChPoints(t)$ = {the set of all PA_k and SA_k found in $So(t)$}
We can refine our adaptive process which must upgrade our normal adaptive model $M(t)$ only through PA_k and SA_k. Thus we have

Definition 2.6. The normal adaptive model $M(t)$ at time t is given by the union of the normal adaptive model $M(t-1)$ at time $(t-1)$ and the set of all change points found in $So(t)$:

$$M(t) = M(t-1) \cup ChPoints(t).$$

3 Experiments and Results

3.1 Data Collection and Preprocessing

The source data set is formed by four successive weeks of firewall logs of our Department, from May 31st to June 27th, 2004. A transaction is a single logged packet. The target dataset is built by reconstructing the entire connections from single packets. In our analysis we have used only accepted ingoing connections and we have a file for every day that contains all the connections opened and closed in that day.

Since the goal is to create the daily description of our connections, we have chosen to work with few but fundamental attributes, namely, *Proto* (nominal) which can have only two values: udp and tcp; *StartHalfHour* (integer), the time field of a packet; *Dst* (integer) which represents the IP address of department public servers; *Service* (nominal): the requested service (http, ftp, smtp and many other ports); *NumPackets* (integer): the number of connection packets; *Length* (integer): the time length of the connection; *NationCode* (nominal): the two digit code of the nation the source IP belongs to.

In the clustering step we used k-means [12] while the rules, which provide an intensional description of clusters, are generated by means of the algorithm PART [13]. Rules are transformed into symbolic objects [15].

3.2 Building the Normal Adaptive Model and Evaluating It

The rules we selected to test the methodology are the most prevalent aspects in our network traffic and can contain only change points i.e. points which represent a natural change of the network traffic.

The normal adaptive model ***M(t)*** we build is very simple but we must evaluate its performance i.e. we have to demonstrate it is able to represent significantly the network traffic and its evolution.

It is not possible to use techniques like ROC curves to evaluate its performance because we know nothing about data. We can obtain only statistical profiles about data to know what was the network traffic behavior along the 4 analyzed weeks.

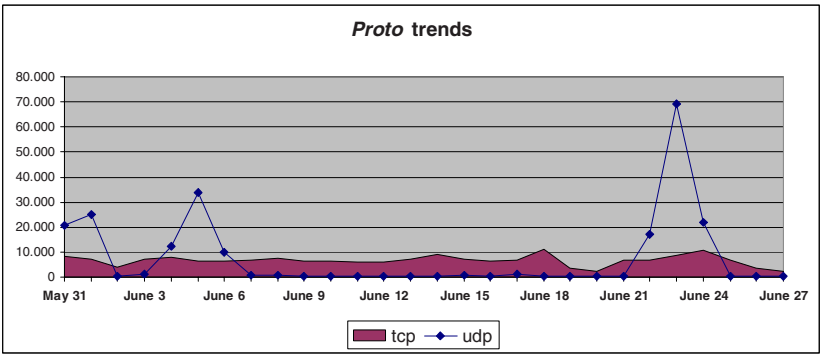


Fig. 1. In the graph, the trend of the *Proto* attribute is represented; on the y-axis there is the number of instances while on the x-axis day labels are reported

In Fig.1 we have represented the statistical profiles of the attribute *Proto*. In the graph, we have the tendencies of the two only values the attribute can have: udp (the line) and tcp (the solid). It is evident that we can aggregate the 28 days in two big groups according to their statistical profiles: group A formed by {31, 1, 4, 5, 6, 22, 23, 24} and group B all the remaining days belong to. The different behaviour is ought to the p2p connections explosion we have in group A days.

The idea to evaluate our methodology is to identify a system-dependent set (or range) of three threshold values for $\boldsymbol{T}$, $\boldsymbol{T_{mean}}$, $\boldsymbol{T_{min}}$, such that the normal model they build is able to differentiate these two days groups.

The proposed methodology works properly if new SOs are added to the normal adaptive model only when the statistical profile of the analyzed day is different from that one of the previous days. The model is not modified when we face days whose statistical profile is already known. We expect that the model will be modified strongly in the first week while new SOs are less and less added as days go on. Further, the network traffic behavior of the days with an already known profile, has to be represented mainly by means of SOs generated in similar previous days.

In the experiments, the three thresholds vary in these intervals: $\boldsymbol{T}$ = [2..5], $\boldsymbol{T_{mean}}$ =[3..8], $\boldsymbol{T_{min}}$=[1..4]. The choice of these ranges of values was not random but we looked for system-dependent values by computing the mean (5.0) and minimum (2.0) of all dissimilarity quantities as pivot values.

M(0), i.e. the start-up model, is initialized to the entire set of SOs generated for the first day (May 31st). The models obtained for $\boldsymbol{T}$=[3..5] along all the 4 weeks, are not satisfactory because they possess almost the same set of SOs of the starting point,

M(0). The models obtained for T=2 along all the 4 weeks react to the different network traffic behaviors and the results depend on the values of the other two thresholds, T_{mean} and T_{min} but they have the same profile.

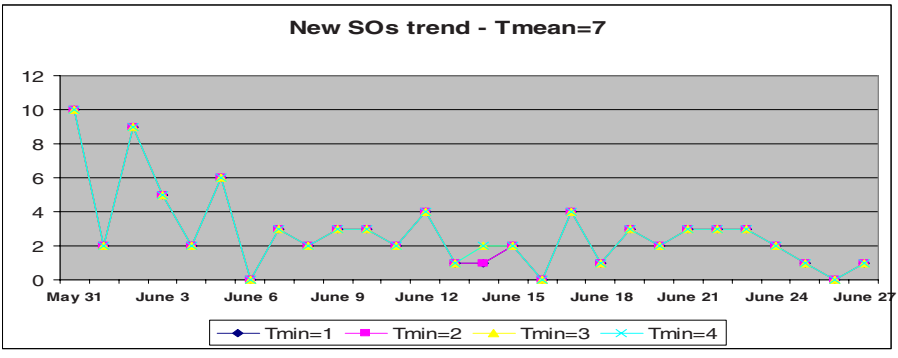


Fig. 2. In the graph, the trend of new SOs for T_{mean}=7 are represented

In Fig.2 we report the number of SOs added to the model along the four weeks for the threshold set (2,7, T_{min}=[1..4]). It is interesting to notice that for T_{mean}= 3 or T_{mean}= 4, the results depend from the value of the third threshold T_{min}; for T_{mean}= 7 or T_{mean}= 8, the results are totally independent from T_{min} (i.e. we obtain always the same set of SOs for T_{min}=[1..4]) while for T_{mean}= 5 or T_{mean}= 6, the results have the same tendencies but they are not totally identical. Further, the model we obtain represents successive group A days by means of group A's SOs and successive group B days by means of group B's SOs. This means that, for example, on the June 24th, the model represents it by means of SOs generated on June 6th while representing June 16th by means of SOs generated on June 2nd.

4 Conclusions and Future Work

The main aim of this work is to investigate the potentialities of a new methodology in order to identify and learn patterns in the network traffic. The experiments have been made on the network traffic of our department but the features used are general and so this methodology can be applied to the traffic of every network active device.

The model obtained with a system-defined range of values as to the three thresholds, successfully identifies already known behavior and, at the same time, identifies anomalies and ranks them in order to understand which are change points, i.e. new but legitimate behavior. Source data represent prevalent aspects of our network traffic and this implies that they only contain legitimate behavior. Our future work will concentrate on two main issues: 1) including new source data in order to study the performance of the methodology in identifying outliers and 2) analyzing the range of the possible thresholds values to characterize them better.

References

1. Ghoting, A., Otey, M.E., Parthasarathy, S.: Loaded: Link-based Outlier and Anomaly detection in Evolving Data Sets. In: Proceeedings of the IEEE International Conference on Data Mining, IEEE Computer Society Press, Los Alamitos (2004)
2. Takeuchi, J., Yamanashi, K.: A Unifying Framework for Identifying Changing Points and Outliers. IEEE Transactions on Knowledge and Data Engineering 18(4) (2006)
3. Wang, K., Stolfo, S.: Anomalous Payload-based Network Intrusion Detection. In: RAID (2004)
4. Knorr, N., Ng, P.: Algorithms for Mining Distance-Based Outliers in Large Datasets. In: VLDB (1998)
5. Breunig, et al.: LOF: Identifying Density-Based Local Outliers. In: KDD (2000)
6. Eskin, E., Arnold, A., Prerau, M., Portnoy, L., Stolfo, S.: A Geometric Framework for Unsupervised Anomaly Detection: Detecting Intrusions in Unlabeled Data (2002)
7. Yamanishi, K.: On-line unsupervised outlier detection using finite mixture with discounting learning algorithms. In: KDD (2000)
8. Mahoney, M., Chan, P.: Learning Nonstationary Models of Normal Network Traffic for Detecting Novel Attacks. In: 8th ACM KDD (2002)
9. Hofmeyr, S., et al.: Intrusion Detection using Sequences of System Calls (1997)
10. Tandon, G., Chan, P.: Learning Rules from System Call Arguments and Sequences for Anomaly Detection. In: Workshop on Data Mining for Computer Security. In: ICDM (2003)
11. Wang, K., Stolfo, S.: One Class Training for Masquerade Detection. In: Workshop on Data Mining for Computer Security. ICDM (2003)
12. Jain, A.K., Murty, M.N., Flyn, P.J.: Data Clustering: a Review. ACM Computing Surveys 31(3) (1999)
13. Witten, I., Frank, E.: Generate Accurate Rule Sets Without Global Optimisation. In: Machine Learning: Proceedings of the 15th International Conference, Morgan Kaufmann Publishers, San Francisco (1998)
14. Gowda, K.C., Diday, E.: Symbolic Clustering Using a New Dissimilarity Measure. In Pattern Recognition 24(6), 567–578 (1991)
15. Caruso, C., Malerba, D., Papagni, D.: Learning the daily model of network traffic. In: Hacid, M.-S., Murray, N.V., Raś, Z.W., Tsumoto, S. (eds.) ISMIS 2005. LNCS (LNAI), vol. 3488, pp. 131–141. Springer, Heidelberg (2005)

A Dedicated Hardware for Fingerprint Authentication

G. Danese, M. Giachero, F. Leporati, G. Matrone, and N. Nazzicari

Dip. di Informatica e Sistemistica – Università di Pavia – Via Ferrata 1, 27100 Pavia, Italy
Tel.: +39 0382 985350; Fax.: +39 0382 985373
{gianni.danese, mauro.giachero, francesco.leporati,
nelson.nazzicari}@unipv.it, giuliamatrone01@ateneopv.it

Abstract. Biometric systems are defined as systems exploiting automated methods of personal recognition based on physiological or behavioural characteristics. Among these, fingerprints are very reliable biometric identifiers. Trying to fasten the image processing step makes the recognition process more efficient, especially concerning embedded systems for real-time authentication. In this work we propose an FPGA-based architecture that efficiently implements the high computationally demanding core of a matching algorithm.

Keywords: Authentication and Security, Fingerprints, Image Processing, Hardware Accelerators, FPGA, Phase-only correlation.

1 Introduction

The use of biometric techniques applied both to control people in the access to data and locations or to authenticate them, is ousting traditional knowledge or token based systems like password, PINs, smartcards and electronic or mechanic keys which can be easily lost, stolen or discovered. Most diffused biometric technologies employ fingerprints, hand geometry, iris or face featuring as indicators, since they exploit more safety, usability and efficiency than other solutions although each one has to be chosen on the basis of the application characteristics.



Fig. 1. A typical fingerprint

Indeed, the identification based on fingerprints can be applied in the design of automatic authentication systems due to their uniqueness among subjects and persistence throughout the life. The ever increasing diffusion of embedded systems, into which low cost miniaturized circuits and sensors are integrated, extends application benefits, while raising the need for reducing the size of databases containing suitable biometric templates and the time required for answering to queries. Proper matching and classification algorithms can be found in literature, able to restrict the database search to a subset from which fingerprint templates can be extracted and compared with those for whom the authentication is required. These algorithms, unfortunately, show

B. Apolloni et al. (Eds.): KES 2007/ WIRN 2007, Part II, LNAI 4693, pp. 117–124, 2007.
© Springer-Verlag Berlin Heidelberg 2007

real time requirements which often cannot be provided without resorting to powerful computing systems. However, the introduction of dedicated architectures can improve the performance both in terms of size and in terms of processing times, since the computational core often consists of few instructions repeatedly executed and a special purpose CPU can easily overcome a general purpose one. In particular, accelerators designed and implemented on FPGA (Field Programmable Gate Array) can be one of the most appealing solutions in terms of costs, flexibility and performance, sometimes (but not always) bringing to development of highly specialised ASIC circuits for industrial production. In this work we present an FPGA based system for fingerprint matching based on the well known Phase Only Correlation (POC) technique and able to outperform modern general purpose processors.

2 Fingerprints

Typical fingerprints are characterised by an alternation of *ridges* and *furrows* (dark and light areas in fig. 1) and differ not only among subjects but also among same subject's fingers. They are definitively grown up in a 7 months fetus and they are not affected by surface abrasions, burns and cuts since their original pattern is reproduced as the skin regenerates.

The different arrangement and shape of dermic fibers give origin to several ridges and to a precise *papillar* layout which is used as subject's identifier. This papillar configuration does not change during the subject's life and since it's different among individuals, it can be used for a systematic classification.

The efforts made in making automatic the matching process based on digital representation of fingerprints led to the development of *Automatic Fingerprint Identification Systems* (AFIS) [1]. Typically, there are millions of fingerprint records in a database which needs to be entirely searched for a match, to establish the identity of an individual. In order to provide a reasonable response time for each query nowadays special hardware solutions implement matching and/or classification algorithms in a really efficient way.

3 Classification

A fingerprint can be thoroughly described by two kind of different parameters, *global* and *local*. The former identify a general property of the ridges and farrows weft and is used for classification, while the latter are typically extracted from a restricted portion of a digital fingerprint representation and allow the unique identification of the fingerprint. As for the first aspect, the pattern shows regions into which the ridges assume a greater curvature together with frequent endings: within these regions singular points are present called *core* and *delta* (fig. 2) with the first acting as fingerprint centre, while



Fig. 2. Singular points

the second being the point where two ridges diverge after being parallel. Since these points show high steadiness and invariance to rotation and scale variation, they are widely employed in the classification algorithms till now developed. Following the so-called "Galton-Henry classification scheme", fingerprints are basically divided in five different classes: *arch, tented arch, right loop, left loop, whorl* (fig. 3). We refer the reader to [2], for details about the features of the 4 classes. For what concerns local investigation, nearly 150 different fingerprint characteristics have been identified (*minutiae*) corresponding to irregularities in ridges, but their visibility is unfortunately highly influenced by the surface conditions. Different types of minutiae can be observed leading to different classifications: for example, while the ANSI institute proposes a four classes distinction (ending, bifurcation, crossovers and undetermined) the FBI considers only ending and bifurcations.

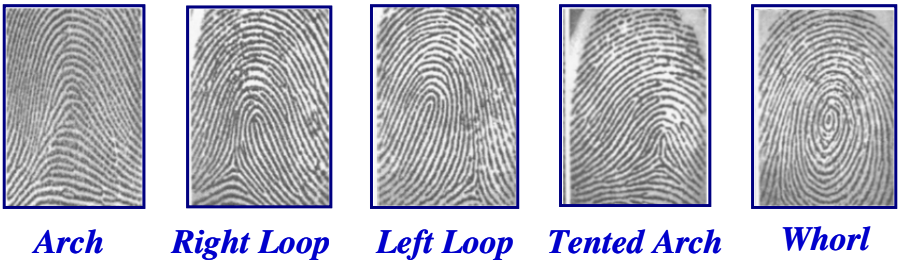


Fig. 3. Galton-Henry classification scheme

Although several approaches have been conceived to achieve automatic fingerprints classification, in practice they can be grouped in the following categories:

- *Model based*, which ascribes the fingerprint to one of the 5 mentioned classes on the basis of the *core* and *delta* points displacement*;*
- *Structure based*, which makes use of the *orientation field parameter* calculated from the acquired image of the fingerprint itself (although affected by noise);
- *Frequency based*, which evaluates the spatial spectrum of the acquired image;
- *Syntactical*, which uses a formal grammar to describe and classify fingerprints.

Also hybrid models have been developed, mixing two or more of the previous approaches. Some classification methods are based on the evaluation of the ridge global orientation and then try to identify singularities in the fingerprint to trace it back to one of the five categories. This approach reproduces that one usually employed by experts when performing a "manual classification" and consists of the following three steps:

- *Estimation of the local ridge orientation:* the normalised image is divided in blocks to which is assigned a proper orientation term (*orientation field*) by evaluating the gradients of the grey tones in the block;
- *Search of singularities:* a point in the orientation field is classified as ordinary, core or delta, by evaluating the Poincaré index along a little closed curve around

the point itself: the index is the sum of the orientations met by going along the curve counterclockwise.

- *Classification of the fingerprint:* by establishing a proper set of rules.

4 The Matching Algorithm

The performance of matching algorithms implemented in fingerprint recognition systems is closely related to the fingerprint representation itself and to the quality of the image. The majority of these methods are feature-based and consequently highly influenced by fingertip surface conditions. In order to avoid such problems, the matching algorithm we worked on, proposed by K. Ito et al. in [3], [4] employs a global measure of correlation between the input image and the template one, so as to achieve robust matching also with low-quality fingerprints. This new matching technique is based on the *Phase-Only Correlation* (POC) function computation, obtained from the phase component of the two images cross-correlation function spectrum in the frequency domain.

Let us consider the input fingerprint image and the registered template, represented as discrete two-dimensional functions, respectively $g(n_1,n_2)$ and $f(n_1,n_2)$. If we compute the 2D Discrete Fourier Transforms (DFTs) of these two images, denoted as $G(k_1,k_2)$ and $F(k_1,k_2)$, then the cross-spectrum $R_{FG}(k_1,k_2)$ of $G(k_1,k_2)$ and $F(k_1,k_2)$ is defined as:

$$R_{FG}(k_1,k_2) = F(k_1,k_2)\overline{G}(k_1,k_2). \tag{1}$$

where $\overline{G}(k_{1,}k_2)$ represents the complex conjugate of $G(k_1,k_2)$. Thus, the ordinary cross-correlation function $r(n_1,n_2)$ is the 2D Inverse DFT (IDFT) of $R_{FG}(k_1,k_2)$.

By applying the 2D IDFT only to the normalized cross spectrum $\hat{R}_{FG}(k_{1,}k_2)$ instead, we obtain the POC function:

$$\hat{R}_{FG}(k_{1,}k_2) = \frac{F(k_{1,}k_2)\overline{G}(k_{1,}k_2)}{\left|F(k_{1,}k_2)\overline{G}(k_{1,}k_2)\right|} = \mathrm{e}^{j\theta(k_{1,}k_2)}. \tag{2}$$

$$\hat{r}(n_1,n_2) = \frac{1}{N_1N_2}\sum_{k_1=-M_1}^{M_1}\sum_{k_2=-M_2}^{M_2}\hat{R}_{FG}(k_1,k_2)\mathrm{e}^{j\frac{2\Pi k_1n_1}{N_1}}\mathrm{e}^{j\frac{2\Pi k_2n_2}{N_2}}. \tag{3}$$

$$\theta(k_{1,}k_2) = \theta_F(k_{1,}k_2) - \theta_G(k_{1,}k_2). \tag{4}$$

with: $n_1 = -N_1 \ldots N_1$, $n_2 = -N_2 \ldots N_2$ and $N_1 = 2M_1+1$, $N_2=2M_2+1$,

where $\hat{R}_{FG}(k_{1,}k_2)$ contains the phase information given by $R_{FG}(k_1,k_2)$, the two discrete images dimensions are $N_1 \times N_2$, $\theta_F(k_1,k_2)$ and $\theta_G(k_1,k_2)$ represent respectively the phase component of $F(k_1,k_2)$ and $G(k_1,k_2)$.

Usually, the ordinary cross-correlation function resulting from low-quality images can be hardly interpreted; moreover, noise and non-linear distortions remarkably

decrease the correlation value, also working with two fingerprints coming from the same fingertip. On the other hand, a good index of similarity between the two images could be represented by the height of the POC function maximum peak: the more similar they are the more sharp and distinct this peak is (fig. 4). By using such a measure of the likeness degree, we achieve robustness and also high accuracy.

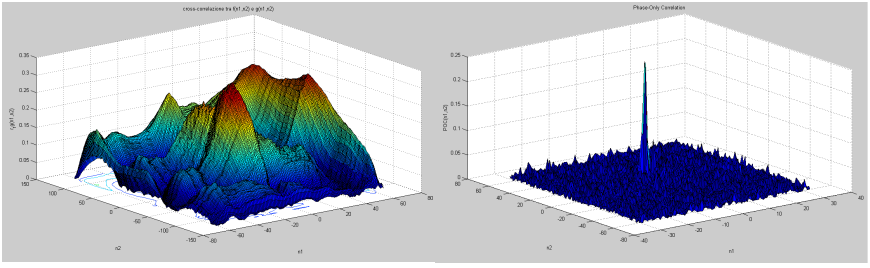


Fig. 4. A comparison between ordinary correlation and POC

Since the POC-based approach is sensitive to image rotation and deformation, a pre-processing of the fingerprints should be performed first [4]. To eliminate the differences between the two images, first of all the *core* point of both fingerprints should be localized by using a *Poincaré index*-based algorithm [1] [2], and afterwards it should be targeted as a reference point to align them.

Then, a set of POC functions should be computed between the registered fingerprint image and every single rotated version of the input image over the angular range $-\theta_{max} \le \theta \le \theta_{max}$ with an angle spacing of 1°, where θ_{max} is experimentally determined. The rotated image to be chosen for going on with the following processing steps is the one giving the highest peak of correlation.

If both fingerprints have a *core* point, the translational displacement between the two images can be simply obtained by computing the distance between the mentioned singularities. Otherwise, the displacement entity is given by the peak location of the POC function between the rotation-normalized image and the template.

Finally, the two images are aligned and subsequently their projection on vertical and horizontal axis are analysed in order to extract their intersection area, also called overlapped region. The aim of this operation is to eliminate all those areas going to produce uncorrelated noise which could affect the information provided by the POC function. All the pre-processing steps described above create the optimal working conditions for the chosen matching algorithm. Despite the fact that the POC function itself manages to give a good evaluation of the similarity degree between the two compared fingerprints, it often contains noisy and meaningless components in the band of high frequencies.

To improve matching performances, the cross-phase spectrum $\hat{R}_{FG}(k_1, k_2)$ could be computed by redefining the frequency ranges to be considered. While using the POC function, the parameters k_1 and k_2 vary respectively in the bands $-M_1 \ldots M_1$ and $-M_2 \ldots M_2$, now the two amplitude spectrums of both images should be projected onto the axes k_1 and k_2 in order to evaluate the new frequency superior limits K_1 and K_2 [4].

So, we can redefine the ranges as follows:

$$k_1 = -K_1 ... K_1 \quad k_2 = -K_2 ... K_2 \,. \tag{5}$$

where the new band dimensions are:

$$L_1 = 2\mathrm{K}_1 + 1 \quad L_2 = 2\mathrm{K}_2 + 1 \,. \tag{6}$$

The POC function will be now redefined as *Band-Limited* POC (BLPOC) function, according to the following equation:

$$\hat{r}^{K_1 K_2}(n_1, n_2) = \frac{1}{L_1 L_2} \sum_{k_1=-K_1}^{K_1} \sum_{k_2=-K_2}^{K_2} \hat{R}(k_1, k_2) e^{j\frac{2\Pi k_1 n_1}{L_1}} e^{j\frac{2\Pi k_2 n_2}{L_2}} \,. \tag{7}$$

With $n_1 = -K_{1.} .. K_1 , \quad n_2 = -K_{2.} .. K_2$.

Generally, BLPOC discriminatory abilities are superior to the POC ones; using this new function, the peak height is always normalized to 1 but usually the peak isn't unique, due to fingerprint elastic deformation phenomena during image acquisition.

Taking also this problem into account, a global index of similarity, whose value is called *matching score*, could be obtained by summing up the most significant peak heights, so that the matching algorithm is given even more robustness.

5 Our Implementation

We implemented the POC algorithm on Altera Stratix II FPGA (EPS2S180F1020C4 model) and decided to work with 128x64 pixel 256 grey tones images. The selected algorithm consists on several steps.

5.1 Two Dimensions Discrete Fourier Transform (2d-DFT)

Our starting point was a one dimension DFT core (1d-DFT) supplied by Altera that implements a Radix 2/4 decimation in frequency (DIF) FFT algorithm [5]. To obtain this result, we had to pass through two steps: first we calculated the DFT on each row. We obtained a complex matrix with the same number of rows and columns, but doubling storage space (we needed to memorize both real and imaginary parts). Then, we calculated again DFT, but this time column by column. In the final matrix, low indexes address low frequency information.

5.2 Phase Difference Calculation

After calculating DFT of input and template image, we have obtained two 2D cartesian notation complex number matrices. To achieve the phase difference $(\theta_1 - \theta_2)$ for every correspondent cell of the two matrices we need to calculate the following expression:

$$\theta_1 - \theta_2 = \arctan\left(\frac{\Im_1}{\Re_1}\right) - \arctan\left(\frac{\Im_2}{\Re_2}\right). \tag{8}$$

where $\Im_1, \Re_1$ and $\Im_2, \Re_2$ are the two correspondent Cartesian values. That formula is heavy to be directly computed – division and exact arctangent are slow operations. Luckily, fingerprint acquisition introduces a lot of noise (finger can be dirt, cold, scarred and so on) so we don't need the exact phase difference result: an approximate one can be enough. We implemented the arctangent operation using a look up table addressed by the 8 most significant bits of every couple of imaginary and real data. We needed to throw away LSbs because both real and imaginary parts are 16 bit wide data: 32 bit addressing 10 bit words mean 5 GB of storage space, clearly too much for an embedded device. Before arctangent calculation, data are normalized to reach the maximal precision: since we are only looking for ratio it's possible to left shift both data to eliminate leading zeros and minimize significant data waste.

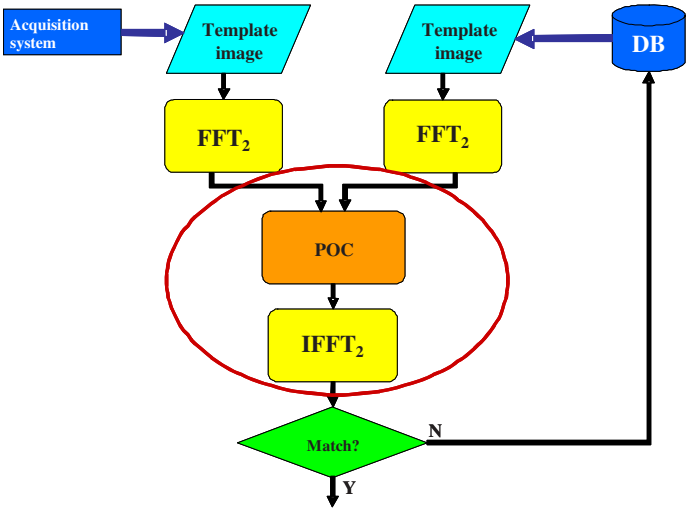


Fig. 5. The matching algorithm

5.3 Inverse Transform

To convert a complex number from polar to Cartesian representation using only phase information is equivalent to normalize data on modules, as shown below:

$$e^{j\theta} = \frac{\Re + j\Im}{\rho} = \cos\theta + j\sin\theta, \quad Polar : \rho e^{j\theta}, \quad Cartesian : \Re + j\Im \,. \tag{10}$$

We implemented sinus and cosine operations through two different look-up tables addressed by phases difference (10 bit wide 0.35 degree precision addressing) sampled with constant step and storing 8 bit wide data (precision equal to 0.0078). To get the cross-correlation matrix we performed an inverse 2d-DFT using the same approach used with the direct one. Fig. 5 summarizes the steps involved.

6 Performance

In this first implementation Quartus II reports a maximum clock frequency of 96.99 MHz (after a timing simulation) and resource occupancy nearly equal to 25%. Assuming a 90 MHz oscillator (easier to be found) our circuit needs nearly 590 μs to elaborate two images. We compared that result with a C routine compiled with the *gcc* compiler using -O3 optimization option and running on a 2 GHz AMD Athlon 64, using the FFTW transformation library (FFTW is considered one of the fastest FFT software implementations [6]). On that CPU the same elaboration took 3856 μs, so our circuit gives an effective speed-up factor of seven.

7 Future Developments

There are a lot of possible improvements for this project:

- implementing BLPOC: it eliminates part of the noise and reduces latency time because decreases the amount of data to be processed;
- parallelizing the implementation: the input image is compared with a lot of template images and every comparison is independent;
- reducing the amount of memory used: a lot of on-board RAMis used to save intermediate results. A memory reduction will bring to a more compact design;
- implementing a direct 2d-DFT without using the Altera 1d-DFT block. That implementation could be optimized to out purposes and result quicker and smaller.

References

[1] Maio, D., et al.: Handbook for Fingerprint Recognition. Springer, Heidelberg (2005)
[2] Jain, K., Pankanti, S.: Fingerprint Classification and Matching. In: Handbook for Image and Video Processing, Academic Press, London (2000)
[3] Ito, K., et al.: A Fingerprint Recognition Algorithm Using Phase-Based Image Matching for Low-Quality Fingerprints. In: IEEE Conf. on Image Proc., Genova 2005 vol. 2, pp. 33–36 (2005)
[4] Ito, K., et al.: A Fingerprint Matching Algorithm Using Phase-Only Correlation. IEICE Trans. Fundamentals E87-A(3), 682–691 (2004)
[5] http://www.altera.com/products/ip/dsp/transforms/m-ham-fft.html
[6] http://www.fftw.org

Belnap's Logic in Sharing Protected Web Resources

Sylvia Encheva[1] and Sharil Tumin[2]

[1] Stord/Haugesund University College, Bjørnsonsg. 45, 5528 Haugesund, Norway
sbe@hsh.no
[2] University of Bergen, IT-Dept., P.O. Box 7800, 5020 Bergen, Norway
edpst@it.uib.no

Abstract. In this paper we present a model that can prevent conflict situations caused by applying both positive and negative authorizations for access to a resource. Such conflict situations may occur if an organization has decentralized administration, and/or several collaborating organizations have access to one resource and some of them apply positive authorizations while others apply negative authorizations. The proposed solution involves Belnap's logic.

Keywords: Collaboration, positive and negative authorization.

1 Introduction

A number of computer-based access control systems apply positive authorizations. Such authorizations define those accesses that are going to be allowed [1]. Authorization models supporting positive authorization apply closed policy, i.e. access is allowed to a few users only.

A serious problem with this approach is that a particular user without a given authorization from this closed policy can obtain an authorization from a different resource manager. Suppose a resource at an educational institution is accessible by lecturers and students. This institution has decentralized administration, i.e. lecturers and students are managed by different resource managers. Then a user who is a student in one subject and a lecturer in another one can lack authorization to access the resource as a student but might be able to obtain the desired authorization as a lecturer.

Authorization models supporting negative authorization apply open policy, i.e. accesses are to be allowed to all but a few users. Examples of negative authorizations are an approval of a bank card transaction in which the customer's account number is compared against a list of cancelled accounts numbers and verification systems in which only poor credit risks are noted in the credit check. A credit that is not negatively reported is assumed to be verified. Other applications are discussed in [6].

B. Apolloni et al. (Eds.): KES 2007/ WIRN 2007, Part II, LNAI 4693, pp. 125–132, 2007.
© Springer-Verlag Berlin Heidelberg 2007

Negative authorizations are often used because they give opportunities to include exceptions [4]. Without involvement of exceptions one should considerably increase the number of authorizations, and thus make the management of authorizations more complicated. Following the example with lecturers and students we consider a case where all lecturers are granted an authorization to a resource with the exception of the one who is also a student. Without negative negative authorizations, one should satisfy the requirement by giving a positive authorization to each lecturer except one. With help of negative authorizations, this situation can be resolved by granting a positive authorization to the group of lecturer and a negative authorization to the one who is a student.

We present a model that prevents conflicts generated by applying positive and negative authorizations to users accessing resources in a large networked system.

The rest of the paper is organized as follows. Related work, basic terms and concepts are presented in Section 2. The management model is described in Section 3. The system architecture is discussed in Sections 4. The paper ends with a description of the system implementation in Section 5 and a conclusion in Section 6.

2 Background

A formal model of role based access control (RBAC) is presented in [8]. Permissions in RBAC are associated with roles, and users are made members of appropriate roles, thereby acquiring the roles' permissions. The RBAC model defines three kinds of separation of duties - static, dynamic, and operational. Separation of duties was discussed in [4].

While RBAC provides a formal implementation model, Shibboleth [11] defines standards for implementation, based on OASIS Security Assertion Markup Language (SAML). Shibboleth defines a standard set of instructions between an identity provider (Origin site) and a service provider (Target site) to facilitate browser single sign-on and attribute exchange.

The semantic characterization of a four-valued logic for expressing practical deductive processes is presented in [3]. In most information systems the management of databases is not considered to include neither explicit nor hidden inconsistencies. In real life situation information often come from different contradicting sources. Thus different sources can provide inconsistent data while deductive reasoning may result in hidden inconsistencies. The idea in Belnap's approach is to develop a logic that is not that dependable of inconsistencies. The Belnap's logic has four truth values 'T, F, Both, None'. The meaning of these values can be described as follows: an atomic sentence is stated to be true only (T), an atomic sentence is stated to be false only (F), an atomic sentence is stated to be both true and false, for instance, by different sources, or in different points of time (Both), and an atomic sentences status is unknown. That is, neither true, nor false (None).

A user is defined as a valid domain identity at a particular organization O_i. A group is a set of users. A resource defines a set of protected Web objects.

A permission defines a right of a user to perform an action on a resource. An authorization gives a set of permissions to a user to execute a set of operations on a specific set of resources.

A billatice is a set equipped with two partial orderings $\leq_t$ and $\leq_k$. The t partial ordering $\leq_t$ means that if two truth values a, b are related as $a \leq_t b$ then b is at least as true a. The k partial ordering $\leq_k$ means that if two truth values a, b are related as $a \leq_k b$ then b labels a sentence about which we have more knowledge than a sentence labeled with a.

A *closed policy* permits specification of only positive authorizations and allows only those accesses that are explicitly authorized. An *open policy* permits specification of only negative authorizations and allows only those accesses that are not explicitly denied [5].

3 Management Model

3.1 Two Collaborating Organizations

Suppose an organization has two groups of users who are interested to access a particular resource. The two groups are managed by two different resource managers where one of them is applying closed policy and the other is applying open policy. Conflicts of access permits may occur if a user belongs to both groups (see Section 3.2). Such conflicts can be avoid if four-valued logic is applied.

If two collaborating organizations have groups described as above and wish to avoid conflicts related to access permits than sexteen-valued logic should be applied. Obviously further increase of the number of collaborating organizations would lead to higher number for the many-valued logic needed. In Section 3.2 we propose a possible solution that needs only four-valued logic.

3.2 Several Collaborating Organizations

This management model refers to collaborating organizations using resources hosted by some of these organizations.Suppose a resource at one organization can be accessed by two groups A_1 and B_1 of members of organization O_1 and two groups A_2 and B_2 of members of another organization O_2. Suppose these four groups are administered by four resource managers, two at organization O_1 and two at organization O_2. Assume the resource managers of groups A_1 and A_2 apply closed policy and the resource managers of groups B_1 and B_2 apply open policy.

Some of the conflict situations that may occur are - a user belongs to group A_1 or group A_2 and at the same time belongs to group B_1 or group B_2, or another user may be affiliated with two organizations and belong to three or four groups. In order to avoid such conflicts we propose use of four valued logic. All groups $A_i, i = 1, ..., n$ are considered as one group A and all groups $B_i, i = 1, ..., n$ are considered as one group B with respect to the resource.

Based on the truth table for Belnap's logic [3] we propose the following:

- A user belongs to group A and does not belong to group B. The user is authorized to access the resource.
- A user belongs to both groups A and B. The user is not authorized to access the resource before his/her membership is considered by the corresponding resource managers.
- A user is neither a member of group A nor of group B. The user is authorized to access the resource, provided he/she belongs to at least one of the organizations applying open policy.
- A user does not belong to in group A and belongs to group B. The user is not authorized to access the resource.

Let signify the resource owner organization as the publisher organization (PO) and an organization that uses the shared resources as a subscriber organization (SO).

Each subscriber organization administers its own domain users group membership for a particular subscribed resource. Two groups provide access control to each subscribed resource, namely, group A that applies close policy and group B that apples open policy to that particular resource.

To a publisher organization, the two groups defined at the subscriber organizations, will be referred to as a list of pairs {SO-domain, URI}. The domains are the subscriber organizations' domain names of the Domain Name System (DNS), for example 'uib.no' and 'hsh.no'. The Uniform Resource Identifiers (URIs) are the subscriber organizations' Web application addresses from which the publisher organization's access controller can do queries about group membership and thus access rights to a particular resource, of a user from a particular subscriber organization.

The resource owner needs to maintain a database containing subscriber organization table (SOT) for each subscriber organization subscribing to his/her resources. A particular SOT entry can be for example,

```
['uib.no', 'http://gatekeeper:letmein@siam.uib.no/groups'].
```

In addition to SOT, the publisher organization also needs to maintain a resource names and resource policies table (RPT). For example a resource 'math-1' with a default closed policy will have an entry in the RPT as ['math-1', 'A'].

Each subscriber organization will provide local groups management and groups database containing access control table (ACT) for each resource they wish to subscribe. Each resource is referred to by a pair name, PO-domain and will have two groups, referred by a triplet {A, name, PO-domain} for close policy and {B, name, PO-domain} for open policy. For example, a user at a particular subscriber organization 'uib.no', with a user identifier 'edpst' is a member of the closed policy group A of a resource named 'math-1' published by domain 'hsh.no', will then have an entry in ACT, ['edpst', 'A', 'math-1', 'hsh.no']. Data entries in ACT will determine weather a user has access right to a particular resource.

For each shared resource the resource owner, at the publisher organization needs to communicate the name of all the published resources to all potential subscriber organizations. This can be accomplished by providing descriptions of all resources in an Extensible Markup Language (XML) document at an URI collectively agreed before hand by all partners, for example,

```
'http://subscriber:gogetit@mat.hsh.no/resources'.
```

All subscribers can consult this XML file at any time for information pertinence to all published resources from a particular publisher organization, which then can be used to initialize or update information in subscriber organizations' databases represented in ACT.

Within this model, both the publisher organizations and the subscriber organizations need to provide Web services for each other in order to communicate user identities and authorizations to control access on shared Web resources. There are many ways of providing these services, where among most common ones are, Java based remote method invocation (RMI), XML remote procedure (XML-RPC) and Simple Object Access Protocol (SOAP). We proposed a simpler mechanism inspired by Representational State Transfer (REST).

Compared to other mechanisms, REST is much simpler because REST does not define its own transport protocol. REST depends on generic Web interface of HTTP GET, POST, PUT and DELETE. In order to provide and utilize Web services one needs only to deploy Web server and client that support XML documents formatting and parsing.

4 System Architecture

To support REST type Web services participating organizations need to install and maintain a so-called three-tier architecture of Presentation Tier, Application Tier, and Data Tier.

The presentation tier provides interfaces to the system. These interfaces are referred to by applications' clients using some published URIs. A client can query a service of an application by addressing its URI and using HTTP GET parameters, for example,

```
'http://gatekeeper:letmein@siam.uib.no/groups?user=edpst\&resource=math-1'.
```

We propose using Apache Web server for implementing the presentation tier.

The application tier implements system processing logic. Data submitted by client queries trigger processing events in the application tier. Depending on URI, programs or scripts in the application tier will be executed. The GET parameters will be used as input parameters to these programs that will provide a response or an error message. We propose using 'mod_python' to the Apache Web server. The application server will be programmed using Python scripting language. The module 'mod_python' supports the Apache capabilities by incorporating Python runtime interpreter within the Web server itself. All the

Apache's application programming interfaces (API) are directly connected to the Python runtime environment, which makes Web application programming simpler.

The data tier implements data store for the system. Information is stored and retrieved from a relational database management system (RDBMS). We propose using Oracle XE, a free small RDBMS from Oracle. It will then be easy to upgrade the database to enterprise level, if that need arises. The RDBMS is chosen since it should support Structured Query Language (SQL) standard. The database stores long-term persistence information about users, groups, resources and applications states. Data stored in the database will affect the behavior of the whole system. Applications initialize and modify information in the database. Data is written, replaced and deleted by application programs. There are many Python Oracle modules that integrate the application tier to the data tier in the system. We propose using DCOracle2 module.

By providing Web services architecture using simple REST style we believe that the system will be easy to implement and maintain. Any Web tools can be used, if they collectively support HTTP, XML and SQL.

We propose using the basic authentication scheme when a client connects to a server. The simple mechanism is a part security measure to the system. The basic authentication is implemented by the Web server and is easy to deploy. In addition of host address access control, this provide us with good enough security measure for the system mentioned in this paper. Thus, any client that knows the basic authentication parameters and belongs to clients-list of a server can make use of the services provided by the server.

5 System Implementation

The subscriber organizations provide a portal to their local users. By using cookies and redirect, an authenticated user can be transferred from local portal to a shared Web resource. The central issue in implementing the system is on how the XML responses from the server look like. We propose providing XML response containing security information together with the reply. Our proposal was based on this request/response scenario:

- Client uses basic authentication to send query of a particular service to a server.
- A server parses GET parameters and executes appropriate program if parameters are valid and the client belongs to the list of valid clients.
- A server formats the reply block in XML containing message identifier, timestamp and data blocks. The server then calculates a hash value for this reply block. The hash is then encrypted using symmetric key encryption using a shared secret key that both the server and the client know. A complete XML is sent back to the client.
- Client collects the reply block and verifies that the hash is valid. The client parses the XML and executes appropriate program with the reply provided by the server.

A typical example of XML reply document is as follows:

```
<?xml version="1.0" encoding="utf-8"?>
<zebra rows="2" reply="user\# hash="537851618991316fae7bd23b0b03adc4"
cipher="blowfish">
    <id>Ver2.06 GetAgent@mat.hsh.no \#12</id>
    <ts>20070304170853</ts>
    <user>
        <id>edpst</id>
        <domain>uib.no</domain>
        <resource>math-1</resource>
        <group>A</group>
    </user>
    <user>
        <id>sbe</id>
        <domain>uib.no</domain>
        <resource>alg-2</resource>
        <group>A</group>
    </user>
</zebra>
```

The client needs to trust that the response reply comes from the server and that the XML message has not been tempered with during transit. Hash function and symmetric key encryption provide this trust.

6 Conclusion

A problem arises if a user is affiliated with an organization applying both positive and negative authorization managed by different resource managers or with several organizations at the same time. This problem is difficult to solve and may not be a major issue if a conflict of access rights can be resolved applying many-valued logic. We propose a model that simplifies user access to resources in a large networked system.

References

1. Al-Kahtani, M., Sandhu, R.: Rule-based RBAC with negative authorization. In: 20th Annual Computer Security Applications Conference, Arizona (2004)
2. Belnap, N.J.: How a computer should think. In: Contemporary Aspects of Philosophy. Proceedings of the Oxford International Symposia, Oxford, GB, pp. 30–56 (1975)
3. Belnap, N.J.: A useful four valued logic. In: Dunn, J.M., Epstain, G. (eds.) Modern uses of multiple-valued logic, pp. 8–37. D. Reidel Publishing Co., Dordrecht (1977)
4. Bertino, E., Bonatti, P.A., Ferrari, E.: TRBAC: A temporal Role-Based Access Control model. ACM Tr. on ISS 3(3), 191–223 (2001)
5. Bertino, E., Jajodia, S., Samarati, P.: A Flexible Authorization Mechanism for Relational Data Management System. ACM Transactions on Information Systems 17(2), 101–140 (1999)

6. Bhatti, R., Bertino, E., Ghafoor, A., Joshi, J.B.D.: XML-based specification for Web services document security. IEEE Computer 37(4) (2004)
7. Davey, B.A., Priestley, H.A.: Introduction to lattices and order. Cambridge University Press, Cambridge (2005)
8. Ferraiolo, D., Cugini, J., Kuhn, D.R.: Role-Based Access Control (RBAC): Features and motivations. In: 1995 Computer Security Applications Conference, pp. 241–248 (1995)
9. Ferraiolo, D., Kuhn, D.R., Chandramouli, R.: Role-Based Access Control. Artech House, Computer Security Series (2003)
10. Schwoon, S., Jha, S., Reps, T., Stubblebine, S.: On generalized authorization problems. In: Proc. 16th IEEE Computer Security Foundations Workshop, Asilomar, Pacific Grove, CA, June 30-July 2, 2003, pp. 202–218. IEEE Computer Society Press, Los Alamitos (2003)
11. `http://shibbolethinternet2.edu`

IDS Based on Bio-inspired Models

Paolo Gastaldo[1], Francesco Picasso[1], Rodolfo Zunino[1],
Álvaro Herrero[2], Emilio Corchado[2], and José Manuel Sáiz[2]

[1] Dept. of Biophysical and Electronic Engineering (DIBE), Genoa University
Via Opera Pia 11a, 16145 Genoa, Italy
{paolo.gastaldo, francesco.picasso, rodolfo.zunino}@unige.it
[2] Department of Civil Engineering, University of Burgos
C/ Francisco de Vitoria s/n, 09006 Burgos, Spain
{ahcosio, escorchado, jmsaiz}@ubu.es

Abstract. Unsupervised projection approaches can support Intrusion Detection Systems for computer network security. The involved technologies assist a network manager in detecting anomalies and potential threats by an intuitive display of the progression of network traffic. Projection methods operate as smart compression tools and map raw, high-dimensional traffic data into 2-D or 3-D spaces for subsequent graphical display. The paper compares three projection methods, namely, Cooperative Maximum Likelihood Hebbian Learning, Auto-Associative Back-Propagation networks and Principal Component Analysis. Empirical tests on anomalous situations related to the Simple Network Management Protocol (SNMP) confirm the validity of the projection-based approach. One of these anomalous situations (the SNMP community search) is faced by these projection models for the first time. This work also highlights the importance of the time-information dependence in the identification of anomalous situations in the case of the applied methods.

Keywords: Unsupervised Learning, Projection Methods, Auto-Associative Back-Propagation, Computer Network Security, Intrusion Detection, Visualization.

1 Introduction

Intrusion Detection Systems (IDS's) [1] monitor traffic in computer networks and take, or suggest, defensive actions when they detect suspect activities. IDS's are common elements in modern infrastructures to enforce network policies. Today's commercial systems typically rely on a knowledge base of rules to discriminate normal from malicious traffic. The set of rules, however, is susceptible to inconsistencies, and continuous updating is required to cover previously unseen attack patterns.

Human beings seem to be able to learn without explicit supervision. One aim of unsupervised learning is to mimic this aspect of human learning and hence this type of learning tends to use more biologically plausible methods than those using error descent methods. Both projection and unsupervised methods [2], [3], [4], [5] can

B. Apolloni et al. (Eds.): KES 2007/ WIRN 2007, Part II, LNAI 4693, pp. 133–140, 2007.
© Springer-Verlag Berlin Heidelberg 2007

successfully apply to the visual based IDS's technology. This paper analyzes some of those technologies, and compares the performances of two actual implementations of such paradigms: Cooperative Maximum Likelihood Hebbian Learning (CMLHL) [6] and Auto-Associative Back-Propagation (AABP) neural networks [7].

2 Visual Inspection of Traffic in Modern IDS's

A visual-based IDS could be organized as follows (Fig. 1):

- packets traveling across the network are intercepted by a capture device;
- traffic is coded by a set of features spanning a multidimensional vector space;
- the compression component operates on feature vectors and yields a two or three-dimensional representation of the network traffic;
- the outcome of the projection component is presented to the network manager.

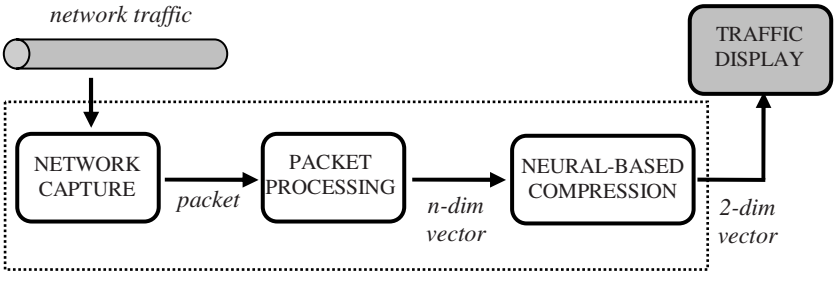


Fig. 1. A general schema of a visual-based Intrusion Detection System

The compression component processes an n-dimensional vector that has been previously assembled by a "packet processing" module, which extracts numerical features associated with each network packet. Hence, unlike other models which operate at the connection level [8], [9], the proposed IDS operates at the packet level. The design of the feature set is a crucial issue that has been thoroughly addressed in the literature [10]. It has been proved that timestamp, source and address port, and protocol uniquely identify a connection [11]. When dealing with Transmission Control Protocol (TCP) traffic, additional features may be required (e.g. to track connection state [12]); instead, User Datagram Protocol (UDP) traffic can be effectively characterized by a reduced feature set [4].

The compression component clearly is the core of the overall IDS, and is designed to yield an effective representation of network traffic. Its overall function is to support the network supervisor at detecting traffic anomalies. A connectionist approach seems to best fit the anomaly-detection problem mainly because it allows a system to learn empirically the input-output relationship between raw traffic data and its subsequent interpretation. Indeed, the outlier-detection method does not require any a-priori analytical formulation of the underlying phenomenon. Section 3 illustrates two different projection approaches to the dimensionality reduction mapping task of traffic data.

3 Nonlinear Projection Methods for Dimensionality Reduction

3.1 Neural Implementation of Exploratory Projection Pursuit

The classical statistical method of Exploratory Projection Pursuit (EPP) [3] provides a linear projection of a data set onto a set of basis vectors which best reveal the interesting structure in data; interestingness is usually defined in terms of how far the distribution is from the Gaussian distribution. Maximum Likelihood Hebbian Learning (MLHL) [13] is a neural implementation of EPP: it identifies interestingness by maximising the probability of the residuals under specific probability density functions which are non-Gaussian.

The Cooperative Maximum Likelihood Hebbian Learning (CMLHL) model [6] extends the MLHL paradigm by adding lateral connections [6], which have been derived from the Rectified Gaussian Distribution [14]. The resultant net can find the independent factors of a data set but does so in a way that captures some type of global ordering in the data set. Considering a D-dimensional input vector ($\mathbf{x}$), and a Q-dimensional output vector ($\mathbf{y}$), with W_{ij} being the weight (linking input j to output i), then CMLHL can be expressed [6] as:

1. Feed-forward step:

$$y_i = \sum_{j=1}^{D} W_{ij} x_j \; ; \; i = 1,...,Q \cdot \tag{1}$$

2. Lateral activation passing:

$$y_i(t+1) = \left[y_i(t) + \tau\left(b - Ay\right) \right]^{+} ; \; i = 1,...,Q \cdot \tag{2}$$

3. Feedback step:

$$e_j = x_j - \sum_{i=1}^{Q} W_{ij} y_i \; ; \; j = 1,...,D \cdot \tag{3}$$

4. Weight change:

$$\Delta W_{ij} = \eta . y_i . sign\left(e_j\right) | \, e_j \, |^{p-1} ; \; i = 1,...,Q; \; \; j = 1,...,D \cdot \tag{4}$$

Where: η is the learning rate, τ is the "strength" of the lateral connections, b is a bias parameter, p a parameter related to the energy function [6] and A a symmetric matrix used to modify the response to the data [6]. The effect of this matrix is based on the relation between the distances separating the output neurons.

3.2 Auto-associative Back-Propagation Networks

An Auto-Associative Back-Propagation (AABP) network involves a Multi-Layer Peceptron (MLP) structure [15] which typically includes at least one "hidden" layer. It associates an input vector, $\boldsymbol{x} \in \Re^D$, with an output vector, $\boldsymbol{y} \in \Re^Q$, computed as:

$$y_q(\mathbf{x}) = w'_{q,0} + \sum_{u=1}^{N_h} \left[w'_{u,q} \cdot \sigma\left(w_{u,0} + \sum_{k=1}^{D} w_{u,k} x_k \right) \right]; \; q = 1,...,Q \cdot \tag{5}$$

where, N_h is the depth of the sigmoid series expansion, and W represents the coefficients of the weights for the interconnections between the two upper layers. Empirical fitting drives the weights, W, so that the network best reproduces (in an Euclidean sense) the desired ($\boldsymbol{x}$, $\boldsymbol{y} \equiv \boldsymbol{x}$) mapping over a given training set.

Auto-associativity in the AABP model stems from imposing the output target equal to the input stimulus, hence $\boldsymbol{y} \equiv \boldsymbol{x}$ and $Q = D$; compression instead results from imposing that $N_h < D$. The overall goal of this unsupervised approach is a reduction in dimensionality by forcing the network to replicate the training sample distribution. At run-time, an AABP network associates each input vector with the "coding" values computed by the hidden neurons. Since the hidden layer is typically smaller than the input/output ones, these mapping outputs support the (lossy) transformation from the input space into a lower-dimensional representation. A three-layer AABP network implements a mapping that is, in fact, affine to linear Principal Component Analysis (PCA).

Non-Linear Principal Component Analysis (NLPCA) [7] was designed to circumvent the limitations of linearity inherent in the PCA model (Fig. 2). The output layer still imposes the auto-associative constraint $\boldsymbol{y} \equiv \boldsymbol{x}$, and a hidden layer supports dimensionality reduction. Both the compression and reconstruction sections, however, include an additional layer of neurons, thus yielding a five-layer network. At run-time after completion of training, the mapping outputs of the middle "coding" layer provide the low-dimensional representation of each input vector. The increase in power of representation conveyed by the NLPCA augmentation is remarkable. Typically, the "reconstruction" section and the compression section will always be symmetrical and will therefore always yield equivalent, universal (inverse) mapping capabilities.

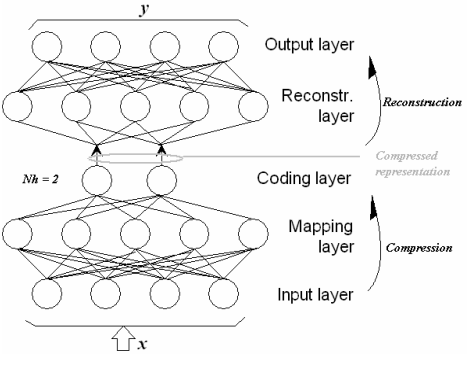


Fig. 2. The Non-Linear AABP network

The main advantage is that the compressed representation does not relate to any linear model but stems from a mostly general representation that is learned empirically. NLPCA techniques fit those domains in which 1) a non-linear representation best encompasses the observed empirical phenomenon, and at the same time, 2) a considerable number of empirical samples are available.

4 Using Projection Methods for Visual Based IDS's

In principle, any unsupervised method applies to the involved representation process; Self-Organizing Maps [1] and Vector Quantization-based methods [5] have had a considerable success in supporting IDS's technology. As compared with those models, projection based approaches offer the crucial advantage of combining compression ability and the support for graphical, intuitive representation. The following sections discuss the performance of the CMLHL model and the AABP model in the monitoring and detection of anomalous events in a real scenario.

4.1 Identification of SNMP Community Search

For simplicity and without loss of generality, the present IDS is targeted to detect traffic anomalies within the Simple Network Management Protocol (SNMP), which is a part of the Transmission Control Protocol/Internet Protocol (TCP/IP) protocol suite. SNMP is an application layer protocol that supports the exchange of management information between network devices. This protocol enables network administrators to manage network performance and is used to control network elements such as routers, bridges and switches. This property makes SNMP data quite sensitive and liable to potential attacks. Indeed, an attack based on the SNMP may severely compromise system security, as reported by CISCO [16].

In order to test the above mentioned models under the IDS's field, this research addresses some different types of attacks relying on the SNMP:

- SNMP port sweep: a port scan (or sweep) is an attempt to count the services running on a machine (or a set of machines) by probing each port for a response.
- SNMP community search: the community string can be seen as the SNMP password for versions 1 and 2. The intruder sends SNMP queries to try and determine the SNMP community string.
- MIB information transfer: the Management Information Base (MIB) collects information on each managed device, including sensitive data such as network and router information (e.g. IP and MAC addresses and VLan configuration). As specified by the Internet Activities Board, the SNMP is used to access MIB objects; thus, protecting a network from malicious MIB information transfer is crucial.

Two different data sets have been used in the present work:

– Data set 1 (utilized in [4]): it contains three consecutive sweeps on some machines. Port numbers 161, 162 and 3750 have been probed for SNMP. This is followed by an MIB information transfer (for a known or default community name).
– Data set 2: it differs from the previous one in that the attacker probes for 3 different community names on 3 different ports. That is, 3 different hosts try to check 3 different community strings ("public", "private" and "aab") for SNMP on ports 161, 1161 and 2161 at the same time. This attack is set between the port sweeps and the MIB transfer. After trying to find whether and where (hosts and port numbers) SNMP is running, the community string is requested in order to access the information contained in the MIB. When the community string has been found, the MIB information is read. It is worth highlighting that the projection models face the SNMP community search (included in this data set) for the first time.

Both data sets contained network packets related to normal and anomalous situations. They were captured from UDP traffic, as SNMP uses UDP as the transport protocol. Hence, the data sets included only packets that use UDP as transport layer, IP as network layer and some protocols (SNMP among them) as application layer.

4.2 Feature Extraction

The eventual network-based IDS for the detection of SNMP anomalous traffic is structured as shown in Fig. 1. The "packet processing" component generates feature vectors by working out information contained in the packet header. In the present research, network packets are characterized by using the set of features that already proved to be effective for detection of anomalous SNMP traffic [4]. The set of five features that are extracted from packets contribute to build up the neural-network input vector, $\mathbf{x} \in \Re^5$. These features can be listed as follows: protocol ID (the protocol of the packet), source port (the port number of the device that sent the packet), destination port (the port number of the host to which the packet was sent), size (packet size in Bytes) and time (the time when the packet was sent).

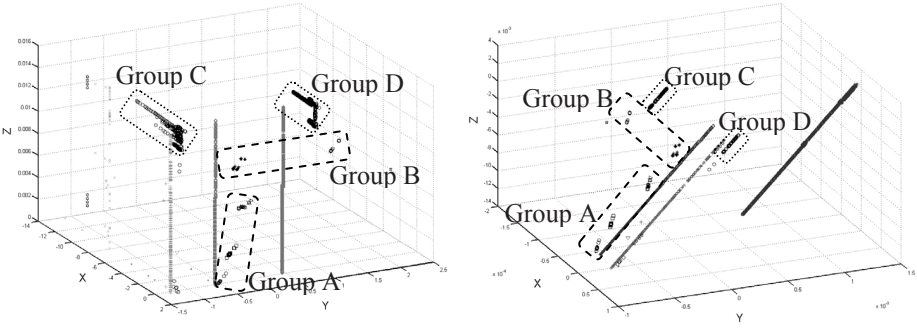
As such, at the output of the "packet processing" module the network traffic is mapped in a five-dimensional feature space. According to the set-up discussed in Section 2, the compression module exploits the projection method (either CMLHL or AABP) to generate a two-dimensional representation of the network traffic by starting from the original five-dimensional space. Thus, first an offline training phase uses empirical data to set the configuration of weight quantities for the neural network. Then, the eventual neural system is used to process the feature vectors generated at run-time and to feed the visual display.

4.3 Identifying Different Anomalous Situations

The following experimental verification shows the performance of three projection models (linear PCA, CMLHL and AABP) to analyse their response to the data sets described above. In these experiments, the configuration of the AABP network included a number of 30 nodes in the hidden layers (coding and reconstruction), while of course the number of neurons in the middle layer was N_h=2.

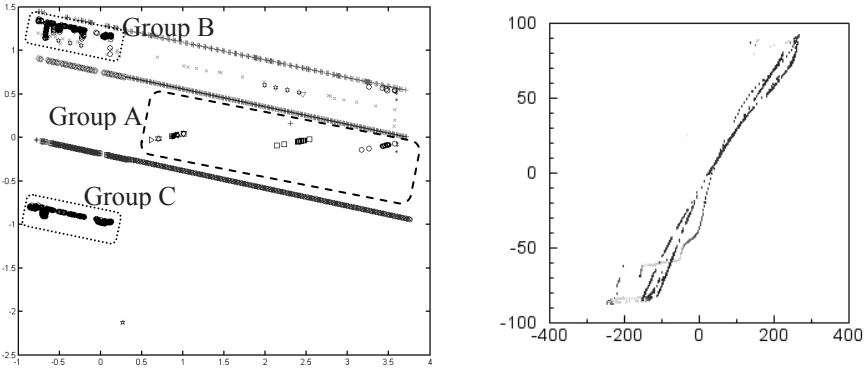
As can be seen in Fig. 3.a, the CMLHL projections manage to identify all the anomalous situations contained in Data set 2. The three sweeps (Group A) and the community search (Group B) are identified due to their non-parallel direction to the normal one. The MIB information transfer (Groups C and D) is identified due to its orthogonal direction with respect to the normal traffic (vertical and parallel straight lines) and to the high density of packets. Linear PCA projection was tested on the same data for comparison (Fig. 3.b). PCA could identify the port sweeps (Group A) and the community search (Group B), but failed at detecting the MIB information transfer: in the graph the MIB packets (Groups C and D) evolve in a direction parallel to the "normal" one.

It can be seen in Fig. 4.a how CMLHL is able to identify the anomalous situations (Groups A, B and C) included in Data set 1 as it has been explained before. This outcome shows the intrinsic robustness of CMLHL, which is able to properly respond to a complex data set that includes time as a variable. In the case of AABP, this issue is not clear at all when using the time as one of the input variables. Including a linear



a) Tridimensional CMLHL projection b) Tridimensional PCA projection

Fig. 3. Experimental verification of the projection methods for Data set 2: a) 3-D CMLHL projection; b) 3-D PCA projection



a) Bidimensional CMLHL projection b) Bidimensional AABP projection

Fig. 4. Bidimensional projections for Data set 1 (including time in the input variables): a) CMLHL projection; b) AABP projection

growing variable (time) forced the two inner neurons to learn a linear path, which is reflected in Fig. 4.b.

The comparative approach supports the view that projection methods are an interesting and powerful tool in the identification of anomalous situations. This study shows for the first time how the CMLHL is capable of identifying the anomalous situation related to an SNMP community search. Another interesting issue is the capability of CMLHL to process time information as one of the data variables. In contrast, the AABP network can not deal with the time variable, even though it obtains similar results to CMLHL by plotting its results against time [17]. The model can be extended by the application of multiagent systems in an architecture akin to the MObile VIsualization Connectionist Agent-Based IDS (MOVICAB-IDS) [18].

Acknowledgments. This research has been partially supported by the MCyT project TIN2004-07033.

References

1. Laskov, P., Dussel, P., Schafer, C., Rieck, K.: Learning Intrusion Detection: Supervised or Unsupervised? In: Roli, F., Vitulano, S. (eds.) ICIAP 2005. LNCS, vol. 3617, pp. 50–57. Springer, Heidelberg (2005)
2. Hertz, J.A., Krogh, A., Palmer, R.G.: Introduction to the Theory of Neural Computation. Addison-Wesley, Redwood City, CA (1991)
3. Friedman, J.H., Tukey, J.W.: A Projection Pursuit Algorithm for Exploratory Data-Analysis. IEEE Transactions on Computers 23(9), 881–890 (1974)
4. Corchado, E., Herrero, A., Sáiz, J.M.: Detecting Compounded Anomalous SNMP Situations Using Cooperative Unsupervised Pattern Recognition. In: Duch, W., Kacprzyk, J., Oja, E., Zadrożny, S. (eds.) ICANN 2005. LNCS, vol. 3697, pp. 905–910. Springer, Heidelberg (2005)
5. Zheng, J., Hu, M.: An Anomaly Intrusion Detection System Based on Vector Quantization. IEICE - Trans. Inf. Syst. E89-D(1), 201–210 (2006)
6. Corchado, E., Fyfe, C.: Connectionist Techniques for the Identification and Suppression of Interfering Underlying Factors. Int. J. Patt. Recogn. Artif. Intell. 17(8), 1447–1466 (2003)
7. Kramer, M.A.: Nonlinear Principal Component Analysis Using Autoassociative Neural Networks. AIChE Journal 37(2), 233–243 (1991)
8. Lippmann, R., Haines, J.W., Fried, D.J., Korba, J., Das, K.: Analysis and Results of the 1999 DARPA Off-Line Intrusion Detection Evaluation. In: Debar, H., Mé, L., Wu, S.F. (eds.) RAID 2000. LNCS, vol. 1907, pp. 162–182. Springer, Heidelberg (2000)
9. Sabhnani, M., Serpen, G.: Application of Machine Learning Algorithms to KDD Intrusion Detection Dataset within Misuse Detection Context. In: 2003 Int. Conference on Machine Learning, Models, Technologies and Applications, pp. 623–630 (2003)
10. Lee, W., Xiang, D.: Information-theoretic Measures for Anomaly Detection. In: Proc. IEEE Symposium on Security and Privacy (S&P 2001), pp. 130–143. IEEE Computer Society Press, Los Alamitos (2001)
11. Lee, W., Stolfo, S.J., Mok, K.W.: Mining in a Data-Flow Environment: Experience in Network Intrusion Detection. In: Proc. 5th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, ACM Press, San Diego, California (1999)
12. Lee, W., Stolfo, S.J., Mok, K.W.: Adaptive Intrusion Detection: A Data Mining Approach. Artificial Intelligence Review 14(6), 533–567 (2000)
13. Corchado, E., MacDonald, D., Fyfe, C.: Maximum and Minimum Likelihood Hebbian Learning for Exploratory Projection Pursuit. Data Mining and Knowledge Discovery 8(3), 203–225 (2004)
14. Seung, H.S., Socci, N.D., Lee, D.: The Rectified Gaussian Distribution. Advances in Neural Information Processing Systems 10, 350–356 (1998)
15. Rumelhart, D.E., McClelland, J.L.: Parallel Distributed Processing. MIT Press, MA (1986)
16. Cisco Secure Consulting: Vulnerability Statistics Report (2000)
17. Herrero, A., Corchado, E., Gastaldo, P., Zunino, R.: A Comparison of Neural Projection Techniques Applied to Intrusion Detection Systems. In: Sandoval, F., Prieto, A., Cabestany, J., Graña, M. (eds.) IWANN'2007. LNCS, pp. 1138–1146. Springer, Heidelberg (2007)
18. Herrero, A., Corchado, E., Sáiz, J.M.: MOVICAB-IDS: Visual Analysis of Network Traffic Data Streams for Intrusion Detection. In: Corchado, E., Yin, H., Botti, V., Fyfe, C. (eds.) IDEAL 2006. LNCS, vol. 4224, pp. 1424–1433. Springer, Heidelberg (2006)

Network-Based Anomaly Intrusion Detection Improvement by Bayesian Network and Indirect Relation

ByungRae Cha and DongSeob Lee

[1] Dept. of Computer Eng., Honam Univ., Korea
chabr@honam.ac.kr
[2] Dept. of Information & Communication Eng., Honam Univ., Korea
dslee@honam.ac.kr

Abstract. In this paper, Network-based anomaly intrusion detection method using Bayesian Networks was estimated probability values of behavior contexts based on Bayes theory and Indirect relation. The contexts of network-based FTP service was represented Bayesian Networks of graphic types. We profiled concisely network-based FTP behaviors using behavior context by prior, posterior and Indirect relation. And this method be able to visualize behavior profile to detect/analyze anomaly behavior. We achieve simulation to translate audit data of network into Bayesian network which is network-based behavior profile for anomaly detection.

1 Introduction

In particular, as Unix operation system, which is one of main components of the Internet network, and TCP/IP have many problems in terms of information protection, all the networks connected to the Internet have been seriously exposed to hackers' attacks. Authorization and access control alone are not enough to solve security problems, and intrusion detection system to protect information has been developed as the 2nd line defence. In order to perform anomaly detection, it is very important how behavior profiling is constructed. The network-based detection system detects intrusion by collecting packet data on network. It uses IP Address, port, TCP which are header information of network packet. Research of Bayesian technique in intrusion detection field is in the beginning stage. Bayesian technique to detect intruders which pretend to be normal users was presented by Mehdi Nassehi[1] as a technical report, and Steven L. Scott[2] applied this technique to perceive patterns of network intrusion detection and to detect intrusion and presented the values obtained into probability values, and described relations between data and intrusion model. This study applies Bayesian Framework for production of profile to represent behaviors in network area. This study overcomes uncertainty of intrusion detection by applying the Bayesian Network and indirect relation. And it provides graphic representation of new intrusion patterns and modified intrusion patterns for detection/classification information and analysis.

B. Apolloni et al. (Eds.): KES 2007/ WIRN 2007, Part II, LNAI 4693, pp. 141–148, 2007.
© Springer-Verlag Berlin Heidelberg 2007

2 Related Works

In 1987, a general intrusion detection model was presented. Intrusion detection models include Denning model[3], Shieh model[4] and Kumar model[5]. Kumar model is based on Colored Petri Net by Jensen, intrusion behavior is expressed as Colored Petri Net, and more than one initial status and a single final status are used to define matching in model. In particular, context is stored as a color of token, and conditions are expressed by guard expression and behaviors are expressed by behaviors of status.

Bayesian theory is one of methods to deal with uncertainty such as probability, Dempster-Shafer Theory and Fuzzy Theory. First, uncertainty means lack of necessary information to make a decision, and reasons of uncertainty include existence of partial information due to loss of information, collision between information, lack of confidence in information and limit to languages to present knowledge. Therefore, this study is to apply Bayesian theory for anomaly(new and modified) intrusion detection. Bayesian network expresses conditional independence of Bayesian theory as a graph of network type. That is, it expresses actual knowledge as given directed acyclic graphs. The Bayesian network is called a causal network or a belief network. Interference is possible through knowledge expressed by Bayesian network. The types of inferences are Causal Inference, Diagnostic Inference and Mutual Causal Inference. Research on the Bayesian technique in intrusion detection is still in the beginning. [1,2,6,7,8] ADAM[6] used Naive Bayes classifier that classifies packets using preliminary information on network anomaly detection. To detect intruders which pretend to be normal users, a detection system using the Bayesian technique was presented as a technical report by Mehdi Nassehi[1] and Steven L. Scott[2] applied the Bayesian technique for intrusion pattern recognition and anomaly detection and presented the relations with audit data and intrusion models as probability values. Cha[7,8,9] carried a Bayesain Framework simulation with UNM data through definition and application of the Bayesian technique for program behavior anomaly detection. And network-based anomaly intrusion detection system collects packet data on network to detect anomaly intrusion. The detection systems such as ADAM[6], NIDES, and SPADE detect anomaly behaviors using IP address, ports, and TCP status as header information of packet as audit data. Matthew made network anomaly detection system with PHAD and ALAD to detect anomaly on the packet header data and application layer.[10,11]

3 Network-Based Behavior Profiling of Bayesian Technique

We need methods to express behaviors for anomaly intrusion detection. The Bayesian framework model presented by this study can express network-based FTP service and provide graphic information for easy analysis. This study proposes how to notate using string to express behavior of behavior profile and Bayesian framework using Bayesian network of graphic technique. For intrusion

detection, audit data have to be processed from log data of network packet data. To express behaviors, we have to define a concept called session. Session indicates that many events are collected to express a behavior. Network-based behavior defines session as a collection of packet data through IP address, port number, and TCP communication procedure.

3.1 Analysis of FTP Service of DARPA 2000 Intrusion Data

For construction of network normal behavior profiles, we need a technique that describes one behavior. To express session, we need notation of behavior pattern to be defined. As signs to describe a behavior used in this study, "<" and ">" mean the beginning and end of each session. And "-" means division between packets information. For representation of network behaviors, Figure 1 presents behavior pattern notation of Network-based FTP service session using DARPA 2000 NT Data[12] of MIT Linclon University for construction of network behavior profiles. When the Bayesian network is prepared with a use of an example of network behaviors in Figure 1, it becomes a behavior to FTP and FTP-data service as in Table 1.

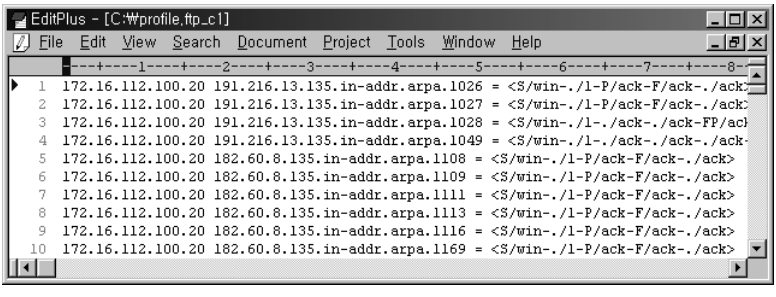


Fig. 1. Behavior pattern notation of Network-based FTP service

Table 1. Classification of FTP and FTP-data session

FTP and FTP-data session of Inside Tcpdump Data : 716					
FTP-data : 599		FTP : 117			
Packets of Min	5 > 3	Packets of Min		1 > 0	
Packets of Max	1508 > 231	Packets of Max		122 > 106	
Completion Session : 746		Incompletion Session : 20			
	Reset	Reset	Reset-Oneway	Oneway	
745	1	9	1	4	6

3.2 Apply to Bayesian Network and Indirect Relation

For network anomaly detection, first, the Bayesian network is applied and normal behaviors of packet data are profiled. The header information of packet is used for detection of network intrusion and misuse. This study classifies behaviors using

packet header information of network on TCP/IP service according to different services and detects anomaly by profiling normal behaviors according to network services. Most of the network intrusion detection systems detect anomaly by anomaly of TCP/IP packet and various characteristics of packets in respect to intrusion. This study is to classify network anomaly intrusion more clearly by applying restrictions to specific services in header information of packets. This study profiles normal behaviors in FTP services using the Bayesian network. Action of control port and data port are profiled by the Bayesian network using FTP service packet data. For FTP service, if GET or PUT commands are executed in control port, data port is opened and data are transmitted. In this case, branch of the Bayesian network is presented in Definition 2[7]. When transmission of data is completed, data port is closed and merge process of Definition 3[7] is presented. Here, The control and data ports have a indirect relation.

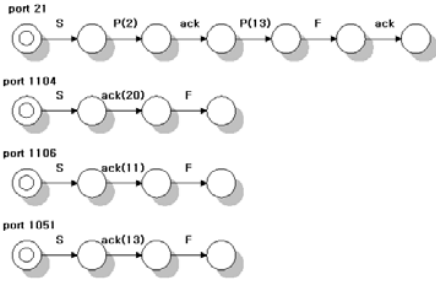


Fig. 2. Profiling of FTP Service using Bayesian Network

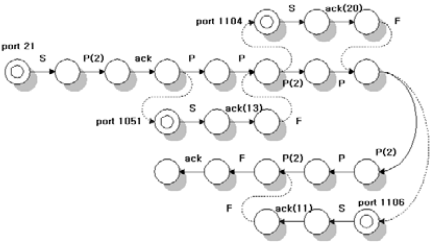


Fig. 3. Profiling of FTP Service using Bayesian Network and indirect relation, CASESEN attack

When an indirect relation is found during FTP service of control port and data port at each state, probability value that corresponds to a demerit of each state is applied. Demerit δ is defined as $1+(1-\beta)^2$, and β is Bayesian probability value.

Collected control port and data port behaviors of FTP are profiled. A action involved in FTP and FTP-data service whose source IP is 172.16.112.100, whose port No. is 20, whose destination IP is 172.16.113.204, whose ports are 1052, 1104, and 1106 respectively is profiled as in Figure 2. From a destination IP point of view, control port No. 21 and data port No. 1052, 1104, and 1106 have a indirect relations. Figure 3 shows indirect relation of control port and data port by dot lines. Figure 4 shows sequence alignment processing for NTIS attack using MSA algorithm. Figure 5 shows NTIS profiling by Bayesian network.

4 Simulation of Anomaly Detection

This simulation that is designed to detect anomaly with the Bayesian technique applied uses 2000 NT network attack data in DARPA Intrusion Detection Data

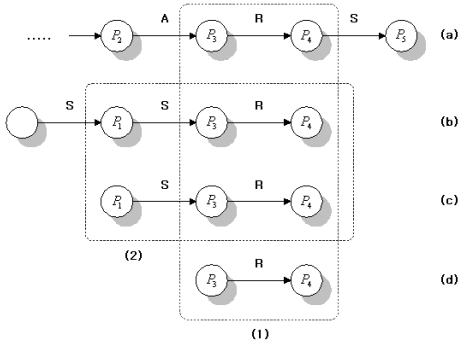


Fig. 4. Sequence alignment process for NTIS attack

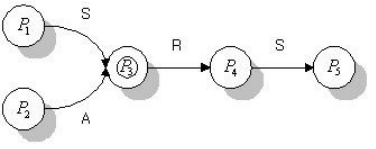


Fig. 5. NTIS Profiling by Bayesian Network

collections to detect attacks to FTP service. The simulation for anomaly intrusion detection detects CASESEN attack, SECHOLE attack and NTIS attack included in FTP and FTP-Data. The CASESEN attack as U2R intrusion makes bad use of object directory that is sensitive to NT system. For the SECHOLE attack, intruders pretend to be normal users and lock systems by the SECHOLE program. Matthew[10] carried anomaly detection using 33 specific data of subnetworks. This study carries anomaly detection using classification information, sessions, and flags of packet headers that are high-ranked network data. For network anomaly detection, this study profiles network services using prior information and indirect relation of behaviors in each service and the Bayesian network.

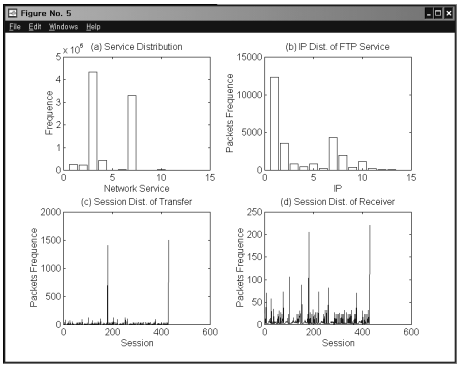


Fig. 6. Prior information for Network Anomaly Detection

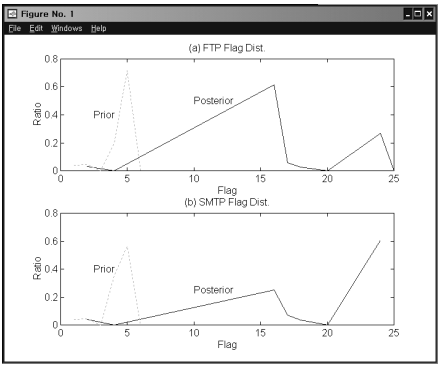


Fig. 7. Prior and Posterior Prob. Dist. of FTP and SMTP by Flags

Figure 6 presents prior information and preprocessing using packet header information of network. As a preprocessing, this study classifies networks according

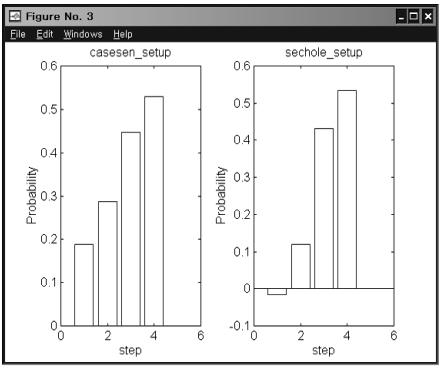


Fig. 8. Posterior Prob. of CASESEN and SECHOLE attack

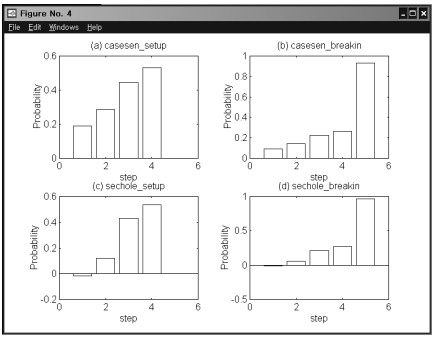


Fig. 9. Posterior Prob. of CASESEN and SECHOLE attack by Indirect Relation

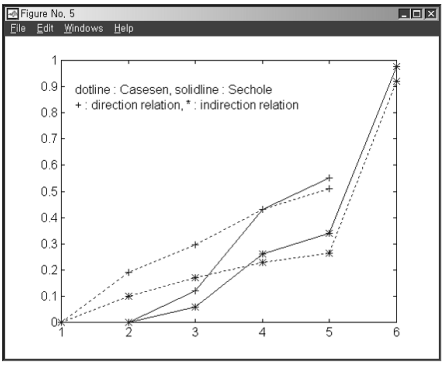


Fig. 10. Prob. values change by Indirection Relation

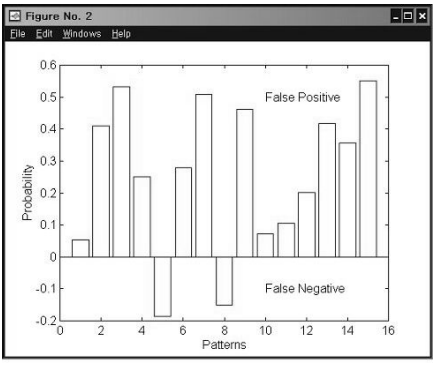


Fig. 11. Anomaly Detection Rate of NTIS sessions

to services, sessions are classified according to IP, port number, connection of TCP protocol, data transmission, services such as cancel of connection involved in FTP service and transmitted and received packet header information in each session and obtains information by TCP flag. Figure 7 presents prior probability and posterior probability distribution by network FTP service and flags of SMTP service. Compared to other services, it is found that FTP service has a dense distribution of PUSH and Ack flag.

100% of the CASESEN attack and the SECHOLE attack to normal behavior profiles are detected by the Bayesian posterior probability values as seen in Figure 8. For the CASESEN attack, as intrusion progresses, probability values of anomaly detection increases step by step. The SECHOLE attack is not detected at all in the first session, but when the expanded Bayesian network is used, detection probability value increases gradually.

Figure 9 shows an indirect relation to intrusion detection by the Bayesian network is applied for the CASESEN attack and the SECHOLE attack, probability values are remarkably expanded as in Figure 9 (b) and (d) compared to Figure 9 (a) and (c), and their patterns are clearly detected. Clear classification of patterns are provided as expanded probability values for the classification of intrusion patterns and detection of intrusion pattern by indirect relation technique. Figure 9 (c) and (d) don't detect first session as four sessions exist in the SECHOLE attack. However, probability values of the SECHOLE attack is scaled by indirect relation and probability value in Figure 9 (c) is expanded as much as that of Figure 9 (d). Clear detection of attack is not achieved by individual sessions in FTP service, but as detection information is provided by indirect relation, attacks are detected by probability value as attack progresses. Figure 10 shows probability values change of CASESEN attack(dot line) and SECHOLE attack(solid line) by Direction relation(+) & Indirection relation(*). Figure 11 shows anomaly intrusion detection rate 87% of NTIS attack and doesn't detect fifth and eighth sessions. The anomaly intrusion detection rate of NTIS attack has 13% of False Negative rate.

If packet data of normal behavior sessions in the FTP service are not enough, calculated value to obtain prior and posterior information is not much, but anomaly detec-tion by the Bayesian probability value is hard. Finally, it is found that there is much left to be detected. However, when only one session is profiled with the Bayesian network, there is much left to be detected, but non-detection rate can be decreased by expansion and indirect relations of the Bayesian network. Conversely, as the number of packet data that constitute sessions in one session increases, calculation amount of probability values increases, but intrusions are clearly detected and classified.

5 Conclusion

This study applies the Bayesian technique to solve uncertainty of anomaly intrusion detection. For anomaly detection, behavior profile should be constructed. By comparing constructed normal behavior profiles with new behaviors, normal profiles are separated from abnormal ones. Prior to anomaly intrusion detection, it is very important to identify how behavior profiling is constructed. It suggests how to profile network behaviors according to services using the Bayesian network and Indirect Relation for high anomaly detection rate. Network behaviors are classified according to network services, and prior and posterior relations of normal packets through TCP/IP Flag are profiled as probability values with the Bayesian network. And this paper suggests a graphic-based Bayesian Framework to describe and analyse behaviors in network-based anomaly detection. Network-based profiling is graphically constructed using Bayesian network. In simulation, The anomaly intrusion detection that uses DARPA intrusion data collections detects the CASESEN attack, SECHOLE attack and NTIS attack.

When packet data of normal behavior session in FTP service are scanty, calculated values by prior and posterior information are small and detection by

the Bayesian probability value is difficult. Finally, non-detection rate increases. However, if one session is profiled by the Bayesian network, intrusions are undetected, but by the expansion and indirect relations of the Bayesian network, non-detection rate can be decreased. Conversely, as the number of packet data that constitutes sessions to a session increases, calculated amount of probability values increases, but anomaly intrusions are clearly detected and classified.

References

1. Nassehi, M.: Characterizing Masqueraders for Intrusion Detection. Computer Science/Mathematics (1998)
2. Scott, S.L.: A Bayesian Paradigm for Designing Intrusion Detection Systems, Computational Statistics and Data Analysis (June 20, 2002)
3. Denning, D.E.: An Intrusion-Detection Model. IEEE Transaction on Software Engineering SE-13(2), 222–232 (1987)
4. Shieh, S.-P., Gligor, V.D.: On a Pattern-Oriented Model for Intrusion Detection. IEEE Transaction on knowledge and Data Engineering 9(4) (July/August 1997)
5. Kumar, S., Spafford, E.H.: An Application of Pattern Matching in Intrusion Detection, Technical Report CSD-TR-94-013 (June 17, 1994)
6. Barbara, D., Couto, J., Jajodia, S., Popyack, L., Wu, N.: ADAM: Detecting Intrusions by Data Mining. In: Proceedings of the 2001 IEEE Workshop on Information Assurance and Security (2001)
7. Cha, B.R.: The Prototype of Bayesian framework based on XML for System Call Profiling, GESTS Int'l. Trans. Computer Science and Eng. 15(1) (2005)
8. Cha, B.R.: FTP Anomaly Detection Improvement by Indirection Relation and BF-XML Profiling. Journal of Convergence Information Technology 1(1) (December 2006)
9. Cha, B.R.: Improvement of Anomaly Intrusion Detection Performance by Indirect Relation for FTP Service. In: IWANN 2007. LNCS, vol. 4507, pp. 895–902. Springer, Heidelberg (2007)
10. Mahoney, M.V., Chan, P.K.: Learning Nonstationary Models of Normal Network Traffic for Detecting Novel Attacks (2002)
11. Mahoney, M.V., Chan, P.K.: PHAD: Packet Header Anomaly Detection for Identifying Hostile Network Traffic, Florida Institute of Technology Technical Report CS-2001-04 (2001)
12. `http://www.ll.mit.edu/IST/ideval/data/data_index.html`

Security Policies Tuning Among IP Devices

Simone Ferraresi[1], Emanuele Francocci[2], Alessio Quaglini[2], and Francesco Picasso[3]

[1] ELSAGDATAMAT
V. Laurentina, 760 - 00143 Roma, Italy
simone.ferraresi@elsagdatamat.com
[2] INFOCOM Dept.- Sapienza Università di Roma
Via Eudossiana, 18 - 00184 Roma, Italy
[3] DIBE Dept.- Univ. of Genova
Via All'Opera Pia 11a - 16145 Genova Italy
francesco.picasso@unige.it

Abstract. Packet filtering is a core operation in network security management. As networks and services become more complex, managing access-list rules becomes an error-prone task. Conflicts among policies can cause holes in security and prevent the load optimization in the network. Here we propose an algorithm to automatically tune up the configurations of IP level security devices. Our aim is to keep the external behaviour (end-to-end) unchanged, while removing distributed inconsistencies all over the network. The algorithm is described in full details and results of the software implementation are shown.

Keywords: IP, Policies, Rules, Conflicts, Packet Filtering.

1 Introduction

A key feature of secure systems, including networked ones, is the management of security policies, from those at high level down to the platform specific implementation. Security policy defines constraints, limitations and authorization on data handling and communications. In a network environment many problems may also arise due to the inconsistency of policies implemented by firewalls and secure access gateways interconnected over an insecure network (e.g. in a S-VPN context). As distributed secure systems increase in complexity, policy configuration and maintenance tasks become increasingly prone to errors.

Today policy definition is often performed manually. This process is error prone and can give rise to holes in security of the entire system, which are hardly detectable before suffering an attack. For this reason a verification phase is needed to check whether the global behaviour of the network architecture and of the single devices are consistent with the aim of the network administrator. The same verification phase should be performed either when a policy is defined or after any modification or integration.

In general a policy conflict occurs when desired effects of a policy are ambiguous or not clear [2]. In this article we will focus on determining consistent policies in order to obtain a global conflict-free network configuration.

B. Apolloni et al. (Eds.): KES 2007/ WIRN 2007, Part II, LNAI 4693, pp. 149–158, 2007.
© Springer-Verlag Berlin Heidelberg 2007

2 Modelling of Rules Relation

Firewall rules specify a filtering condition $\mathcal{C}$ and an action $\mathcal{A}$. The condition $\mathcal{C}$ aims at selecting those IP packets to which applies the corresponding action. Packet filtering is based on the values of five header fields: IP destination and source addresses (*dst_ip*, *src_ip*), destination and source port numbers (*dst_port*, *src_port*) and the IP header protocol type (*prot*). For each of these selectors we need to specify a value or a range of values. To each predicate listed in the policy file we associate a rule R defined as a five-tuple of selector variables ranges:

$$R = \{S_{prot}, S_{src_ip}, S_{src_port}, S_{dst_ip}, S_{dst_port}\}$$

where $S_{sel} = ANY$ or $S_{sel} = [sel_{min}, sel_{max}]$
for $sel \in \{prot, src_ip, src_port, dst_ip, dst_port\}$, ANY being a "don't care" flag. Equivalently we could replace ANY by the entire possible range of the corresponding selector. In the following we refer to the five selectors with a variable $i \in F = \{$prot, src_ip, src_port, dst_ip, dst_port$\}$. We denote the i-th selector range associated to a given rule R by $R[i]$.

We can now formally state the basic relations among rules for the conflict analysis.

Definition 1. *Rules* R_x *and* R_y *are* ***Disjoint*** *if the intersection of at least one field in* R_x *and the corresponding field in* R_y *is the empty set.*

Definition 2. *Rules* R_x *and* R_y *are* ***Exactly Matching*** *if every field in* R_x *is equal to the corresponding field in* R_y*.*

Definition 3. *Rules* R_x *and* R_y *are* τ***-Partially Matching*** *if there are* τ *fields in* R_x *that are partially overlapped with the corresponding fields in* R_y *and there are exactly (5 -* τ*) fields in* R_x *that are either subsets or equal to the corresponding fields in* R_y*.*

Definition 4. *Rules* R_x *and* R_y *are* ***Correlated*** *if some fields in* R_x *are subsets of or equal to the corresponding fields in* R_y *and the rest of the fields in* R_x *are supersets of the corresponding fields in* R_y*.*

Definition 5. *Rules* R_x *and* R_y *are* τ***-Partially Correlated*** *if there are* τ *fields in* R_x *that are partially overlapped with the corresponding fields in* R_y *and some fields in* R_x *are subsets of or equal to the corresponding fields in* R_y *and the rest of the fields in* R_x *are supersets of the corresponding fields in* R_y*.*

3 Conflict Scenario

In this section we will try to analyse two simple scenarios in order to fully understand what our problem is.

In the first scenario, represented in Fig. 1, we consider two devices - respectively device 1 and device 2 - in the same path.

Definition 6. *A **path** is a directed sequence of devices which connects two networks.*

Definition 7. *Taken a device in the path as a reference, we refer to any preceding device in the path as **upstream device** and to any following device in the path as **downstream device**.*

Definition 8. *Two networks connected by a path are referred to as **net rule terminals.***

Let us imagine that both the devices only have one rule in their rule lists. The rule in the first device allow traffic from any host in network 151.1.0.0/16 to any host in network 152.1.0.0/16, using any source port, any destination port and any IP protocol. The second device, however, only allows host 151.1.1.2 in the source network to communicate with any host in the destination network (152.1.0.0/16) using any source port, any destination port and any protocol.

Examining this scenario, it is very easy to understand that the traffic matched by the rule in the second device is a subportion of the traffic matched by the rule of the first device: it means that some traffic allowed by the first device will be blocked by the second device, resulting in having spurious traffic.

Definition 9. *A traffic flow is referred to as **spurious traffic** when it is allowed by a device, but blocked by any downstream device along the path.*

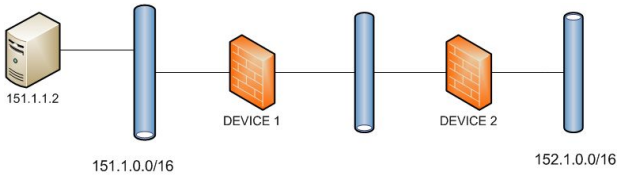


Fig. 1. The first scenario. One path composed by two devices connects two net rule terminals.

If we swapped the rules of the two devices, we would obtain a situation in which the first device has the most specific rule and the second device the most general one; this time, some traffic matched by the rule in the second device would never get to this device since it would be blocked by the first one; this traffic will be called eclipsed traffic.

Definition 10. *A traffic flow is referred to as **eclipsed traffic** when it is allowed by a device, but blocked by any upstream device along the path.*

Although this scenario is a very simple one, it can be useful to understand our goal: what we want to determine is a set of rules which would suit any device in the same path, i.e. we want to determine a set of rules which identifies the portion of traffic that should flow through the path in order to substitute this set of rules in all the rule lists of the devices in the path; we will call this set of rules path policy.

Definition 11. *A **path policy** is a high level policy described by a rule list which identifies the portion of traffic that flows through a path.*

In the second scenario we will have to deal with another problem: given that we have determined some paths and one path policy for each of them, it is likely that some devices belong to more than one paths. These devices will be called common nodes.

Definition 12. *A **common node** is a device which belongs to more than one path.*

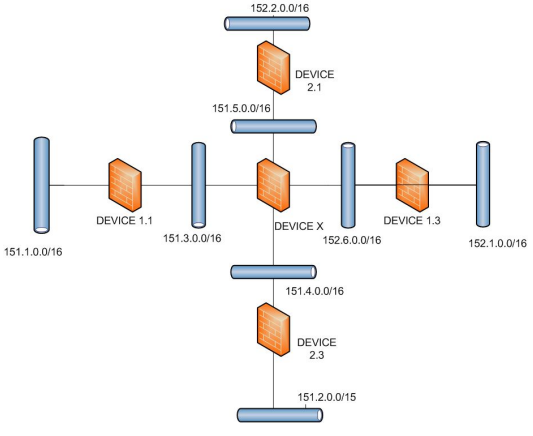


Fig. 2. The second scenario. Two paths composed by three devices each run through the same common node.

Let us consider the simple situation shown in Fig. 2: two paths, each of them composed by three devices and both of them sharing the same device in the middle (i.e. the second device of the first path coincide with the second device of the second path). This device will be our common node. It is easy to understand that we will have to determine a new policy for this device taking into account both the first and the second path policy. More in general, a common node policy will have to be consistent with the policies of all the paths running through it.

4 The Algorithm

4.1 Phase One: Data Import

The input data needed by the algorithm consists of:

- the addresses of the networks in the architecture,
- the IP address and mask of devices interfaces to the networks,
- the rules contained in each device rule list.

4.2 Phase Two: Rules Classification

In this phase a classifier reads one by one the rules in the rule lists and determines the relations between each rule and all the rules previously analysed. This information will be used to build a data structure called *pseudo tree*:

Definition 13. *A **pseudo tree** is a hierarchically organized data structure in which the relation between any couple of rules is made explicit*

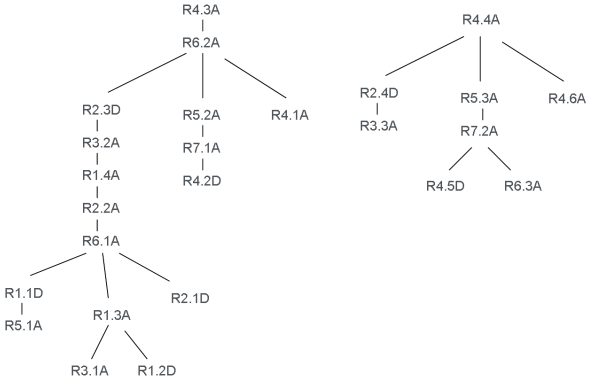


Fig. 3. The Pseudo tree. Rm.nX means that the node represents the n-th rule of the m-th device. X can be either A if the action is ALLOW or D if the action is DENY.

Algorithm 1. Phase 2: Classifier

```
1: Create new Complete Pseudo Tree CPT
2: foreach rule r in Ruleset do
3:     foreach ClassifiedRule cR in CPT do
4:         classify(r, cR)
5:         if r is to be fragmented then
6:             fragment(r, classifiedRule)
7:             remove(r, pseudoTree)
8:             insert fragments at the bottom of the Ruleset
9:         else
10:            insert(r, CPT)
```

A pseudo tree might be formed by more than one tree. Each tree node represents a rule of the rule list. The relation parent-children in the trees reflects the inclusion relation between the portions of traffic matched by the selectors of the rules: a rule will be a child node if the traffic matched by its selectors is included in the portion of traffic matched by the selectors of the parent node. Any two rules whose selectors match no overlapping portions of traffic will not be related by any inclusion relation. In each tree there will be a root node which represents a rule that includes all the rules in the tree and there will be one or more leaves which represent the most specific rules.

When the classifier finds a couple of rules which are not inclusively matching, exactly matching or disjoint, it will split one of them into two or three pieces in order to obtain fragments which can be classified in the pseudo tree.

4.3 Phase Three: Paths Detection

Examining the source and destination addresses of the rules in the pseudo tree, we can identify a large number of net rule terminals. It can be demonstrated that, in our analysis, it is sufficient to take into consideration only the net rule terminals corresponding to the rules which are leaves in the pseudo tree. In other words, in this phase our algorithm takes the most specific rules of the architecture, i.e. the leaves of the pseudo tree, and gets the source and destination addresses associated with these rules. Using these net rule terminals, our algorithm locates them in the network architecture and discovers all the possible paths from the source network to the destination network.

Algorithm 2. Phase 3: Paths detection

```
1: foreach Leaf leaf in Complete Pseudo Tree CPT do   ▷ the leaves of the CPT are the most specific rules
2:     get source and destination networks of the rule
3:     locate source and destination network in the architecture
4:     path[]=pathFinder(sourceNet,destinationNet)
5:     save path[] in ArrayList paths[][]   ▷ It saves the directed sequence of IPdevices and the source and destination networks associated with the path
```

The output of this phase consists of lists of paths: each list refers to one couple of net rule terminals and contains all the possible paths connecting these terminals.

4.4 Phase Four: INTRA Semplification

In this phase some operations upon the single rule lists of the devices in the architecture is performed. The aim of these operations is to restrict the number of rules in every rule list without changing the external behaviour of the device. We ought to take into consideration some data structures, called *device pseudo tree*, one for each device, obtained from the pseudo tree by considering only the rules belonging to one device. Chances are that one rule might have the same action as a rule which directly includes it, i.e. the former is a child to the latter. This means that the child rule is in a way redundant because, if it were not in the rule list, the same portion of traffic would be matched by the parent rule which has got the same action. Therefore, our algorithm checks such redundancies in every device pseudo tree and deletes the child rules.

4.5 Phase Five: INTER Elaboration

This phase can be considered as the core phase of the algorithm; its goal is to analyse all the rules in the rule lists of the devices in a path and to determine a

Algorithm 3. Phase 4: INTRA Simplifier

```
1: foreach Node node in Complete Pseudo Tree CPT do
2:     get the id of the device the rule belongs to
3:     if exists Device Pseudo Tree.id==id then
4:         insert(node, devicePseudoTree)
5:     else
6:         create(Device Pseudo Tree DPT, id)
7:         insert(node, DPT)
8: foreach DevicePseudoTree DPT do
9:     foreach Rule rule in DPT do
10:        if rule.action==rule's father.action then
11:            delete(rule)
12: update CPT
```

policy, namely the path policy. Putting aside so far the problem of the common nodes that will be analysed in the following phase, these path policies can be considered as the sets of rules which should replace the actual sets of rules in each device of the path. By doing this, each device of one path will have the same set of rules as the other devices in the path. This fact means that there will not be anymore either spurious or eclipsed traffic in the path because the rules of each device will exactly match.

In order to operate this elaboration, the algorithm will build one pseudo tree for each path we detected in the architecture and will populate these trees only with the rules belonging to the devices of the path to which the tree is associated. At the end of this operation, we will have as many pseudo trees as the number of the paths we are considering. In each of these trees we will find the rules of the devices that constitute the path; subsequently the algorithm will operate an elaboration on these trees, called *INTER elaboration*, erasing some of the rules of each tree. This elaboration follows two simple rules:

- a portion of traffic will be allowed only if all the devices in the path allow that portion of traffic;
- a portion of traffic will be blocked if at least one device in the path blocks that portion of traffic.

Algorithm 4. Phase 5: INTER Elaboration

```
1: foreach Path path in paths[][] do
2:     Create a new Path Pseudo Tree PPT
3:     foreach Rule rule in Complete Pseudo Tree CPT do
4:         get the id of the device the rule belongs to
5:         if device belongs to path then
6:             insert node in PPT
7: foreach Path Pseudo Tree PPT do
8:     Path Policy pathPolicy=INTERElaboration(PPT)
```

4.6 Phase Six: COMMON NODE Elaboration

At this point, we have determined all the path policies associated with the paths in the network architecture. The next problem to deal with is that some devices are likely to belong to more than one path.

Algorithm 5. Phase 6: COMMON NODE Elaboration

```
1: foreach Common Node commonNode do
2:     Create a new Common Node Pseudo Tree CNPT
3:     foreach Rule rule in Complete Pseudo Tree CPT  do
4:         get the id of the device the rule belongs to
5:         if device belongs to a path running through commonNode then
6:             insert node in CNPT
7: foreach CommonNodePseudoTree CNPT do
8:     Node Policy nodePolicy=COMMONElaboration(CNPT)
```

In this situation it is necessary to find a way to determine a new set of rules for each device which will be consistent with the path policies of the paths passing through this device. In order to achieve this goal, the algorithm builds one new pseudo tree for each of the common nodes including in it all the rules belonging to the path policies corresponding to that common node. We will obtain as many pseudo trees as the number of common nodes in the architecture. Subsequently the algorithm operates an elaboration on these trees, called *COMMON NODE elaboration*, erasing some of the rules from each tree. This elaboration follows two simple rules:

- a portion of traffic will be blocked only if all the path policies of the common node block that portion of traffic;
- a portion of traffic will be allowed if at least one path policy of the common node allows that portion of traffic.

5 Results and Software Implementation

Several important results have been achieved with this algorithm.

The first one is that, assuming that the rules inside the device are already ordered, the application of the algorithm to a network architecture which shows a certain behaviour returns a network architecture which shows the identical behaviour it showed before. This means that if a certain portion of traffic was able to flow through the network, the same portion of traffic will be able to reach its destination after the application of the algorithm as well. Symmetrically, if some traffic was blocked somewhere in the architecture and was not able to reach its destination, it will continue being blocked after the application of the algorithm. The key point is that once the algorithm runs, the resulting device rule lists are fully consistent, i.e. neither spurious nor eclipsed traffic can be found anymore. This is achieved under the constraint of not changing the overall behaviour of the architecture.

The second result concerns the load optimization of the network. Let us consider an example: some packets originated by a source network and destined to a network could flow through some devices and then, at a certain point, be blocked by a firewall before reaching their destination. After the application of the algorithm, all the devices in the architecture will be tuned up and this means that if a packet is blocked by a device in its path, then it would be also blocked by all the other devices in the path. The packet in question would be instantly

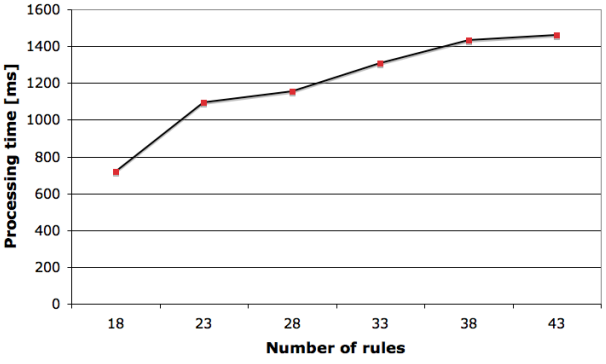


Fig. 4. Overall processing time for intra and inter device conflicts resolution

blocked by the first device in the path, disburdening the network of the useless load of some devices processing a packet that could not reach its destination.

The third result concerns security issues; let us consider the same situation described in the previous paragraph; saying that a packet is not able to reach its destination means that some devices in the path do not want this packet to reach its destination. After the application of the algorithm, this packet would be stopped at the beginning of the path, killing a traffic flow of no use except for being exploited for security attacks.

Analysis and resolution methodologies have been implemented in a software tool delevolped in C#, the object oriented Microsoft programming language. The tool is named I^2DCR.

The algorithm has been tested on several architectures and scenarios which allowed to verify the actions undertaken every time. In Fig. 4, we show the results of some tests executed on a plausible architecture changing the total number of rules. We plotted the required processing time of the software as a function of the total number of rules. The relationship between number of rules and processing time appears to be linear. Obviously, both the processing time and the percentage of time needed by every phase are deeply influenced by the network architecture and by the types of conflicts detected among the rules.

In all the scenarios we tested, the software performed the modifications in the devices rule sets as expected, with negligible running times (in the order of seconds on a PC Pentium IV 1.5 GHz) considering that this kind of processing is not supposed to be a real-time processing.

References

1. Hare, C., Siyan, K.: Internet Firewalls and Network Security. New Riders (1996)
2. Ferraresi, S., Pesic, S., Trazza, L., Baiocchi, A.: Automatic Conflict Analysis and Resolution of Traffic Filtering Policy for Firewall and Security Gateway. In: ICC'07 (2007)

3. Westerinen, et al.: Terminology for Policy-Based Management. RFC-3198
4. Al Shaer, E., Hamed, H.: Discovery of Policy Anomalies in Distributed Firewalls. In: IEEE INFOCOM'04 (March 2004)
5. Al Shaer, E., Hamed, H.: Modeling and Management of Firewall Policies. IEEE eTransactions on Network and Service Management 1-1 (April 2004)
6. Al Shaer, E., Hamed, H.: Firewall Policy Advisor for Anomaly Detection and Rule Editing. In: Proceedings of IEEE/IFIP Integrated Management Conference (IM2003) (March 2003)
7. Hari, H., Suri, S., Parulkar, G.: Detecting and Resolving Packet Filter Conflicts. In: Proceedings of IEEE INFOCOM 2000, IEEE Computer Society Press, Los Alamitos (2000)
8. Ioannidis, S., Keromytis, A., Bellovin, S., Smith, J.: Implementing a Distributed Firewall. In: Proceedings of 7th ACM Conference on Computer and Comminications Security (CCS'00) (November 2000)
9. Cheswick, W., Bellovin, S.: Firewalls and Internet Security. Addison-Wesley, Reading (1995)

A Novel Image Edge Detection Using Fractal Compression

Liangbin Zhang and Lifeng Xi

School of Computer Science and Information Technology, Zhejiang Wanli University
Ningbo, Zhejiang 315100, P.R. China
zlb@zwu.edu.cn

Abstract. Image edges are the foundation of image texture and shape figure extraction. In this paper we propose a novel edge detection method based on the self-similarity of fractal compression. We point out that the mean-square-error distance (MSE) of fractal compression can be used to extract edge of fractal image effectively. The self-similarity coefficient between the local range block and the searching domain block is centered at the current pixel being processed, and near-center self-affine transform is applied in local searching process, finally a binary operator is used to threshold its magnitude and produce the edge map of the image. The results of experiments show that the proposed new algorithm for image edge detection is valid and effective. It also has good anti-noise performance.

Keywords: image edge detection, fractal coding, MSE, self-similarity.

1 Introduction

The extraction of edge information is important in many fields such as image analysis and computer vision. One conventional approach to the problem of edge detection employs high spatial frequency enhancement threshold algorithms. These algorithms use spatial operators to enhance the original image, forming an edge enhancement strength map. A threshold is then applied to the edge strength map to determine the presence of edge pixels. But many objects in nature that are so complex and so erratic that it appears hopeless to use just the familiar objects from classical geometry for their modeling.

The simplest and classic edge detection algorithms process every pixel on original image, such as gray level gradient, Sobel, Laplacian, Prewit, Robert and Rosenfeld threshold operators[1], which all are designed using the gray level variation around edge area and make judgment about edge existence by first or second direct derivative. All these edge detection algorithms are simple to compute and run rapidly, but sensitive to image noise disturbance, which is lead to low effectiveness of edge detection.

Many efforts have made to improve the effectiveness of image edge detection. Marr has proposed the Gaussian-Laplacian edge detection filter, which firstly smooth original image using two-dimension Gaussian weighted function, then detect the zero

B. Apolloni et al. (Eds.): KES 2007/ WIRN 2007, Part II, LNAI 4693, pp. 159–166, 2007.
© Springer-Verlag Berlin Heidelberg 2007

cross point of second direct derivative as edge point[2]. Hueckel□Prewitt□Haralick et al utilize best-simulated curve plane to approximate the original processing image, and detect edges on the curve plane[3,4].

In recent years, the theory of fractals provides a compact description for such real-life objects or images and has been used effectively in characterizations and segmentation of natural visual imagery. Fractal image compression, which is based on the IFS(iterated function system) proposed by Barnsley[5], is a novel approach to image coding. Its performance relies on the presence of self-similarity between the regions of an image. Since most images process a high degree of self-similarity, fractal compression contributes an excellent tool for compressing. Using Jacquin's algorithms [6-7], we noticed that for a shade block or a midrange block, the decoded image has an overall very good fidelity to the original. Most of the artifacts visible in reconstructed image are likely to occur with either edges or strongly contrasted textures. From the experimental data we discovered that for a shade block or a midrange block, mean-square-error distance (MSE) is a very small value, but for a edge block, MSE is very large. So MSE can be used to extract edge of fractal image effectively. In this paper we point out a novel edge detection method using fractal compression. We choose MSE for extracting image edge detection, the self-similarity coefficient between the local range block and the searching domain block is centered at the current pixel being processed, and near-center self-affine transform is applied in local searching process, finally a binary operator is used to threshold its magnitude and produce the edge map of the image. Experiment result shows that our image edge detection algorithm is valid and effective especially for noisy image.

The balance of the paper is organized as follows: theoretical foundations of fractal image compression is stated in Section 2. Algorithm for edge detection using fractal compression is described in Section 3. Experimental results and analyses is reported in Section 4. Conclusion is included in Section 5.

2 Theoretical Foundations of Fractal Image Compression

2.1 The principle of IFS and the Collage Theorem

The theory of IFS and the Collage Theorem constitute the broad foundation of FIC [8].

IFS: let (F, p) be a complete metric space, where p is a given metric and F is the space of the digital images. Let $w: F \to F$ be a map. If there is a real number s (0<s<1), such that $p(w(x), w(y)) < sp(x, y), x, y \in F$, then w can be called the contractive transformation. An IFS consists of a complete metric space (F, p) and a contractive transformation set $W = \bigcup_{i=1}^{n} w_i = w_1 \cup w_2 \cup ..... \cup w_n$ defined on F.

The collage theorem: if W is a contractive mapping as defined above, and $f \in F$, then $d(A, f) \le (1-s)^{-1} d(w(f), f)$, where d is the Hausdorff Distance and A is the attractor of IFS.

2.2 Jacquin Coding Method

Fractal image coding makes good uses of image self-similarity in space by ablating image geometric redundant. Fractal coding process is quite complicated but decoding process is very simple, which makes use of potentials in high compression ratio. The main theory of fractal image coding is based on iterated function system, attractor theorem, and collage theorem. Regard original compressible image as attractor, how to get LIFS parameters is main problem of fractal coding. Fractal block coders are essentially based on the work done by Jacquin in [6-7], with several improvements and variations presented in [8-10]. Jacquin method is as following:

1. A given image I is divided into non-overlapping M range blocks of size $B \times B$ and into arbitrarily located N domain blocks of size $2B \times 2B$. The range blocks are numbered from 1 to M, and represented by $R_i (1 \le i \le M)$. Similarly, the domain blocks are from 1 to N, and represented by $D_j (1 \le j \le N)$.

2. For each range block R_i, the best matched domain $D_k (1 \le K \le N)$ and an appropriate contractive affine transformation τ_{ik} which satisfy the following equation are found through

$$d(R_i, \tau_{ik}(D_k)) = \min d(R_i, \tau_{ij}(D_j)) \tag{1}$$

$$d(R_i, \tau_{ij}(D_j)) = \frac{\sum_{0 \le N \le B} (x_{l,k} - \bar{x}_{l,k})^2}{B \times B}, (0 \le l, k \le B-1) \tag{2}$$

Where τ_{ij} is an contractive affine transformation from the domain block D_j to the range block R_i; the distortion measure $d(R_i, \tau_{ij}(D_j))$ is the Mean Square Error (MSE) between the range block R_i and the contractive domain block $\tau_{ij}(D_j)$.The contractive affine transformation τ_{ij} is composed of two mappings ϕ_j and θ_{ij} as follows:

$$\tau_{ij} = \theta_{ij} \circ \phi_j \tag{3}$$

The first mapping ϕ_j is the transformation of domain-block size to the same size as range blocks. This transformation can be described as follows: The domain block D_j is divided into non-overlapping unit blocks of size 2×2; and each pixel value of the transformed block $\phi_j(D_j)$ is an average value of four pixels in each unit block in D_j. The second mapping θ_{ij} consists of two steps: The first step transforms the block $\phi_j(D_j)$ by one of the following eight transformations: rotation around the center of the block $\phi_j(D_j)$, through 0^0,$+90^0$,$+180^0$, and $+270^0$, and each rotation after orthogonal reflection about mid-vertical axis of the block $\phi_j(D_j)$.Those eight transformations are called isometries. The second step is the transformation p_{ij} of pixel values of a block obtained by the first step. This transformation p_{ij} is defined as

$$p_{ij}(v) = s_{ij}v + h_{ij} \tag{4}$$

where v is a pixel value of the block obtained by the first step, and the parameters s_{ij} and h_{ij} are computed by the least square analysis of pixel values of the range block R_i and the block obtained by the first step. We call the parameters s_{ij} and h_{ij} a scaling coefficient and an offset, respectively. The LIFS parameters listed below are encoded: (1)Parameters to indicate a location of the best matched domain block; (2)A parameter to indicate an isometric on the best matched domain block; (3) A scaling coefficient and an offset. The proposed method quantizes these LIFS parameters [8].

3 Algorithm for Edge Detection Using Fractal Compression

3.1 MSE for Extracting Edge Detection

According to fractal compression theory, the edges of object or its components possess fractal self-similarity character, so we know that the pixel on image edge should have great gray level gradient and self-similarity coefficient. With the help of experiments, we discovered that for a smooth block or a shade block, the Mean Square Error as $d(R_i, \tau_{ij}(D_j))$ is zero or a very small value, but for block containing edges, MSE is very large. Hence MSE can enhance edges effectively.

A digital image is regarded as a gray level curve plane Z=f(x,y) in three-dimension space, where Z denotes the gray level value at on pixel point and (x,y) is the point position. For any pixel point on image, we can compute the self-similarity coefficient within the image local area centered on the processing pixel. Fig.1 shows the range block and the domain block centered on an image processing pixel point. If the position of an image pixel point is (x0, y0), and B×B denote the range block centered on the pixel point, where B is the value of 2, 3, 4 etc. The domain block size centered on the (x0, y0) image pixel point is D×D, usually D is the value of 2B or 3B. In Fig.1 the value of B is 3, and the value of D is 3B.

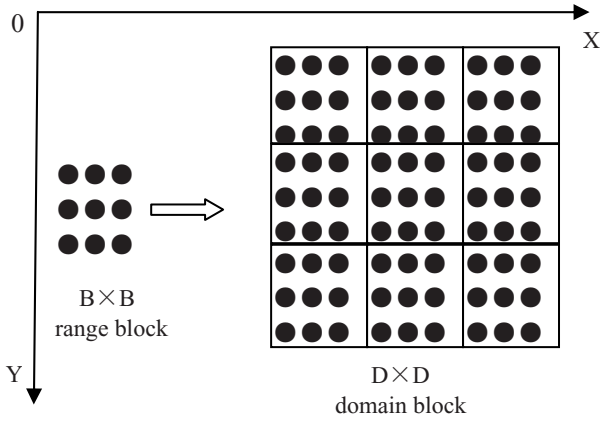


Fig. 1. Range and domain block centered on an image processing pixel point

Based on the self-similarity character of fractal compression, we know that certain similarity at some level should exit between the image range block B×B and the domain block D×D in the shapes of three dimension gray level curve planes. To evaluate the similarity between the two gray level curve planes, pixel numbers of the range block and the domain block should be equal. So, sampling or average compressing process should be necessary for the D×D image domain block. In sampling process, we can sample pixel gray level average value of every neighboring range block within D×D image domain block. For example, Fig.1 let (b_1,b_2,….b_6) represent the gray level value of centered range block pixels, and the sampled pixel gray level values of domain block pixels are represented as (d_1,d_2,……d_6) respectively. If there exists the self-similarity between the range block and domain block, then the mean value of square sum of all corresponding pixels' differences should have minimum value, that is the value of E in form(5) should be the minimum value.

$$E = \frac{1}{k}\sum_{i=1}^{k}(d_i - b_i)^2 \tag{5}$$

Where K is equal to $B \times B$. Evidently E is the Euclidian distance between two k-dimension samples in pattern recognition theory. The less the distance is, the similarity character between the two samples is much closer. In the pattern recognition and classification theory, in addition to the distance, there are similar coefficient S and coherent coefficient R to evaluate the similarity between two patterns. The similar coefficient S is defined as (6) and coherent coefficient R is defined as (7):

$$s = \frac{\sum_{l=1}^{k} d_{ni} b_{ni}}{(\sum_{i=1}^{k} b_{ni}^2 \sum_{i=1}^{k} d_{ni}^2)^{1/2}} \tag{6}$$

Where d_{ni} and b_{ni} are the normalized values of d_i and b_i . Here $b_{ni} = (b_i - b_{\min})/(b_{\max} - b_{\min})$, $d_{ni} = (d_i - d_{\min})/(d_{\max} - d_{\min})$ $(i = 1,2,...,k)$;where $b_{\max}$ and $b_{\min}$ are the maximum and minimum values of $b_i (i = 1,2,.....k)$;and $d_{\max}$ and $d_{\min}$ are the maximum and minimum values of $d_i (i = 1,2,.....k)$.

$$R = \frac{\sum_{i=1}^{k}(b_i - \bar{b})(d_i - \bar{d})}{(\sum_{i=1}^{k}(b_i - \bar{b})^2 . \sum_{i=1}^{k}(d_i - \bar{d})^2)^{1/2}} \tag{7}$$

Here, $\bar{b} = \frac{1}{k}\sum_{i=1}^{k} b_i$, $\bar{d} = \frac{1}{k}\sum_{i=1}^{k} d_i$. Apparently, the larger the value of S or R is , the less the value of E is. So from statistical aspect we can use the value of R or S as the index to evaluate the similarity level between the range block $B \times B$ and the local block $D \times D$ centered on any image pixel.

3.2 Our Edge Detection Algorithm

The image is composed of low frequency of smooth area and high frequency of edge area and random noise. The smooth area is the area of background or objects and the edge area is the area of object's or its components' edges. Self-similarity character should exist in edge area and the pixels on the image edges should have great gray level gradient and self-similarity coefficient[11-13]. Based on this idea, the steps of the image edge detection algorithm, which combined the self-similarity character of fractal image edge is as following:

(1) input original image f(i, j) of N×N.
(2) for every processing pixel g(i, j), determine the range block B×B and corresponding domain block D×D, e.g. B=3, D=3B.
(3) compute the relative standard deviation in the given image. Each pixel g(i, j) is taken as the center of square block R_i including 3×3 pixels, a suitable domain-to-range block transform is made in a similar way to the fractal coding scheme, and its edge detection value can be computed according to Eq.(2), and find the maximal $d(R_i, \tau_{ij}(D_j))$ as MSE.
(4) If $d(R_i, \tau_{ij}(D_j)) < T_1$, where T_1 is an threshold, then the processing pixel is not taken as edge pixel, and the output pixel value will be g(i, j)=0, otherwise go to next step (5).
(5) Compute the value of R or S between center range block and domain block.
(6) If the value of R or S is less than T_2, where T_2 is another threshold, then the processing pixel is not taken as edge pixel, let g(i, j)=0, otherwise edge pixel is detected and output pixel value g(i, j)=1.
(7) Return to the step (2) and process the next image pixel.

With the help of experiments, we discovered the fact that most of the range blocks R_i possess short-distance piecewise self-similarities. All that is needed to do is to search R_i's match D_i just among a few of the blocks whose center is near the center of R_i. This process is called as near-center self-affine transform method for convenience. So we improved the third step of the algorithm as following: Because of short-distance piecewise self-similarity, a suitable domain-to-range block transformation is made in near-center self-affine transform method and searching process is in local domain block rather than in all domain blocks.

4 Experimental Result and Analyses

All the experiments are carried out on a computer with Intel 2.5Ghz and 512M RAM in the Win2000 professional operating system and Matlab7.0 language is used. Original image is classical 128×128 grey-level Lena face coded with 8 bits per pixel.

Utilizing the proposed algorithm, the experiments of edge detection on image "lena" is implemented. Fig. 2 shows the result of edge detection experiments.



(a) Lena image with additive Gaussian white noise



(b) edge image by Roberts operator



(c) edge image by our algorithm

Fig. 2. The result of edge detection experiments

The result shows the edge detection method is effective. Since the new algorithm for edge detection avoids the gradient operation, it is insensitive to noise. Fig. 2 (a) is lena image which contain additive Gaussian white noise. Fig. 2 (b) is the edge image by using Roberts operator and Fig. 2(c) is the edge image by using our algorithm. Comparing Fig. 2 (b) with Fig. 2 (c), we can find that our scheme has better noise immunity than Roberts operator and our method is very suitable for the object detection in natural scenes based on edges.

In addition, we should point out that the edge detection by our method is related with range block size. If we choose a larger rang block than 3×3, the edges detection by our method is thicker. By using the small size of range block, we can get correct edges.

5 Conclusion

In this paper we have proposed a new algorithm for edge detection using fractal compression, and the MSE is used successfully in edge detection which has better noise immunity than Roberts operator. What's more, applying near-center self-affine transform method in local searching process improved algorithm. Thus our algorithm provides an effective way to extract good edges from noisy images.

Acknowledgments. This research is supported by National Natural Science Foundation of China (Grant No. 10671180), and Research Project of Education Department of Zhejiang Province of China (No. 20060183).

References

1. Xu, J.H.: Image Process And Analysis, pp. 45–100. Academic Press, Beijing (1994)
2. Marr, D., Hildreth, E.: Theory of Edge Detection. Proc. R. Soc. Lond. B207, 187–217 (1980)
3. Hueckel, M.H.: A local Visual Operator Which Recognizes Edges And Lines. Journal of the Association for Computing Machinery 20, 634–647 (1973)

4. Haralick, R.M.: Digital Step Edges from Zero Crossing of Second Directional Derivative. IEEE Trans. PAMI-6, 58–68 (1984)
5. Barnsley, M.F.: Iterated Function Systems and the Global Construction of Fractals. Proc. Roy. Soc. London, 243–275 (1985)
6. Jacquin, A.E.: A Novel Fractal Block-coding Technique for Digital Image. In: Proceedings of ICASSP IEEE International Conference on ASSP, vol. 4, pp. 34–42 (1990)
7. Jacquin, A.E.: Fractal Image Coding a Review. Proceeding of the IEEE 81, 212–223 (1993)
8. Shouji, C., Liming, Z.: Fractal And Image Compression. Shanghai Science And Technology Education Publishing House (1998)
9. Fisher, Y.: Fractal Image Compression. Fractals 2, 25–36 (1994)
10. Fisher, Y.: Fractal Image Compression: Theory and Application. Springer, Heidelberg (1995)
11. Motallebi, P., Ogunbona, P.O.: Edge Image Description Using Fractal Interpolation. IEEE Tench-Speech and Image Technologies for Computing and Telecommunications, 553–556 (1997)
12. Ida, T.: Image Segmentation and Contour Detection Using Fractal Coding. IEEE Transactions on Circuits And Systems For Video Technology 8, 968–975 (1998)
13. Qiong, L., Jun, G., Long, G., Huoming, D.: A Novel Edge Detection Method Based On Fractal Theory. In: IEEE int. Conf. Neural Network & Signal Processing, pp. 1105–1108 (2003)

A Rapid Algorithm for Computing ST Reliability of Radio-Communication Networks

Fei Gao, Xuejun Liu, and Hanzhong Liu

School of Computer Science and information technology, Zhejiang Wanli University
Ningbo, Zhejiang, 315100, P.R.China

Abstract. A new formula for computing ST reliability of radio-communication networks (RCN) from Source to Terminal t was presented. The new formula contains terms which correspond one by one to a class special ST path (SPRT). For a given networks, the terms of the new formula are fewer than those corresponding Inclusion-Exclusion principle. An algorithm for computing ST reliability was presented, it computed ST reliability or produced a ST reliability expression by enumerating a class of special networks of given networks. Because the structure of this class of new networks which need to be enumerated was relativity small, the new algorithm's performance was better than Inclusion-Exclusion principle algorithm .Finally an example illustrates our conclusion.

Keywords: Reliability, Reliability expression, Inclusion Exclusion principle, Radio-communication network.

1 Introduction

Along with Radio-communication networks (RCN)apply from extensive to extensive, Radio-communication networks (RCN) have become importance. it not only inexpensive, but also suited to some especial occasion ,for example: Mobile networks, Adhoc networks and Data collection system etc. The people attach importance to The reliability of RCN, the reliability of networks have been lucubrated, but the research on the reliability of Radio-communication networks (RCN) is less. AboEIFotoh and Colbourn refer to a probability graph model, prove the problem computing the reliability of Radio-communication networks (RCN) is a NP hard problem, give an approximate algorithm. Some special RCN have been studied[2,3] ,Fanjia-Kong[4,5] give a Factoring algorithm of K-terminal reliability of RCN.

We assume the transmitting medium of RCN is perfect, and doesn't have physics obstacle, the sensitivity of every work station is same, the send power and style of antenna is same. so the model of RCN is a probability graph G(V,E), every workstation is a node, every node have 2 states (working and failure), the probability of node working have been known and independence each other , if have a edge node v_1 connect node v_2 then node v_1 and node v_2 can communicate each other, every edge

B. Apolloni et al. (Eds.): KES 2007/ WIRN 2007, Part II, LNAI 4693, pp. 167–174, 2007.
© Springer-Verlag Berlin Heidelberg 2007

always working. There is 2 special nodes s&t in the nodes set of G(V,E), the reliability of G is the probability source s and terminal t can communicate each other.

2 Definition and Notation

2.1 Definition

Pa_{ij} : A path from i to j.

SPa_{ij} : A path from i to j, and the set of nodes on the path hasn't subset that can build up Pa_{ij}, we name the simplest path Pa_{ij}.

$SPRT$: The root node of the tree is s, leaf nodes are t, middle nodes are the others nodes of graph G, every path form s to t in the tree corresponding one by one the SPa_{ij} of G.

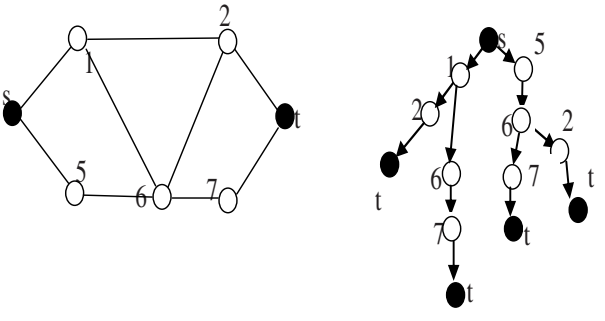


Fig. 1. A RCN networks and its SPRT tree

2.2 Notion

$G(V,E)$: RCN, V is the set of the nodes of G, E is the set of the edges of G.

s,t : The source node of G and the terminal node if G.

m: The number of Spa(s,t) of G.

S_i : The set of the sons of node i.

K_P : The set of the nodes in path p.

w_i : A node in the $SPRT$.

$v_i(\overline{v_i})$: The state of node is working(not working).

$p_i(\overline{p_i})$: The working (failure) state probability of node i.

$R(G)$: The ST reliability of RCN.

3 The New Topological Formula and Rapid Algorithm for Reliability of RCN

We consider a Radio-communication networks (RCN) $G(V,E)$, $V=(1,2,3,...n)$, $E=x_1,x_2,x_3,...x_e$, $n=|V|$ $e=|E|$. p_i is the probability of the node i working,=1- p_i is the probability of the node i failure.

The definition of the reliability of RCN networks [1]:

$R(G)$ the probability that have a Pa(s,t) in G that every nodes in the $Pa(s,t)$ is working

$$R(G) \quad \Pr(\sum_{i=1}^{n} P_i) \tag{1}$$

According to the Boolean algebra axiom $P(x+xy)=P(x)$, if $Pa(s,t)$ isn't the simple path $SPa(s,t)$, then K_{pa} (the nodes set of $Pa(s,t)$) have a subset K_{spa} that can make up a $SPa(s,t)$, $K_{spa} \subset K_{pa}, P(Pa+Spa)=P(SPa)$ hence:

Theorem 1. $R(G)=P(\sum_{i=1}^{m} SPa_i), i=1,2,3,...m$, m is the number of $SPa(s,t)$ of G.

Theorem 2. If Spa is a st simple path, then it is one by only confirmed by its nodes set K_{Spa}.

Prove: The length of Spa is $|Spa|=|K_{Spa}|-1$, suppose $Spa=sw_1w_2...w_{k-1}t$, because there isn't parallel edge in RCN networks, if Spa isn't exclusive, suppose the other $Spa_2=sw_{i_1}w_{i_2}...w_{i_{k-1}}t$, there is at least one the node that its locality in Spa_2 move along relative to in Spa, suppose it is w_i,the father of w_i is w_m $m<i-1$, w_i is neighbor to w_m in graph G, then $sw_1w_2...w_mw_iw_{i+1}...w_{k-1}t$ is a path $Spa(s,t)$. Its length is less than k-1, it's conflicting to Spa is a simple path, so Spa is a st simple path, it is one by only confirmed by its nodes set K_{Spa}.

Theorem 3. A tree T include all SPa path in G, it's a $SPRT$ tree if and only if the node in a discretionary path isn't its ancestor or the brother of its ancestor.

Prove: If a node is its ancestor then there is a circle in the tree, it is conflicting to the definition of tree. If a node is a brother of its ancestor, see Fig. 2: s→v→w→t is a

$Pa(s,t)$ path, its length is less than the length of s→v→a(w)→w→t, it is conflicting to the definition of $SPa(s,t)$ path. If the tree have a path that it isn't $SPa(s,t)$ path, according to the definition of $SPa(s,t)$ path, there is node in the path that it is neighbor to its ancestor in graph G, then it is a brother of its ancestor in the tree, it is conflicting to the condition of theorem. Hence the theorem is correct.

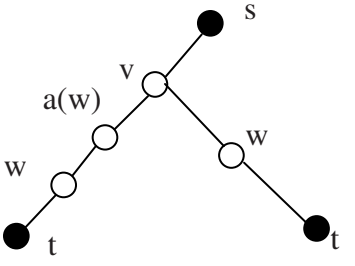


Fig. 2. A SPRT tree

Theorem 4. The reliability of G , $R(G) = \sum_{i=1}^{m} P(SP_i)(1-R(G_i)$,The G_i in formula can made up like this: the source node s of G is the source node of G_i , the terminal node t of G is the terminal node of G_i , every middle node in $SPa_k(k<i)$ of the $SPRT$ tree of G ,

1. If $s \to w_j$ path in the SPa_i path, then the path amalgamation to the source *s* of G_i .
2. If w_j is the middle node of SPa_k path and it's a node in the SPa_i path too, then delete the node and link the father and the son of the node w_j .
3. If $w_j \to t$ path in the SPa_i path, then the path amalgamation to the terminal *t* of G_i .
4. If the nodes set of a path is a subset of another path node set, then delete the path. There isn't node in graph G_i that it's a node in the SPa_i path except *s&t* node.
5. Graph G_i is make up of these path.

Prove: We suppose $SPRT$ have $SP_1, SP_2, ...SP_m$, $A_1, A_2, ...A_m$ express the event working corresponding the SPa_i path.

$$
\begin{aligned}
&R(G) = P(A_1 + A_2 + \ldots + A_m) \\
&= P(A_1 + \overline{A_1}A_2 + \ldots + \overline{A_1}\,\overline{A_2}\ldots\overline{A_{i-1}}A_i + \ldots + \overline{A_1}\,\overline{A_2}\ldots\overline{A_{m-1}}A_m) \\
&= P(A_1) + P(\overline{A_1}A_2) + \ldots + P(\overline{A_1}\,\overline{A_2}\ldots\overline{A_{i-1}}A_i) + \ldots + P(\overline{A_1}\,\overline{A_2}\ldots\overline{A_{m-1}}A_m) \\
&= \mathrm{Pr}(A_1) + P(\overline{A_1}A_2) + \ldots + P((\overline{A_1}A_i)(\overline{A_2}A_i)\ldots(\overline{A_{i-1}}A_i)) \\
&+ \ldots + P((\overline{A_1}A_m)(\overline{A_2}A_m)\ldots(\overline{A_{m-1}}A_m))
\end{aligned} \tag{2}
$$

We suppose $SPa_i = Q_1Q_2Q_3Q_4Q_5, SPa_j = Q_1Q_6Q_3Q_7Q_5$ (see Fig. 3)

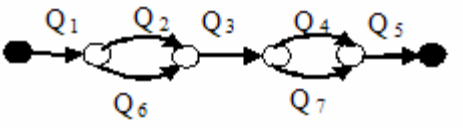


Fig. 3. Two SP path in the SPRT tree

$$
\begin{aligned}
\overline{A_i}A_j &= \overline{Q_1Q_2Q_3Q_4Q_5} \cap (Q_1Q_6Q_3Q_7Q_5)) \\
&= (\overline{Q_1} + \overline{Q_2} + \overline{Q_3} + \overline{Q_4} + \overline{Q_5}) \cap (Q_1Q_6Q_3Q_7Q_5) \\
&= \overline{Q_2}(Q_1Q_6Q_3Q_7Q_5) + \overline{Q_4}(Q_1Q_6Q_3Q_7Q_5) \\
&= \overline{Q_2Q_3}(Q_1Q_6Q_3Q_7Q_5) \\
&= \overline{Q_2Q_3}A_j
\end{aligned} \tag{3}
$$

This Q_2Q_4 is a $SPa(s,t)$ path of G_i, then

$$
\begin{aligned}
R(G) &= P(A_1) + P(\overline{G_1}A_2) + \ldots + P(\overline{G_i}A_i) + \ldots + P(\overline{G_m}A_m) \\
&= P(A_1) + (1 - P(G_1))P(A_2) + \ldots \\
&+ (1 - P(G_i))P(A_i) + \ldots + (1 - P(G_m))P(A_m)
\end{aligned} \tag{4}
$$

Algorithm:

s: initialization state node(source node);

OPEN table: the table storing node that not extended;

CLOSED table: the table storing node that extended;

1. Input probability graph G and the source node s, the terminal node t, $R(\Phi) = 0, i = 1$;
2. Construct a search graph only including *s*, put *s* into the OPEN table;
3. Construct a CLOSED table and intercalate it as empty table;
4. If the OPEN table is a empty table, then output $R(G)$, exit;
5. Choice the first node in the OPEN table, move it out from the OPEN table and put into the CLOSED table, mark the node as v;

6. If the node v is the terminal node t then find a simple path($SPa(s,t)$),If every $SPa(s,t)$ in G_1 haven't same node compare the other $SPa(s,t)$ in G_1, then

$$R(G) = R(G) + (1 - \prod_{all\ \ SPa\ \ in\ \ G_1} P(SPa(G_1)_k))P(SPa_i),$$

otherwise, recursive compute $R(G_1)$ on the theorem 4,

$$R(G) = R(G) + (1 - R(G_1))P(SPa_i)$$;i=i+1,goto 4;

7. Extend the node v, make a nodes set that the nodes aren't its predecessor and a brother of its predecessor, mark the set as S(v),if S(v)=φ,then delete the node v from the CLOSED table, go to(4),otherwise insert the nodes in S(v) into the search graph as subsequence node v;
8. The nodes in S(v) set a pointer point to its father node v, and insert it into the OPEN table;
9. Arrange the nodes in the OPEN table according to DFS policy, go to 4;

We use the Fig. 4 explain the algorithm.

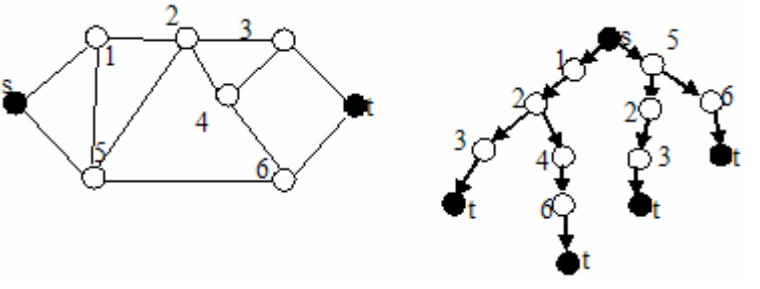


Fig. 4. a RCN networks and its SPRT tree

Table 1. Spa path of RCN networks and Gi reduced by Spa path

I	SPa_i	A_i	G_i
1	S123t	123	Φ
2	s1246t	1246	s 3 t
3	s523t	523	s 1 t
4	s56t	56	s 2 3 t; 1 2 4

This is a recursion algorithm, for the first time cycle, search a simple path (Sp_1), $G_1=\Phi$, $R(G_1)=0$. For the second time cycle, $G_3=v_3$

$$R(G)=P(v_1v_2v_3)+(1-R(G_2))P(v_1v_2v_4v_6)=p_1p_2p_3+\overline{p_3}p_1p_2p_4p_6$$

Keep up search like this, create SP_3, SP_4, gained G_3, G_4

$$\begin{aligned} G_3 &= x_1 \\ R(G) &= R(G)+(1-R(G_3))P(v_5v_2v_3) \\ &= R(G)+\overline{p_3}p_2p_3p_5 \end{aligned} \tag{5}$$

G_4 have 2 simple path and they have a same node , we recursively compute

$$\begin{aligned} G_4 &= v_2v_3+v_1v_2v_4 \\ R(G) &= R(G)+(1-R(G_4))P(v_5v_6) \\ &= R(G)+\overline{p_2(p_3+\overline{p_3}p_1)}p_5p_6) \end{aligned} \tag{6}$$

The OPEN table is an empty table here, then the expression is:

$$\begin{aligned} R(G) &= P(v_1v_2v_3+\overline{v_3}v_1v_2v_4v_6+\overline{v_1}v_2v_5v_6+\overline{v_2(v_3+\overline{v_3}v_1)}v_5v_6) \\ &= p_1p_2p_3+\overline{p_3}p_1p_2p_4p_6+\overline{p_1}p_2p_5p_6+\overline{p_2(p_3+\overline{p_3}p_1)}p_5p_6 \end{aligned} \tag{7}$$

The number of the simple paths (Spa) is less than the number of the *st* paths, G_i is a simple graph, $R(G_i)$ can easy computed. The workload at first hand apply Inclusion Exclusion principle or enumerate *st* path is very big, so the expression of $R(G)$ is very long. Compare our algorithm the workload is less, and the expression is shorter. This shows the advantage of algorithm.

References

1. AboEIFotoh, H.M., Colbourn, C.J.: Computing 2-terminal Reliability for Radio-broadcast Networks. IEEE Trans. on Reliability. 38(5), 538–555 (1989)
2. AboEIFotoh, H.M., Colbourn, C.J.: Efficient Algorithm for Computing the Reliability of Permutation and Interval Graph. Networks 20(6), 883–899 (1990)
3. AboEIFotoh, H.M.: Algorithm for Computing Message Delay for Wireless Networks. Networks 29(1), 117–124 (1997)
4. Fanjia, K., Baiyou, Q., Guangxing, W.: An Algorithm for Reliability of Radio Broadcast Networks. Journal of Northeastern University (Natural Science) 19(5), 492–494 (1998)

5. Zhao, L.C., Kong, F.J.: A New Formula and an Algorithm for Reliability Analysis of Networks. Microelectron and Reliability 37(3), 511–518 (1997)
6. Satyaranayanna, A., Prabhakar, A.: New Topological Formula and Rapid Algorithm for Reliability Analysis of Complex Networks. IEEE Trans. on Reliability 27(1), 82–100 (1978)
7. Satyaranayanna, H.J.N.: A New Algorithm for Reliability Analysis of Multiterminal Network. IEEE Trans. on Reliability 30, 25–33 (1981)
8. Satyaranayanna, A.: Unified Formula for Analysis for Some Networks Reliability Problems. Trans. on Reliability 31(1), 23–32 (1982)

A Second-Order Upwind Difference Scheme for a Singularly Perturbed Problem with Integral Boundary Condition in Netural Network

Zhongdi Cen[1] and Xin Cai[2]

[1] Institute of Mathematics, Zhejiang Wanli University, Ningbo 315100, P.R. China
[2] School of Sciences, Jimei University, Xiamen 361021, China

Abstract. In this paper we consider a first order singularly perturbed quasilinear boundary value problem with integral boundary condition which arises in netural network. The problem is discretized using a hybrid upwind difference scheme on a Shishkin mesh. Applying the discrete maximum principle and barrier function techniques we show that the scheme is almost second order convergent, in the discrete maximum norm, independently of singular perturbation parameter. Numerical experiments support these theoretical results.

Keywords: Singular perturbation; upwind difference scheme; Shishkin mesh; uniform convergence.

1 Introduction

This article is prompted by a recent publication in [1] where the authors consider the following singular perturbation boundary value problem with integral boundary condition which arises in netural network

$$\varepsilon u'(t) + f(t,u) = 0, \quad t \in I = (0,1], \tag{1}$$

$$u(0) = \mu u(1) + \int_0^1 b(s)u(s)ds + d, \tag{2}$$

where $0 < \varepsilon << 1$ is the perturbation parameter, μ and d are given constants. $b(t)$ and $f(t,u)$ are assumed to be sufficiently continuously differentiable functions in $\bar{I} = I \cup \{t = 0\}$ and $\bar{I} \times R$ respectively and moreover

$$\frac{\partial f}{\partial u} \geq \alpha > 0 \quad \text{and} \quad |\mu| + \|b\|_\infty < 1. \tag{3}$$

Under these assumptions, problem (1)-(2) has a uniqueness solution $u(x)$, in general, displays a boundary layer of width $O(\varepsilon)$ at $t = 0$.

Differential equations with integral boundary conditions have been considered for many years. For a discussion of existence and uniqueness results and for applications of problems with integral boundary conditions see, [3,4,8] and references therein.

B. Apolloni et al. (Eds.): KES 2007/ WIRN 2007, Part II, LNAI 4693, pp. 175–181, 2007.
© Springer-Verlag Berlin Heidelberg 2007

The numerical analysis of singular perturbation cases has always been far from trivial because of the boundary layer behavior of the solution. Such problems undergo rapid changes within very thin layers near the boundary or inside the problem domain [6,7]. Recently, Amiraliyev et al [1] gave a finite difference scheme on a Shishkin mesh for problem (1)-(2). They proved that the scheme is only almost first-order convergent in the discrete maximum norm.

In this paper we present a hybrid upwind difference scheme on a Shishkin mesh. We shall show that the scheme is almost second order convergent, in the discrete maximum norm, independently of singular perturbation parameter.

An outline of the paper is as follows. In the next section we present some analytical results of the boundary value problem (1)-(2). The numerical method is described in section 3. In section 4 we analyze the convergence properties of the scheme. Finally numerical example is presented in section 5.

Notation. Throughout the paper, C will denote a generic positive constant (possibly subscripted) that is independent of ε and of the mesh. Note that C is not necessarily the same at each occurrence. To simplify the notation we set $g_i = g(x_i)$ and $g_{i-1/2} = g((x_{i-1} + x_i)/2)$ for any function $g(x)$, while g_i^N denotes an approximation of $g(x)$ at x_i.

Assumption 1. Throughout the paper we shall assume that $\varepsilon \le CN^{-1}$ as is generally the case in practice.

2 Properties of the Exact Solution

For constructing layer-adapted meshes correctly, we need to know the asymptotic behavior of the exact solution. This behavior will be used later in the analysis of the uniform convergence of the finite difference approximations defined in section 3.

Lemma 1. Let $u(t)$ be the solution of problem (1)-(2). Then we have

$$|u^{(k)}(t)| \le C(1+\varepsilon^{-k}e^{-\alpha t/\varepsilon}) \quad \text{for } t \in \bar{I},\ k = 0,1,2,3. \tag{3}$$

Proof. See [1] for a proof with $k = 0,1$; the argument there works also for $k = 2,3$.

3 Discretization and Mesh

For our discretization we use a Shishkin mesh. Let N, be an even positive integer. Let τ denote a mesh transition parameter defined by

$$\tau = \frac{2\varepsilon}{\alpha}\ln N. \tag{4}$$

The piecewise uniform mesh divides each of the interval $[0,\tau]$ and $[\tau,1]$ into $N/2$ equidistant subintervals. The mesh width of subinterval in $[0,\tau]$ is h. We use the notation H for the mesh width in $[\tau,1]$. These mesh widths satisfy

$$h=\frac{4\varepsilon}{\alpha}N^{-1}\ln N,\quad N^{-1}\le H\le 2N^{-1}. \tag{5}$$

We propose the following hybrid upwind difference scheme for approximating problem (1)-(2):

$$T^N u_i^N=0,\qquad \text{for } i=1,2,\cdots,N, \tag{6}$$

$$u_0^N=\mu u_N^N+\sum_{i=1}^{N}\frac{b_{i-1}u_{i-1}^N+b_i u_i^N}{2}h_i+d, \tag{7}$$

where

$$T^N u_i^N=\begin{cases}\varepsilon\dfrac{u_i^N-u_{i-1}^N}{h_i}+f(t_{i-1/2},\dfrac{u_{i-1}^N+u_i^N}{2}) & i=1,2,\cdots,N/2,\\ \varepsilon\dfrac{u_i^N-u_{i-1}^N}{2}+f(t_i,u_i^N) & i=N/2+1,\cdots,N.\end{cases}$$

4 Analysis of the Method

To investigate the convergence of the method, note that the error function $z_i^N=u_i^N-u_i\ (0\le i\le N)$ is the solution of the discrete problem

$$\varepsilon\frac{z_i^N-z_{i-1}^N}{h_i}+f(t_{i-1/2},\frac{u_{i-1}^N+u_i^N}{2})-f(t_{i-1/2},u_{i-1/2})=\varepsilon(u'_{i-1/2}-\frac{u_i^N-u_{i-1}^N}{h_i})\quad 1\le i\le\frac{N}{2},$$

$$\varepsilon\frac{z_i^N-z_{i-1}^N}{h_i}+f(t_i,u_i^N)-f(t_i,u_i)=\varepsilon(u'_i-\frac{u_i^N-u_{i-1}^N}{h_i}),\quad \frac{N}{2}+1\le i\le N-1,$$

$$z_0^N-\mu z_N^N-\sum_{i=1}^{N}\frac{b_{i-1}z_{i-1}^N+b_i z_i^N}{2}h_i=\sum_{i=1}^{N}\frac{b_{i-1}u_{i-1}+b_i u_i}{2}h_i-\int_0^1 b(s)u(s)ds.$$

For $i=1,2,\cdots N/2$ we use Taylor expansion for f about $t_{i-1/2}$ to obtain

$$L^N z_i^N\equiv\varepsilon\frac{z_i^N-z_{i-1}^N}{h_i}+a_i\frac{z_{i-1}^N+z_i^N}{2}=R_i, \tag{8}$$

where

$$a_i=\frac{\partial}{\partial u}f(t_{i-1/2},u_{i-1/2}+\rho\xi_i),\ 0<\rho<1,\ \xi_i=\frac{u_{i-1}^N+u_i^N}{2}-u_{i-1/2},$$

$$R_i=\varepsilon(u'_{i-1/2}-\frac{u_i-u_{i-1}}{h_i})+\frac{\partial}{\partial u}f(t_{i-1/2},u_{i-1/2}+\rho\xi_i)(u_{i-1/2}-\frac{u_{i-1}+u_i}{2}) \tag{9}$$

For $i=N/2+1,\cdots,N$ we use Taylor expansion for f about t_i to obtain

$$L^N z_i^N \equiv \varepsilon \frac{z_i^N - z_{i-1}^N}{h_i} + a_i z_i^N = R_i, \tag{10}$$

$$a_i = \frac{\partial}{\partial u} f(t_i, u_i + \sigma\eta_i),\ 0 < \sigma < 1,\ \eta_i = u_i^N - u_i, R_i = \varepsilon(u_i^{'} - \frac{u_i - u_{i-1}}{h_i}). \tag{11}$$

For $i = 0$ we have

$$L_0^N z_0^N \equiv z_0^N - \mu z_N^N - \sum_{i=1}^{N} \frac{b_{i-1} z_{i-1}^N + b_i z_i^N}{2} h_i = R_0, \tag{12}$$

where

$$R_0 = \sum_{i=1}^{N} \frac{b_{i-1} u_{i-1} + b_i u_i}{2} h_i - \int_0^1 b(s)u(s)ds. \tag{13}$$

It is easy to verify that the matrix associated with L^N is an M-matrix. Hence we obtain the following discrete maximum principle.

Lemma 2. The operator L^N defined by (8),(10),(12) satisfies a discrete maximum principle, i.e. if v_i and w_i are mesh functions that satisfy $L^N v_i \le L^N w_i$ for $i = 0,1,\cdots,N$, then $v_i \le w_i$ for all i.

Next lemma gives the truncation error estimate.

Lemma 3. Under the assumptions (3) the truncation error R_i satisfies

$$|R_i| \le \begin{cases} C\{\int_{t_{i-1}}^{t_i} (1 + \varepsilon^{-1} e^{-\alpha s/(2\varepsilon)}) ds\}^2 & i = 0,1,\cdots,N/2, \\ C\int_{t_{i-1}}^{t_i} (\varepsilon + \varepsilon^{-1} e^{-\alpha s/\varepsilon}) ds & i = N/2+1,\cdots,N. \end{cases}$$

Proof. For $i = 0$ we have

$$\begin{aligned} |R_0| &= \sum_{i=1}^{N} \left| \frac{b_{i-1}u_{i-1} + b_i u_i}{2} h_i - \int_{t_{i-1}}^{t_i} b(s)u(s)ds \right| \\ &\le \sum_{i=1}^{N} \int_{t_{i-1}}^{t_i} \left| \frac{b_{i-1}u_{i-1} + b_i u_i}{2} - b(s)u(s) \right| ds \le \frac{3}{2} \int_{t_{i-1}}^{t_i} |(b(s)u(s))^{"}| (s - t_{i-1}) ds \\ &\le C \int_{t_{i-1}}^{t_i} (1 + \varepsilon^{-2} e^{-\alpha s/\varepsilon})(s - t_{i-1}) ds, \end{aligned} \tag{14}$$

where we have used Lemma 1.

For $i = 1,2,\cdots,N/2$ we use Taylor expansion for u and f about t_i to obtain

$$\begin{aligned} |R_i| &\le \varepsilon \left| u_{i-1/2}^{'} - \frac{u_i - u_{i-1}}{h_i} \right| + \left| \frac{\partial}{\partial u} f(t_{i-1/2}, u_{i-1/2} + \rho\xi_i) \right| \cdot \left| u_{i-1/2} - \frac{u_{i-1} + u_i}{2} \right| \\ &\le \frac{3\varepsilon}{2} \int_{t_{i-1}}^{t_i} |u^{'''}(s)| (s - t_{i-1}) ds + \frac{3}{2} \left| \frac{\partial}{\partial u} f(t_{i-1/2}, u_{i-1/2} + \rho\xi_i) \right| \end{aligned}$$

$$\cdot \int_{t_{i-1}}^{t_i} |u''(s)|(s-t_{i-1})ds \le C\int_{t_{i-1}}^{t_i}(1+\varepsilon^{-2}e^{-\alpha s/\varepsilon})(s-t_{i-1})ds, \tag{15}$$

where we have used (3) and Lemma 1.

For $i = N/2+1,\cdots,N$ we also use Taylor expansion for u and f about t_i to obtain

$$|R_i| \le \varepsilon |u_i' - \frac{u_i - u_{i-1}}{h_i}| \le \varepsilon \int_{t_{i-1}}^{t_i} |u''(s)|ds \le C\int_{t_{i-1}}^{t_i}(\varepsilon + \varepsilon^{-1}e^{-\alpha s/\varepsilon})ds. \tag{16}$$

To bound the right-hand side of (14)-(15) we use the inequality

$$\int_{t_{i-1}}^{t_i} g(s)(s-t_{i-1})^{(k-1)}ds \le \frac{1}{k}\{\int_{t_{i-1}}^{t_i} g(s)^{1/k}ds\}^k,$$

which holds true for any positive monotonically decreasing function g on $[t_{i-1}, t_i]$ and for arbitrary $k \in \mathrm{N}^+$; see [2]. From this we complete the proof of the lemma.

For the Shishkin-type mesh we have the following result.

Lemma 4. On the Shishkin mesh we have

$$\int_{t_{i-1}}^{t_i}(1+\varepsilon^{-1}e^{-\alpha s/\varepsilon})dt \le CN^{-1}\ln N \quad \text{for } i = 0,1,\cdots,N/2,$$

$$\int_{t_{i-1}}^{t_i}(\varepsilon+\varepsilon^{-1}e^{-\alpha s/\varepsilon})dt \le CN^{-2} \quad \text{for } i = N/2+1,\cdots,N.$$

Proof. For $i = 0,1,\cdots,N/2$ we have

$$\begin{aligned}\int_{t_{i-1}}^{t_i}(1+\varepsilon^{-1}e^{-\alpha s/\varepsilon})ds &\le h_i - \frac{1}{\alpha}e^{-\alpha s/\varepsilon}\Big|_{t_{i-1}}^{t_i} = h_i - \frac{1}{\alpha}(e^{-\alpha t_i/\varepsilon} - e^{-\alpha t_{i-1}/\varepsilon}) \\ &= h_i - \frac{1}{\alpha}e^{-\alpha t_{i-1}/\varepsilon}(e^{-\alpha h/\varepsilon} - 1) \le CN^{-1}\ln N,\end{aligned} \tag{17}$$

For $i = N/2+1,\cdots,N$ we have

$$\int_{t_{i-1}}^{t_i}(\varepsilon+\varepsilon^{-1}e^{-\alpha s/\varepsilon})ds \le \varepsilon h_i - \frac{1}{\alpha}e^{-\alpha s/\varepsilon}\Big|_{t_{i-1}}^{t_i} \le C\varepsilon N^{-1} + \frac{2}{\alpha}e^{-\alpha t_{i-1}/\varepsilon} \le CN^{-2} \tag{18}$$

by the choice of the transition point τ and Assumption 1.

Combining (17) and (18), we are done.

Now we can get our main result for the upwind difference scheme.

Theorem 1. Let u be the solution of problem (1),(2) and u^N be the solution of finite difference scheme (6)-(7) on the Shishkin mesh. Then we have the following error estimate

$$|u_i - u_i^N| \le CN^{-2}\ln^2 N \quad \text{for } i = 0,1,\cdots,N.$$

Proof. Combining Lemmas 3 and 4 yields the following result

$$|L^N z_i^N| = |R_i| \le CN^{-2}\ln^2 N \quad \text{for } i = 0,1,\cdots,N.$$

Applying the discrete maximum principle, with the barrier function

$$S_i = CN^{-2}\ln^2 N \quad \text{for } i = 0,1,\cdots,N,$$

we can obtain

$$|z_i^N| \le CN^{-2}\ln^2 N \text{ for } i = 0,1,\cdots,N.$$

From this we complete the proof.

5 Numerical Results

In this section we verify experimentally the theoretically results obtained in the preceding section. We study the test problem

$$\varepsilon u' + 2u - e^{-u} + t^2 = 0, \quad 0 < t \le 1, \tag{19}$$

$$u(0) = \frac{1}{2}u(1) - \frac{1}{4}\int_0^1 e^{-s}u(s)ds + 1. \tag{20}$$

For our tests we take 10^{-8} which is a sufficiently small choice to bring out the singularly perturbed nature of the problem. The exact solution of our test problem is not available. Therefore, we use the double mesh principle to estimate the errors and compute the experiment rates of convergence in our computed solution. We measure the accuracy in the discrete maximum norm

$$e_u^N = \max_i |u_i^N - u_{2i}^{2N}|,$$

the convergence rate

$$r^N = \log_2 \frac{e_u^N}{e_u^{2N}}.$$

Table 1. Numerical results for problem (19), (20)

N	Error	Rate
16	1.2086e-2	1.445
32	4.4377e-3	1.510
64	1.5578e-3	1.572
128	5.2394e-4	1.617
256	1.7082e-4	1.735
512	5.1327e-5	-

The numerical results (Table 1) are clear illustrations of the convergence estimate of Theorem 1. They indicate that the theoretical results are fairly sharp.

Acknowledgments. The work was supported by National Natural Science Foundation (Grant No. 10671180) of China and Research Project (20060182) of Department of Education of Zhejiang Province of China and the Natural Science Foundation of Fujian Province of China for contract A0610025.

References

1. Amiraliyev, G.M., Amiraliyeva, I.G., Kudu, M.: Anumerical treatment for singularly perturbed differential equations with integral boundary condition. Appl. Math. Comput. 185, 574–582 (2007)
2. de Boor, C.: Good approximation by splines with variable knots. In: Meir, A., Sharma, A. (eds.) Spline Functions and Approximation Theory. In: Proceedings of Symposium held at the Unoversity of Alberta, Edmonton, May 29-June 1, 1972, Birkhäuser, Basel (1973)
3. Jankowski, T.: Differential equations with integral boundary conditions, J. Comput. Appl. Math. 147, 1–8 (2002)
4. Jankowski, T.: Extensions of quasilinearization method for differential equations with integral boundary conditions, Math. Comput. Model. 37, 155–165 (2003)
5. Kellogg, R.B., Tsan, A.: Analysis of some difference approximations for a singular perturbation problem without turning points, Math. Comp. 32, 1025–1039 (1978)
6. Linß, T.: Layer-adapted meshes for convection-diffusion problems. Comput. Methods Appl. Mech. Eng. 192, 1061–1105 (2003)
7. Miller, J.J.H., O'Riordan, E., Shishkin, G.I.: Fitted Numerical Methods for Singular Perturbation Problems. World Scientific Publishing Co., Singapore (1996)
8. Samoilenko, A.M., Ronto, N.I., Martynyuk, S.V.: On the numerical-analytic method of problems with integral boundary conditions. Dokl. Akad. Nauk. Ukrain. SSR 4 (in Russian) 4, 34–37 (1991)

Application of Genetic Algorithm to Hausdorff Measure Estimation of Sierpinski Carpet

Qili Xiao and Lifeng Xi

School of Computer Science and Information Technology, Zhejiang Wanli University, Ningbo, Zhejiang, 315100 P.R. China
{xql, xilf}@zwu.edu.cn

Abstract. In this paper, we discuss the realization of the genetic algorithm on calculating the Hausdorff measure of the Sierpinski gasket with compression ratio 1/2 in detail, mainly including the encoding and decoding method, generation of the initial population, and fitness computation. The experimental results prove that the genetic algorithm is an effective and universal method to calculate Hausdorff measure.

Keywords: Genetic algorithm; Sierpinski gasket; Hausdorff measure; binary string; initial population.

1 Introduction

The genetic algorithm, which is a self adaptive and global optimizing probability search algorithm and is inspired by evolution, is introduced and investigated by John Holland (1975) and by students of Holland (e.g., DeJong, 1975). The genetic algorithm provides a universal method of solving complex problem. An implementation of the genetic algorithm begins with a population of chromosomes. One then evaluates these structures and allocates reproductive opportunities in such a way that those chromosomes which represent a better solution to the target problem are given more chances to "reproduce" than those chromosomes which are poorer solutions. The "goodness" of a solution is typically defined with respect to the current population. The implementation of the genetic algorithm is what Fig. 1 shows.

The genetic algorithm is robust, and the range of problems to which the genetic algorithm have been applied is quite broad [1], [2], [3]. Especially for the large complex nonlinear system, the genetic algorithm has particularly predominant capability.

The Hausdorff measure computation of fractals is well known as a problem, even for Sierpinski gasket. Recently, the genetic algorithm has been shown as a good approach [4], since it is able to handle the large space search more efficiently. However, for a different fractal, more works have to be done. One of these is how to encode the individual. For the Sierpinski gasket, the simplest method is just using a binary string. Second is how to generate the initial population for a particular problem. Third is how to evaluate individuals.

B. Apolloni et al. (Eds.): KES 2007/ WIRN 2007, Part II, LNAI 4693, pp. 182–188, 2007.
© Springer-Verlag Berlin Heidelberg 2007

In this paper, we will focus on the Sierpinski gasket with compression ratio 1/2. Section 2 outlines the Sierpinski gasket knowledge. Then we will discuss the encoding and decoding method, generation of the initial population, and fitness computation in detail in Section 3. Finally, the experimental results and the future work will be given in Section 4.

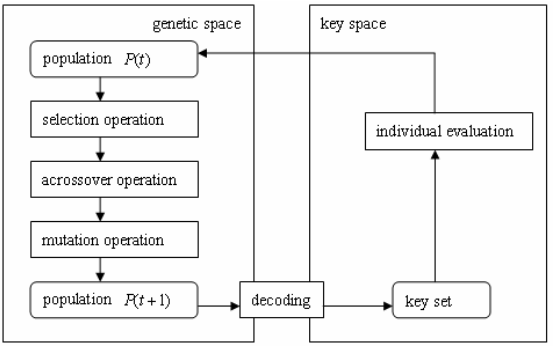


Fig. 1. The implementation of Genetic algorithm

2 On Sierpinski Gasket

Take a unit equilateral triangle in the Euclidean plane R^2 and denote it by F_0. If the midpoints of the three sides of F_0 are linked, four equal equilateral triangles are got. Removing the interior of the middle one (inverted), we get a set denoted by F_1. If the above procedure is repeated for each equilateral triangle in F_1, we get a set denoted by F_2. Repeating the above procedure infinitely (such as Fig. 2), we have $F_0 \supset F_1 \supset \cdots \supset F_k \supset \cdots$, and the non-empty set $F = \bigcap_{m=0}^{\infty} F_m$ is called the Sierpinski gasket yielded by F_0. By [4], [5] and [6], the Hausdorff dimension s of F is $\log_2^3$, and the Hausdorff measure $H^s(F)$ of F is $\lim_{m\to\infty} \inf_{U\in F_m'} \frac{|U|^s}{\mu(U)}$, where there are three mappings given respectively by Eq(1), Eq(2) and Eq(3), and where $F_{i_1 i_2 \ldots i_m} = S_{i_1} \circ S_{i_2} \circ \ldots \circ S_{i_m}(F_0)$, $F_m' = \{ U : U$ is a union of some small equilateral triangles $F_{i_1 i_2 \ldots i_m}$ in the m-th structure$\}$.

$$S_1(x) = x/2 \tag{1}$$

$$S_2(x) = x/2 + (1/2, 0) \tag{2}$$

$$S_3(x) = x/2 + (1/4, \sqrt{3}/4) \tag{3}$$

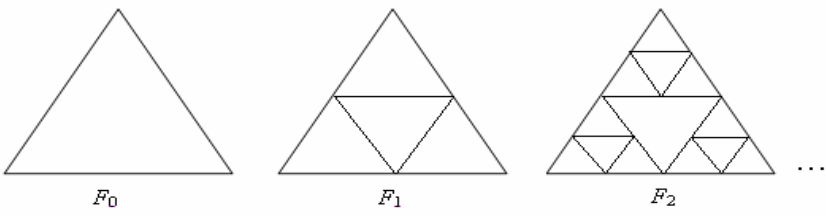


Fig. 2. The structure of the Sierpinski gasket

3 Genetic Algorithm

3.1 Encoding

During the process of calculating the Hausdorff measure, the method of binary code with fixed length is used. The encoding method is as follows: if $F_{i_1 i_2 \ldots i_m}$ is contained in the set $U \in F_m^{'}$, then its corresponding code is 1, otherwise its corresponding code is 0. In this way, a code list of 3^m length can show whether each one of the 3^m equal equilateral triangles in the m-th structure is chosen or not. In order to solve the problem easily, the origin is regarded as the starting point here, and every small equilateral triangle is encoded according to the anticlockwise order.

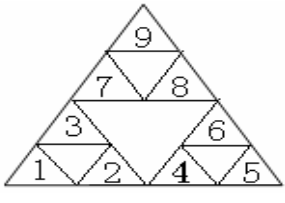


Fig. 3. Small equilateral triangles of the second structure

Now let's state the detail encoding method with the second structure. In the second structure, there are $3^2 = 9$ equal small equilateral triangles altogether. If we regard the origin as the starting point and encode every small equilateral triangle according to the anticlockwise order, the marks of the 9 equal small equilateral triangles are as shown in Fig. 3. We choose every one of the 9 small equilateral triangles one by one according to their marks from small to large. If a small equilateral triangle is chosen, we use 1 to represent on the corresponding position of the individual code, otherwise use 0. For example, 001 010 100 represents that three equilateral triangles of the 9 small equilateral triangles in the second structure are chosen and their marks are 3, 5 and 7 respectively (such as Fig. 3).

3.2 Decoding

During the calculation of the Hausdorff measure, an individual represents a choice. Then it is the decoding method that any binary digit 1 of the individual represents which of the 3^m small equilateral triangles is chosen, namely that the corresponding small equilateral triangle is the result which functions act on F_0.

The detail of decoding is as follows:

- Firstly, divide the whole code into three equal length groups, every small equilateral triangle that is represented by 0 or 1 in the first group is inside the equilateral triangle $S_1(F_0)$, and every small equilateral triangle that is represented by 0 or 1 in the second group is inside the equilateral triangle $S_2(F_0)$, and every small equilateral triangle that is represented by 0 or 1 in the third group is inside the equilateral triangle $S_3(F_0)$.
- Secondly, respectively divide every group above into three equal length groups, every small equilateral triangle that is represented by 0 or 1 in the first group of every group above is respectively inside the equilateral triangle $S_1 \circ S_1(F_0)$ and $S_1 \circ S_2(F_0)$ and $S_1 \circ S_3(F_0)$, and every small equilateral triangle that is represented by 0 or 1 in the second group of every group above is respectively inside the equilateral triangle $S_2 \circ S_1(F_0)$ and $S_2 \circ S_2(F_0)$ and $S_2 \circ S_3(F_0)$, and the rest may be deduced by analogy.
- Finally, continue dividing every group of the second step into three equal length groups until there is only one digit 0 or 1 in every group.

3.3 Generation of the Initial Population

The population is the foundation of evolution of Genetic algorithm. The character of the population decides the search capability of Genetic algorithm. And the astringency of the genetic algorithm is determined by the astringency of the population.

By analyzing the character of the structure of the classical Sierpinski gasket, we find the following:

- When the random $m > 1$ and there are $3^t (0 < t < m)$ equal small equilateral triangles non-chosen, which lie on the three vertexes of F_0 equally, the corresponding individual has the maximum fitness.

For example, in the third structure, if there are $3^2 = 9$ equal small equilateral triangles which are not chosen in some individuals, the individual, which the non-chosen triangles lie on the three vertexes of F_0 equally (such as Fig. 4), is the best individual with maximal fitness.

- When random $m > 3$ and $0 < t < m$ there are $(m-1)$ individuals altogether which have 3^t equal equilateral triangles non-chosen, lying on the three vertexes of F_0 equally. By calculating, we have found that the individual's fitness increases gradually with $t(t \le m-3)$ increasing, and that the individual's fitness decreases gradually with $t(t \ge m-3)$ decreasing.

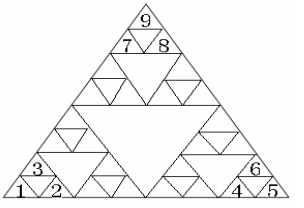


Fig. 4. Small equilateral triangles of the third structure

To improve the constringency pace of Genetic algorithm, the optimal searching subspace is the space, in which there are $r \in (3^{m-2}, 3^{m-4})$ equal small equilateral triangles non-chosen, lying on the three vertexes of F_0 equally.

3.4 Calculation of Fitness

Here fitness function is $F(U) = \dfrac{\mu(U)}{|U|^s}$, then finding minimum value is transformed into finding maximum value, where $\mu(U)$ represents the value of dividing the number of 1's in the individual code by 3^m, and $|U|$ represents the maximum distance of two random vertexes of six vertexes of two random small equilateral triangles. In order to calculate the maximum distance of any set U, a construction function series of every equilateral triangle in U must be found first, and then the coordinates of all vertexes of all small equilateral triangles in U can be calculated. Therefore the time used in calculating individual fitness increases rapidly. The following method is very effective, and the time of calculating individual fitness would be decreased greatly.

Step 1: calculate the coordinates of all vertexes of all small equilateral triangles. The detail calculation is as follows:

- Work out the construction function series of all small equilateral triangles as introduced above.
- Calculate the coordinates of the left vertex of all small equilateral triangles. The detail is as follows:

Suppose A, B and C are three small equilateral triangles in the m-th structure, and they are inside the equilateral triangle E, which is one of the equilateral triangles in the $(m-1)$-th structure with side $l = \frac{1}{2^{m-1}}$ (such as Fig. 5). If the coordinates of the left vertex of the equilateral triangle E is (x_0, y_0), the coordinates of the left vertex of small equilateral triangles A, B and C are given respectively by Eq(4), Eq(5) and Eq(6).

$$(x_A, y_A) = (x_0, y_0) \tag{4}$$

$$(x_B, y_B) = (x_0, y_0) + (\frac{1}{2} * l, 0) \tag{5}$$

$$(x_C, y_C) = (x_0, y_0) + (\frac{1}{4} * l, \frac{\sqrt{3}}{4} * l) \tag{6}$$

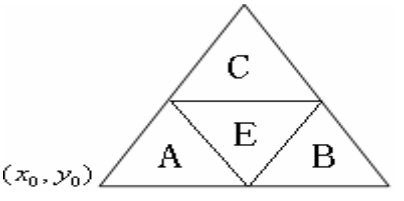


Fig. 5. Equilateral triangle E and three small equilateral triangles of the next structure

- Calculate the coordinates of the other vertexes of all small equilateral triangles. The detail is as follows: If a small equilateral triangle is one of equilateral triangles in the m-th structure and the coordinates of the left vertex is (x, y), the coordinates of other vertexes in anticlockwise order are $(x, y) + (\frac{1}{2^m}, 0)$, $(x, y) + (\frac{1}{2^m} * \frac{1}{2}, \frac{1}{2^m} * \frac{\sqrt{3}}{2})$.

Step 2: Calculate the individual fitness. The detail calculation is as follows:

- Find the coordinates of the small equilateral triangles chosen in the current individual according to the result of step 1.
- Calculate the diameter $|U|$ and the number of 1's in the individual code.
- Calculate the corresponding fitness.

4 Experimental Results and Analysis

During the calculation of the Hausdorrf measure, the main crossover method is the one-point crossover, and the mutation method is the gene mutation. The values of the key parameters are the following: the crossover probability $p_c = 0.9$, the mutation probability $p_m = 0.001$, the population size $popsize = 50$, the generation number $gen = 100$. According to above-mentioned ideas, we have programmed. The result of the experiment is what Table 1 shows.

Table 1. Values of $H^s(F)$ based on Genetic algorithm

m	$H^s(F)$	approximate computation times (unit: second)
2	0.95075	3
3	0.91041	10
4	0.869775	16
5	0.841718	28
6	0.831492	151
7	0.829821	1004
8	0.825065	10550
9	0.821064	110945

The main conclusions of this paper are the following:

- The Hausdorff measure of the Sierpinski gasket with compression ratio 1/2 $H^s(F) \le 0.821064$.
- In theory, the Genetic algorithm, which keeps the best individual, can calculate the optimal Hausdorff measure. And when $m \le 7$, the Hausdorff measure can be obtained quickly. But when $m > 7$, the calculating time increases sharply. Therefore, the future work is to improve the method of calculating the fitness further.

References

1. Ming, Z., Shudong, S.: Theory and Application of Genetic Algorithm. National Defence Industry Press, Beijing (2002)
2. Guoliang, C., Xufa, W.: Genetic Algorithm and Its Application. People Post Publishing Company, Beijing (2001)
3. Xiaoping, W., Liming, C.: Theory, Application and Software Realization of Genetic Algorithm. Publishing Company of Xi'an Jiaotong University, Xi'an (2002)
4. Lifeng, X., Zhongdi, C.: Some Front Problems of Fractal—Computation of Hausdorff Measure. Journal of Zhejiang Wanli University 14, 1–3 (2001)
5. Zuoling, Z.: Hausdorff Measure of Sierpinski Gasket. Science in China (A) 40, 1016–1021 (1997)
6. Zuoling, Z.: Hausdorff Measure of Self-similar set. Science in China (A) 28, 103–107 (1998)

Decomposability of Implication Operators*

Zhudeng Wang

Institute of Mathematics, School of Computer Science and Information Technology, Zhejiang Wanli University, Ningbo 315100, People's Republic of China
zhuden.w@pub.yc.jsinfo.net

Abstract. In this paper, we study the decomposability of S-implications, R-implications, n-reciprocal R-implications and n-reciprocal QL-implications.

Keywords: Fuzzy logic; t-norm; implication; direct decomposition.

The concept of the direct product of t-norms on a product lattice was introduced by De Baets and Mesiar [1]. They characterized $\vee$-distributive t-norms on a product lattice that are direct product of t-norms. In [2], Karacal discussed the direct decomposability of S-implications.

This paper is a continuation of [3,4,5,6]. We characterize the implication operator on a product lattice that is a direct product of two implication operators and study the direct decomposability of S-implications, R-implications, n-reciprocal R-implications and n-reciprocal QL-implications on a product lattice.

The knowledge about lattices required in this paper can be found in Birkhoff [7]. The terminology and notions about t-norms and pseudo-t-norms employed in the sequel are in agreement with ones in [3,4,5,6,8,9].

Throughout this paper, unless otherwise stated, L, L_1 and L_2 always represent any given complete Brouwerian lattices with maximal element 1 and minimal element 0; J stands for any index set.

The direct product $L_1 \times L_2$ of L_1 and L_2 is the set of all couples (a, b) with $a \in L_1$ and $b \in L_2$, partially ordered by the rule that $(a_1, b_1) \leq (a_2, b_2)$ if and only if $a_1 \leq a_2$ in L_1 and $b_1 \leq b_2$ in L_2. It is easy to see that the direct product $L_1 \times L_2$ is also a complete Brouwerian lattice (see [7]) and

$$(a_1, b_1) \vee (a_2, b_2) = (a_1 \vee a_2, b_1 \vee b_2),$$
$$(a_1, b_1) \wedge (a_2, b_2) = (a_1 \wedge a_2, b_1 \wedge b_2),$$

where (a_1, b_1) and (a_2, b_2) are any two elements of $L_1 \times L_2$.

Moreover, for any $(a, b) \in L_1 \times L_2$, we define $l((a, b)) = a,\ r((a, b)) = b$.

Definition 1 (Wang and Yu [5])**.** An implication operator (or an implication) I on L is a binary operation on L, satisfying two conditions:

* **Foundation item:** Ningbo Natural Science Foundation and Science Foundation of Zhejiang Wanli University.

B. Apolloni et al. (Eds.): KES 2007/ WIRN 2007, Part II, LNAI 4693, pp. 189–193, 2007.
© Springer-Verlag Berlin Heidelberg 2007

$(I1)$ $I(1,b)=b,\ I(0,b)=1\ \forall b\in L$;
$(I2)$ $a,b,c\in L,\ b\le c\ \Rightarrow\ I(a,b)\le I(a,c)$.

An implication operator I on L is called infinitely $\wedge$-distributive if it satisfies the following condition:

$(I_\wedge)$ $a,b_j\in L(j\in J)\ \Rightarrow\ I(a,\bigwedge_{j\in J}b_j)=\bigwedge_{j\in J}I(a,b_j)$.

Definition 2 (De Baets and Mesiar [1], Wang and Yu [6])**.** Let $T_1\in L_1^{L_1\times L_1}$ and $T_2\in L_2^{L_2\times L_2}$. A binary operation T on $L_1\times L_2$, defined by

$$T((a_1,b_1),(a_2,b_2))=(T_1(a_1,a_2),T_2(b_1,b_2))\ \ \forall(a_1,b_1),(a_2,b_2)\in L_1\times L_2,$$

is called the direct product (or Cartesian product) of T_1 and T_2, and written as $T=T_1\times T_2$.

In [2], Karacal studied the direct decomposability of S-implication operators on a product lattice. In this paper, based on Karacal [2], we study the implication operator on a product lattice that is a direct product of two implication operators and discuss the direct decomposability of S-implications, R-implications, n-reciprocal R-implications and n-reciprocal QL-implications on a product lattice.

Firstly, we characterize the implication operator on a product lattice that is a direct product of two implication operators.

Theorem 1. Let I be an implication operator on $L_1\times L_2$. Then, I is a direct product of two implication operators if and only if I satisfies the following conditions:

$(I3)$ $l(I((a,b),(c,d)))=l(I((a,1),(c,1)))\ \ \forall a,c\in L_1,\ b,d\in L_2$;
$(I4)$ $r(I((a,b),(c,d)))=r(I((1,b),(1,d)))\ \ \forall a,c\in L_1,\ b,d\in L_2$.

Proof. Straightforward.

Theorem 2. If I is an implication operator on $L_1\times L_2$ and satisfies the condition:

$(I5)$ $(a_1,b_1)\le(a_2,b_2)\ \ \Rightarrow\ \ I((a_1,b_1),(c,d))\ge I((a_2,b_2),(c,d))\ \ \forall a_1,a_2,c\in L_1,b_1,b_2,d\in L_2$,

then I is a direct product of two implication operators if and only if it satisfies the condition:

$(I6)$ $I((a,b),(c,d))=I((a,1),(c,1))\wedge I((1,b),(1,d))\ \ \forall a,c\in L_1,b,d\in L_2$.

Proof. If there exist two implication operators I_1 on L_1 and I_2 on L_2 such that $I=I_1\times I_2$, then, for any $a,c\in L_1,\ b,d\in L_2$,

$$\begin{aligned}&I((a,1),(c,1))\wedge I((1,b),(1,d))=(I_1(a,c),I_2(1,1))\wedge(I_1(1,1),I_2(b,d))\\&=(I_1(a,c),1))\wedge(1,I_2(b,d))=(I_1(a,c),I_2(b,d))=I((a,b),(c,d)).\end{aligned}$$

Conversely, if I satisfies the conditions $(I5)$ and $(I6)$, then

$$r(I((a,b),(c,d))) = r(I((a,1),(c,1))) \wedge r(I((1,b),(1,d))) \quad \forall a,c \in L_1, b,d \in L_2.$$

Noting that $r(I((a,1),(c,1))) \geq r(I((1,1),(c,1))) = r((c,1)) = 1$, we see that

$$r(I((a,b),(c,d))) = r(I((1,b),(1,d))) \quad \forall a,c \in L_1, b,d \in L_2.$$

Similarly, we have that

$$l(I((a,b),(c,d))) = l(I((a,1),(c,1))) \quad \forall a,c \in L_1, b,d \in L_2.$$

Thus, it follows from Theorem 1 that I is a direct product of two implication operators. □

Below, by Theorem 2, we study the direct decomposability of S-implications, R-implications, n-reciprocal R-implications and n-reciprocal QL-implications on a product lattice.

Example 1. S-implication (see Dubois and Prade [10], Fodor [11]) on $L_1 \times L_2$ is given by

$$I_S((a,b),(c,d)) = S(n(a,b),(c,d)) \quad \forall (a,b),(c,d) \in L_1 \times L_2,$$

where S is a t-conorm and n a strong negation on $L_1 \times L_2$.

It is easy to see that I_S satisfies the condition $(I5)$. Thus, I_S is a direct product of two implication operators if and only if it satisfies the condition $(I6)$.

When $I_S = I_1 \times I_2$, $I_1 = l(S(n(a,1),(c,1))$ is an S-implication operators on L_1 and $I_2 = r(S(n(1,b),(1,d))$ is an S-implication operators on L_2 (see Corollary 2 in [2]).

Example 2. R-implication (see Dubois and Prade [10], Fodor [11]) on $L_1 \times L_2$ is defined by using the concept of residuation as follows:

$$I_R((a,b),(c,d)) = \bigvee\{(x,y) \mid T((a,b),(x,y)) \leq (c,d)\} \quad \forall (a,b),(c,d) \in L_1 \times L_2,$$

where T is a t-norm on $L_1 \times L_2$.

We can see that I_R satisfies the condition $(I5)$ and hence I_R is a direct product of two implication operators if and only if it satisfies the condition $(I6)$.

When T is an infinitely $\vee$-distributive t-norm (i.e., left-continuous t-norm (see [8]) on $L_1 \times L_2$ and $I_R = I_1 \times I_2$, it follows from Theorem 3.1 in [6] that $T = T(I_1 \times I_2) = T(I_1) \times T(I_2)$ is a direct product of two infinitely $\vee$-distributive t-norms, and $I_1 = I(T(I_1))$ and $I_2 = I(T(I_2))$ are two infinitely $\wedge$-distributive R-implication operators.

Example 3. n-reciprocal R-implications (see Fodor [11]) on $L_1 \times L_2$ are of the form

$$I_R^n((a,b),(c,d)) = \bigvee\{(x,y) \mid T(n(c,d),(x,y)) \leq n(a,b)\} \quad \forall (a,b),(c,d) \in L_1 \times L_2,$$

where T is a t-norm and n a strong negation on $L_1 \times L_2$.

It is easy to verify that I_R^n satisfies the condition $(I5)$ and hence I_R^n is a direct product of two implication operators if and only if it satisfies the condition $(I6)$.

Clearly, $I_R^n((a,b),(c,d)) = I_R(n(c,d), n(a,b))$.

When n is the direct product of a strong negation n_1 on L_1 and a strong negation n_2 on L_2 (see [2]), we have that

$$I_R^n((a,b),(c,d)) = I_R((n_1(c), n_2(d)), (n_1(a), n_2(b))).$$

Moreover, if T is an infinitely $\vee$-distributive t-norm on $L_1 \times L_2$ and $I_R = I_1 \times I_2$, then it follows from Example 2 that I_1 and I_2 are two infinitely $\wedge$-distributive R-implication operators and

$$I_R^n((a,b),(c,d)) = (I_1(n_1(c), n_1(a)), I_2(n_2(d), n_2(b))) = (I_1^{n_1}(a,c), I_2^{n_2}(b,d)),$$

i.e., I_R^n is the direct product of two n-reciprocal R-implications $I_1^{n_1}$ on L_1 and $I_2^{n_2}$ on L_2.

Example 4. n-reciprocal QL-implications (see Fodor [11]) on $L_1 \times L_2$ are introduced by

$$I_{QL}^n((a,b),(c,d)) = S((c,d), T(n(a,b), n(c,d))) \quad \forall (a,b),(c,d) \in L_1 \times L_2,$$

where T is a t-norm, S a t-conorm and n a strong negation on $L_1 \times L_2$.

Similarly, we can show that I_{QL}^n satisfies the condition $(I5)$ and hence I_{QL}^n is a direct product of two implication operators if and only if it satisfies the condition $(I6)$.

When S, T, and n are, respectively, the direct products of two t-conorms S_1 and S_2, two t-norms T_1 and T_2, and two strong negations n_1 and n_2,

$$\begin{aligned} &I_{QL}^n((a,b),(c,d)) = S((c,d), T(n(a,b), n(c,d))) \\ &= (S_1(c, T_1(n_1(a), n_1(c))), S_2(d, T_2(n_2(b), n_2(d)))) = (I_{QL}^{n_1}(a,c), I_{QL}^{n_2}(b,d)). \end{aligned}$$

Thus, I_{QL}^n is the direct product of two n-reciprocal QL-implications $I_{QL}^{n_1}$ on L_1 and $I_{QL}^{n_2}$ on L_2.

Finally, we illustrate that the condition $(I5)$ in Theorem 2 is indispensable by following example.

Example 5. We define a binary operation I on $[0,1]^2$ as follows:

$$I((a,b),(c,d)) = \begin{cases} (c,d), & (a,b) = (1,1), \\ (1,1), & (a,b) = (0,0), \\ (1,d), & (a,b) = (0,1), \\ (c,1), & (a,b) = (1,0), \\ (0,0), & \text{otherwise.} \end{cases}$$

It can be shown that I is an implication on $[0,1]^2$ and satisfies the condition $(I6)$. But,

$$0 = l[I((1, \frac{1}{2}), (1,0))] \neq l[I((1,1),(1,0))] = 1,$$

and hence it follows from Theorem 1 that I isn't a direct product of two implication operators $[0,1]$. Noting that

$$(0,0) = I((0,\frac{1}{2}),(1,1)) < I((0,1),(1,1))] = (1,1),$$

we see that I doesn't satisfy the condition $(I5)$.

References

1. De Baets, B., Mesiar, R.: Triangular norms on product lattices. Fuzzy Sets and Systems 104, 61–75 (1999)
2. Karacal, F.: On the direct decomposability of strong negations and S-implication operators on product lattices. Information Sciences 176, 3011–3015 (2006)
3. Wang, Z.D.: Generating pseudo-t-norms and implication operators. Fuzzy Sets and Systems 157, 398–410 (2006)
4. Wang, Z.D.: Notes on the pseudo-$t$norms and implication operators on a complete Brouwerian lattice (I). Fuzzy Systems and Mathematics 20(1), 107–111 (2006)
5. Wang, Z.D., Yu, Y.D.: Pseudo-t-norms and implication operators on a complete Brouwerian lattice. Fuzzy Sets and Systems 132, 113–124 (2002)
6. Wang, Z.D., Yu, Y.D.: Pseudo-t-norms and implication operators: direct products and direct product decompositions. Fuzzy Sets and Systems 139, 673–683 (2003)
7. Birkhoff, G.: Lattice Theory. Amer. Math. Soc. Coll. Publ., Rhode Island (1967)
8. Ma, Z., Wu, W.M.: Logical operators on complete lattices. Information Sciences 55, 77–97 (1991)
9. Wang, Z.D.: On L-subsets and TL-subalgebras. Fuzzy Sets and Systems 65, 59–69 (1994)
10. Dubois, D., Prade, H.: Fuzzy sets in approximate reasoning, part 1: inference with possibility distributions. Fuzzy Sets and Systems 40, 143–202 (1991)
11. Fodor, J.C.: Contrapositive symmetry of fuzzy implications. Fuzzy Sets and Systems 69, 141–156 (1995)

Fault Detecting Technology Based on BP Neural Network Algorithm

Ran Jin, Kun Gao, Zhigang Chen, Chen Dong, Yanghong Zhang, and Lifeng Xi

School of Computer Science And Information Technology, Zhejiang Wanli University
Ningbo, Zhejiang 315100, P.R. China
ran.jin@163.com

Abstract. This paper describes an automatic online detecting system. In the system, digital image processing technology is used to preprocess X-ray images of the products, and neural network algorithm is applied to diagnose faults. The fault recognition model adopts an improved back-propagating neural network, which is trained by a series of standard X-ray images of correctly assembled products. The detecting system combines digital radiography technology with digital image processing, and applies the back-propagating neural network algorithm in the fault recognition process. The system improves the speed and reliability of fault detection and has application in the field of industrial nondestructive detection.

Keywords: nondestructive detection; fault diagnosis; neural networks; image processing.

1 Introduction

It is effective to detect inner parts of complex mechanical products, which can ensure the product quality, and prevent accident occurrence. The mostly mechanical products are composed by many spare parts, they fold and stagger in the space position. The mechanical products are commonly sealed by non-transparent outer covering, and internal component condition is not viewed directly, so it is difficult to detect many faults. The conventional detection uses artificial methods, such as sample detection or devastating test. For these methods, not only the expense is high, but also the consume is large, moreover belong to the afterwards analysis. Because the human factor may create missing or sentencing by mistake, which can greatly reduce the accuracy to detect and recognize, so these methods are not satisfied with the real-time production need. If the environment is bad or product itself has the certain risk, these conventional methods also are unsafe. This paper describes an automatic online detecting system which can detect inner parts of complex mechanical products. In the system, digital image processing technology is used to preprocess X-ray images of the products, and neural network algorithm is applied to diagnose faults. The fault recognition model adopts an improved back-propagating neural network, which is trained by a series of standard X-ray images of correctly assembled products. During

B. Apolloni et al. (Eds.): KES 2007/ WIRN 2007, Part II, LNAI 4693, pp. 194–201, 2007.
© Springer-Verlag Berlin Heidelberg 2007

the process of detection, two images of objects in different directions are capable of acquiring the status of the essential parts. After comparing the features of the two preprocessed images and standard images, the network can estimate different types of faults of the essential parts[1][2].

2 Examination System Structure and Principle of Work

The entire detecting system is composed by the hardware and software. The hardware mainly includes X-ray fountain, the worktable for circumgyrating detection, radial visible light transforming screen, digitized equipment, mechanical and electrical controller, computer and the peripheral equipments such as image memory, display, output[3][4][5]. The software is an application about image processing and pattern recognition based on operating system.

2.1 System Structure

System composition as Fig.1. The entire system is composed by imaging and recognition. When system detection, the radial from X-ray fountain beats on radial transforming screen through penetrating the detected product on worktable, then transforms into visible light image, and obtains this image by CCD digital camera, finally, the image is put into computer. The electromechanical controller of detecting system can gain original radial image about different directions of detected product through controlling step-advance electrical machine to drive the worktable revolving Through image pretreatmen , feature extraction and analysis, disposal and recognition to component by computer, the system can realize automatic detection to the product inner component condition, at last the detecting results are exported by export equipment[6][7].

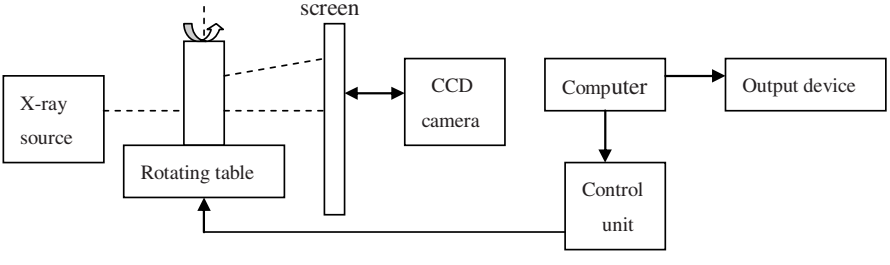


Fig. 1. Configuration of NDT system

2.2 Fault Recognition Principle

The commonly used pattern recognition method have the template match, the statistical pattern recognition, the syntax structure recognition, the fuzzy pattern recognition, pattern recognition based on the knowledge, pattern recognition based on the model and the nerve network method and so on[8].

The neural network, based on the cerebrum nervous system structure and the functional simulation, has auto-adapted, auto-organization and auto-learning capability, which may enhance the quantity and the quality memory in the neural network through unceasing study to standard samples. Especially, the nerve network can extract the comparability from similar samples and the difference from different category samples, and materialize in connecting power values adjustment between neural cells of neural network[8]. Even if environment information is not completeness, it still was allowed to obtain a more satisfying answer through computing, thus opened a new way for designing and developing the fault diagnosis system. The domestic and foreign literature indicated that, at present many researchers have used neural network models in the fault detection and diagnosis domain, and obtains a better effect.

System this paper adopted as Fig.2 shows, which can divide into train and recognition.

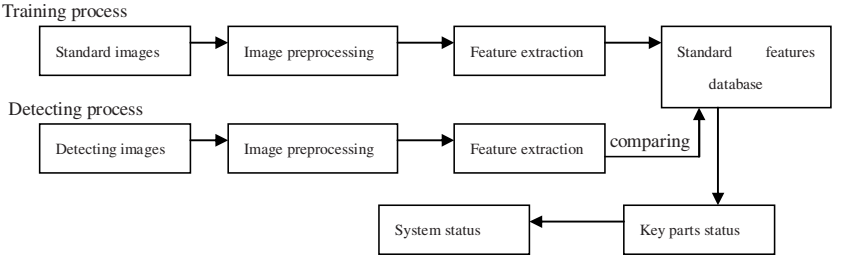


Fig. 2. NN training process and detecting process

The learning process of neural network model is divided into two steps:

(1) The work table circles the rotation axis to revolve, each the certain angle obtains an image, then can obtain each directional primitive image after revolving a week. For instance angle gap is ten degrees, then may obtain nineteen images. The step advance machine controlled by the computer can adjust the angle gap to satisfy the different detecting request.
(2) Extracting the gradation characteristics of essential spare parts after preprocessing for primitive images, and establishing s n r s e characteristic storehouse.

First of all is to preprocess for sample images which have been obtained. The preprocess of images include the filter goes chirp, the background division, the image enhancement, then according the product characteristics to orient essential parts and extract the corresponding gradation characteristics. Finally establishing characteristics storehouse and training neural network.

Underside are recognition processes:

Shoot an image for detected product and extract the gradation characteristics about corresponding parts after preprocessing, then match with standard sample images to find a sample image close to standard sample images.

(1) Under assign an erroneous scope condition, compare characteristics with standard sample images, and obtain the comparative results, then can use the image analytic theory to judge the part with the dislocation .
(2) If certain essential spare parts are unable to judge, revolves the product an angle(commonly is ninety degrees), turn to (1) step.
(3) Using reliable theory can calculate system condition After all condition of essential parts is confirmed.
(4) The detected results which are exported by computer can display the product system condition (normal or fault). If system has fault, then must export part condition that causes the system failure, so can record each product condition in the database to prepare repairs a second time. In this way, system can realize the product automatic detection and the diagnosis[9][10].

In the ordinary circumstances, two images which have difference of ninety degrees can judge the product condition.

3 Image Preprocess

Preprocessing to standard images and detected images before recognition. The aim of image preprocessing is to wipe off noise, and stand out product component information. This paper uses the preprocessing methods mainly have the filter, the target background division and the image enhancement[11].

3.1 Filter Noise Reduction

Apart from the presence of quantum fluctuant noise, there is few impulse noise in X-ray images. This paper uses median filter to remove noise and not make the edge blur.

3.2 Background Division

In actual detection, detected target is only part of the whole image, and the rest is background. Due to the impact of noise and uneven ray images, the accuracy of feature extraction is generally reduced by background factor. Therefore it is necessary to divide up target and background to remove background which can eliminate the disadvantage infection to characteristics extraction. Fig.3(a) is a histogram about original image, it belongs to double apexes distribution seeing from this Histogram, and trough is more broad. If iterative valve value choosing algorithm is adopted, background and object can be detached as well as simply and quickly. Underside are concrete steps:

(1) Choosing initial threshold value as average of image intensity;
(2) Dividing image into two groups through using valve value, $R1$ and $R2$;
(3) Computing the average of region $R1$ and the average of $R2$, and obtain the average is respectively $\mu 1$ and $\mu 2$;
(4) Choosing a new valve value;
(5) Repeating (2)—(4), till $\mu 1$ and $\mu 2$ are no longer change[12].

3.3 Image Enhancement

In ray imaging, the dynamic range of radial images is restricted from the imaging Technology, then dynamic range is often smaller, and image contrast is not high. This paper introduces an amendatory method to histogram that can realize image enhancing and stand out the features information of the image.

Fig.3(a) is an original product image and opposite histogram, Fig.3(b) is an image after preprocessing and opposite histogram[13].

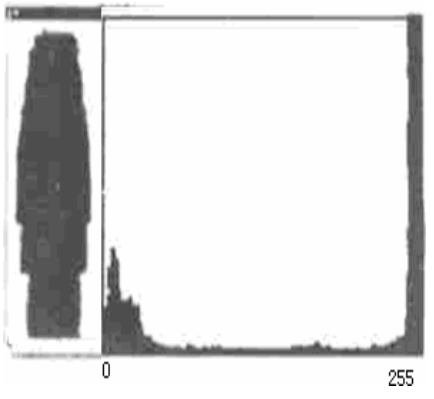


Fig. 3(a). Original image and its histogram

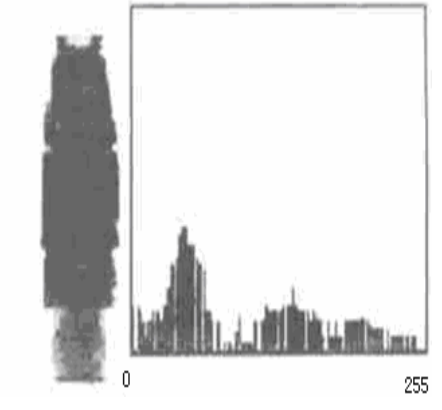


Fig. 3(b). Preprocessed image and its histogram

3.4 Characteristic Extraction

The noise of radial image is great, product structure is complex, inner components overlap and stagger each other. Due to the location of imaging components at least on one dimension direction is uncertain, therefore characteristics extraction and recognition is difficult if recognizing barely by image structure. The extracted characteristics require to reflect the greatest possible location of components, and must be easily manageable[14][15].

4 Neural Network Model

Back-propagating network(BP network) is a type of neural network, which is the most clear clarity and the most widely used. When positive direction spread, the imported model disposes layer by layer by way of hidden units from the input layer and sent to the output layer, neural state of each layer only affect neural state of the next layer. If the expected output can not be obtain in the output layer, so must transfer to back propagation, and let error signal back along the original link pathway, the error signal can be became least through amend the values of each nerve cell. The advantages of BP network mainly represents in pattern matching, pattern classification, pattern recognition and pattern analysis. In these aspects, BP network have the strongly input-output mapping ability, strongly information processing capacity and strongly robust[16].

4.1 Network Topology

Multilayer BP network consist of a imported layer, a output layer and a hidden layer at least.

4.2 BP Network Using

Considering followings in model design:

(1) Using three layer BP network

Research indicate that can approach using a hidden layer for a continuous function in any closed interval, then a three BP network can achieve mapping from arbitrary N dimensions to M dimensions. The gradient will become more unstable and convergence rate dropped sharply if we adopt a number of hidden layer, so we adopt a three-layer network structure having a hidden layer in this paper.

(2) Choosing sigmoid non-linearity transferring function

$$f(\xi) = \frac{1}{1 + \exp(-\xi - \theta)} \tag{1}$$

In formula, θ—threshold windage

(3) Mending of error function

Using cross entropy function can reduce about half of the study time.

$$E = \sum_{j} [-y_j \ln(V_j^m) - (1 - y_j) \ln(1 - V_j^m)] \tag{2}$$

(4) The selection of hidden units

The selection of hidden units is a very complex issue. Analytic formula has expressed no good can be said that the number of hidden units are directly related with problematic request, the number of input and output modules. Generally speaking, the hidden network modules too few, might not be a good network training or network not "strong", so this network fails to recognize not previously studied samples, and fault-tolerant is bad; but too many hidden units also make learning a long time, and not necessarily the smallest error, so exist a optimal hidden units. In experiment, we ensure hidden units refer to as follow formula:

$$h = \sqrt{(N + M)} + \beta \tag{3}$$

In formula, M is output units, N is input units, β is a constant between 1 to 10.

(5) Mending algorithm

BP algorithm has few problems such as slow convergence and local minimum. Model adopts Variable Stepping algorithm to improve its convergence, and step adopts dynamic mode to adjust. If error increases after one training, then the next step should reduce properly; if error decreases after one training, then the next step should increase properly. The method can learn and train automatically, and reduce the number of training. Amending formula as follow:

$$\Delta w_{ij}(t+1) = -\eta d_j^k V_j^{k-1} + \alpha \Delta w_{ij}(t) \tag{4}$$

In formula, α —Adjusted parameter.

5 Conclusion

Experimental product has two steel balls, their diameter is 3mm—4mm, and the steel balls are pivotal parts. Figure.4 shows characteristic fitting curve about established standard samples, small dots are characteristics of a sample. Abscissa correspond different angle, and vertical coordinate is characteristic quantum that be normalized. In a certain area that the error allows, when the imaging relevant characteristics exceed the scope of characteristic values, these values belong to the matched samples, so we judge the steel balls are not put into the product. The stars in Fig.4 are characteristic values of product that two separated by 90 degrees image when lacking one steel ball. When we choose 0.03 as allow error, then the result of recognition is leakage loaded with steel ball, the other essential pieces of similar results. The detecting system can judge correctly the good products, and quickly recognise the fault products that the internal components leak and fall off. Choosing rational allow error can achieve the misjudgment rate to least according to testing results.

The detecting method put by this paper combines digital radiography technology with digital image processing, and applies the neural network algorithm in characteristic matching and recognition. In recognition, we adopt an improved back-propagating neural network, which is trained by a series of standard X-ray images of correctly assembled products. Feature space is compressed a simple gray information of one dimensional distribution, which has a good detection performance. training times and detecting times is short(a neural network with nineteen samples trains within thirty minutes), network can be trained to recognise many faults such as leakage and dislocation, which is suitable for complex product line measurement of the internal components.

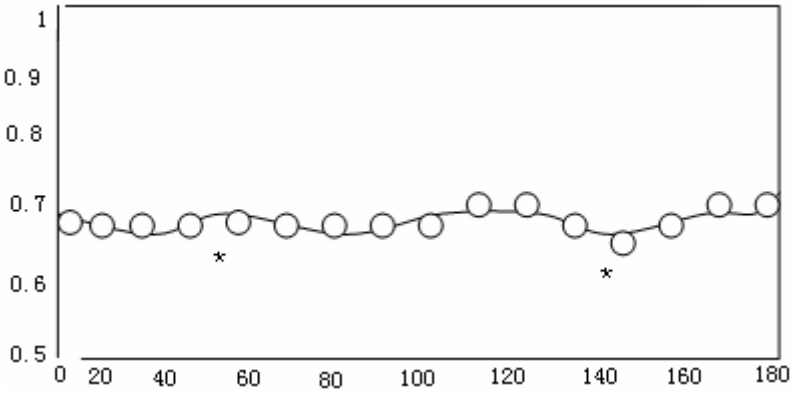


Fig. 4. Original image and its histogram

References

1. Leger, R.P., Garland, W.-M.J., Poehlman, W.-F.S.: Fault Detection and Diagnosis Using Statistical Control Charts and Artificial Neural Networks. Artificial Intelligence in Engineering (1998)

2. Gomm, J.B., Weerasinghe, M., Williams, D.: Diagnosis of Process Faults with Neural Networks and Principal Component Analysis. Proc.Inst.Mech.Eng., 131–143 (2000)
3. Simani, S., Fantuzzi, C.: Fault Diagnosis in Power Plant Using Neural Networks. Information Sciences, 125–136 (2000)
4. Weerasinghe, M., Barry, G.J., Willams, D.: Neural Networks for Fault Diagnosis of A Nuclear Fuel Processing Plant at Different Operating Points. Control Engineering Practice, 281–289 (1998)
5. Zhuang, Z.: ANN Based Fault Testing on Roolling Bearing. Chinese Journal of Scientific Instrument, 73–741 (2000)
6. Liu, L.Z.: The Study on Feasibility of Adopting Neuron Network in Diagnosis of Coolant Pump System in Nuclear Heating Pile. Process Automation Instrumentation, 4–71 (1999)
7. Burch, S.: Digital Enhancement of Video Images for NDT. Ndt International, 51–56 (1987)
8. Faruqi, A.R., Tyrell, G.C.: Evaluation of Gadolinium Oxy-sulphide Phosphor Used in CCD Detectors for Electron Microscopy. Ultramicroscopy (1999)
9. Omer, D., Erik, L.R.: Noise Reduction in X-ray Microtomography Images by Anisotropic Diffusion of The Scanner Projection Images. Spie, 690–697 (1999)
10. Sinha, R.K.: Backpropagation Artificial Neural Network Detects Changes in Electro-encephalogram Power Spectra of Syncopic Patients. Jmed. Syst., 63–68 (2007)
11. Neelamegam, P.: An Approach to Measure The Densities of Solids Using An Artificial Neural Network. Instrum. Sci. Technol., 189–199 (2007)
12. Ferentinos, K.P.: Neural network-based detection of mechanical, sensor and biological faults in deep-trough hydroponics. Comput Electron Agr, 65–85 (2003)
13. Zhang, B., Yan, J.: Application of BP Neural Net in Fault Diagnosis of Excavator Electric system. Construction Machinery (2006)
14. Chen, J.C.: Machine Learning for Information Retrieval: Neural Networks, Symbolic Learning, and Genetic Algorithms. Jasis, 193–216 (1995)
15. Zhang, L.: Expert System Based on Neural Network for Compressor Fault Diagnosis. Tractor. Farm Transporter (2006)
16. Sun, X.: Application of BP Neural Networks to Fault Diagnosis of Turbocharging System of Diesel Engine. Application Research of Computers (2006)

Image Information Hiding Encryption Using Chaotic Sequence

Zhen Liu and Lifeng Xi

School of Computer Science and Information Technology, Zhejiang Wanli University
Ningbo, Zhejiang 315100, P.R. China
{Liuzhen,xilf}@zwu.edu.cn

Abstract. In this paper, we propose a new image watermark algorithm of information hiding scheme based on chaotic encryption. Chaotic sequence is regarded as a noise-like spread sequence, and the secret signal to be transmitted is modulated and hidden in the carrier image as white noise to implement secure communication. The encryption principle of new algorithm is firstly presented in this paper, and the merits of new scheme are quantitatively analyzed by defining integrality and imperceptibility of image information. The results of computer simulation are described in the end.

Keywords: Information hiding, chaotic sequence, encryption algorithm, wavelet transform.

1 Introduction

Chaotic system is a kind of complicated, nonlinear dynamic system. It has perfect security with the following properties: good pseudo-randomicity, orbital inscrutability, extreme sensitivity about the initial value and the control parameter, and etc. In the recent years, chaotic encryption communication has become a hot research topic. With the increased attention on the multi-media information security, the image encryption technology is widely applied, especially of the numerical images encryption. Recently, traditional data encryption is valid only to two dimension data and text information, but image data has the following properties: large capacity and the high redundancy degree, therefore the traditional data encryption has some limitation in the numerical image encryption[1]. Currently some special methods are used in the information encryption. In this paper, we propose a new algorithm of image information encryption scheme based on chaotic sequence. Multiplying several pieces of secret images with the respective chaotic sequence, we get the spread signal, and then add the signal to the transform domain of the carrier image. In this way we can implement networks encryption communication with high security and large capacity.

2 Chaotic System Principle

Chaos is an aperiodic, non-convergent process, which sensitively depends on the initial state. It is determined but stochastic-like and exists in the nonlinear

B. Apolloni et al. (Eds.): KES 2007/ WIRN 2007, Part II, LNAI 4693, pp. 202–208, 2007.
© Springer-Verlag Berlin Heidelberg 2007

dynamic system. A discrete temporal nonlinear dynamic system can be described as follows:

$$x_{k+1} = \theta(x_k) \tag{1}$$

Where x_k P k=0,1,2,3…… they are states. While θ P → P is a mapping, which links the current state x_k the next state x_{k+1}. So if we start with the initial state x_0, and do mapping[2][3] θ repeatedly, we will get a sequence $\{x_k; k=0,1,2,3\ldots\ldots\}$. This sequence is the track of the discrete system.

One simple but widely used dynamic system is a logistic mapping. It is defined as the following formula:

$$x_{k+1} = \mu x_k (1 - x_k) \tag{2}$$

Where $0 \le \mu \le 4$ is branch parameter and $x_k \in (0,1)$. Research on Chaos dynamic system show that when $3.5699456... < \mu \le 4$, logistic mapping will be in chaos condition. That is, with the initial state x_0, the state sequence $\{x_k;k=0,1,2,3\ldots\ldots\}$ generated by logistic mapping is an aperiodic, nonconvergent course, which sensitively depends on the initial state.

There is another simple mapping, Chebyshev mapping, with the parameter of order, k order Chebyshev mapping is like:

$$\tau(x_{k+1}) = \cos(n(\cos^{-1} x_k)) \tag{3}$$

Where x_k is in the range of -1 1 . In fact, logistic mapping can also be defined in the region of -1 1 with the help some transform. Then we will get

$$x_{k+1} = 1 - \lambda x_k^{\,2} \quad \text{where} \quad \lambda \in [0,2] \tag{4}$$

With the condition of full mapping λ =2, logistic mapping is topologically conjugated with Chbeshev mapping. They have the same PDF,

$$\rho(x) = \left\{ \begin{array}{l} \dfrac{1}{\pi\sqrt{1 - x^2}}, -1 < x < 1 \\ 0, else \end{array} \right\} \tag{5}$$

If μ =4, with the logistic mapping of 2 , PDF will be as

$$\rho(x) = \left\{ \begin{array}{l} \dfrac{1}{\pi\sqrt{x(1-x)}}, 0 < x < 1 \\ 0, else \end{array} \right\} \tag{6}$$

With the help of $\rho(x)$, We can see some statistic properties of logistic mapping[4] generated chaos sequence. For example, the temporal average of x is:

$$\bar{x} = \lim_{N\to\infty} \frac{1}{N} \sum_{i=0}^{N-1} x_i = \int_0^1 x\rho(x)dx = 0 \tag{7}$$

We can also get the correlation function for the sequences with the initial state of x_0 and y_0,

$$c(l) = \lim_{N\to\infty} \frac{1}{N} \sum_{i=0}^{N-1} (x_i - \bar{x})(y_{(i+l)} - \bar{y}) = \int_0^1 \int_0^1 \rho(x,y)(x-\bar{x})(\tau'(y)-\bar{y})dxdy = 0 \tag{8}$$

Where PDF $\rho(x, y) = \rho(x) * \rho(y)$

Then the auto correlation function ACF is $\delta(l)$, that is what we want, from the discussion above, we can see that, in spite of the determinability of the chaos dynamic system, the statistic properties show the simplicity of format and sensitivity on initial value, in common with the white noise.

3 Image Information Hiding Method

The method of image information hiding based on chaotic encryption is to multiply the original secret image with the chaotic sequence and get a spread frequency signal with pseudo-random noise, then add this signal to the transform domain of carrier image. In this case, a high security and large volume networks communication is accomplished.

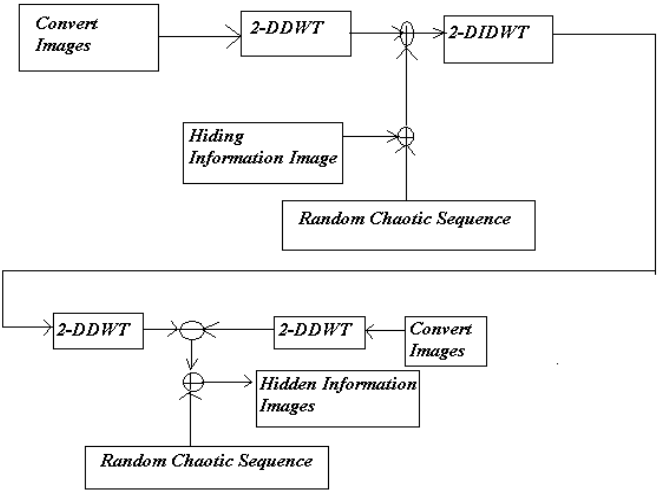


Fig. 1. Multi-image information hiding scheme

In this method, wavelet transform is applied. Wavelet transform is hot method, which divides the carrier image into several levels or parameters, as to hide multi secret images in different level or parameters. Please note that there could be disturbance between the secrete images when then hidden in a level. To reduce this, we can make use of the properties of function $\delta(l)$. As mentioned above, Chebyshev mapping is typical chaos system and its chaos sequence show high auto correlation performance, and the chaos sequence matrix is close to Hadamard matrix, so less disturbance between the secrete images will be produced.

Figure 1. shows the schematic diagram of multi-image Information Hiding based on wavelet transform. It includes two main parts: information embedding and extraction.

(1) Information Embedding

For gray image, a two dimensional matrix is used to record the value for the pixels of the image. A denotes the original carrier image, when it is processed with 2D-DWM , four components are generated. We mark it as B,

$$B=[cA,cH,cV,cD]=dwt2(A) \tag{9}$$

Take four pieces of images as an example, $C_i(i=1,2,\ldots,n)$ is for the secret image information. R_i is chaos random matrix. As simplified, we assume C_i and R_i are the same size matrixes.

$$D_i = C_i * R_i \tag{10}$$

Then Di and B will have the randomicity of Ri

$$E = [cA + K * D_1, cH + K * D_2, cV + K * D_3, cD + K * D_4] \tag{11}$$

In the formula (11), K is weighted factor of information hidden. F is the 2D-DWM of E,

$$F = idwt2(E) \tag{12}$$

Then discrete the value into 8 bits gray levels, you will get the hiding image.

$$G = uint8(F) \tag{13}$$

So Ci is included in G. and G will be transmitted to the receiver through the public networks. Normally, you can not distinguish it from the original one.

(2) Information Extraction

After the image is received, another 2D-DWM is required,

$$H = dwt2(G) = [cA', cH', cV', cD'] \tag{14}$$

Afterward, subtract B from H and include the weighted factor,

$$I = \begin{cases} (cA' - cA)/K \\ (cH' - cH)/K \\ (cV' - cV)/K \\ (cD' - cD)/K \end{cases} \tag{15}$$

I multiplies with R_i', the embed information J_i is obtained.

$$J_i = I * R_i' \tag{16}$$

To display it as gray image,

$$Q_i = u\,\mathrm{int}\,8(J_i) \tag{17}$$

Q_i is the recovery of C_i. It can be seen above that only the precondition below is satisfied, Qi can be close to Ci.

$$\begin{aligned} &< R_i', R_j >= \sum_j (R_j * R_i') = G^2 * \delta_{ij} \\ &G^2 \quad R_i * R_i'(i = 1, 2, ...) \quad \delta_{ij} = \begin{cases} 1 \\ 0 \end{cases} \end{aligned} \tag{18}$$

4 Results and Discussion

In the computer simulation process, 4 pieces of hidden secret images are shown in Figure 2, and the original covert image and the new covert image are shown in Figure 3. And then simulation results are made in discussion.



(a)C_1 (b) C_2 (c) C_3 (d) C_4

Fig. 2. 4 pieces of hidden secret images

Imperceptibility is one of the important performance index[5] in information hiding technology. In the algorithm, it is the requirement that the difference of the original covert image and the secret images is not found by computer or people. The image's correlation coefficient γ is the index.

$$\gamma = \frac{\sum_m \sum_m (A_{mm} - \bar{A})(G_{mm} - \bar{G})}{\sqrt{(\sum_m \sum_m (A_{mm} - \bar{A})^2)(G_{mm} - \bar{G})^2)}} \tag{19}$$

Where, $\bar{A}$ and $\bar{G}$ are respectively the average values of image A and image G. When the weighting coefficient K equals 0.01, the correlation coefficient is 0.99952, it means the hidden information is highly imperceptible. And we can not tell the difference. At the same time, we can see the original image and the main information is not lost.



(a) original image (b) new carrier image(k=0.01)

Fig. 3. Original image and new carrier image

5 Conclusion

According to the theory of chaos and wavelet transform, we propose a new method of image information hiding scheme in the paper. Multiple secret images are hidden in the original covert image with small noise form, and the chaotic sequence plays the dual roles of key and makes secrete signal into white noise form. Compared with traditional method, the new method is improved greatly in the security, capacity and the robustness of the information hiding, and these are verified through computer simulation results. For the future, with the development of spectrum techniques[6], the algorithm could be used to hide and encrypt images in it.

References

1. Kocarev, L.: Chaos-based Cryptography: A Brief Overview. IEEE Circuits and Systems Magazine 1, 6–21 (2001)
2. Habutsu, T., Nishio, Y., Sasase, I.: A Secret Cryptosystem by Iterating a Chaotic Map. In: Davies, D.W. (ed.) EUROCRYPT 1991. LNCS, vol. 547, pp. 127–140. Springer, Heidelberg (1991)
3. Fridrich, J.: Symmetric Ciphers Based on Two-dimensional Chaotic Maps. Int. J. Bifurcation Chaos 8, 1259–1284 (1998)

4. Jakimoski, G., Kocarev, L.: Chaos and Cryptography: Block Encryption Ciphers Based on Chaotic Maps. IEEE Trans. Circuits Syst. 48, 163–169 (2001)
5. Weeks, M., Bayoumi, M.: Discrete Wavelet Transform: Architecture, Design and Performance Issues. The Journal of VLSI Signal Processing 35, 155–178 (2003)
6. Cox, J., Kilian, J., Leighton, T., Shamoon, T.: Secure Spread Spectrum Water Making for Multimedia. IEEE Trans on Image Processing 6, 1673–1687 (1997)

Improvement of an Image Encryption Algorithm Based on Combined Multidimensional Chaotic Systems

Chen Dong[1], Jifang Li[1], Lifeng Xi[1], Jie Hao[2], and Ran Jin[1]

[1] School of Computer Science and Information Technology, Zhejiang Wanli University, Ningbo, Zhejiang 315100, P.R. China
[2] Department of Computer Teaching , Zhejiang Wanli University, Ningbo, Zhejiang 315100, P.R. China
{dongchen,jifangl,xilf,haojie1978,ran.jin}@zwu.edu.cn

Abstract. In this paper, we propose a chaotic image encryption algorithm, which adopts the mechanism of shuffling the positions and changing the grey values of image pixels. Firstly, the discrete Kolmogorov flow is used to shuffle the positions of the image pixels in the spatial-domain. The second, the Liu chaotic system is adopted to generate integer key for the image encryption, and the shuffled image is encrypted by the integer key. The third, we show that the advantages of large key space due to its improved security and the maturity of the technology using chaotic image encryption algorithm.

Keywords: Chaotic, Image encryption, Kolmogorov flow, Liu chaotic system.

1 Introduction

With the development of information technology and increasing frequency of information exchange, the research on information security engineering is becoming more and more important. As a novel technology in cryptography, digital chaotic cipher has been widely conducted a lot of research [1][2]. Chaos, as a kind of nonlinear phenomenon, has many remarkable features: such as pseudo-random, unpredictability of evolution of its orbits, properties of mixing and sensitivity to initial conditions and system parameters. These features are in accordance with the requirements of cryptography. Research shows that there are some relations between the traditional cipher and chaos. As compared with the traditional cryptography that has many challenging due to the bulk data capacity, high redundancy of widely-used patterns and multi-media information, the chaotic cipher has a great superiority over the encryption of multi-media information [3].

Since 1980s, the idea of using chaotic systems to design new ciphers has attracted more and more researchers. The utilization of chaos in cryptography depends on the natural relationship between the two research areas: strong dynamical properties of chaotic systems imply strong crypto graphical properties of cryptosystems, and the basic way to design cryptosystems containing good strong crypto graphical properties implies quasi-chaos [4][5]. The boom of research on chaotic ciphers occurred from 1989; two chaotic ciphers schemes proposed by Matthews [6], Pecora[7] and

B. Apolloni et al. (Eds.): KES 2007/ WIRN 2007, Part II, LNAI 4693, pp. 209–216, 2007.
© Springer-Verlag Berlin Heidelberg 2007

Carroll[8], and many other researchers are made to get many promising achievements[9][10]. Although a number of recently proposed chaotic ciphers have been crypt analyzed, many others have not been effectively attacked yet. FASS curve, Arnold transforms and life-game has put forward many new digital images scrambling algorithm, but all of them are limited. For example, Periodicity on some algorithms are determined, on other algorithms are independent of key. At the same time, some systematic ways to design chaotic ciphers have been proposed with careful considerations on the security of algorithms [11][12].

In order to improve the security of information, this paper presents improvement of an image encryption algorithm based on a combined multidimensional chaotic system. This paper is organized as follows. In section 2, we review Kolmogorov flow is used to scramble image, which fully utilizes the chaotic sequence with a simple form, initial conditions, the sensitivity of the statistical features etc. In section 3, we propose the sequence values of the Liu chaotic system is adopted to generate integer key for the grayscale image encryption, which makes the Encryption key is highly sensitive. Finally in section 4, we verify the effectiveness of our proposed solutions through analysis and computer simulation.

2 Displacement Encryption Scheme

A discrete use of the Kolmogorov flow transforms image replacement. Key features of this approach are a large space, such as the image of 256 × 256, choice for the key space 2^{209}; the image of 512 × 512, choice for the key space 2^{418}. Also, such a method has good confused and scattered. In this paper, methods for expansion, in order to meet the N × M image, the application of this algorithm. The specific algorithm is described as follows [13][14]:

Regarding N × M image, supposes an integer sequence:

$$\delta = \left(n_1, n_{2,} \ldots n_k\right), 0 \langle n_i \langle N \tag{1}$$

Then

$$\sum_{i-1}^{k} n_i = N \tag{2}$$

If q_s=M/n_s determined q_s (s = 1,2,…,k).

Here all q_s are positive integer, N_s indicated contains replacement (x, y) left side of vertical climatic zone, formula as follows:

$$N_S \begin{cases} 0 & s = 1 \\ n_1 + \ldots + n_{s-1} & s = 1, \ldots, k \end{cases} \tag{3}$$

Regarding $(x, y) \in [N_S, N_S + n_s] \times [0, M]$

Definition mapping replacement:

$$T_{n,\delta} : [0, N] \times [0, M] \to [0, N] \times [0, M]$$

Then $T_{n,\delta}$

$$T_{n,\delta}^{-1}(x,y)=(q_s(x,y)+(yMODq_s),(yDIVq_s)+N_S) \tag{4}$$

Definition mapping the counter- replacement:

$$T_{n,\delta}:[0,N]\times[0,M]\to[0,N]\times[0,M]$$

Then $T_{n,\delta}^{-1}$

$$T_{n,\delta}^{-1}(q_s(x-N_S)+(yMODq_s),(yDIVq_s)+N_S)=(x,y) \tag{5}$$

3 Liu Chaotic System and Image Picture Encryption

Chongxin Liu proposed a new chaos system - Liu system, its mathematical expression as follows:

$$\begin{cases} x=a(y-z) \\ y=bx-kx^2 \\ z=cz+hx^2 \end{cases} \tag{6}$$

System parameters are a, b, k, c, h, when a=10, b=40, k=1, c=2.5, h=4, equation (6) is at the chaotic state. Takes at the beginning of the system (2.2, 2.4, 38), the time sector takes [0, 70], use changes length of stride Runge-Kutta-Felhberg algorithm, solves this chaos system differential equation, may obtain system mode evolution path (Fig. 1). The condition evolution path was composed by many separate spots, has unfolded this non-linear chaos system complex chaos characteristic [15][16].

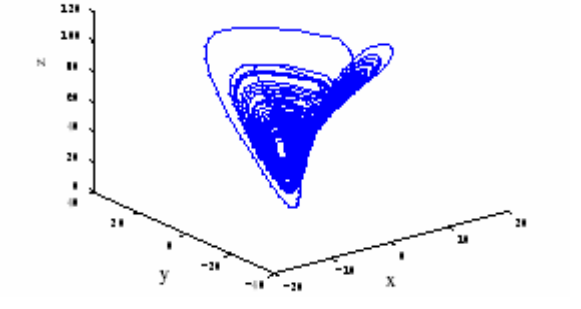


Fig. 1. Liu chaos system three dimensional condition

In the scheme, we can obtain three sequence vales from the Liu chaos system, the encryption formula as follows [17][18]:

$$s_i^{'} = \left(s_i + \left[\frac{key_i}{999} \times 256 \right] \right) MOD256 \tag{7}$$

s_i, $s_i^{'}$ respectively set at around i ,which point scrambling image pixel value, key_i is three figures integer key. Encryption process essential step as follows:

(1) We put the scrambling image according to the grating scanning sequence to one dimensional sequence . $S = \{s(i), i = 1,2,\cdots, N \times N\}$, $N \times N$ is the picture size.

(2) We change length of stride fourth order five level of Runge-Kutta -Felhberg algorithm, solve the chaos system differential equation. Then, we obtain three lengths to be bigger than $\lfloor N \times N/3 \rfloor$ sequence of real numbers $\{x(i), y(i), z(i), i = 1, \ldots \lfloor N \times N/3 \rfloor \ldots\}$.

(3) After takes three real numbers chaos sequences the section length for [N×N/3] subsequence, we take in various sequences each real number sequence value decimal point 5th after three figures, obtains three key sequence which is composed by three figures integers: $\{I_x(i), I_y(i), I_z(i), i = 1, \ldots, \lfloor N \times N/3 \rfloor\}$

Let

$$I_O(i) = \lfloor o(i) \times 10^5 - \lfloor o(i) \times 10^5 \rfloor \rfloor \times 10^3$$
$$o = \{x, y, z\}$$
$$i = 1,2, \ldots, \lfloor N \times N/3 \rfloor$$

(4) we scramble image pixel encryption:

$$s_u^{'} = \left(s_u + \left[\frac{I_x(i)}{999} \times 256 \right] \right) MOD \ \ 256$$
$$s_v^{'} = \left(s_v + \left[\frac{I_y(i)}{999} \times 256 \right] \right) MOD \ \ 256$$
$$s_w^{'} = \left(s_w + \left[\frac{I_z(i)}{999} \times 256 \right] \right) MOD \ \ 256$$

Put

$$u = 3 \times (i-1) + 1$$
$$v = 3 \times (i-1) + 2$$
$$w = 3 \times (i-1) + 3$$
$$i = 1,2 \ldots, \lfloor N \times N/3 \rfloor$$

(5) We put encrypt image pixel sequence to transform as the two dimensional matrix, which obtain the final cipher text image.

4 Results

Uses the Matlab7.0, we will get the results as follows:

(1) We take 256×256 the Lena gradation image to carry on the experiment, which take the Liu chaos system starting value is: $x_0 = [2.2,2.4,38]$, the system time sector takes $[0,360]$; $p = 3, q = 5, n = 15$. Fig. 2.shows the encryption effect.



Fig. 2. Lena primitive image



Fig. 3. Encryption image

(2) We takes 128×128 the man image to carry on the experiment, which take the Liu chaos system starting value $x_0 = [2.0,2.2,36]$. We show the encryption effect by taking $[0,360]$; $p = 2, q = 4, n = 12$ Fig. 3. shows the encryption effect.



Fig. 4. Man primitive image



Fig. 5. Encryption image

(3) Lana statistical characteristic analysis:

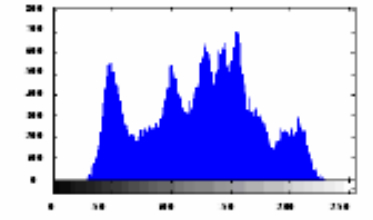


Fig. 6. Lena original image histogram, shows the Lena original image pixel histogram

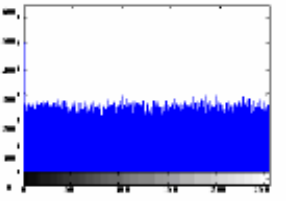


Fig. 7. Encrypting image histogram ,shows Lena encrypting image pixel histogram

Obviously, the encryption process changes the primitive image pixel value non-uniform distribution to the image pixel value even distribution; make the pixel value to be equal in $[0,255]$. The statistical characteristic of plain text is completely broken, causes the relativity between plain text and cipher text greatly reduces.

(4) Man statistical characteristic analysis:

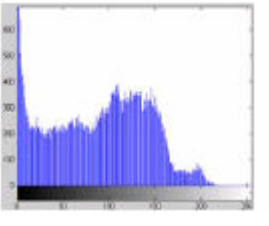


Fig. 8. Man original image histogram, shows man original image pixel histogram

Obviously, the encryption process changes the primitive image pixel value non-uniform distribution to the image pixel value even distribution; makes the pixel value to be equal in $[0,255]$. The statistical characteristic of plain text is completely broken, causes the relativity between plain text and cipher text greatly reduces.

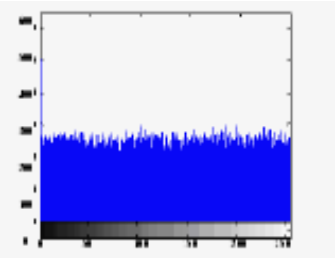


Fig. 9. Encrypting image histogram, shows man encrypting image pixel histogram

5 Conclusion

This article proposed a new algorithm of image pixel scrambling and pixel value encryption, which has the advantages of the chaos system. This method enhances the data encryption security greatly, and has following main merits. We overcome the one dimension chaos shortcoming that the system cannot resist phase space reconstruction attacking. Then we took the three dimensional chaos system real numbers' initial value as the key, it dramatically developed the key space and thus this makes the algorithm have the ability to resist the exhaustive attack.

The experiments and analysis show that this model can solve the some problems. The chaos system has the complex non-linear evolution characteristic and thus it causes that the scrambled text to the key has a high sensitivity; meanwhile, because the key is stochastically produced, the algorithm has one by one system feature. The experiments show that proposed approach is very effective.

However, we have just studied the small part of it. There are many actual problems to need the study and solving. To study better method of image pixel scrambling and encryption are our future work.

References

1. Diffie, W., Hellman, M.: New Directions in Cryptography. IEEE Transactions on Information Theory 22(6), 644–654 (1976)
2. Fridrich, J.: Image Encryption Based on Chaotic Maps. In: Proc. IEEE Conf. on Systems. Man and Cybernetics, Orlando, Florida, pp. 1105–1110. IEEE Computer Society Press, Los Alamitos (1997)
3. Milanovic, V., Mona, E Z: Synchronization of Chaotic Neural Networks for Secure Communications. In: IEEE International Symposium on Circuits and Systems, pp. 28–31. IEEE Computer Society Press, Los Alamitos (1996)
4. Alattar, A.M, Ghassan, I.A.: Improved Selective Encryption Techniques for Secure Transmission of MPEG Video Bit-streams. In: Proc. IEEE ICIP'99, Kobe Japan, vol. 4, pp. 256–260 (1999)
5. Guo, D., Cheng, L.: A New Symmetric Probabilistic Encryption Scheme Based on Chaotic Attractors of Neural Networks. Applied Intelligence, 71–84 (1999)
6. Dang, P., Chen, M.: Image Encryption for Secure Internet Multimedia Application. IEEE Trans Consumer Electronics 46(3), 395–403 (2000)

7. Lee, J., Jae-cheol, R.: Strong User an Authentication in IEEE 802.11 Wireless LAN[C]. In: Chung, C.-W., Kim, C.-k., Kim, W., Ling, T.-W., Song, K.-H. (eds.) HSI 2003. LNCS, vol. 2713, pp. 638–643. Springer, Heidelberg (2003)
8. Beldhouche, F., Qidwai, U.: Binary Image Encoding Using ID Chaotic Maps[C]. In: Proc. Of IEEE Annual Technical Conference, pp. 39–43. IEEE Computer Society Press, Los Alamitos (2003)
9. Chen, G., Mao, Y: Chaos-based Image Encryption, Handbook of Computational Geometry for Pattern Recognition [M]. Springer, Berlin (2003)
10. Katygiannis, T., Owens, L.: Draft Wireless Network Security[R].National Institute of Standards and Technology, Special Publication 800-48, US Department of Commerce (2002)
11. IEEE. Standards for Local and Metropolitan Area Networks: Standard for Port-based Network Access Control[S] IEEE Draft P802.1X/D11 (2001)
12. Yi, D., Zhou, J., Han, X.L.: A Kind New Sets Transforms Randomly and in Pictorial Information Hiding Application [J]. Chinese Science 30(5), 440–447 (2000)
13. Ying, W., Yan, W., Qing, D.: Based on Sets Randomly and Applies with the Fusion Digital Image Hideaway Technology[J]. Chinese image graph journal (A) 5(8), 644–649 (2000)
14. Wen, Z., Li, T., Zhang, Z.: An Image Encryption Technology Based on Chaotic Sequences [J]. Computer Project 31(10), 130–132 (2005)
15. Liu, C., Liu, T., Liu, L.: A New Chaotic Attractor [J]. Solitons and Fractals 22(5), 1031–1038 (2004)
16. Zhu, C., Li, L., Chen, Z.: New Image Encryption Algorithm Based on Combined Multidimensional Chaotic Systems [J]. Computer Project 33(2), 142–144 (2007)
17. Chen, G., Mao, Y., Charles, K.: A Symmetric Image Encryption Scheme Based on 3D Chaotic Cat Maps [J]. Chaos, Solitons and Fractals 3(21), 749–761 (2004)
18. Guan, Z., Huang, F., Guan, W.: Chaos-based Image Encryption Algorithm[J]. Physics Letters A 346(1-3), 153–157 (2005)

Investigation on Fitting Graph Based on Fractal Dimension's Pretreatment

Min Jin, Qin Wang, and Lifeng Xi

School of Computer Science and Information Technology , Zhejiang Wanli University, Ningbo 315100, P.R. China
{jmlch,qinwang,xilf}@zwu.edu.cn

Abstract. In this paper, we investigate an algorithm, which integrates the Box dimension's pretreatment with self-affine fractal interpolation function (AFIF). As a result of our experiments, we conclude that the algorithm allows us to fit graphs of various complexities.

Keyword: Fractal, interpolation, Box dimension, AFIF, fitness.

1 Introduction

Interpolation polynomial is a highly important method for approximation of function, numerical analysis and data fitting. The function is very stationary, so it is suitable for fitting stationary data and smooth curve. Its applications are restricted in limited data fitting. In 1986, Fractal interpolation function was elaborated by Barnsley [1]. A special class of fractals, affine fractal interpolation function(AFIF), can simulate not only the graph of smooth curve data, but also can effectively and accurately fit rough curves and vibrating data, such as mountain range outlines, electrocardiograms ...etc [2]. As for the general theory of fractal interpolation function and affine fractal interpolation function, the reader is referred to [1-6].

Although AFIF can fit rough curves data, AFIF have same fractal dimension or same complex degree at each location. However, in the practical application, the graphic complex degrees and the sensitivities may be absolutely different when data respond to time in the disparate time periods. Thus, it is obvious that the fitting may not be effective if we directly use AFIF, which has the same complexity everywhere.

The chief aim of the present work is to investigate an algorithm, which integrates the Box dimension's pretreatment with self-affine fractal interpolation function, to fit different complexity. In this paper, we present experimentation and analysis, which are based on the algorithm. Clustering analysis is adopted according to the each sub-graph's Box dimension. Then the sub-graphs are reconstructed together according to adjacent Box dimension, and fractal interpolation is adopted separately. Finally, it is readjusted and resumed.

The paper is organized as follows. Section 2 outlines relative definitions of fractal dimension and AFIF. Section 3 discusses experimental method and results based on the algorithm to provide reference.

B. Apolloni et al. (Eds.): KES 2007/ WIRN 2007, Part II, LNAI 4693, pp. 217–224, 2007.
© Springer-Verlag Berlin Heidelberg 2007

2 Fractal Interpolation

2.1 Fractal Dimension

Definition 1. Let E be a compact subset of $\mathbf{R}^2$. Given $\delta > 0$, let $N_\delta(E)$ be the smallest cardinality of family of solid squares with side length δ such that the union of these squares covers *E*.

If the limit $\lim_{\delta \to 0} \dfrac{\log N_\delta(E)}{-\log \delta}$ exists, then

$$\dim_B E = \lim_{\delta \to 0} \frac{\log N_\delta(E)}{-\log \delta}. \tag{1}$$

is called the Box dimension of *E*.

Example 1. Smooth curve
The graph of any smooth function has Box dimension 1. This result means that polynomials and splines are too smooth to approximate the graph with high complexity.

Example 2. C×C

Let $C = \{\sum_{i=1}^{\infty} a_i 3^{-i} : a_i = 0 \text{ or } 2 \text{ for each } i\}$, which is called standard Cantor set.

Then $C \times C$ can be covered by 4^n square of side 3^{-n}, and we can check that

$$N_{3^{-n}}(C \times C) = 4^n.$$

Therefore,

$$\dim_B (C \times C) = \lim_{n \to \infty} \frac{\log N_{3^{-n}}(E)}{-\log 3^{-n}} = \frac{\log 4}{\log 3}.$$

Example 3. Graph of continuous function

For the graph *E* of a continuous function f, we let $\delta = 2^{-n}$, and

$$M_\delta(E) = \sum_{k=1}^{2^n} [2^n \omega(f, [\frac{k-1}{2^n}, \frac{k}{2^n}])],$$

where $\omega(f,[a,b]) = \max_{x_1 \in [a,b]} f(x_1) - \min_{x_1 \in [a,b]} f(x_2)$ is the oscillation of f restricted on the interval [*a*,*b*], and [*x*] denotes the smallest integer greater than or equal to x, for example, [3]= [2.5]=3. Then there is a constant $C > 1$ such that $C^{-1} M_\delta \le N_\delta \le C M_\delta$, and thus

$$\dim_B E = \lim_{\delta \to 0} \frac{\log M_\delta(E)}{-\log \delta}. \tag{2}$$

It is easy to show that for any smooth function, we have $D^{-1}\delta \le M_\delta \le D\delta$ for any δ, where $D>1$ is a constant, therefore the graph has Box dimension 1.

2.2 Affine Fractal Interpolation Function

Given points $\{(x_i, y_i)\}_{i=0}^N$ in the plane, we suppose $\{\omega_1, \omega_2, \ldots, \omega_N\}$ is an iterated function system satisfying

$$\omega_i \begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} a_i & 0 \\ c_i & d_i \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix} + \begin{pmatrix} g_i \\ h_i \end{pmatrix}, \tag{3}$$

$$\omega_i \begin{pmatrix} x_0 \\ y_0 \end{pmatrix} = \begin{pmatrix} x_{i-1} \\ y_{i-1} \end{pmatrix}, \ \omega_i \begin{pmatrix} x_N \\ y_N \end{pmatrix} = \begin{pmatrix} x_i \\ y_i \end{pmatrix}, \tag{4}$$

where $|d_i| < 1$ and $a_i \in (0,1)$ for any i with $1 \le i \le N$. By (3) and (4), we notice that $\{\omega_i\}_{i=1}^N$ is determined by $\{(x_i, y_i)\}_{i=0}^N$ and. We always call $\{d_i\}_{i=1}^N$ vertical factors and $\{(x_i, y_i)\}_{i=0}^N$ interpolation points respectively.

Definition 2. Suppose $f(x)$ is a continuous function on the interval $[x_0, x_N]$. Let $\Gamma = \{(x, f(x)) : x \in [x_0, x_N])\}$ be the graph of $f(x)$. We say that $f(x)$ is an affine fractal interpolation function(AFIF), if

$$\Gamma = \bigcup_{i=1}^N \omega_i(\Gamma). \tag{5}$$

Example 4

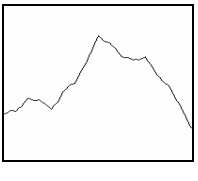


Fig. 1. This graph is an example of AFIF. The position of initial interpolation points are (0,0.3), (0.5,0.8) and (1,0.2). Vertical factors are -0.4 and 0.3.

For AFIF defined by (3)-(5), the dimension $\dim_B \Gamma$ of the graph Γ satisfies the following dimension formula ([1]):

$$\sum_{i=1}^{N} | d_i | \cdot | a_i |^{\dim_B \Gamma - 1} = 1. \tag{6}$$

Remark 1: Formula (6) holds when the interpolation points do not lie in a line simultaneously and $\sum_{i=1}^{N} | d_i | > 1$. Any connected part of the graph Γ of AFIF has the same dimension $\dim_B \Gamma$.

For Example to draw fig. 1, we should use Visual Basic.Net program as follows.

```
Private Sub Button4_Click(ByVal sender As System.Object,
ByVal e As System.EventArgs) Handles Button4.Click
   ax = 0:ay = 0.3:bx = 0.5:by = 0.8:cx = 1:cy = 0.2
   For i = 1 To 5 ' Iterate five times
     j = 0
     For k = 1 To 2 ^ (i + 1) - 1 Step 2
        x = k / 2 ^ (i + 1)
        If 0 < x And x < 0.5 Then
           If i = 1 Then
              y = f1y(bx, by)
           Else
              y = f1y(p1(i - 1, (k + 1) / 2).x, p1(i - 1,
(k + 1) / 2).y)
            End If
          Else
             If i = 1 Then
                y = f2y(bx, by)
              Else
                 j = j + 1
                 y = f2y(p1(i - 1, j).x, p1(i - 1, j).y)
               End If
             End If
             p1(i, (k + 1) / 2).x = x
             p1(i, (k + 1) / 2).y = y
           Next
        Next
End Sub

Function f1y(ByVal x As Double, ByVal y As Double) As Double
   Dim c1 As Double, d1 As Double, h1 As Double
   d1 = -0.4
   c1 = (ay - by - d1 * (ay - cy)) / (ax - cx)
   h1 = ay * (1 - d1) - c1 * ax
   f1y = (ay - by - d1 * (ay - cy)) / (ax - cx) * x + d1
* y + ay * (1 - d1) - c1 * ax
End Function
```

```
Function f2y(ByVal x As Double, ByVal y As Double) As Double
    Dim c2 As Double, d2 As Double, h2 As Double
    d2 = 0.3
    c2 = (by - cy - d2 * (ay - cy)) / (ax - cx)
    h2 = cy * (1 - d2) - c2 * cx
    f2y = (by - cy - d2 * (ay - cy)) / (ax - cx) * x + d2
* y + cy * (1 - d2) - c2 * cx
End Function
```

In this program, the position of point A is (0, 0.3), B is (0.5, 0.8) and C is (1, 0.2). Vertical factor d for function f1y is -0.4, for function f2y is 0.3. Once those factors are confirmed, by formula (3) (4), we get factor c and h. Through using five times iteration, we should have 65 points, which are saved in array p, to draw the curve.

3 Experimental Method and Results on the Algorithm

Step 1: We divide the interval into several subintervals and thus obtain some sub-graph.
Step 2: By using formula (2) to estimate the dimension of each sub-graph.
Step 3: Clustering the sub-graphs according to their dimensions, we reconstruct some new graph $\Gamma_1, \Gamma_2, \cdots \Gamma_k$, each of which is composed of some sub-graphs with adjacent values of Box dimension. In the process of reconstruction, we should make translation for each sub-graph along y-axis to ensure the *connectedness* of graph.
Step 4: For each new graph Γ_i in Step 3, we use AFIF to approximate it.
Step 5: Reconstruct these AFIF graphs to obtain an approximation of the original graph.

According to above method, we initiated experiment to provide reference.

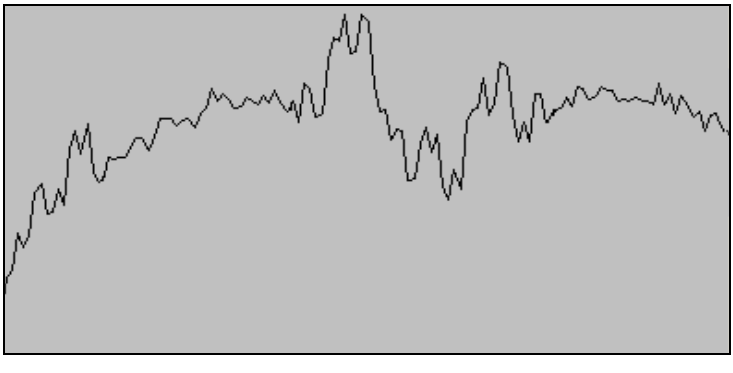


Fig. 2. The graph is a piece of mountain range outlines. We use it as an example for AFIF based on Box dimension's pretreatment.

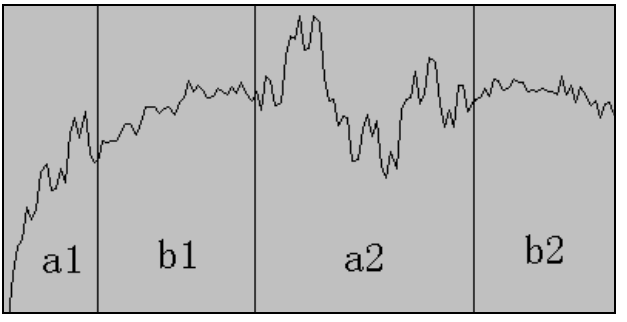


Fig. 3. Here we give a partition of the original graph to obtain 4 parts: a1, b1, a2 and b2

By formula (2), we get $\dim_B$(Part a1) ≈0.37, $\dim_B$ (Part a2)≈0.38 and $\dim_B$ (Part b1)≈0.14 , $\dim_B$ (Part b2)≈0.13.

And thus we reconstruct graphs as follows.

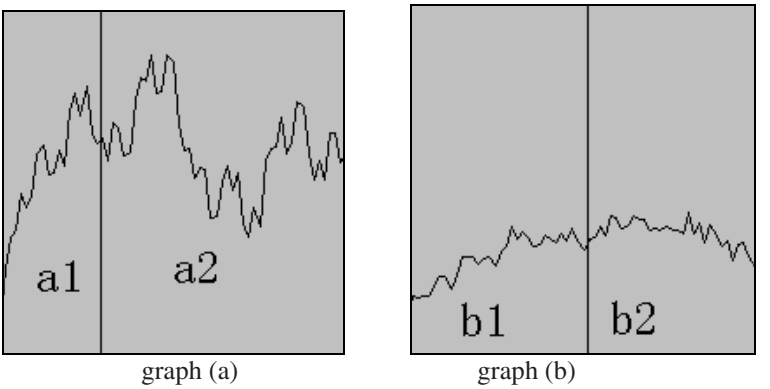


Fig. 4. Reconstruction of graphs according to Box dimension

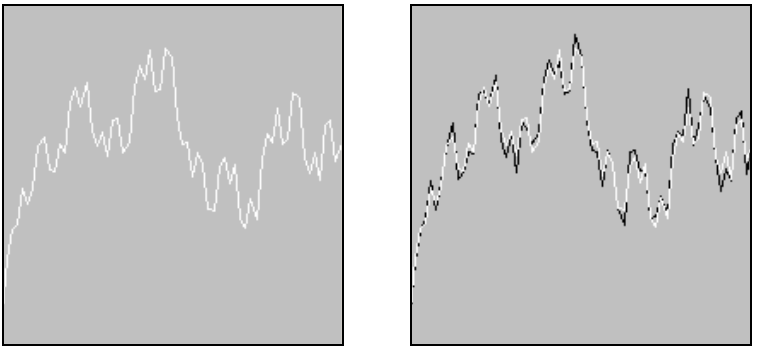


Fig. 5. The left white curve graph is an AFIF (a') relative to graph (a). The position of initial interpolation points are (0,0.1), (0.5,0.85) and (1,0.6). Vertical factors are 0.6 and -0.7. The right white and black curve graph is effect of fitting.

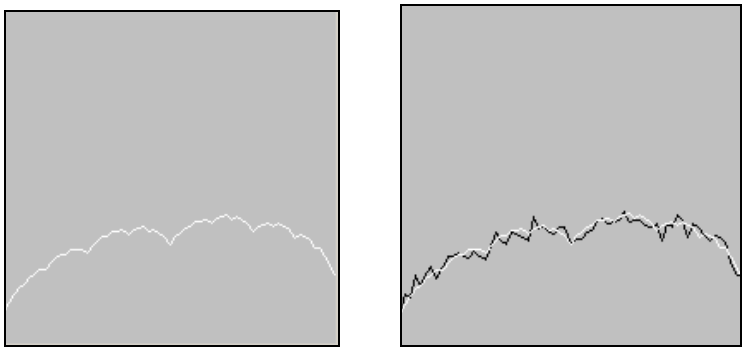


Fig. 6. The left white curve graph is an AFIF (b') relative to graph (b). The position of initial interpolation points are (0,0.1), (0.5,0.3) and (1,0.2). Vertical factors are 0.5 and 0.6. The right white and black curve graph is effect of fitting.

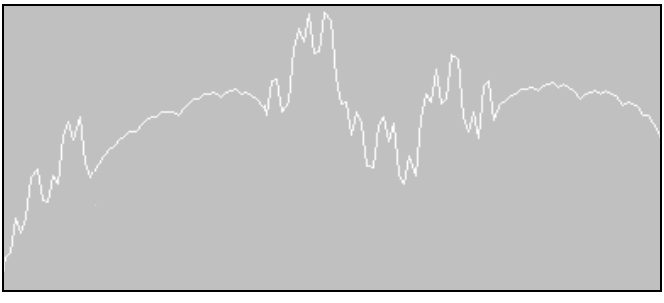


Fig. 7. Reconstruction of AFIF (a') and (b')

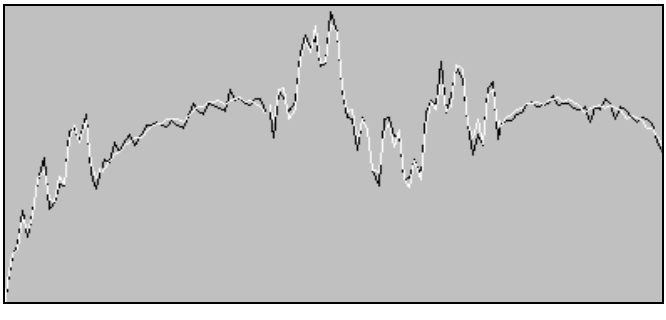


Fig. 8. Original graph (black) and graph of AFIF graphs with reconstruction (white)

As a result of our experiments, we concluded that integrating the Box dimension's pretreatment with AFIF is successful. Through the method, we could break through restrictions on single complexity and supply wild application on fitting.

References

1. Barnsley, M.F.: Fractal functions and interpolation. Constr. Approx. 2, 303–329 (1986)
2. Barnsley, M.F., Elton, J., Hardin, D., Massopust, P.: Hidden variable fractal interpolation functions. SIAM J. Math. Anal. 20, 1218–1242 (1989)
3. Barnsley, M.F., Elton, J.H., Hardin, D.P.: Recurrent iterated function systems. Constr. Approx. 5, 3–31 (1989)
4. Barnsley, M.F., Harrington, A.N.: The calculus of fractal interpolation functions. J. Approx. Theory. 57, 14–34 (1989)
5. Dalla, L., Drakopoulos, V.: On the parameter identification problem in the plane and the polar fractal interpolation functions. J. Approx. Theory. 101, 202–289 (1999)
6. Mazel, D.S., Hayes, M.H.: Using iterated function system to model discrete sequences. IEEE Trans. on Signal 40, 1724–1734 (1992)

Practice of Video Encryption Algorithms Based on Chaotic Sequence and Scan Pattern

Qian Li

School of computer science and information technology, Zhejiang Wanli University
Ningbo, Zhejiang 315100, P.R. China
liqian@shsafe.com

Abstract. In this paper, we propose a new method of digital video encryption based on the SCAN methodology that is a formal language-based two-dimensional spatial accessing methodology, which can generate very large number of scanning paths or space filling curves. The SCAN-based video encryption is first performed by detecting the difference between adjacent frames and compressing these frame differences, and finally the encrypted video is obtained by encrypting the above frame differences. The main characteristics of the proposed methods are image encryption, which can be bounded pixelwise by a user specified value, very large number of encryption keys, and the ability to encrypt large blocks of any digital data. The simulation with standard color test videos (Tennis) indicates that the proposed encryption algorithm can achieve satisfying results. In the end, the characteristics and effects of the algorithm are summarized and the aspects of the subsequent work are prospected.

Keywords: Video encryption; chaotic sequence; scan methodology; difference matrix; difference-compressed matrix.

1 Introduction

In recent years, with the increasingly popularization and application of multimedia data in network and daily life, video stream security is attracting increasing attention , as the specific video stream applications. Among all multimedia data, it goes without saying that video image data is the most important, so two major research orientations for video image processing are established in multimedia computer technology application: first, compressed coding of video image; second, real-time processing of video image and protection of video information.Among all video image compression coding algorithms, MPEG Compression Algorithm [1] and H.263 Compression Algorithm, due to advantages such as high compression ratio and ability to provide high quality image, etc, have been accepted as international standards and well supported in most operation platforms with great compatibility. The protection of video image data, mainly including data encryption technology and information hiding, is used to prevent sensitive information from being easily stolen, interpolated, illegally copied and distributed.

B. Apolloni et al. (Eds.): KES 2007/ WIRN 2007, Part II, LNAI 4693, pp. 225–232, 2007.
© Springer-Verlag Berlin Heidelberg 2007

Due to the wide application of MPEG[1] video standard, most of the encryption algorithms are directed toward encrypting MPEG video. For example, Qiao, et al proposed a video encryption algorithm named VEA[3], which divide the plain text block into two parts. The odd part of the block is encrypted with DES algorithm to provide half of the encrypted text and the other half is the result of bit-by-bit XOR operation between even part and odd part of the plain text .This algorithm reduced encryption complexity to approximately half yet maintained relatively high security. But due to the changed data format by this algorithm, the data can't be easily cropped and pasted. Yen and Guo [6] proposed a way of encrypting motion vector in MPEG video stream with random sequence generated from chaotic mapping, emphasizing rapidity and no change of original data format in this method, yet merely encrypting motion vector can't meet the demand of security. Tang has come up with a way to encrypt by scrambling DCT data [5], with all 64 DCT coefficient completely scrambled, yet by doing this, the compression ratio after entropy coding varies greatly due to violating energy sorting order in Z scanning. Qiao, et. al's analysis[4] shows that by replacing Z-scanning with random scrambling, the compression ratio is greatly reduced and the password system can't resist known plaintext attack. To sum, et. al improved this method[7] by dividing all 64 DCT coefficients into 3 segments according to frequency band so that different segments can be scramble-encrypted separately according to different security and compression ratio requirement, thus achieving higher compression ratio but at the cost of reduced security. In Modified Huffman Tree (MHT) [8] algorithm , different nodes of Huffman tree are assigned different code words, thus the tree is varied, and encryption can be carried out with several trees, but this algorithm requires extra space to store Huffman tree and its key space depends on the structure of the tree, so it can't ensure sufficient security. It's obvious that these algorithms have their shortcomings in aspects such as security and compression ratio.

Based on the above literature [1-10], this article put forward a new video sequence encryption algorithm based on chaotic sequence and SCAN-pattern, which, by using different two-dimensional storage space access methods, generates very large number of scanning paths or space filling curves and defines SCAN-pattern in a normal language. The process of video encryption is first performed by finding the difference matrix between adjacent frames in the video sequences and compressing this matrix, and finally encrypting the compressed matrix with SCAN-pattern and chaotic sequences. Based on chaotic sequence and SCAN-pattern replacement and substitution operation, non-linear coupling structure with good password nature is attained. Base on chaotic sequence and SCAN-pattern replacement and substitution operation, non-linear coupling structure with good password nature is formed. Besides, the key blending operation in each iteration process can greatly enhance the security of encryption algorithm. Simulation with standard color test video (Tennis) indicates that the proposed encryption algorithm can achieve a more reasonable trade-off between security and compression ratio as well as more rapid speed, thus providing good encryption security and efficiency and a better application prospect.

In the end, the characteristics of the algorithm are summarized and the prospects of the subsequent work are proposed.

2 Mechanism of Chaotic Sequence Generation

At present some research has shown that chaos may occur within certain controlled parameter ranges in a nonlinear dynamic system, producing chaotic sequence with natures such as certainty, pseudo-random, aperiodicity, non-astringency and highly sensitive dependency upon initial value [10]. These natures are of great significance to video encryption as the initial value can be set as part of the key and the generated chaotic sequence can become a shift parameter set H after proper treatment.

A one-dimension discrete time nonlinear dynamic system can be defined as follows [10]:

$$x_{k+1} = \tau(x_k) \tag{1}$$

In it $x_k \in V, k = 0,1,2\cdots$, which is called status. $\tau : V \to V$ is a mapping which maps current status x_k to next status x_{k+1}. Apply formula (1) repeatedly from initial value x_0 and we can get sequence $\{x_k \mid k = 0,1,2,\cdots\}$, which is called a path in discrete time dynamic system.

As a kind of very simple yet widely studied dynamic system, Logistic mapping is defined as follows: $x_{k+1} = \mu x_k(1 - x_k)$. The statistical nature show that[6]: The traverse statistic nature of chaotic sequence is the same as white noise, which is required in video encryption. With simple variable replacement, Logistic mapping can be defined within section(-1, 1) in the following form:

$$x_{k+1} = 1 - \lambda \cdot x_k^{\,2} \tag{2}$$

And $\lambda \in [0,2]$, when $\lambda = 2$, it is called full mapping.

3 Video Encryption and Decryption Algorithm

The core principle of the video encryption algorithm proposed in this article is to first find the difference matrix of adjacent frames, compress the matrix and then encrypt the matrix. The algorithm of finding difference matrix between adjacent frames will lead to loss video compression, but the compression ratio between original and decrypted video sequence can be user-customized. The whole process of video encryption is as follows:

Find difference matrix of adjacent frames $G = (G_1, G_2, \cdots, G_N)$.

input: video file $V=(F_1,F_2,\cdots,F_N)$ output: difference matrix $G=(G_1,G_2,\cdots,G_N)$.

Define frame number of the video sequence as N, $G_1=F_1$, $G_1=F_1$, threshold matrix $T=F_1$, with threshold value TH designated by the user. Carry out the following operation to every two adjacent frames: if $|F_k[i,j]-T[i,j]|\leq TH$, then $G_k[i,j]=-1$, or else $G_k[i,j]=0$; and $1\leq i\leq h$, $1\leq j\leq w$, $2\leq k\leq N$; w,h are respectively the width and height of a single frame image.

Compress above-mentioned difference matrix G, and difference-compressed matrix X.

Input: difference matrix G output: difference-compressed matrix X.

Define $X=\Phi$, $Count=0$, break down G into 4×4 windows. Calculate each windows W, if in W there is at least one element not equal to –1, $X=Append(X,EncodeWindow(W))$, $Count=Count+1$; Then assign value $X=Append(B_{Count,m+n-4},X)$. Here expressions m,n are used to respectively indicate the digits of x-coordinate x and y-coordinate y, expression $B_{p,q}$ means describing integer p with q bits of binary number, $C=Append(A,B)$ means the operation of amalgamating matrix A, B to get C.

Function $EncodeWindow(W)$ is calculated as follows: Define $U=\Phi$, (x,y) are the coordinates of left top W, $U=Append(U,B_{x/4,m-2})$, $U=Append(U,B_{y/4,n-2})$, access each element p in W in raster scan sequence. If $p=-1$, $U=Append(U,0)$, or else $U=Append(U,1)$, $U=Append(U,B_{p,k})$.

Encrypt video sequence $V=(F_1,F_2,\cdots,F_N)$ and get encrypted video sequence difference-compressed matrix.

Input: video sequence $V=(F_1,F_2,\cdots,F_N)$; output: encrypted difference-compressed matrix X' and encrypted first frame F'_1.

Encryption keys $K = \{k_1, k_2, n, x_0\}$ are made up of four parts, with k_1, k_2 as scan encryption key, n as the number of encryption iteration, x_0 as initial value of chaotic sequence, 0< x_0 <1, and Logistic Mapping is used in generating code stream. The encryption algorithm of function $J = Encrypt(I, K)$ is as follows: input matrix A, with the user-input initial key K and iteration number n and initial value x_0, with formula(2) and k_1 chaotic sequence $\{x'_k \mid k = 0,1,2,\cdots 3\cdot(w+h)\cdot n-1\}$ is generated. Define matrix A, D, G as a two-dimension matrix with the size $w \times h$; define B, C, E, F as one-dimension vector with the length l= $w \times h$; $A = chaotic_encrypt(I, x'_k)$; $B = scan_encrypt(A, x_{k_1})$; $D = scan(C, s0)$, $E = scan(D, d0)$; $2 \le j \le w \times h$, $2 \le j \le w \times h$ $C[1] = B[1]$, $C[j] = \left(B[j] + ((C[j-1]+1)R[j]) \bmod 256\right) \bmod 256$; $F[1] = E[1]$, $F[j] = \left(E[j] + ((F[j-1]+1)F[j]) \bmod 256\right) \bmod 256$; $G = scan_decrypt(F, x_{k_2})$; $J = G$; carry out multiple times of iteration encryption to the resulting matrix J, with the operation times determined by the user-input key iteration number n, and map this matrix to the range 0-255, the matrix after many times of encryption can be gained as J.

Encrypt initial video sequence V in step (1) and (2) to get difference-compressed matrix, and then encrypt X and F_1 with encryption key K to get $X' = Encrypt(X, K)$, $F_1' = Encrypt(F, K)$. So with this video encryption algorithm the encrypted video difference-compressed matrix X' and encrypted first frame F'_1 can be gained.

Video decryption process is the inversion of encryption process, in which the receiving user must input valid key K (chaotic sequence initial value x_0 and encryption iteration times n, and scan keys k_1, k_2), and use the following formula for decrypting:

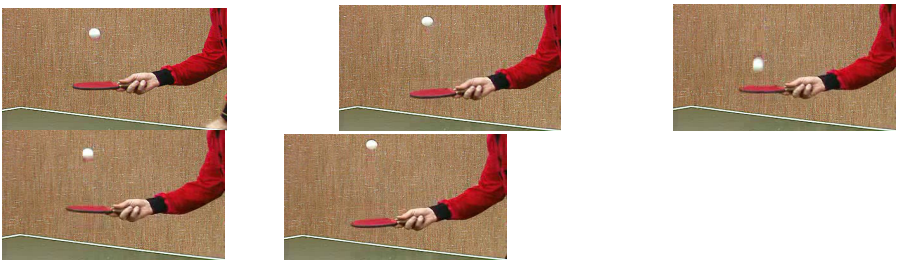
$E[1] = F[1]$, $E[j] = \left(F[j] - ((F[j-1]+1)F[j]) \bmod 256\right) \bmod 256$, $2 \le j \le w \times h$;

$B[1] = C[1]$, $B[j] = \left(C[j] - ((C[j-1]+1)R[j]) \bmod 256\right) \bmod 256$, $2 \le j \le w \times h$;

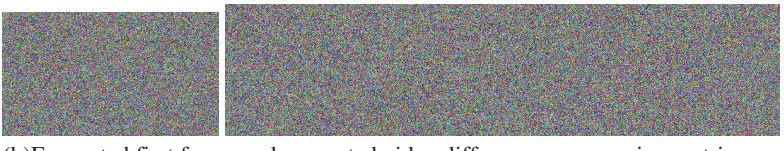
In this way the correct decrypted compressed video difference matrix J'' and the decrypted first frame F_1'' can be educed.

4 Experiment Result and Analysis

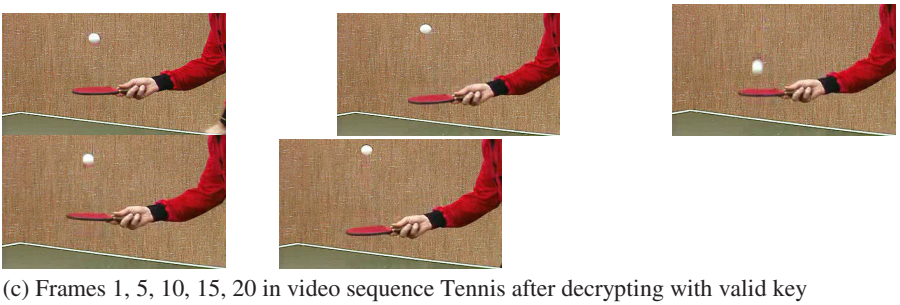
In order to verify the validity of video encryption and decryption algorithm proposed in this book, an algorithm emulation experiment is carried out, using the video encryption algorithm based on chaotic sequence and chaotic sequence on a computer (PIII1.6GHz□512RAM) in VC++, with the following result:



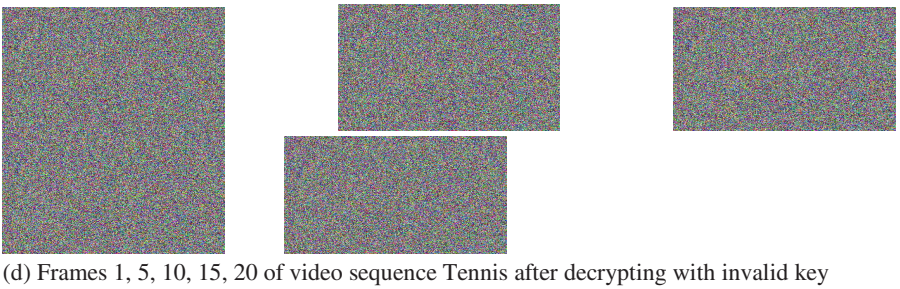
(a) Frames 1, 5, 10, 15, 20 of original Tennis video sequence



(b)Encrypted first frame and encrypted video difference compression matrix



(c) Frames 1, 5, 10, 15, 20 in video sequence Tennis after decrypting with valid key



(d) Frames 1, 5, 10, 15, 20 of video sequence Tennis after decrypting with invalid key

Fig. 1. Results of encrypting and decrypting video sequence Tennis (352×240×20)

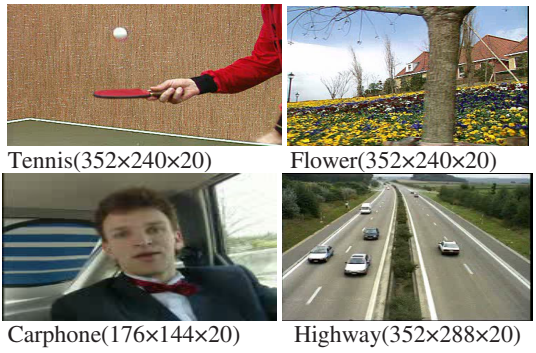


Fig. 2. Four standard colored testing video sequence used in the experiment

Fig. 1 is the testing result of encrypting and decrypting Tennis, a 352*240*20 standard colored video sequence (w=352, h=240, frames=20). The scan encrypting keys are B2(x0 y3 s5 r4), c4, times of encryption iteration n=10, $x_0 = 0.73152836$. In (a) are frames 1, 5, 10, 15, 20 of the original Tennis video in (b) are encrypted first frame and encrypted video difference compression matrix; in (c) are frame 1, 5, 10, 15, 20 in video sequence Tennis after decrypting with valid key and in (d) are Frame 1, 5, 10, 15, 20 of video sequence Tennis after decrypting with invalid scan decrypting key (B2(x0 y3 s5 r3), c4).

Fig. 2 shows the 4standard colored test video sequences, i.e. Tennis, Flower, Carphone and Highway, with 20 continuous frames in each segment.

It can be seen from the simulation that the minimum difference between initial values of scan encryption key will make it impossible to correctly decrypt video sequence, which conforms to the above analysis, thus proving the feasibility of the algorithm proposed in this article.

5 Conclusion and Prospects

This article put forward a new video sequence encryption algorithm based on chaotic sequence and SCAN-pattern, which, by using different two-dimensional storage space access methods, generates very large number of scanning paths or space filling curves and define SCAN-pattern in a normal language. The process of video encryption is first performed by finding the difference matrix between adjacent frames in the video sequences and compressing this matrix, and finally encrypting the compressed matrix with SCAN-pattern and chaotic sequences. Based on chaotic sequence and SCAN-pattern replacement and substitution operation, non-linear coupling structure with good password nature is formed. Base on chaotic sequence and SCAN-pattern replacement and substitution operation, non-linear coupling structure with good password nature is formed. Besides, the key blending operation in each iteration process can greatly enhance the security of encryption algorithm by eliminating insecurity in traditional algorithm which only encrypts using XOR operation to change the value of the pixels. As code streams throughout this algorithm are generated through chaotic mapping and scan-pattern, the output of replacement and

substitution depends heavily on selection of scan key, thus the correct decryption output can only be gained when the user knows the exact key, so this algorithm is of great security.

In theory, the room for key in this algorithm is quite great, or infinite, so one-key-one-time key system requirements can be met. The results of video encryption experiment in Fig. 1 and Fig. 2 have proved the feasibility of the algorithm proposed in this article. Besides, this algorithm is also reliable, attack-resistant and interference-free. The result of simulation test with standard test video sequence including Tennis, Flower, Carphone and Highway show that this algorithm has good encryption security and efficiency. In future studies, the virtues of SCAN-pattern and chaotic sequence in encryption can be further studied and research can also be carried out on its use in secure and effective encryption of video data in combination with other applied cryptographic technology[2] and MPEG video compression technology[5, 6, 7].

References

1. Committee Draft Standard ISO 11172: ISO/MPEG90/176. Coding of moving Fig.s and associated audio
2. ITU-T Rec. H. 263, Video coding for narrow telecommunication channels at< 64kbit/s
3. Qiao, L., Nahrstedt, K.: A new algorithm for MPEG video encryption. In: CISST'97, 1997th edn. Proceeding of the First International Conference on Imaging Science, Systems and technology, Las Vegas, Nevada, pp. 21–29 (July 1997)
4. Qiao, L, Nahrstedt, K, Tam, I.: Is MPEG encryption by using random list instead of zigzag order secure. In: IEEE International Symposium on Consumer Electronics, Singapore, pp. 226–229. IEEE Computer Society Press, Los Alamitos (December 1997)
5. Tang, L.: Methods for encrypting and decrypting MPEG video data efficiently. In: ACM Multimedia'96. Proceedings of the Fourth ACM International Multimedia Conference, Boston, MA, pp. 192–230. ACM Press, New York (1996)
6. Yen, J.C., Guo, Y.J.: A new MPEG encryption system and its VLSI architecture. In: IEEE Work shop on Signal Processing Systems, Taipei, pp. 430–437. IEEE Computer Society Press, Los Alamitos (1999)
7. Ali, S.T., Feng, W.-C.: Efficient multi-layer coding and encryption of MPEG video streams. In: IEEE International Conference on Multimedia and Expo, New York, 30, July-2, Aug, vol. 1, pp. 119–122. IEEE Computer Society Press, Los Alamitos (2000)
8. Wu Chun-Ping, C., Jay Kuo, C.: Efficient multimedia encryption via entropy code design. In: Proceedings of SPIE International Symposium on Electronic Imaging 2001, SanJose, CA, USA, vol. 4314, pp. 128–138 (2001)
9. Maniccama, S.S., Bourbakisa, N.G.: Image and video encryption using SCAN patterns. Pattern Recognition 37, 725–737 (2004)
10. Chang, L.: Large encrypting of binary images with higher security. Pattern Recognition Lett. 19(5), 461–468 (1998)

Research on WebGIS Based on Multi-agent

Taowei Wang, Liangbin Zhang, and Aimin Yang

School of Computer Science and Information Technology, Zhejiang Wanli University
Ningbo, Zhejiang 315100, P.R. China
{nbwtw,zlb,yyh}@zwu.edu.cn

Abstract. Web-based Geographic Information System(WebGIS) is a key direction towards GIS development. By analysing of the deficiency in the design of conventional WebGIS, this paper proposes an architecture of the intelligent WebGIS based on Multi-Agent, which has been applied in a agricultural WebGIS. In order to solve the key problem of Agent communication, the paper also presents an Agent communication model, which decreases the working capacity to maintain the communication protocl and enhances the system's adaptability and stability.

Keywords: WebGIS, Agent, Multi-Agent system, communication protocol.

1 Introduction

GIS has been an essential tool for enterprise, organizations and governments to do scientific management and make important decisions.With the rapid development of network technologies and the popularity of World Wide Web, GIS has been developed from original isolated and centralized information management model to an open and distributed architecture. However, the inherent limitation of current Internet GIS architectures based on connection makes them incompetent to fulfill various requirements applications. Under this circumstances, users heavily depend on those applications and data and it has been very difficult for users to effectively use the Internet's resources. So in WebGIS research field many studies have been done by some internal and external researchers in recent year. WebGIS based on Agent technology is a new research direction in GIS.

The WebGIS architecture enables effective system development that would enhance legacy investments. A spatial object "Feature" encapsulates both geometries and attributes, which can be designed freely and retrieved from feature storage by employing customized formats for specific applications [1]. Based on the features of Agent, many GIS or WebGIS, Location Based Service and specially in mobile GIS system in which used the multi-Agent system architecture was introduced in [2][3].

The research on the multi-Agent system cooperative problem solving is an important problem. In paper [4], Barber thought that problem solving consists of five relative independence compositions: organization formation, planning formation, task decomposition, planning centralization and execution. In paper [5], the complete

B. Apolloni et al. (Eds.): KES 2007/ WIRN 2007, Part II, LNAI 4693, pp. 233–239, 2007.
© Springer-Verlag Berlin Heidelberg 2007

cooperative problem solving process was represented, which contains four parts: recognition, team formation, plan formation and execution. paper [6] represented two new operators VIS and SEE which can relate the environment and internal mental state, analyzed the environment and internal mental state separately, and discussed the relationships which exist between them.

The structure of this paper is as follows. In Section 2, We present the conceptions about Agent and Multi-Agent system. In Section 3, an architecture of WebGIS based on multi-Agent and an Agent communication model are proposed. Finally, Section 4 briefly concludes this paper and outline the future work.

2 Agent Technology and Multi-Agent System

2.1 Agent Conception

There is no universally accepted definition of the term Agent.However, in this paper, Agent is defined as a computer system that is situated in some environment, and that is capable of autonomous action in this particular environment in order to meet its designed objectives[7].

In the course of the research on Agent, the different researchers endowed Agent with different construction, context and capability in their own system.Here we consider that the Agent has three essential characteristics as follows:

Autonomy: Agent controls its action and internal state by itself when they are in motion on some degree and it does not need man or other mechanics' direct control.

Reactivity: Agent could comprehend the circumastance and draw a real time reaction when there is some change in the circumstance.

Interaction: Agent is able to interact with other Agent in order to accomplish and solve a task through a special protocol.

2.2 Multi-Agent System

Agent is defined as software or an entity, which possesses a certain special function, but a single Agent could not solve the complicated and distributed problem in reality. Therefore, one applied system usually is composed of many Agent, these Agent not only have their own's action goal and problem solving ability, but also have the cooperation ability for common goal through a certain organization strategy and protocol. In other words, Multi-Agent system (MAS) consists of many self-determinative or semi-determinative agents, each Agent performs its own function, and can communicate with other Agents to accomplish the solution of whole task through cooperation. Contrast with single Agent, Agent in MAS has some other characteristics as follows:

Sociality: Agent is in a social environment which consists of many Agents. It can cooperate, negotiate or compete with other Agents by using Agent language.

Self-control: In MAS, when an Agent sends out a request, other Agents can accept it under the condition of having the ability and interest of providing service. That is, an Agent can't compel another Agent to provide service.

Cooperativity: In MAS, every Agent which has its own goal must cooperate and negotiate with each other to complete a task.

3 Application of MAS in WebGIS

3.1 The Architecture of WebGIS Based on Multi-Agent

Under the drive of network globalization and information digitalization, geographic information system based on Internet as a promising research field brings new opportunities to the ways of geographic information's accessing, sharing and disseminating. However, Conventional solutions for building WebGIS application are based on either CGI, COM or ActiveX techniques, which becomes more incompetent to fulfill the various requirements of the coming GIS applications under dynamic, complicated and distributed computing environment. Seeking for new distributed computing modes of networked GIS is an important research area for sharing effectively and real timely the global distributed geographic information. The technology of multi-Agent is an important way for developing distributed intelligent system, which provides a new method to study and develop WebGIS.

In this paper, an architecture model of WebGIS based on multi-Agent is proposed as shown in figure 1, which has been applied in a digital agricultural spatial information management platform in a province of China.

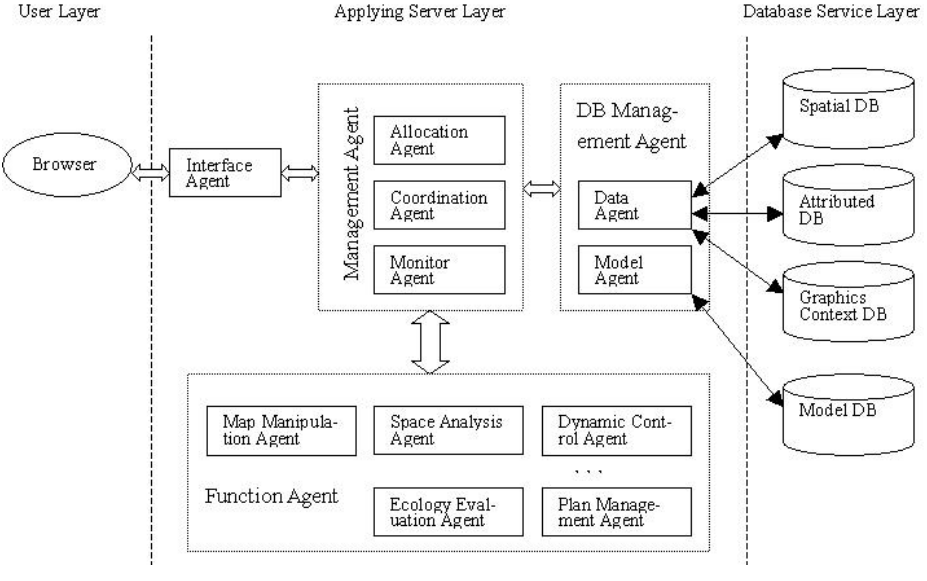


Fig. 1. The architecture of agriculture WebGIS based on Multi-Agent

Fig. 1. shows that the intelligent architecture of agriculture WebGIS which is based on the frame work of agricultural space information system, which is guided through Agent-oriented approach in design and implementation for the system instead of the traditional structured design approach and the object-oriented approach. The system model has three layer structures which are user layer, application server layer and database service layer. These layers can be explained as follows:

a) User Layer: This layer is a man-machine interface of the whole WebGIS. On one hand, Clients use the system's function by using Web browser as the tool, such as searching for some information, customizing the personalization Web side etc. On the other hand WebGIS provides the friendly and simple operating interface to users.

b) Application Server Layer: In this layer, WebGIS fulfills the users' various application and service requests, which is made up of the interface Agent, Management Agent, function Agent and database Agent. Management Agent is the core part of the system, which including the Allocation Agent, Coordination Agent and Monitor Agent. Function Agent realizes some function of the digital agricultural spatial information management system, which includes the map manipulation Agent, space analysis Agent, dynamic control Agent, plan management Agent, ecology evaluation Agent and so on. Database Agent operates the different database which including data Agent and model Agent.

c) Database service Layer: Agricultral WebGIS needs a kind of data which mainly includes the spatial database, attributed database, graphics context database and model database.

3.2 Functions of Each Agent

Interface Agent. The interface Agent undertakes the interface management between people and computers, thus, it acts as a bridge between the users and the Agents' functions. On one hand it submits the request which comes from client to the system, and on the other hand, it feeds back the response results to users. In addition, the Agent can generate different Web sites according to different user's permission.

Management Agent. Management Agent can be divided into three Agents, which are the Allocation Agent, the Coordination Agent and the Monitor agent. These Agents main tasks are shown as follows:

a) These Agents control the system flow and dispatch some or a single Agent to fulfill some tasks.

b) They manage the other Agents in the whole system including Agent's mark , type, function and so on.

c) They organize and coordinate the other Agents and resolve the conflict among the Agents.

d) They accept the task coming from the interface Agent and allocate the work to some Agents. Moreover, they deliver the complete results to interface Agent.

Function Agent. Function Agent includes some Agents which realize the functions of digital agricultural spatial information system management platform such as map manipulation, spatial analysis, dynamic control, plan management, ecology evaluation and decision support etc.

Data Agent. In WebGIS, it is one of the most important functions to manage data. Here, we use the special Agent to manage many kinds of databases such as spatial database, attributed database, graphics context database. On one hand it manages the requests from the management Agent and gain needed data, on the other hand it is responsible for storing the data which is offered from the management Agent. Besides, the Agent supervises and records the system's category information.

Model Agent. The model Agent responses the modelling requests from the management Agent, then searches for the needed agricultural model which mainly includes the disaster prediction model, yield prediction model, economic spatial analysis model, resource evaluation model and so on.

3.3 An Agent Communication Model Based on Dynamic Modification Protocol

Communication language and protocol are the basis to exchange information and knowledge effectively among Agents. KQML [8] is a widely used communication language, which defines the representation style of message and mode of work and supports the knowledge sharing among run-time agents. A KQML-based communication model is proposed where Agent can add new protocol or modify protocol without any modification of the protocol software's source code. The working capacity to maintain the communication protocol can be decreased, thus the adaptablility and stability are enhanced. This communication model has been applied in design of protocol interface for digital agricultural WebGIS.

A communication protocol is considered that it is consists of some CFSM (Communication Finite State Machines) [9] and messages which are exchanged by FIFO(first-in-first-out)queue. Therefore, it can be looked upon as a series of received and dispatched message sequences. Essentially, Agent communication is a state transfer mode where messages are mapped into Agent's state transfer relationship.

Assuming an expression of a state transfer process is defined as follows:

$$T_k(In_k, Proc_k, Re_k)$$

Where In_k is a message, which comes from the other Agent. $Proc_k$ is a process according to the In_k. Re_k is a message responsing to the $Proc_k$.

Agent communication process is based on a standard communication protocol, could be regarded as Agent's state transfer process. Therefore, a standard communication protocol can be mapped into state transfer rule database which is made up of the state transfer condition table and state transfer process table. In real application, we can modify the state transfer condition table and state transfer process table when an Agent communication protocol needs to be added. The inner communication protocol interface of Agent is shown in figure 2 which is composed of five elements, which are controller, evaluation machine, state transfer condition

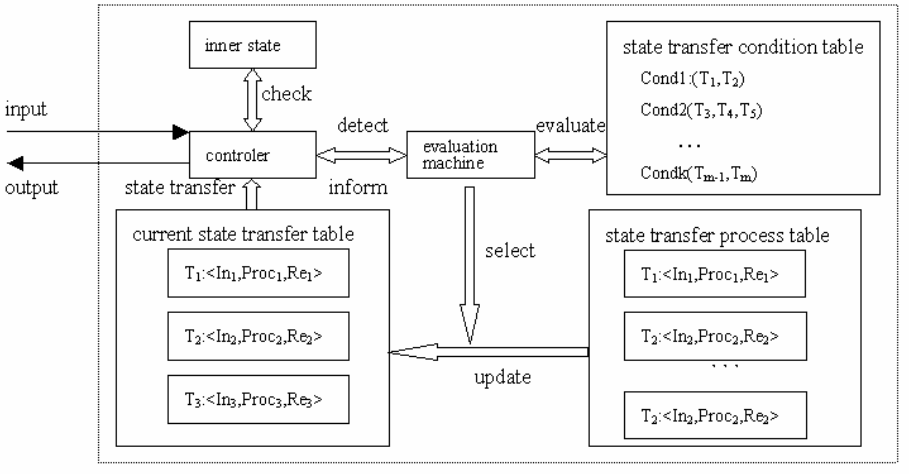


Fig. 2. The inner protocol interface of Agent

table, state transfer process table and current state transfer process table. Their working processes include three steps shown as follows:

Step 1: Controller checks the current state then informs the evaluation machine.

Step 2: Evaluation machine selects a process of the state transfer according to the current state and state transfer condition.

Step 3: Controller triggers off a process of the state transfer according to imported messages and the current state transfer process.

4 Conclusions

Combining multi-Agent technology and WebGIS, this paper proposes an architecture of the intelligent WebGIS based on Multi-Agent. Moreover, discussing the cooperative problem solving in MAS, the paper also gives a communication model where a new protocol is added or modified without changing the protocol software's source code and these models have been are successfully used in a real agricultural WebGIS. Recently, Mobile-Agent is very hot issue in Agent technological field, which has advantages in dealing with and resolving distributed problems in GIS. Many active researchers have done some studies about Mobile-Agent, the next step of the work is to study the Mobile-Agent-based WebGIS.

References

[1] Takino, S., Okamoto, S., Wan, B Y.: The development and application of a new type webGIS engine based on feature concept in Japan. In: Asia GIS2003 Conference, Wuhan China (2003)

[2] Guo, G.Y., Wang, S.K.: A multi-Agent system architecture for geographic information gathering. Journal of Zhejiang University Science, 1367–1373 (2004)
[3] Luo, Y.M., Wang, X.L.: Research on the Geo-Agent. Journal of Computer Research & Development 32, 1504–1511 (2000)
[4] Barber, K.S., Liu, T.H., Han, D.C.: Agent oriented design. TR98-UT LIPS-AGENTS-01 Technical Report of The University of Texas at Austin (1999)
[5] Wooldridge, M., Jennings, N.R.: The cooperative problem-solving process. Journal of Logic Computation 9, 563–592 (1999)
[6] Zhang, S.M., Shi, C.Y.: Modeling Agents combined with the state of environment. Journal of Computer Research and Development 39, 1587–1591 (2002)
[7] Wooldridge, M., Jennings, N.R.: Intelligent Agents: theory and practice. The Knowledge Engineering Review 10, 115–152 (1995)
[8] Finin, T., Mckay, D., Fritzon, R.: An overview of KQML: A knowledge query and manipulation language. Department of Compter Science, Stanford University (1992)
[9] Brand, D., Zafiropulo, P.: On communication finite state machines. Journal of the ACM 30, 323–342 (1983)

Transient Air-Fuel Ratio Estimation in Spark Ignition Engine Using Recurrent Neural Networks

Yanhong Zhang[1], Lifeng Xi[1], and James Liu[2]

[1] School of Computer Science and Information Technology, Zhejiang Wanli University, Ningbo, Zhejiang 315100, P.R. China
[2] BorgWarner Automotive Components (Ningbo) Co.,Ltd, Ningbo, Zhejiang 315104, P.R. China
ailleen1212@126.com, jingliu@borgwarner.com

Abstract. Neural networks is very useful in modeling processes for which mathematical modeling is difficult or impossible. In the present work recurrent neural network (RNN) is used for air-fuel ratio (AFR) estimation in Spark Ignition (SI) Engine. AFR estimation is difficult due to the nonlinearity and dynamic behavior in SI engines. Additionally, delays in engine dynamics limit the performance of engine controller. Estimating AFR a few steps in advance can help engine controller to take care of these. RNN is trained using data from engine simulations in MATLAB/SIMULINK environment. Uncorrelated signals were generated for training and validation. It has been shown that recurrent neural network can predict engine simulations with reasonably good accuracy.

Keywords: Air-fuel ratio, air-fuel ratio estimation, recurrent neural network.

1 Introduction

Air-fuel ratio (AFR) is a key index affecting power performance, fuel economy and exhaust emissions of the spark ignition (SI) engine. The control of AFR in SI engines becomes a key research activity for automobile manufacturers world-wide. There has been a continuing effort to improve the efficiency and exhaust emission of the automobile engine for the last three decades [1],[2]. Due to the strict emission standards spark ignition engines are equipped with three-way catalytic converter, which achieves highest conversion efficiency at stoichiometric ratio of air and fuel; as such AFR should be controlled close to it. A modern engine control modules (ECM) handles this task very well under steady-state conditions by estimating the AFR using an oxygen sensor in the exhaust gas stream. It becomes more difficult to control AFR to its stoichiometric value during transient conditions due to nonlinear and dynamic behaviour of SI Engines [3]. The inherent time delay in the feedback system caused by the oxygen sensor imposes a bandwidth limitation on the closed-loop system. This bandwidth limitation will result in AFR oscillation during fast transients, which in turn will result in increased pollutant emissions and fuel consumption [4].

B. Apolloni et al. (Eds.): KES 2007/ WIRN 2007, Part II, LNAI 4693, pp. 240–246, 2007.
© Springer-Verlag Berlin Heidelberg 2007

As the electronics transitioned to increasingly capable microprocessors, it became practical to entertain the use of complex control methods. The algorithms based on modern control were proposed for automotive engine control. Particularly, with the development of artificial intelligent theory, algorithms based on that have been proposed in [6], [12],[13].These control algorithms also provide a promising method for engine AFR control [5],[6].

In the present work an AFR estimation methodology using RNN is developed. It is able to predict AFR a few steps in advance with reasonable accuracy. This estimation model can be used for indirect control of AFR. Another application of AFR estimation is that it can also act as a virtual sensor. The results using different prediction horizon are evaluated and these predictions were found to be quite accurate as errors are within +/- 0.02 for approximately 95% of the transient.

This paper is organized as follows. Section 2 outlines the principles of recurrent neural network. Section 3 introduces AFR estimation using neural network. Simulation results are given and discussed in section 4. We briefly conclude the work in Section 5.

2 Neural Networks

Neural networks is a relatively new artificial intelligence technique that emulates the behavior of biological neural systems in digital software or hardware. These networks can automatically "learn" complex relationships among data. The neural networks is most suitable for dynamic models in the representation of nonlinear systems/plants. They are capable of identifying the interaction between the inputs and outputs [7],[8],[9].

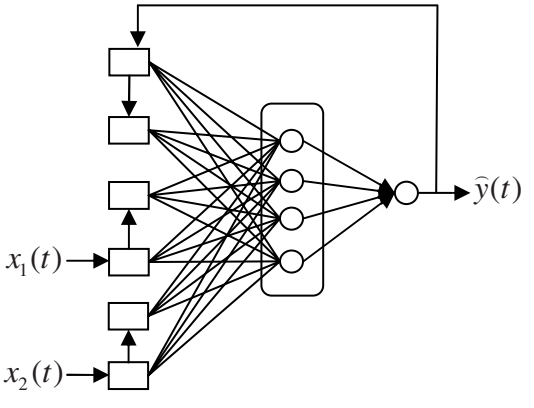


Fig. 1. Recurrent Neural Network

A RNN shown in Fig.1 distinguishes itself from a multilayer feed forward neural network as it has feedback loop. The feed forward connections and feedback loop involve the use of unit-delay elements, which results into a nonlinear dynamic behaviour. The neural network output can be expressed by Eq(1).

$$\hat{y}(t \mid w) = NN[\varphi(t, w), w] \tag{1}$$

Where, w is the weight parameter vector to be identified during the training process and $\varphi(t)$ represents the network input vector given in Eq(2)

$$\begin{aligned} \varphi(t, w) = [\, & \hat{y}(t-1 \mid w), .., \hat{y}(t-p \mid w), x_1(t-q), .., x_1(t-q-m_1+1), ..., \\ & x_j(t-q), .., x_j(t-q-m_j+1)] \end{aligned} \tag{2}$$

Where, j is number of inputs, p and m_j are the number of past outputs as feedback variables and past j^{th} input as feed forward variables, respectively. q is the system inherent time delay. The error signal at the output node is given by Eq(3).

$$\varepsilon(t \mid w) = y(t) - \hat{y}(t \mid w) \tag{3}$$

The cost function given by Eq(4) is the sum of mean-square error (normalized sum of square error [NSSE]) and a regularization term which is most often called weight decay term.

$$\xi(w) = \frac{1}{2} \sum \varepsilon^2(t / w) / N_P + \frac{1}{2} w^T \beta w \tag{4}$$

Where, β is the weight decay. The parameter vector w is optimized by minimizing the cost function $\xi(w)$.

3 AFR Estimation

The accuracy of AFR estimation and control is affected by the Injection and Transport delays. The time between the input signal received by ECU and formation of air-fuel mixture is known as injection delay. T exhaust system is equipped with a UEGO sensor which is utilized to quantitatively sense the AFR and to control the fuel injection [10], [11].

The relative AFR, λ, given by Eq(5)

$$\lambda = \lambda_a / \lambda_d \tag{5}$$

Where λ_a is the actual AFR, λ_d is the desired AFR; λ is the quantity to be controlled and the objective is to maintain it at a target value. The measurement device is a "wide range" UEGO sensor which has a steady state output signal proportional to the actual AFR.

The existence of a significant system time delay, or transport delay between the formation of the AFR mixture and the point of measurement imposes limits on the

performance that can be achieved by using feedback control only. This leads to a general approach for the control of the AFR that uses a scheme with both feedforward and feedback. A feedforward setting for the fuel injection based on the current operating point, and a control signal from an error driven feedback controller, where the error is the difference between the set point λ_d and the measured signal λ_a [12], [13].

In our work the AFR estimation using neural network is carried out as follows:

3.1 Selection of Input Variables

The relevant inputs are decided using the preliminary analysis of nonlinear engine model. AFR obviously depend upon the air mass flow rate and fuel rate. It is a known fact that the air mass flow has strong dependence on load and engine speed; hence they are taken as inputs to neural networks. In present work load is represented by throttle position. The actual fuel flow rate mixing with the air stream is dependent on the fuel dynamics and is controlled by amount of fuel injected, hence it is chosen as another input. EGR flow and temperature influence both steady state and transient responses. Hence they are also chosen as inputs.

3.2 Training

The two important issues in training is selection of stopping criterion and selection of initial weight decay parameter. In the present work, data are generated using a deterministic engine model; early stopping criterion [8] is not used as it can cause over-fitting of the data. The stopping criterion used here is as suggested by [9]:

(1) Maximum number of iterations has exceeded 500, **or**
(2) Cost function ξ is less than 10^{-6}, **or**
(3) Biggest weight change, i.e., $\max\left|\left(w_k^{i+1} - w_k^i\right)\right|$ is less than 10^{-3} **and** change in cost function is less than 10^{-7}.

The weight decay has been found out using the trial and error method. It was found when $\beta = 0.01$, it not only gives the minimum training and generalization error, but also the minimum difference between the two.

4 Validation Results and Selection of Network Structure

Validation of RNN model is carried out using following two techniques:

(1) Estimation of the generalization mean square error.
(2) Visualization of its ability to predict.

In the first technique, the mean-square error given by Eq(6) is evaluated for the generalization data.

$$\xi(w) = \frac{1}{2}\sum \varepsilon^2(t|w) / N_P \tag{6}$$

The estimation of generalization NSSE is useful for validation, but its primary function is for network structure selection, which refers to selection of optimum number of hidden nodes. Neural networks with different topologies are trained. The number of past outputs as feedback variable is kept constant as 3. The number of hidden neurons is varied from 8 to 20.

2-step ahead predictions give error within +/- 0.02 for 95.83% of the transient.
3-step ahead predictions give error within +/- 0.02 for 94.76% of the transient.
4-step ahead predictions give error within +/- 0.02 for 95.17% of the transient.

Further, tests are conducted to check whether the trained RNN has learnt the different dynamics involved in engine model. This is done by first simulating for a constant throttle angle, speed, EGR flow rate and temperature and a square wave signal for fuel injection rate shown in Fig. 2. Fig. 2 shows quite satisfactory results as RNN simulation is closely following engine simulation results with good accuracy. Similarly, simulations are done for a constant fuel injection, flow rate and temperature of EGR and a square wave signal for throttle angle as shown in Fig. 3. Fig. 4 shows that RNN is able to follow engine simulation with good precision too. Lastly, we change EGR flow rate and keep other parameters constant. This allows exciting manifold filling and crankshaft dynamics due to change in EGR flow rate. Again it can be seen that RNN is able to predict engine simulation results with acceptable accuracy.

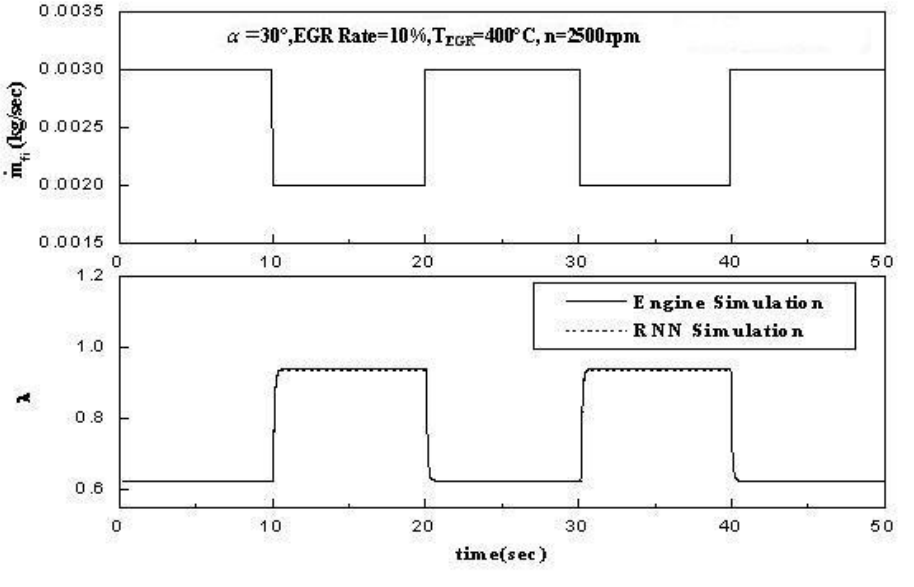


Fig. 2. Engine and RNN simulation for pulse input of m_{fi}

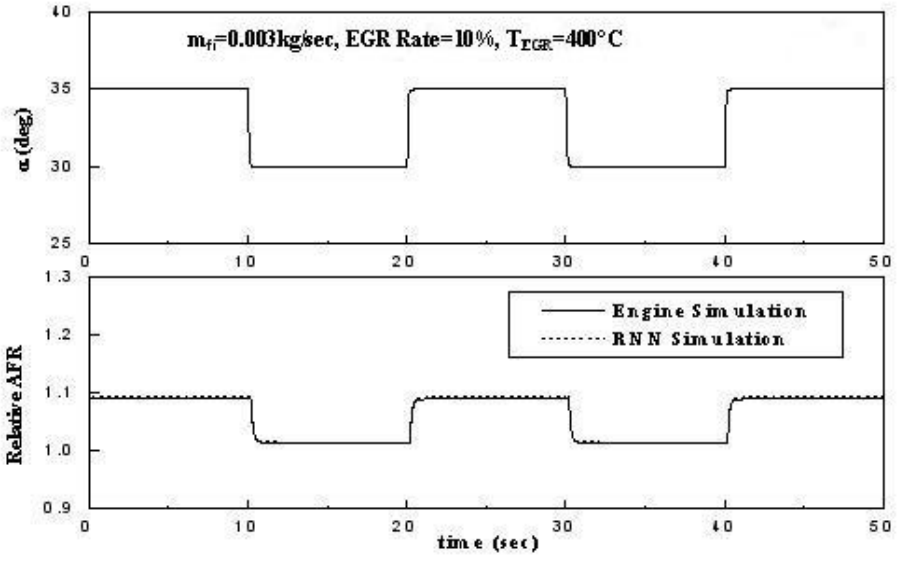


Fig. 3. Engine and RNN simulation for pulse input of α

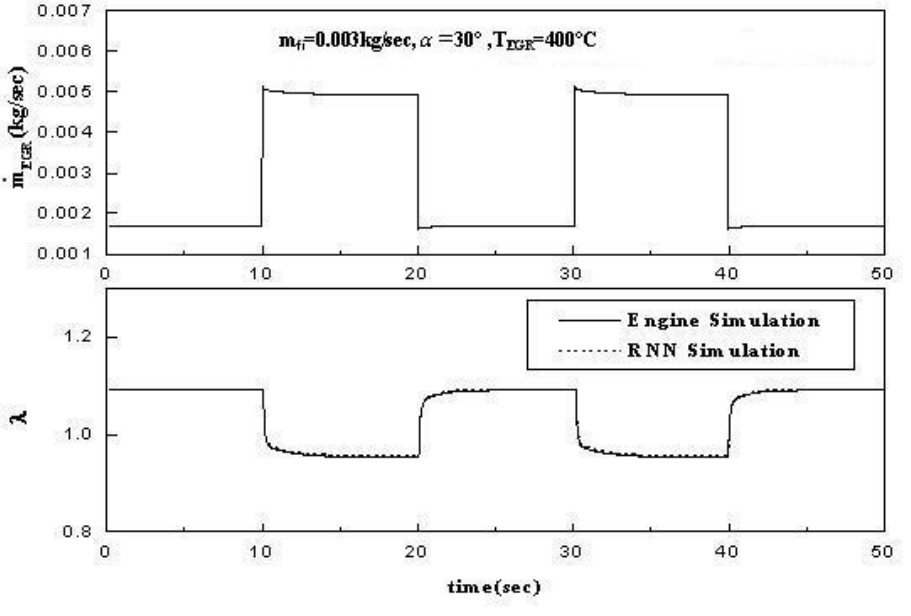


Fig. 4. Engine and RNN simulation for pulse input of m_{EGR}

5 Conclusion

In this paper, RNN estimation model has been presented. It successfully estimates the nonlinear dynamic behavior of a spark ignition engine. These predictions were found to be quite accurate as estimation errors are within +/- 0.02 for approximately 95% of the transient. The results on the multiple sequential injection engine illustrate the excellent AFR estimation which may be difficult to achieve in transient conditions. They also demonstrate that the RNN can provide a promising strategy for engine AFR prediction and control.

References

1. Alippi, C., Russis, C., Piuri, V.: A Neural-Network Based Control Solution to Air-Fuel Ratio Control for Automotive Fuel-Injection Systems. IEEE Transactions on Systems, Man, and Cybernetics-Part C: Applications and Reviews 2, 259–268 (2003)
2. Arsie, I., Pianese, C., Sorrentino, M.A.: Procedure to Enhance Identification of Recurrent Neural Networks for Simulating Air-Fuel Ratio Dynamics in SI Engines. Engineering Applications of Artificial Intelligence 19, 65–77 (2006)
3. Hendricks, E., Sorenson, S.C.: Mean Value Modelling of Spark Ignition Engines. SAE Technical Paper No. 900616 (1990)
4. Hendricks, E., Jensen, M., Kaidantzis, P., Rasmussen, P., Vesterholm, T.: Transient A/F Ratio Errors in Conventional SI Engine Controllers. SAE Technical Paper No. 930856 (1993)
5. Guo-xiu, L., xin, Z.: Transient Air-Fuel Ratio Control in a CNG Engine Using Fuzzy Neural Networks. Journal of Beijing Institute of Technology 1, 100–103 (2005)
6. Copp, D.G, Burnham, K.J., Lockett, F.P.: Model Comparison for Feedforward Air/Fuel Ratio Control. In: IEEE UKACC International Conference on Control'98 (1998)
7. Narendra, S.K., Parthasarathy, K.: Identification and Control of Dynamical Systems Using Neural Networks. IEEE Transactions on Neural Networks 1, 4–27 (1990)
8. Haykins, S.: Neural Networks: A Comprehensive Foundation. Pearson Education, Inc. (2002)
9. Nørgaard, M., Ravn, O., Poulsen, N.L., Hansen, L.K.: Neural Networks for Modelling and Control of Dynamic Systems. Springer, Heidelberg (2000)
10. Brian, A.A., Jones, V.K., Powell, J.D., et al.: Adaptive Air-Fuel Ratio Control of A Spark—Ignition Engine. SAE 940373 (1994)
11. Masahiro, N., Akira, O., Ahinichi, A.: Model-Based Fuel InJection Control System for SI Engines. SAE 961188 (1996)
12. Cesre, A., Cosimo, D.R, Vincenzo, P.A.: Fine Control of The Air-To-Fuel Ratio with Recurrent Neural Networks. In: IEEE Instrumentation and Measurement Technology Conferenee, St.PauI, Minnesota, USA, IEEE Computer Society Press, Los Alamitos (1988)
13. Frith, A.M, Gent, C.R, Beaumont, A.J.: Adaptive Control Of Gasoline Engine Air-Fuel Ratio Using Artificial Neural Networks. In: IEEE Artificial Neural Networks Conference, IEEE Computer Society Press, Los Alamitos (1995)

A Pattern Recognition Algorithm Used in Knowledge-Based Adaptive Control System to Select Strategy

Krzysztof Brzostowski and Jerzy Świątek

Institute of Information Science and Engineering, Wrocław University of Technology
krzysztof.brzostowski@pwr.wroc.pl, jerzy.swiatek@pwr.wroc.pl

Abstract. In the paper a pattern recognition algorithm in knowledge-based adaptive control system has been discussed. Presented approach based on classifiers, as it is shown, is a useful tool to control non-linear plant with uncertainties. In many cases it is not possible to propose control law. Instead of it expert's knowledge may be used. The pattern recognition method which is utilized by controller to make a decision allows to manage the stored knowledge effectively. Moreover, the system based on presented methodology works faster which is desirable in Real Time System. The biomedical application has been discussed.

1 Introduction

Real-world processes are difficult to describe by exact mathematical models. Unfortunately if the control law has to be designed, the precise description of the plant's dynamics is required. If the object can be specify by linear and deterministic formulas it is possible to propose the sufficient control algorithm according to classic control theory [2], [9]. In practical applications most of the plants have complex non-linear dynamic properties. All of them may lead to sophisticated mathematical description which is not desirable in Real Time Systems where fast response is crucial. Non-linearities are not only problems which have to be overcome by theorists and practitioners. The other features of real process which entails the difficulties are the problems with internal interactions in plant and external with environment. In addition the plant output depends not only on the control input which is produced by system's controller but also on unmeasurable disturbances. The problems which were mentioned above have forced researchers to offer alternative control techniques.

The methods based on assumption that for control plants the exact mathematical description is not required. It means that the control law cannot be formulated. Instead of it, to steer the plant expert's knowledge can be used. The most important properties of designed system are fast response and proper accuracy. The problems which were mentioned above have made researchers to inquire new techniques of control real-world objects. To manage that kind of inconveniences the designed algorithms have to be more sophisticated. It helps

B. Apolloni et al. (Eds.): KES 2007/ WIRN 2007, Part II, LNAI 4693, pp. 247–254, 2007.
© Springer-Verlag Berlin Heidelberg 2007

to resolve many issues in which the decision has to be made. To this end it is desirable to make the most of available information as much as possible.

One of the techniques which can be used in control system where the object is non-linear and measurements are disturbed is a knowledge-based pattern recognition approach. This techniques was first proposed by E.H. Bristol [1]. In the space of thirty years many problems have been resolved by using this approach. The most popular are technical applications [10], [11], [12], [17]. The other areas are problems of controlling biological systems [3], [4], [5]. In [6] a survey of some nonstandard practical applications have been presented. Taking advantage of discussed methodology, it is possible to control plant effectively when the classical methods fail.

Many modern control systems can adapt to the changing external conditions and object's internal parameters. Adjustable control is utilized to improve the whole process by updating the parameters of controller's algorithm. It is important especially when the designer of the system cannot predict exact behavior of the real plant. It is worth noticing that self-tuning is natural properties of knowledge-based patter recognitions techniques.

This approach demands to use, in control system, the model of the considering object where the input's influence on internal parameters is comprised. The parameters of relationship have to be estimated by the system's identification methods. Unfortunately, in a real process it is usually impossible to obtain measurements of needed variables directly. It can be caused by complicated nature of the process or other circumstances. This problem may be overcome by employing two-stage identification [13], [15].

2 Problem Statement

In the paper the approach to control complex objects has been presented. The idea bases on two-stage identification and pattern recognition methods. Authors have proposed many different structures where knowledge-based pattern recognition is applied e.g. [8], [17]. This approach is usually implemented to supervised the system's modules. In the paper some modifications of these design solutions have been introduced.

In the control systems where discussed methodology is applied decision is made on estimated plant parameters. In many real-world problems gaining measurements can be impossible or very expensive. To resolve this problem two-stage scheme can be applied (Fig.1).

Next modification which has been proposed is to use knowledge-based pattern recognition algorithm not to supervise control modules, as it was mentioned above. Instead of it, algorithm bases on discussed approach is used by controller to steer plant. It is non-standard task, because the control sequence is discrete and plant is non-linear. The whole proposed adaptive control system based on knowledge is shown on (Fig.1).

In general, the adaptive control of complex system consists of automatic identification approach at the first O_1 and second stage O_2, pattern recognition

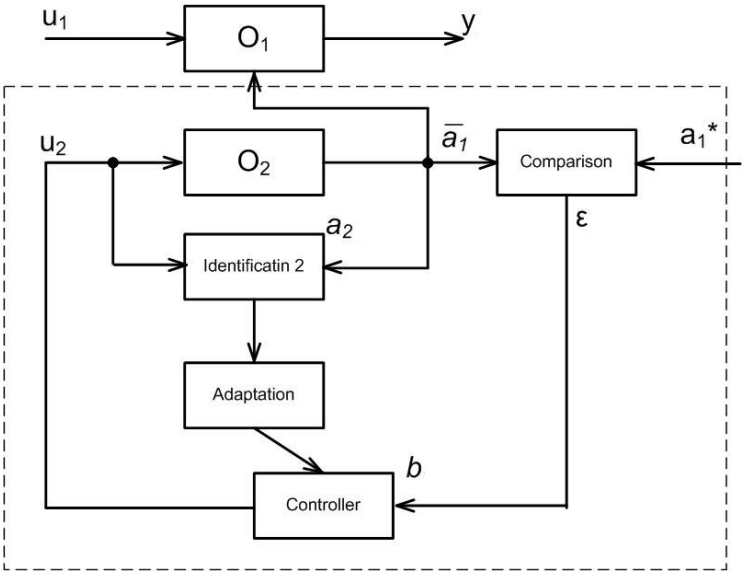


Fig. 1. Knowledge-based adaptive control system with two-stage identification

methods which are used by controller and self-tuning adaptation algorithm. In the considering problem the first stage of the complex object is auxiliary and it is used to compute a_1. In order to estimate the variable a_1 measurement data are required. In this case the series of u_1 and y are needed to determine the unknown values of the parameters. Let us notice that during each experimental trial at the first stage, the variable u_2 which is the input at the second stage remains settled. Identification of a_1 at the first stage has to be repeated for the different values of the input at the second stage. The variable u_2 has to be changed after each experimental test. As a result, the relationship describing the influence object's input u_2 to the internal parameters a_1 is obtained. The variable which is used to describe the relation is defined as a_2.

Collected previous experience of the system can be used in the future to control it. It is related to efficient utilization of the knowledge stored in data base. In the first step the current state of the plant has to be determined (by identification of internal parameter a_1). Next, the best possible strategy is chosen from knowledge base. The methodology usually leads to less accurate algorithm but faster and more robust, which is not meaningless in many real-world problems.

3 Control Via Pattern Recognition

Let us consider the plant described at the first stage as follows:

$$\overline{y} = \Phi_1(u_1, a_1) \tag{1}$$

and at the second (see Fig.1):

$$\overline{a}_1 = \Phi_2(u_2, a_2), \tag{2}$$

where a_1 is internal parameter of the plant, u_2 is the second stage's input and a_2 is a vector of parameters describing relationship between u_2 and a_1. The plant is steered by controller's algorithm:

$$u_2 = \Psi_c(\varepsilon(\overline{a}_1, a_1^*), b), \tag{3}$$

where b is a vector of controller's algorithm parameters and ε is an error function. It was assumed that the value of controller's parameters are settled.

To control, the sequence of $(u_{2k_1}, u_{2k_2}, \ldots, u_{2k_n})$ are taken from finite set, i.e.:

$$U_2 = \{u_{21}, u_{22}, \ldots, u_{2K}\}, \tag{4}$$

where K is the number of possible controls. The series is denoted as π and named *scenario*. The main purpose of the considered controller's algorithm is to find a strategy (also called policy [7]) $\pi = (u_{2k_1}, u_{2k_2}, \ldots, u_{2k_n})$ (n is the number of control sequences) which minimizes a performance criterion $Q_c(\pi, \overline{a}_1)$. The set of all scenarios is defined as follows:

$$\pi \in \Pi = \{\pi_1, \pi_2, \ldots, \pi_l\}. \tag{5}$$

Consequently it leads to solve the following optimization task: find π^* such that:

$$Q_c(\pi^*, \overline{a}_1) = \min_{\pi \in \Pi} Q_c(\pi, \overline{a}_1). \tag{6}$$

The results of optimization task can be described by:

$$\pi^* = \psi_c(\varepsilon(\overline{a}_1, a_1^*)). \tag{7}$$

Practically, calculating the optimal parameters of the algorithm can be expensive and time consuming. In a real-world application it is not desirable because of limitations. As it was mentioned above, the pattern recognition algorithm instead of (7) was proposed. To design the proper control algorithm it is required to design the learning sequence. It leads to prepare a set of parameters as follows: for the sequence of plant parameters $\overline{a}_{11}, \overline{a}_{12}, \ldots, \overline{a}_{1N}$ the respective scenarios are $\pi_1^*, \pi_2^*, \ldots, \pi_{2N}^*$ (N is the number of learning sequence's elements) and they have to be determined by use of (7). It can be denoted such that:

$$X_N = \{(\overline{a}_{11}, \pi_1^*), (\overline{a}_{12}, \pi_2^*), \ldots, (\overline{a}_{1N}, \pi_N^*)\}. \tag{8}$$

Using the learning sequence (8) and the pattern recognition algorithm, for the given parameters $\overline{a}_1$ the respective sequence of $\tilde{\pi}$ may be calculated,

$$\tilde{\pi}^* = \Psi_{PR}(\varepsilon(\overline{a}_1, a_1^*), X_N), \tag{9}$$

where Ψ_{PR} is the proper pattern recognition algorithm.

4 Adaptation Via Pattern Recognition

In previous section the control algorithm has been proposed. In real world application it is not possible to predict all possible behaviors of the plant. It is important, during designing phase, to provide for possibility to learning how the object responses to proper scenario. To this end adaptation can be used. It means that during control the system is monitoring input's influence on object and, if necessary, adjust it to the new condition.

In this section the adaptation algorithm based on pattern recognition has been proposed.

Let us consider the plant which is described by (1), (2) and controlled by algorithm (9). This procedure depends on vector of the controller's parameters $b \in R^S$.

It is necessary to calculate the optimal controller's algorithm settings b^* for which the control performance index $Q_a(b, \overline{a}_2)$ is optimal. In this case the following optimization task has to be solved:

$$Q_a(b^*, \overline{a}_2) = \min_{b \in R^S} Q_a(b, \overline{a}_2). \tag{10}$$

To determine the optimal parameters b^* classical optimization algorithm ψ_a can be used:

$$b^* = \psi_a(\overline{a}_2). \tag{11}$$

As it was mentioned earlier, in connection with control, using exact algorithm in each iteration can be not advisable. To this end a pattern recognition approach may be applied. Before it can be used, the learning sequence has to be prepared. The fixed set is shown below:

$$X_M = \{(\overline{a}_{21}, b_1^*), (\overline{a}_{22}, b_2^*), \ldots, (\overline{a}_{2N}, b_M^*)\}, \tag{12}$$

where M is the number of learning sequences. The algorithm using pattern recognition approach to select parameter $\tilde{b}$ of the controller's algorithm is:

$$\tilde{b}^* = \Psi_{PR}(\overline{a}_2, X_M). \tag{13}$$

In general, determined scenario $\tilde{\pi}^*$ and parameter $\tilde{b}^*$ may differ from optimal value of that which was calculated by the algorithm (7) and (11), respectively. The quality of control sequence depends on pattern recognition method and learning sequence used. In Real Time system, where quick response is crucial it is necessary to reach a compromise between quality of control and fast operating.

5 Simulation Study

For simulation purpose it was assumed that the plant is described by the set of equations. These are used to model immune system's response. To fully emulation of real-world process, measurements are contaminated. The considering system contains three parameters which can be described as vector $a_1 =$

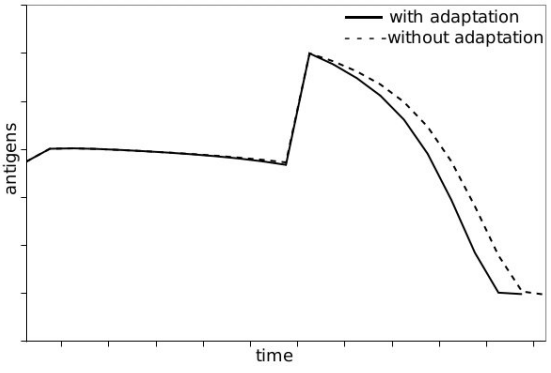


Fig. 2. Adaptation ability of presented system

(a_{11}, a_{12}, a_{13}). During this study it was assumed that they are known because the identification algorithm at the first stage is used to estimate them.

Proposed system describes response of human immune system. Two of parameters which characterise discussing plant do not change during control, but they are used to quantify the properties of immune system (a_{11}) and viruses which has infected human body (a_{12}). The parameter a_{13} is changing during process. The interpretation of it is that this variable characterizes the number of antigens which infected the body. The aim of the control, in discussing problem, is to minimize (even completely eliminate) the number of antigens which infected body. To this end the strategy of treatment has to be designed. It means that number of controls must be determined. In the presented problem injection is input of the plant.

Since determination of optimal sequence (scenario) π^* practically is not possible, to control the chosen system an algorithm based on pattern recognition has been used.

During presented simulation study, simple kNN ([16])algorithm was applied. The learning sequences, which were used, had been collected. In real-world, they can be interpreted as a history of various inmate, who was treated. It can be used to plan curing strategy for other patients. Resting on previous information is a quick and comfortable methodology to treating patients. But it is not possible to contain all useful information in training sequence. In addition, the inmate can be different from one to another. In order to solve the problem, as well as possible, adaptation algorithm has been used. The main idea is to monitor what is the influence of input sequence (π) on internal parameter $\overline{a}_1$ estimated at the first stage.

During experimentation the ability to adapt to the new condition has been tested. As it was mentioned it is not possible to predict exact behavior of the plant, especially when it is non-linear, interacts with environment and the measurement are disturbed.

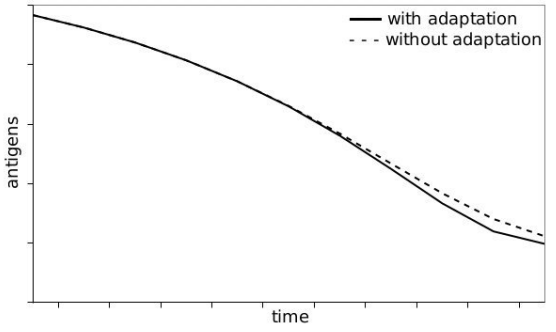


Fig. 3. Adaptation ability of presented system

For the sake of chosen model the experiments have been done in ideal conditions. There was any disturbance which can influence on the quality of identification. It was assumed that the vector of internal parameters $\overline{a}_1$ has been already estimated. On Fig.2 and 3 the results of the experiments have been presented. The second one shows the system's response on the disturbance which, in the presented case, is the new contagion during treatment. As it is seen, adaptation helps the system to overcome the problem with unexpected disturbances.

Fig.3 shows that the system in which adaptation layer has been applied may improve operating properties. During this experiment, the system has to make a decision on the situation which is completely different from that described in learning sequence.

6 Further Works

In the paper experimental study has been presented. It is very important part of the testing algorithm but in many cases it is not sufficient. There is a need to prove the procedure's convergence and its conditions. To this end theory of stochastic control can be applied. Analysis will be extended to the first stage of the complex object which was proposed (see Fig.1).

7 Conclusions

In the paper results obtained by use of algorithm based on pattern recognition has been presented. As it was shown the proposed approach can be useful tool to control the non-linear and complex process. During the training phase the system can learn how to behave in different control situation. Information may be stored and used in the proper moment.

On the other hand, objects can response to the same control sequence in different manner. The best choice in this situation is to apply adaptation at the higher level. Operating in various planes and effective utilization of gained information allows to design the control system based on knowledge which are robust and with ability to overcome the problem of nonlinearities and uncertainties.

References

1. Bristol, E.H.: Pattern Recognition: An Alternative to Parameter Identification in Adaptive Control, Automatica, pp. 197–202 (1991)
2. Bubnicki, Z.: Modern Control Theory. Springer, Heidelberg (2005)
3. Frangu, L., Caraman, S., Ceanga, E.: A Pattern Recognition Controller Applied to the Bioprocess Control. In: CSCS13. Proceedings of The 13-th International Conference on Control Systems and Computer Science, pp. 187–191 (2001)
4. Frangu, L., Ceanga, E., Caraman, S., Boutalis, Y.: A Pattern Recognition Approach to Intelligent Behaviours: Switching the Startegies. In: IEEE International Symposium "Intelligent Systems", pp. 369–372. IEEE Computer Society Press, Los Alamitos (2002)
5. Hamrita, T.K., Wang, S.: Pattern Recognition for Modeling and Online Diagnosis of Bioprocesses. IEEE Transactions on Industry Applications, 1295–1299 (2000)
6. James, J.R., Suski, G.J.: A Survey of Some Implementations of Knowledge-Based Systems for Real-Time Control. In: Proceedings of the 27th IEEE Conference on Decision and Control, pp. 580–585. IEEE Computer Society Press, Los Alamitos (1988)
7. Kushner, H.: Introduction to Stochastic Control, Holt, Rinehart and Winston (1971)
8. Konstantinov, K.B., Yoshida, T.: A Knowledge-Based Pattern Recognition Approach for Real-Time Diagnosis and Control of Fermentation Processes as Variable Structure Plants. IEEE Transactions on Systems, Man, and Cybernetics, 908–914 (1991)
9. Nelles, O.: Nonlinear System Identification - From Classical Approaches to Neural Networks and Fuzzy Models. Springer, Heidelberg (2001)
10. Saridis, G.N.: Application of Pattern Recognition Methods to Control Systems. IEEE Transactions on Automatic Control, 638–645 (1981)
11. Seif, A.A.: On the Adaptive Pattern Recognition Control. In: Proceedings of CompEuro, Computer Systems and Software Engineering, pp. 706–709 (1992)
12. Sobajic, D.J., Pao, Y.-H., Lee, D.T.: Robust Control of Nonlinear Systems Using Pattern Recognition. In: IEEE International Conference on Systems, Man and Cybernetics, 1989. Conference Proceedings, pp. 315–320. IEEE Computer Society Press, Los Alamitos (1989)
13. Świątek, J.: Two Stage Identification and its Technical and Biomedical Applications. Scientific Papers of the Institute of Control and Systems Engineering of the Wrocław Technical University, Wrocław (1987) (in polish)
14. Świątek, J.: Some Problems of Self-Tuning Control via Pattern Recognition. In: Proceeding of the 12th International Conference on Systems Science, Wrocław, vol. 1, pp. 533–538 (1995)
15. Świątek, J.: Two-Stage Estimation Algorithms for Adaptive Control. In: Proceedings of the 12th Conference on Systems Engineering, pp. 670–672 (1997)
16. Theodoridis, S., Koutroumbas, K.: Pattern Recognition. Academic Press, Elsevier (2006)
17. Zhang, B.S., Leigh, I., Leigh, J.R.: Learning Control Based on Pattern Recognition Applied to Vehiclecruise Control Systems. In: Proceedings of the American Control Conference, pp. 3101–3105 (1995)

Alignment-Based Preprocessing of Personal Ontologies on Semantic Social Network

Jason J. Jung[1], Hong-Gee Kim[2], and Geun-Sik Jo[1]

[1] Inha University, Korea
[2] Seoul National University, Korea
j2jung@intelligent.pe.kr, hgkim@snu.ac.kr, gsjo@inha.ac.kr

Abstract. In semantic social network, the relations between users are inferred by measuring the similarity between the corresponding personal ontologies. However, "over-enriched" personal ontologies have caused some difficulties in being discriminated from other personal ontologies. For efficiently annotating resources in their own repositories, people simply append ontology fragments retrieved from standard ontologies and from other neighbors' personal ontologies along to social links. In this paper, we propose a preprocessing method to extract preferential concepts for comparing with social semantics. In order to prune out irrelevant concepts from personal ontologies, alignment-based concept classification process is designed by checking these two main criteria; i) redundancy (e.g., if there already exist semantically identical concepts), and ii) tendency (e.g., if there exist semantically declined concepts). Finally, we want to show an application scenario to demonstrate our contributions.

1 Introduction

Personal ontology had been regarded as an organizational schemes (e.g., directories) for heterogeneous contents in personal information space [1]. However, with increasing concerns about shared metadata and applications (e.g., FOAF[1], SIOC[2], and Orkut[3]), social interoperability has been more emphasized in the semantic web communities.

Especially, we can capture some hidden patterns (e.g., centrality and cohesive subgroups) by employing social network analysis (SNA) methods. On semantic social network (SSN), one can build his own *personal* ontology to describe and conceptualize the resources (e.g., research papers and photo images) in his information repository. Thus, the relationships between users are obtained from mapping the corresponding personal ontologies. It means that the similarities between personal ontologies are play an important role of extracting hidden knowledge on SSN [6].

However, we have some difficulties in aligning two personal ontologies for measuring the similarities between them. The task of constructing personal ontologies still requires much effort and is often carried out in an ad-hoc manner, even if there are some ontology editing tools (e.g., Protégé , SWOOP , and OilEd).

[1] Friend-Of-A-Friend (FOAF). http://xmlns.com/foaf/0.1/
[2] Semantically-Interlinked Online Communities (SIOC)). http://rdfs.org/sioc/
[3] Orkut. https://www.orkut.com/

B. Apolloni et al. (Eds.): KES 2007/ WIRN 2007, Part II, LNAI 4693, pp. 255–262, 2007.
© Springer-Verlag Berlin Heidelberg 2007

Thus, we have to take into account user behaviors, i.e., how people can generate their personal ontologies. Because of several reasons (e.g., diversity or temporal changes of interests), the size of personal ontologies (e.g., the number of classes and properties) is getting increased over time. We refer to it as *over-enriching* problem of semantic social networks, causing not only redundancy of semantically identical concepts, but also, more seriously, entity mismatching of similarity-based alignment algorithms.

In order to solve the problem, we propose ontology segmentation method to analyze the underlying patterns both in a personal ontology and between personal ontologies. Main idea is to reorganize personal ontologies by separating relevant classes from others[4]. Thereby, we apply similarity-based alignment to i) redundancy checking and ii) tendency analysis methods.

2 How Are Personal Ontologies Built

Personal ontologies are built to conceptualize resources in personal repositories. In order to take into account the drawbacks occurred while building personal ontologies, we note the main behaviors as follows.

1. **Simple aggregation.** Users can pile up the ontology fragments (sometimes, the whole ontology) from well-known standard (or upper-level) ontologies, without any corrections. This is just based on "copy-and-paste" task. Thus, the i-th user's personal ontologies $O^i_{\mathcal{P}}$ is represented as $O^i_{\mathcal{P}} = \{o_k | o_k \sqsubseteq \Omega_{\mathcal{O}}\}$ where $\Omega_{\mathcal{O}}$ is a set of standard ontologies indicated as namespaces (e.g., URIs). We assume that there be a "dummy" root concept for connecting the root concepts of ontology fragments.
2. **Manual editing.** Each ontology fragment can be edited by hand. Some concepts $c_j \in o_k$ can be additionally inserted, modified or deleted.
3. **Annotation.** Resources in each user's repository can be described with the concepts in personal ontology. More particularly, in our study, we refer this as *instantiation*[5].
4. **Alignment.** A very few expertized users can declare the relationships between concepts from their existing personal ontologies and new ontology fragments. This explicit mappings can play a important role of extracting semantic preferences.

Most of users take these activities in common to make their own personal ontologies. However, the personal ontologies generated by the tasks are containing not only "essential" concepts but also additional concepts for disambiguation. For example, simple aggregation tasks are collecting the ontology fragments from other ontologies. There might be the irrelevant concepts such as upper-level concepts, and the duplicated concepts between ontology fragments.

As shown in Fig. 1, only the filled parts of ontologies O_A and O_B are a set of concepts related to the corresponding user A and B, respectively. Problem is that the rest of them impede the alignment performance. In this paper, we want to call them as "over-enriched" personal ontologies, and focus on dealing with the difficulties on mapping between these personal ontologies. Basically, similarity-based alignment [7] tries

[4] Recently, they are called *meta noise*.

[5] Generally, it is not a part of ontology building process, but we consider that personal ontology is similar to the folksonomies reflexing user's mental model.

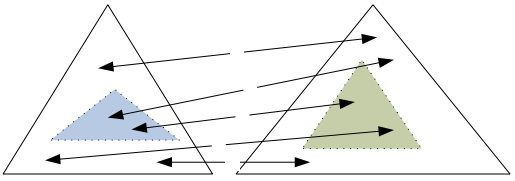


Fig. 1. Alignment between "over-enriched" personal ontologies

to find out the best mapping between two target ontologies among all possible combinations. (Details of this method will be explained in the following section.) Thus, alignment between O_A and O_B is given by

$$Align(O_A, O_B) \longleftarrow \arg_{\{\langle c_A, c_B \rangle\}} \max \sum Sim_C(c_A, c_B) \tag{1}$$

where $c_A \in O_A$ and $c_B \in O_B$. In Fig. 1, the best mapping status between O_A and O_B is composed of five alignments (a_1 to a_5). When we assume that the filled parts of ontologies O_A and O_B are the essential concepts for the users, only alignment a_3 is appropriate. But the rest of them are not useful. More seriously, alignments a_2 and a_4, which are wrong, have taken a chance away from sharing knowledge between these two users.

3 Classification of Concepts in Personal Ontologies

To overcome the problem caused by "over-enriched" personal ontologies, we propose a semantic dimensionality reduction method to classify the concepts in the personal ontologies. This method can discriminate the representative concepts, which are significantly related to user preference, from the other ones. We can classify the concepts of the personal ontologies into three categories, as follows.

1. **Preferential concepts.** Some of concepts in a personal ontology are expressing the user interests. In this paper, we assume that the explicit user activities should be applied to measure the preferential weight of concepts. Such activities are annotating resources and linking to neighbors.
2. **Noisy concepts.** In contrast, as previously explained, these concepts are decreasing the performance of ontology mapping for the collaboration between people, so that they should be ignored when comparing ontologies.
3. **Consensual concepts.** As another category, we can consider consensual concepts. These are a subset of the preferential concepts, which are commonly used within a certain community.

Thus, only the preferential concepts are considered to be aligned with others[6].

The other issue is to discover consensual concepts with respect to the social characteristics of concepts. In fact, a consensus has been regarded as an important challenge

[6] For explaining similarity-based ontology alignment in detail, please refer to [5].

in SNA. It implies the central principles underlying the network [8]. In this context, the consensual concepts in a certain user group most possibly play a role of semantic bridging among people.

3.1 Concept Classification in a Personal Ontology

In this paper, we define a preferential weight p of a concept to measure how closely the concept is related to the user preferences. In other words, this weight means the degree of representative of personal ontologies. Thereby, we exploit two types of analysis methods; redundancy and tendency analysis. More importantly, these methods are considered with the relationships between concepts not only in the same personal ontology, but also in social neighbors' personal ontologies.

For initialization, the concepts in a personal ontology have to be weighted by using the annotation patterns explicitly and manually occurred by users. A set of concepts $C^i_{\mathcal{R}} \subseteq O^i_{\mathcal{P}}$ are assigned with the number of annotated resources. The weight $p_{i,j}$ of j-th concept $c_{i,j} \in C^i_{\mathcal{R}}$ is assigned to $p'_{i,j} = \frac{|res_{i,j}|}{|RES_i|}$ where RES_i is a set of resources annotated by i-th user, and $res_{i,j}$ is a set of resources annotated by $c_{i,j}$. The normalized weight $p_{i,j}$ is given by

$$p_{i,j} = \frac{p'_{i,j}}{\sum_k^{|C^i_{\mathcal{R}}|} p'_{i,k}} \tag{2}$$

where the divider is the summation of weights of all concepts applied to annotation. Hence, when the preferential weight $p_{i,j}$ of concept $c_{i,j}$ is greater than $\tau_{\mathcal{P}}$, we can classify $c_{i,j}$ into preferential concept.

Redundancy checking, $\mathcal{R}$. For the simple aggregation of ontology fragments, we have to detect the duplicated concepts between ontology fragments in a personal ontology. People append a new ontology fragment o_{new} into the existing personal ontology $O_{\mathcal{P}}$. For i-th user, a concept $c_{dup} \in o_{new}$ can be semantically related (e.g., equivalence or subsumption) to the concepts in $O^i_{\mathcal{P}}$. It means that c_{dup} needs not to stay in $O^i_{\mathcal{P}}$ at the same time. Thereby, for $c_{dup} \in o_{new}, c_j \in O^i_{\mathcal{P}}$, we regard a set of pairs of concepts C_{dup} of which similarity $Sim_C(c_j, c_{dup}) \geq \tau_{\mathcal{R}}$ as the semantically identical concepts. Threshold value $\tau_{\mathcal{R}}$ is predefined by users. Based on these duplicated concepts, ontology fragment o_{new} is able to be merged into $O_{\mathcal{P}}$, so that the number of concepts become $|O_{\mathcal{P}}| + |o_{new}| - |C_{dup}|$.

Tendency analysis, $\mathcal{T}$. As next issue, we want to check the tendency of concepts in personal ontologies over time, in order to find out the concepts which is not useful any more. Every time the concepts are applied to resource annotation or the aggregated concepts are duplicated with concept in $C^i_{\mathcal{R}}$, preferential weights of the concepts are adjusted as

$$p_{i,j} \longleftarrow 1/C^i_{\mathcal{R}}, \quad \text{for the newly annotated concept } c_{i,j} \tag{3}$$

$$p_{i,j} \longleftarrow (1 + \Delta) \times p_{i,j}, \quad \text{for the duplicated concept } c_{i,j} \tag{4}$$

$$p_{i,l} \longleftarrow p_{i,l} - \frac{(1+\Delta)\times p_{i,j}}{C^i_{\mathcal{R}} - 1}, \text{ for } l \neq i \tag{5}$$

where Δ is the coefficient of reinforcement.

On the other hand, the weights of the rest of concepts are degraded. The normalization should be conducted right after adjustment process. Consequently, a certain concept might be pruned, when its preferential weight p becomes less than user-specific threshold $\tau_{\mathcal{P}}$.

Example. We want to show a simple example based on two methods for measuring preferential weights. As shown in Fig. 2, a user has to annotate two images) sequentially by using his personal ontology. He takes an activity of simply aggregation by extracting

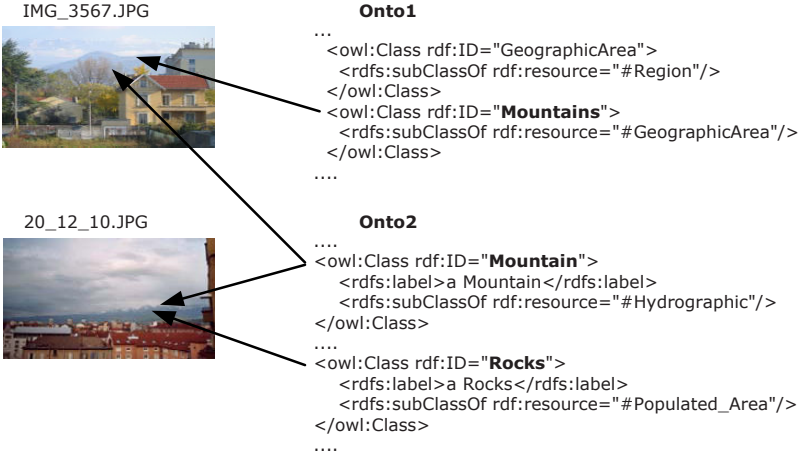


Fig. 2. An example of classifying concepts

an ontology fragment Onto1 from SUMO ontology[7] (or, even, the whole ontology), in order to annotate the first image (e.g., IMG_3567.JPG) with concepts Mountains, StateOrProvince, and ResidentialBuilding (in short, M, SP, RB). These concepts are instantiated to MontBlanc, Rhône-alpes, and YelloHouse, respectively. For initialization, their preferential weight is $p_{\mathsf{M}} = 0.33$, $p_{\mathsf{SP}} = 0.33$, and $p_{\mathsf{RB}} = 0.33$. Then, for second picture (e.g., 20_12_10.JPG), he can take another simple aggregation Onto2 with "space" ontology[8], so that concept Mountain and Rocks (in short, R) can be applied to annotate. From aligning these two fragments, we found out two duplications; i) Onto1:Mountains and Onto2:Mountain, ii) Onto1:ResidentialBuilding and Onto2:Building. Adjustment is computed as

$$p_{\mathsf{M}} = \frac{(1+0.4)\times 0.33}{1+0.25} = 0.462/1.25 = 0.3696, \tag{6}$$

$$p_{\mathsf{SP}} = \frac{0.33-(0.462-0.33)/2}{1.25} = 0.2112, \tag{7}$$

$$p_{\mathsf{RB}} = 0.2112, p_{\mathsf{R}} = 0.25/1.25 = 0.2 \tag{8}$$

[7] http://reliant.teknowledge.com/DAML/SUMO.owl
[8] http://frot.org/space/0.1/

where $\Delta = 0.4$. While second alignment (i.e., ii) detects the duplication for simply reducing the number of concepts, the first alignment (i.e., i) reduces the number of concepts and makes p_{M} enhanced.

3.2 Consensus Concepts on Communities

Here, we want to consider discovering consensus concepts among like-minded people whose preferences are rather closer than others, in order to help people to find the best companies. In other words, an arbitrary user can match his personal ontology with consensual ontologies of user communities (or cohesive subgroups), instead of every individual's personal ontologies.

In this paper, communities from social network can be organized by using social connectivities. We simply assume that authoritative users be playing a role of the center of communities. From [9], the i-th user's authoritative weight is computed by iterative algorithm (i.e., $Hub(u_i) = \sum Auth(u_i')$ and $Auth(u_i) = \sum Hub(u_i')$ where u_i' means the neighbors linking to u_i). When the number of communities is K, we find out the K most centeralized users $Auth^K$ with sorting the authoritative weight by $Auth^K = Top^K_{u \in N}(Auth(u))$ and each community $Comm_i$ is given by

$$Comm_i = \{Auth^i, u_x | Auth^i \in Auth^K, spd(u_x, Auth^i) \geq \delta\} \tag{9}$$

where spd is the shortest path (geodesic) distance, and δ means a radius of the community.

Basically, consensus concepts in a community can be discovered by substructure mining method under *Apriori* assumption. The i-th user's preferential concepts are represented as

$$C^{u_i}_{\mathcal{R}} = \{(c_{u_i,j}, p_{u_i,j}) | u_i \in Comm_\alpha\}, \tag{10}$$

and he is a member of a community $Comm_\alpha = \{u_1, u_2, \ldots, u_{|Comm_\alpha|}\}$. Only $C^i_{\mathcal{R}}$ can be aligned with other users' preferential concepts. This procedure is composed of three steps;

1. initialization of a set of candidate concepts CDT_0

$$CDT_0 = \{(c_{u_i,x}, c_{u_j,y}) | Sim_C(c_{u_i,x}, c_{u_j,y}) \times (p_{u_i,x} + p_{u_j,y}) \geq \tau_{SUP}\}, \tag{11}$$

2. expansion of candidate concepts by join operation $\widetilde{CDT}$, and
3. refinement of the expanded candidate concepts $\widetilde{CDT}$ by evaluation with user-specific minimum support

$$CDT = \{(c_{u_i,x}, \ldots, c_{u_j,y}) | Sim_C(c_{i,j}, c_{x,y}) \times \sum_{u_a} p_{u_a,z} \geq \tau_{SUP}\}, \tag{12}$$

where $\widetilde{CDT}$ and CDT indicate the power sets including the aligned concept sets and the candidate class sets, respectively. The second and third steps are repeated until the constraints such as minimum supports are met. It means that if the number of iteration is T, we can finally get the consensual concepts of which number is T.

4 Application Scenario

In semantic social network, the most important goal of personal ontologies is to provide interoperability between personal information space and search for the like-minded people (or communities). Resources annotated with personal ontologies might be understandable to software modules (e.g., agents) sent from other information space. For example, in order to bookmark and annotate a website "INRIA EXMO", a user can firstly import and aggregate some ontology fragments of the taxonomy ontologies, e.g., ACM computing classification scheme and DMOZ web directory. Then, he perhaps describes it as CS: Research Institutes, Regional: Europe: France, and AI: Knowledge Representation, simultaneously. In order to extract his preferences and compare them with his neighbors by using this "rough and biased" annotations, we have to preprocess a bunch of ontology fragments by using alignment methods and discover the consensus concepts.

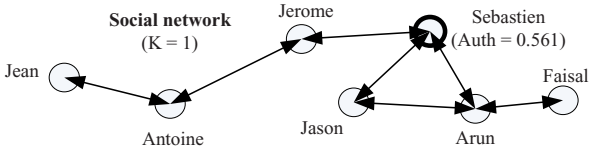


Fig. 3. Socialized semantic space based on aligning personal ontologies

From a social network in Fig. 3, "Sebastian" is discovered as the most authoritative user ($Auth$=0.561), and the user community is organized as { Sebastien, Jerome, Jason, Arun, Faisal } where the radius $\delta = 1$. We can image that by using Eq. 12, the consensus concepts of this subgroup be { AI, CS: Research Institutes }. When Antoine's personal agent embedded with his personal ontology tries to search for the relevant resources, it can easily access to the community first by alignment with consensual concepts.

5 Concluding Remarks and Future Work

In this paper, we have proposed a semantic technique to help people finding like-minded users with similar interests (or preferences), on the basis of matching "over-enriched" personal ontologies. In particular, for efficiently analyzing the user relationships on semantic social network, we emphasize the concept reduction by discriminating representative concepts from duplicated concepts. Moreover, it is also important to ontology evolution. With respect to the activities taken by a user during constructing his personal ontology, ontology alignment-based reduction methods were applied as ontology preprocessing (or cleansing).

As future work, we want to empower the alignment alarming function for dynamic interactions with users, because users are not aware of which concepts are contained in the existing personal ontologies. Also, as extending [4], we have planned to visualize the discovered consensual information.

References

1. Huhns, M.N., Stephens, L.M.: Agents on the web: Personal ontologies. IEEE Internet Computing 3(5), 85–87 (1999)
2. Decker, S., Frank, M.R.: The networked semantic desktop. In: Bussler, C., Decker, S., Schwabe, D., Pastor, O. (eds.) Proc. of the WWW2004 Work. on Application Design, Development and Implementation Issues in the Semantic Web (2004)
3. Jung, J.J.: Collaborative web browsing based on semantic extraction of user interests with bookmarks. J. of Universal Computer Science 11(2), 213–228 (2005)
4. Jung, J.J.: Visualizing recommendation flow on social network. J. of Universal Computer Science 11(11), 1780–1791 (2005)
5. Jung, J.J.: Ontological Framework Based on Contextual Mediation for Collaborative Information Retrieval. Information Retrieval 10(1), 85–109 (2007)
6. Jung, J.J., Euzenat, J.: Towards semantic social networks. In: Franconi, E., Kifer, M., May, W. (eds.) ESWC 2007. Proc. of 4th Euro. Semantic Web Conf, Springer, Heidelberg (2007) (to appear)
7. Euzenat, J., Valtchev, P.: Similarity-based ontology alignment in OWL-Lite. In: de Mántaras, R.L., Saitta, L. (eds.) ECAI'2004. Proc. of 16th Euro. Conf. on Artificial Intelligence, Valencia, Spain, August 22-27, pp. 333–337. IOS Press, Amsterdam (2004)
8. Wasserman, S., Faust, K.: Social Network Analysis. Cambridge University Press, Cambridge (1994)
9. Kleinberg, J.M.: Authoritative sources in a hyperlinked environment. J. of ACM 46(5), 604–632 (1999)
10. Hotho, A., Maedche, A., Staab, S.: Ontology-based text document clustering. Künstliche Intelligenz 16(4), 48–54 (2002)
11. Baumgartner, R., Henze, N., Herzog, M.: The personal publication reader: Illustrating web data extraction, personalization and reasoning for the semantic web. In: Gómez-Pérez, A., Euzenat, J. (eds.) ESWC 2005. LNCS, vol. 3532, pp. 515–530. Springer, Heidelberg (2005)
12. Haase, P., Hotho, A., Schmidt-Thieme, L., Sure, Y.: Collaborative and usage-driven evolution of personal ontologies. In: Gómez-Pérez, A., Euzenat, J. (eds.) ESWC 2005. LNCS, vol. 3532, pp. 486–499. Springer, Heidelberg (2005)
13. Mika, P.: Ontologies are us: A unified model of social networks and semantics. In: Gil, Y., Motta, E., Benjamins, V.R., Musen, M.A. (eds.) ISWC 2005. LNCS, vol. 3729, pp. 522–536. Springer, Heidelberg (2005)
14. Rousset, M.C.: Small can be beautiful in the semantic web. In: McIlraith, S.A., Plexousakis, D., van Harmelen, F. (eds.) ISWC 2004. LNCS, vol. 3298, pp. 6–16. Springer, Heidelberg (2004)

Browser Latency Impact Factors

Andrzej Siemiński

Institute for Applied Informatics
Technical University of Wrocław,
Wybrzeże Wyspiańskiego 27
53-370 Wrocław, Poland
andrzej.sieminski@pwr.wroc.pl

Abstract. The user perceived latency of displaying the Internet pages is one of key factors in his/her assessment of the network. There are many factor that have influence upon the latency. The aim of the paper is to measure the extent of their impact. There are no comprehensive theoretical models to identify such factors. Therefore an experiment was conducted. During the experiment the latency of variety of web pages was measured. The analysis of its results indicates that many popular assumptions are not justified.

Keywords: Internet, browser latency, proxy servers, cacheability, prefetching.

1 Introduction

Users judge the Web sites not only by their content but also by their download time. Previous studies on human cognition [1,2,3] revealed that on the average a man easily accepts a latency lower than 10 seconds. A longer waiting period causes user disorientation and degrades his/her performance. The result corresponds well with the 8 second acceptable latency limit stated in the widely cited report [4]. Unacceptable long latency in the case of e-commerce results in lost revenue.

There are many factors that influence the latency. Due to the complexity and heterogeneity of the Internet the analytical models are inadequate for both researchers and engineers [5]. As a result there is no comprehensive model to predict the latency in real life environment. The constant changes of Internet make its modeling even more difficult. As a result, the previously proposed models focus only on certain aspects of the problem [6].

There are many rule of thumb assumptions about the ways to reduce the latency such as increasing the internet connection bandwidth, promoting cacheability of web objects, using proxy caches or CDN services. The question is whether they are really justified and if so what is their actual influence. To identify the impact factors and to measure their relative importance an experiment was conducted. The download times of pages from servers situated on three continents were recorded and analyzed. The analysis included the following factors:

- Basic connection properties: type of Internet connection, size and number of objects, network distance.

B. Apolloni et al. (Eds.): KES 2007/ WIRN 2007, Part II, LNAI 4693, pp. 263–270, 2007.
© Springer-Verlag Berlin Heidelberg 2007

- Page caching potential: actual or expected time to live time span of objects.
- Types of used buffers: a local and/or a proxy buffer.

The paper is organized as follows. The Section 2 presents the basic network properties of test data set. The various aspects of data caching are discussed in depth in the Section 3 whereas the Section 4 is devoted to the presentation and analysis of the obtained results. The conclusions and some future work areas are discussed in the Section 5.

2 Basic Connection Properties

During the experiment pages were periodically loaded from servers spread all over the world: in the USA, Europe and Australia. The basic properties refer to factors that are outside the control of any user application: type of used Internet connection, the tested set of websites, network distance to original web servers and the size and number of objects making up the pages. All of the tested servers belonged to major newspapers or news services. During the tests the front pages of the websites were periodically loaded for several days. Although the pages changed their content even many times a day their basic properties such as size, number of objects, cacheability of objects remained relatively stable.

2.1 Internet Connection

The experiment was conducted in two environments:

- UNI - a university network having 512Mb/sec bandwidth;
- HOME - a cable connection with a bandwidth of 512 Kb/sec supplied by a local ISP.

A meaningful assessment of the internet connection speed as it is perceived by a user is definitely not an easy task. The bandwidth is always known but it has a direct influence on the latency only for the now obsolete, slow telephone modem connections. Another simple criterion, the throughput is more reliable as it is measured and not calculated. It depends on a great variety of factors e.g. bandwidth, payload size, packet size, network load so on. The obtained results are not easy to interpret. Therefore to describe the connection speed the Surf Speed, one of the tests provided by the Numion Website (www.numion.com) is used. The Surf Speed consists in periodical downloading of a series of objects with varying sizes from servers in specified countries and is regarded as a good quantitative representation of the browser latency. The Table 1 shows the basic statistics the Surf Speed for both environments. There is one column for all measurements and separate columns for the three time periods: Work (from 8 am to 17p.m.), Evening (from 17 p.m. to 23 p.m.) and Night from 23p.m. to 8 a.m.

Table 1. Surf speed in the tested environments in kilobits/sec

Location of Objects	Surf speed	Environment					
		UNI			HOME		
		Work	Evening	Night	Work	Evening	Night
Poland	avg.	791.81	659.05	959.57	419.92	356.71	620.84
	median	674.08	561.80	841.02	202.88	247.60	576.41
	std. dev.	659.25	437.49	765.46	472.90	381.72	416.66
World	avg.	299.31	264.43	351.77	84.48	65.40	97.63
	median	294.92	259.32	345.95	66.75	57.00	69.12
	std. dev.	70.44	61.76	74.11	48.64	30.57	78.55

The above results enable us to draw the following conclusions:

- the immense difference in the throughput of both environments does not correspond with the difference of their Surf Speed;
- In Poland the Internet is used mostly for personal not professional purposes. The heavy workload during the evening hours causes the lowest Surf Speed;
- the difference in Surf Speed is much more apparent in the case of worldwide connections then local (within one country) connections.

2.2 Tested Web Sites and Their Network Distance

The websites tested by the Surf Speed and the websites used in the reported experiment in the paper are different. Therefore the selected servers were tested using the ping and traceroute utilities. The average durations of ping commands are presented in the Table 2.

Table 2. The tested Websites and the ping durations

Code	Url	Avg.Ping (milisec.) UNI	HOME	UniPing/ HomePing	Country
Herald	www.news.com.au/heraldsun	14	18	0.77	Australia
Daily	www.ausdaily.net.au	213	201	1.05	Australia
Cheers	www.cheers.com.au	168	186	0.90	Australia
NYP	www.nypost.com	99	126	0.79	USA
Ohio	www.citynewsohio.com	107	129	0.83	USA
Welt	www.welt.de	77	82	0.93	Germany
Sued	www.sueddeutsche.de	33	39	0.84	Germany
Faz	www.faz.net	34	38	0.89	Germany
Wp	www.wp.pl	11	21	0.52	Poland

The traceroute utility was used to find out the number of hops. Its average number in both environments was similar: 15.2 and 13.8 for UNI and Home environment respectively. In both cases the 4 staring hops were the same. The number of common hops ending the routs was surprisingly small – the average is 3.5. There are a few conclusions that could be drawn from the data:

- Once more again the significant difference in the throughput of the two environments UNI and Home is not reflected by the differences in ping durations.
- Difference in ping times diminishes with the network distance.
- The Australian Herald server behaves in an outstanding manner, different from other Australian servers. This is probably due to the fact that the server uses Content Delivery Network [7] .
- There are substantial differences even for servers located within one country.
- The broadband access to Internet does not guarantee decent ping durations from geographically distant servers.

In the case of remote servers even large difference in the used throughput of Internet connection or the rout taken to fetch the objects does not result in significant differences of the ping durations. The differences are more visible for the near located servers.

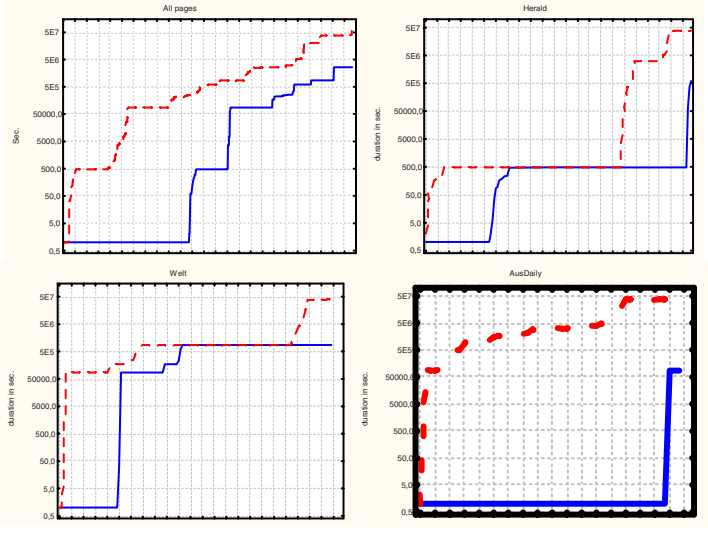


Fig. 1. The distribution of the FP (solid line) and eFP (dashed line) for the selected websites

3 Caching

The paper studies the impact on the browser latency of proxy and local caches – the two most widespread cache types. The proxies were subject of intensive study and numerous paoppers discussed the byte or object hit rate of various algorithms. The metrics refer to the proportion of bytes or objects served from the cache buffer. Every object served from the proxy buffer rather than the WWW server reduces the network traffic. The studies of download time reduction are much more difficult and are less common [8,9].

The HTTP/1.1 protocol specifies in a very detailed manner what data and for how long are considered fresh and could be served from a particular type of a cache [10]. All WWW objects could be divided into:

- Uncacheable, e.g. pages that are generated dynamically.
- Fully cacheable - the HTTP headers define the freshness period (FP). The FP is the time span during which an object could be served safely from a cache.
- Partly cacheable – are objects the without FP but with known age. The eFP (expected Freshness Period) is usually set to 1/5 of objects age.

The proportion of uncacheable WWW objects is high [11,12,13]. The cache control headers are not obligatory and the FP is defined for a relatively small proportion of objects. The age specification is far more popular. The partially cacheable objects could be served from a proxy cache only after a validation - a conditional HTTP request that hopefully takes not as long as a regular, unconditional request.

The distribution of both types freshness periods for all tested objects is shown on the Fig. 1. The dashed lines represent the eFP and the solid lines represent the FP. Note that the Y axis has a logarithmic scale. The graph consists of a number of relatively flat segments which are especially good visible for the FP. The segments represent group of objects having identical or near the same value of the periods. The first segment represents objects with no specified FP. The consecutive segments consist of objects coming from individual websites.

All the websites belonged to the same category. However their webmasters took different approaches to the caching. Only the German websites use extensively the headers. The majority of webmasters do not specify the FP even for image objects that change rare if ever, see the chart for the Daily website. The most cache aware is the webmaster of the Herald website. More then half of all objects have the FP equal approximately to 6 min. This should made make the skimming of the web site convenient and provide fresh data after some time has been spent on reading a page. This facilities local caching but makes proxy buffers useless.

4 Impact Factors

During the experiment the pages of the tested websites we loaded every 20 minutes and the browser latency was recorded. In both environments the four following modes of operation were used: OP (Only Proxy), LP (Local and Proxy), OL (Only Local), NC (No Caching).

The used proxy server forms part of the W3Cache project. The following impact factors were taken into account:

- type of Internet connection (Uni and Home);
- browser mode of operation (OP, OL, LP or NC);
- page size;
- page FP and eFP.

The aim of the experiment was to identify what factors do in fact have influence on the browser latency and what is the extent of their influence.

The Table 3 shows the download durations for all browser operation modes.

Table 3. The average download durations (in sec.)

Env.	Op.Mode	Herald	Daily	Cheers	NYP	Ohio	Welt	Sued	FAZ	WP
	LP	17.2	16.5	3.5	12.5	8.2	6.6	8.8	4.3	2.9
Uni	OL	16.1	15.4	3.3	12.6	7.8	6.2	7.5	3.0	3.3
	OP	17.2	20.4	7.1	19.7	16.7	18,11	9.4	6.1	3.4
	NC	18.9	20.3	6.9	20.2	16.9	18.0	9.5	7.2	3.9
	LP	25.3	11.9	4.7	19.2	9.6	7.3	10.0	7.6	5.0
Home	OL	24.0	11.8	4.5	18.4	9.6	8.5	9.0	6.5	4.4
	OP	25.1	14.8	7.6	27.5	10.3	15.4	11.9	17.0	11.1
	NC	24.9	14.9	8.0	28.2	10.2	15.3	11.8	16.3	10.3

Table 4. Average OL duration analysis, confidence level 0,95

ENV		Herald	Daily	Cheers	NYP	Ohio	Welt	Sued	Faz	Wp
	avg.	16.1	15.4	3.3	12.6	7.8	6.2	7.5	3.1	3.3
Uni	std.d.	4.9	8.9	1.8	7.5	2.8	5.2	3.4	3.7	3.3
	Min	15.3	14.0	3.0	11.1	7.4	5.1	6.8	2.4	2.7
	Max	17.0	17.0	3.6	14.2	8.3	7.2	8.2	3.7	3.8
	avg.	24.8	11.8	4.5	18.4	9.6	8.5	9.0	6.5	4.4
	std.d.	20.3	5.4	4.1	12.8	8.0	9.7	4.1	7.2	5.1
Home	Min	21.4	10.9	3.8	15.9	8.1	6.6	8.2	5.3	3.5
	Max	28.2	12.7	5.2	20.9	11.2	10,.4	9.8	7.7	5.3

Contrary to the widely spread belief the proxy servers are not capable of shortening the latency. There is no statistically significant difference in the values for LP and OL or OP and NC in both environments. On the other hand the introduction of a local buffer is significant. For the Uni environment one can say on the 95% confidence level (see Tables 4 and 5) that the average OL duration is lower than the OP and NC durations for almost all servers. The only exception is the WP server. For the Home environment the same is true for all servers excluding the Herald and Ohio servers.

It is true that the high bandwidth of the Uni Internet connection results in shorter latency than in the case of Home data. On confidence level 0.95 the average OL durations for Uni environment are lower then for the Home environment for all servers excluding only AusDaily, see the Table 4. The reduction of latency is however not spectacular to say the least.

The Herald server is a particular case. The server is the only one that uses the CDN. The data is almost not proxy cacheable. The chart on the Figure 1 shows a flat segment (middle part of the Figure) but it is formed by objects with FP equal to 500 sec. During the test such objects ware not cached as the pages were fetched every 20 minutes that is 1200 seconds. This could explain why it is the only server for which there no statistically significant differences in the durations of all operation modes. Despite extremely short ping duration the browser latency is surprisingly high.

Further dependences were analyzed using the Spearman correlation coefficient. The data is first normalized. To eliminate the influence of different page sizes the duration of the download is replaced by the transfer rate - the total size of a page (in Kb) divided by the download duration in seconds. The susceptibility to caching is represented by the coefficient FP20 which is equal to the proportion of objects with

the FP exceeding 20 minutes. 20 minutes was the server access rate used in the experiment. In a similar way the coefficient eFP20 relates to the eFP. The Table 5 shows the values of the Spearman correlation coefficient for the Uni environment. Similar results were obtained for the Home environment.

Table 5. Values of the Spearman correlation coefficient for the Uni Environment

	uPing	FP(20)	eFP(20)	uLP	uOL	uOP	uNC
uPing	1,0	-0,7	0,7	-0,6	-0,5	-0,8	-0,8
FP(20)	-0,7	1,0	-0,2	0,7	0,7	0,7	0,7
eFP(20)	0,7	-0,2	1,0	-0,1	-0,1	-0,5	-0,5
uLP	-0,6	0,7	-0,1	1,0	1,0	0,7	0,7
uOL	-0,5	0,7	-0,1	1,0	1,0	0,7	0,7
uOP	-0,8	0,7	-0,5	0,7	0,7	1,0	1,0
uNC	-0,8	0,7	-0,5	0,7	0,7	1,0	1,0

The download transfer rates are correlated with the FP(20) but not with the eFP(20). This is true even for the OL and LP operation modes. The degree of correlation is slightly higher for the Home test environment. Surprisingly no correlation was detected between the Ping duration and download transfer rate in all operation modes. The single most influential factor is the usage of a local buffer.

5 Conclusions

Users demand short browser latencies that not exceeds 8 seconds. The figure has a psychological background and is therefore to large extent not negotiable. The aim of the paper was to identify factors what influence the latency. The results of the experiment contradict some widespread convictions. It turned out e.g. that the widening of the bandwidth of an Internet connection does not guarantee decent latency reduction. What is even more interesting is that the proxy caches, although useful in many areas, do not influence the download time in a significant way. The use of a CDN service results in a stable but far from brilliant performance. This result is so surprising that is calls for further studies, which are planned for the near future. The analysis of caching properties of WWW objects shows that the web masters usually do not exploit the plentiful cache control methods provided by the HTTP protocol. This unfortunate as it has an adverse impact the performance of proxy caches and could not be remedied in any other way. The single most influential factor is the local buffer.

In order to reduce significantly the browser latency the simplistic algorithms controlling the local buffer used should be replaced by algorithms that take advantage of object prefetching and are highly personalized. Prefetching is a technique that attempts to predict the pages a user want to read and then loads such pages in advance [14]. Local prefetching algorithms can fully exploit the benefits offered by the so called user dividend, as it is described in [15]. The work on such user profile based prefetching algorithms for local buffers is currently underway.

References

1. Bhatti, N., Bouch1, A., Kuchinsky, A.: Integrating User-Perceived Quality into Web Server Design. Computer Networks (Amsterdam, Netherlands) 33(16), 1–16 (1999)
2. Card, S.A., Moran, T.P., Newell, A.: The psychology of Human-computer Interaction. Lowrence Erlbaum Associates, NJ (1983)
3. Ramsay, J., Barbesi, J., Preece, J.: Psychological Investigation of Long Retrieval Times an the World Wide Web. Interacting with Computers 10(1) (1998)
4. Zona Research: The economic impacts of unacceptable Web Site download speeds, White paper, http://www.zonaresearch.com/deliverables/while-papers/wp17/index.html
5. Cairano-Gilfedder, C., Clegg, R.G.: A decade of Internet research – advances in models and practices. BT Technology Journal archive 23(4), 115–128 (2005)
6. Siemiński, A.: The impact of proxy caches on browser latency. Int. J. Comput. Sci. Appl. 2(2), 4–21 (2005)
7. Valali, A., Pallis, G.: Content Delivery Networks: Status and Trends. IEEE Internet Computing 11, 68–74 (2003)
8. Feldman, A., Caceres, R., Douglis, F., Glass, G., Rabinovich, M.: Performance of Web proxy caching in heterogeneous bandwidth environments. In: Proceedings of INFOCOM, pp. 107–116 (1999)
9. Kroeger, T., Long, D., Mogul, J.: Exploiting the bounds of Web latency reduction from caching and prefetching. In: Proceedings of USENIX Symposium on Internet Technologies and Systems, pp. 13–22 (1997)
10. Gourley D., Totty, B.: HTTP: the Definite Guide, O'Reilly & Associates (2002)
11. Sieminski, A.: Changebility of Web Objects. In: ISDA'05. 5th International Conference on Intelligent Systems Desin and Implementation, Wrocław (2005)
12. Wessels, D.: Web Caching, O'Reilly and Associates, Inc. (2001)
13. Zhang, X.: Cachability of Web Objects, Technical Report 2000-019, Boston U., CS Department (2000), citeseer.ist.psu.edu/zhang00cachability.html
14. Davison, B.D.: The design and Evaluation of WEB Prefetching and Caching Techniques. The State University of New Jersey, Dissertation, Rutgers (2002)
15. Sieminski, A.: The Potentials of Client Oriented Prefetching in Intelligent Technologies for Inconsistent Knowledge Processing, Advanced Knowledge International (2004)

Global Adaptive Request Distribution with Broker

Leszek Borzemski[1], Anna Zatwarnicka[2], and Krzysztof Zatwarnicki[2]

[1] Institute of Information Science and Engineering,
Wroclaw University of Technology, Wroclaw, Poland
Leszek.Borzemski@pwr.wroc.pl
[2] Faculty of Electrical Engineering, Automatic Control and Computer Science,
Opole University of Technology, Opole, Poland
{Anna.Zatwarnicka, KZatwarnicki}@po.opole.pl

Abstract. This paper presents the application of fuzzy logic and neural networks to HTTP request dispatching performed within a geographically distributed Web system. Web sites serve as a global content delivery system where each Web server can respond to the client request. We propose broker-based system architecture with a global request dispatching algorithm called GARDiB. The algorithm uses the fuzzy-neural decision-making mechanism to assign each incoming request to the server with the least expected response time. The response time include the transmission time over the network, both for the request and for the response, as well as the time elapsed on the server responding to the request. We demonstrate through the simulations that our algorithm is more effective than popular global dispatching policies Round-Robin and Weighted Round-Robin.

Keywords: Intelligent systems, Fuzzy-neural control, HTTP request distribution, Web service quality, Dispatching algorithms, Cluster-based architectures, Distributed systems, Client/server.

1 Introduction

Modern Web sites are deployed based on cluster-based server configurations. Web servers in a cluster work collectively as a single Web resource in the network. Web cluster architecture includes a dispatcher that distributes user requests among cluster's Web servers. Further improvement of the Web performance needs distributed Web systems composed by multiple Web sites located at different geographical locations in the Internet (like so called Content Distribution Networks). In such systems global request routing mechanisms and algorithms are needed to dispatch the incoming requests to different Web sites according to the given performance index. In this paper we present a global request routing broker-based mechanism incorporating a fuzzy-neural dispatching algorithm aiming to optimize the request response time.

Fuzzy and neural systems have many significant applications in deployment of intelligent systems [9]. Fuzzy logic controllers can efficiently deal with the uncertainties that exist in physical systems, as well as with intrinsic nonlinearities of controlled systems [8]. By applying of neural networks into fuzzy systems we can

B. Apolloni et al. (Eds.): KES 2007/ WIRN 2007, Part II, LNAI 4693, pp. 271–278, 2007.
© Springer-Verlag Berlin Heidelberg 2007

achieve learning and adaptation characteristics. By such hybrid solutions we can support robust and adaptive control in uncertain and noisy environments. Especially, by neural networks, the fuzzy control parameters can be tuned to achieve better control [8]. In our previous work we have developed the FNRD algorithm for fuzzy-neural adaptive HTTP request distribution in a local Web cluster system [2]. FNRD policy may be useful in the design of the content-aware Web switch for local Web sites [3, 4]. It outperforms other competitive dispatching policies including the state-of-the-art content-aware "reference" algorithms like CAP (Content Aware Policy) [7], LARD (Locality Aware Request Distribution) [11], and RR (Round Robin) policy [6].

In this paper we contribute with the application of fuzzy and neural networks approaches to the request routing done for the distributed Web systems (i.e. at global level) to achieve a response time effective HTTP request dispatching policy. The Web sites of a system under consideration are developed as locally clustered Web systems with fully replicated content where each local Web server can prepare the response to the request. Our global request dispatching system is called GARDiB2 (Global Adaptive Request Distribution with Broker V2). It uses a fuzzy-neural decision-making model at global distribution level as well as at local sites. Our solution, takes into account the standpoint of each individual request while dispatching (due to minimization of expected response time of individual request) and is a step towards providing quality based Web services [5].

The reminder of the paper is organized as follows. Section 2 presents the concept of fuzzy-neural global dispatching. Section 3 shows the simulation model and obtained results. Section 4 concludes our research.

2 Global Adaptive Request Distribution with Broker

Our request distribution system has the overall architecture as shown in Fig. 1. It consists of several local Web clusters with the same local request dispatching algorithm – we assume that all local network dispatchers use FNRD algorithm [2].

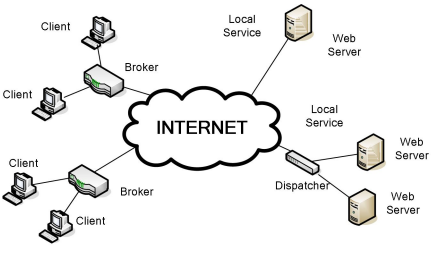


Fig. 1. Architecture of broker based Web distribution system

The clients send their request to the nearest (in the sense of network distance measure) broker which routes the request using fuzzy-neural mechanism aiming at the minimization of the request response time. The request response time includes the

transmission times of request and response through the Internet, and the servicing time of the request on the Local Service. More specifically, the request response time is estimated as the elapsed time between the beginning of sending of the first byte of the request by the broker and the end of receiving of the last byte of the request response by the broker. The key components of the system are the brokers which are established at as many Internet nodes as required. They are placed near the "local community" of clients supported by a given broker, e.g. based on the geographic distance [1]. As other intermediary network appliances they are to extend the Web functionality and performance without violating the principles employed in the design of the Internet. The broker routes each incoming request to the Web server in the Local Service (LS) which offers the best request response time.

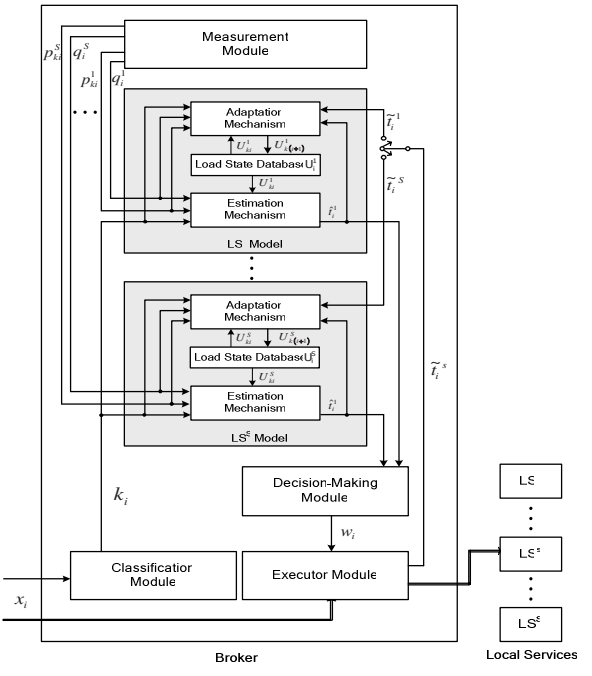


Fig. 2. Architecture of the broker

The general architecture of the broker is shown in Fig. 2. It has the following main components: *Local Service Models* (LSMs), *Decision-Making Module* (DMM), *Executor Module* (EM) and *Measurement Module* (MM). Each broker supports S LSMs LSM^1, LSM^2,…,LSM^s,…,LSM^S. The s-th LSM^s calculates the *estimated request response time* $\hat{t}_i^s$ for its corresponding s-th Local Service LS^s, s=1, 2, …, S for given x_i request. Each request can be classified to one of the k-th class, k=1,2,…, K. LSM^s takes into account the class of the request and vector P_i^s, $P_i^s = [p_{1i}^s,..., p_{ki}^s,..., p_{Ki}^s]$ which includes the estimated local servicing times for all request classes for s-th LS (this data is collected from Local Services every seconds), and the transmission time $\bar{t}_i^s$ of the

probe object requested from LS[s] by the MM based on appropriate monitoring scenario. The vector P_i^s is determined locally by the FNRD algorithm employed in every local Web switch. The values of $\hat{t}_i^s$ for all sites are compared and the MDM chooses the LS w_i, $w_i \in \{1,...,S\}$ with the minimum value of the estimated request response time. Then the EM redirect the request x_i to chosen *s*-th LS. The requested object is sent by LS[s] to the broker which returns it to the client. The EM measures the *actual value of request response time* $\tilde{t}_i^s$ of the request x_i, as well. $\tilde{t}_i^s$ is used by the LSM[s] to update its knowledge about current processing conditions on *s*-th LS.

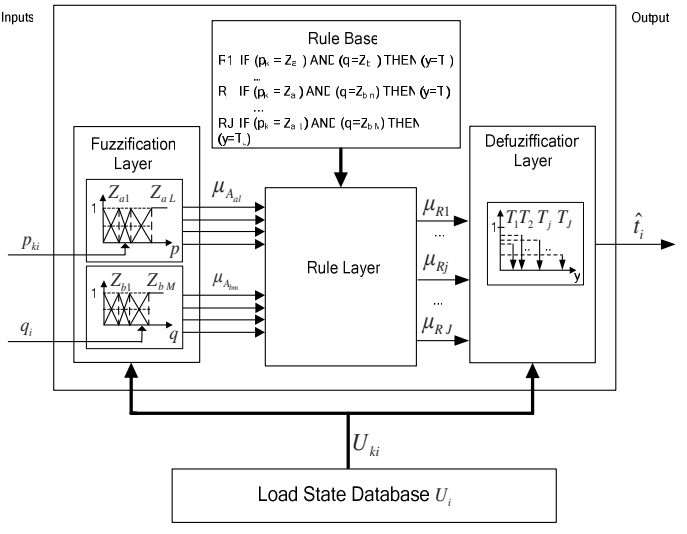


Fig. 3. LS model

LSM internal architecture is shown in Fig. 3. For presentation clarity reasons, we omit in this figure and in further text the *s*–th index of the LS. The LSM has *Classification Mechanism* (CM), *Estimation Mechanism* (EM), *Load State Database* (LSD), and *Adaptation Mechanism* (AM) components. The CM recognizes the class k_i of x_i request on the basis of object size and its category; $k \in \{1,...,K\}$ – we consider as in [2] both the static and dynamic objects. The EM and AM components are built on the basis of a neuro-fuzzy approach according to the Mamdami's model [9] (Fig. 3). The system has three layers: *Fuzzification Layer*, *Rule Layer* and *Defuzzification Layer*.

The EM determines the estimated request response time for x_i based on the k_i, p_i^s, q_i^s and U_{ki} from $U_i = \lfloor U_{1i},...,U_{ki},...,U_{Ki} \rfloor$ database which contains up-to-date information about the load state of the LS. q_i^s is the probe transmission time from *s*-th site at *i*-th moment. Based on the measured actual request response time $\tilde{t}_i$ the AM

updates U_i database in such a way that U_{ki} for the k_i-th class is updated to new value $U_{k(i+1)}$.

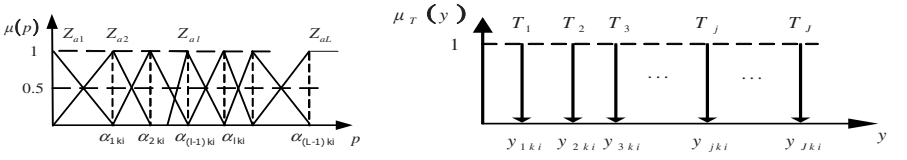


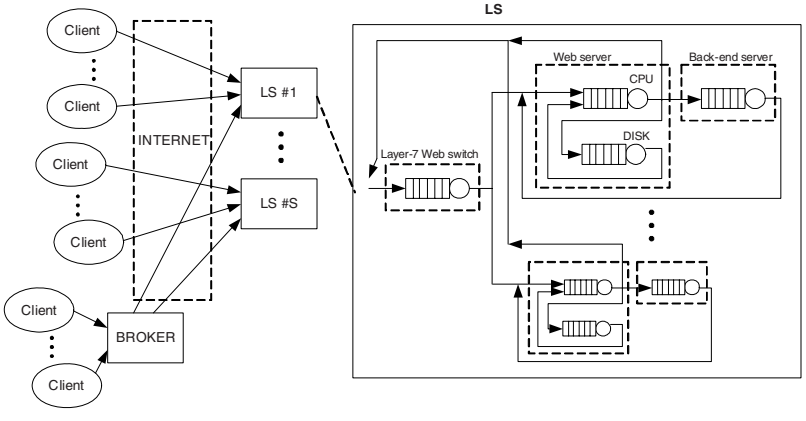
Fig. 4. Fuzzy sets for inputs and outputs

Based on the semantics of both inputs we define L membership functions $Za_0,\ldots,Za_L$ for p_i and M membership functions $Zb_0,\ldots,Zb_M$ for q_i. Both membership functions are triangular and feature a partition of unity. Their shaping is shown in Fig. 4 (more specifically, for p_{ki}), where parameters α_{lki}. The similar parameters β_{mki} define the shaping of the membership functions for q_i. The fuzzy sets of outputs $T_1,\ldots,T_J$ are assumed to be the singletons pointing at $y_{1ki},\ldots, y_{jki}, \ldots, y_{Jki}$ values (Fig. 4). The values of α_{lki}, β_{mki} and y_{jki} are stored in $U_i = [U_{1i},...,U_{ki},...,U_{Ki}]^T$, where a single vector includes information about the k-th class at the moment of incoming of the i-th request to the broker. The parameters are used as the weights of neurons and are tuned during the adaptation process. $U_{ki} = [A_{ki}, B_{ki}, Y_{ki}]$ contains information about k-th class at the moment of incoming of i-th request to the broker where $A_{ki} = \lfloor \alpha_{1\,ki},...,\alpha_{l\,ki},...,\alpha_{(L-1)\,ki} \rfloor$, $B_{ki} = \lfloor \beta_{1\,ki},...,\beta_{m\,ki},...,\beta_{(M-1)\,ki} \rfloor$ and $Y_{ki} = \lfloor y_{1\,ki},...,y_{j\,ki},...,y_{J\,ki} \rfloor$. PROD fuzzy operator is used as AND connective, thus the degree μ_{Rj} of the membership of antecedent of rule Rj is equal to $\mu_{R_j}(p^*, q^*) = \mu_{Z_{al}}(p^*) \cdot \mu_{Z_{bm}}(q^*)$. PROD operator is used for the implication function. The result of the fuzzy inference is a fuzzy value that has to be retransformed (i.e. defuzzified) into a crisp value. The "height" method, which favors the rule with the greatest output value, is used in that operation. $\tilde{t}_i$ is used to tune the weights of the neurons $\alpha_{\varphi kg}$, $\beta_{\theta\,ki}$ and $y_{j\,ki}$ which are modified using the back propagation method each time after the completion of the request. φ and θ are the indexes of the membership functions for inputs p_{ki} and q_i, respectively, $\varphi = 1,2,...,L-1$ and $\theta = 1,2,...,M-1$. The mean square error $E_i = 0.5(e_i)^2$ is calculated for the neuron in the output layer. Then the new weight values are calculated as $\alpha_{\phi k\,(i+1)} = \alpha_{\phi k\,(i)} + \eta_a (\tilde{t}_i - \hat{t}_i) \cdot y_{j\,k\,i} \cdot \sum_{j=1}^{J} (\mu_{Z_{bmi}}) \cdot \sum_{l=1}^{L} \left(\frac{\partial \mu_{Z_{al}i}}{\partial \alpha_{\phi k\,(i)}} \right)$, $y_{j\,k(i+1)} = y_{j\,k(i)} + \eta_y (\tilde{t}_i - \hat{t}_i) \cdot \mu_{R_j i}$, $\beta_{\,k\,(i+1)} = \beta_{\,k\,(i)} + \eta_b (\tilde{t}_i - \hat{t}_i) \cdot y_{j\,kg} \cdot \sum_{j=1}^{J} (\mu_{Z_{al}i}) \cdot \sum_{m=1}^{M} \left(\frac{\partial \mu_{Z_{bm}i}}{\partial \beta_{\,k\,(i+1)}} \right)$. The values of η, L, M were determined in the preliminary simulations: $\eta_a = \eta_t = 0.01$, $\eta_y = 0.5$, $L = 4$, $M = 4$ (thus $J = 16$).

3 Simulation Results

The model of a distributed Web system used in the simulation is shown in Fig. 5. The main delay in request servicing is assumed to be introduced by the Web and database servers. The network delay influences the request response time, as well. We simulated the Web site of the total size of 200 MB. We evaluated the performance using the workload model incorporating the most recent research on Web load.

The simulator runs with almost the same assumptions concerning CPU speed, amount of memory, number of disks and other parameters as in [11] where the processing costs were calculated for Pentium II 300 MHz PC with FreeBSD 2.2.5. Due to the new server and network developments we decreased all time parameters by the factor of ten. We also reduced database service times. The connection establishment and teardown costs were set as 14.5 μs of CPU time each, while transmit processing incurred 4.0 μs per 512 bytes. Disc costs were the following: reading a file from disk has a latency of 2.8 ms, the disk transfer time is 41.0 μs per 4 KByte. For files larger than 44 KBytes, an additional 1.4 ms (seek plus rotational latency) was charged for every 44 Kbytes of file length in excess of 44 KBytes. The LRU cache replacement policy was used; with files of a size of more than 500 KB never cached. The total



b) Static requests

Category	Distribution	Parameters
Requests per session	Inverse Gaussian	μ=3.86, λ=9.46
User think time	Pareto	α=1.4, k=1
Objects per page	Pareto	α=1.33, k=2
HTML object size	Lognormal Pareto	μ=7.630, σ=1.001 α=1, k=10240
Embedded object size	Lognormal	μ=8.215, σ=1.46

Dynamic requests (database)

Type	Mean service time [msec]	Frequency
High intensive	5	0.85
Medium intensive	10	0.14
Low intensive	20	0.01

Fig. 5. (a) A simulation model; (b) Workload model parameters

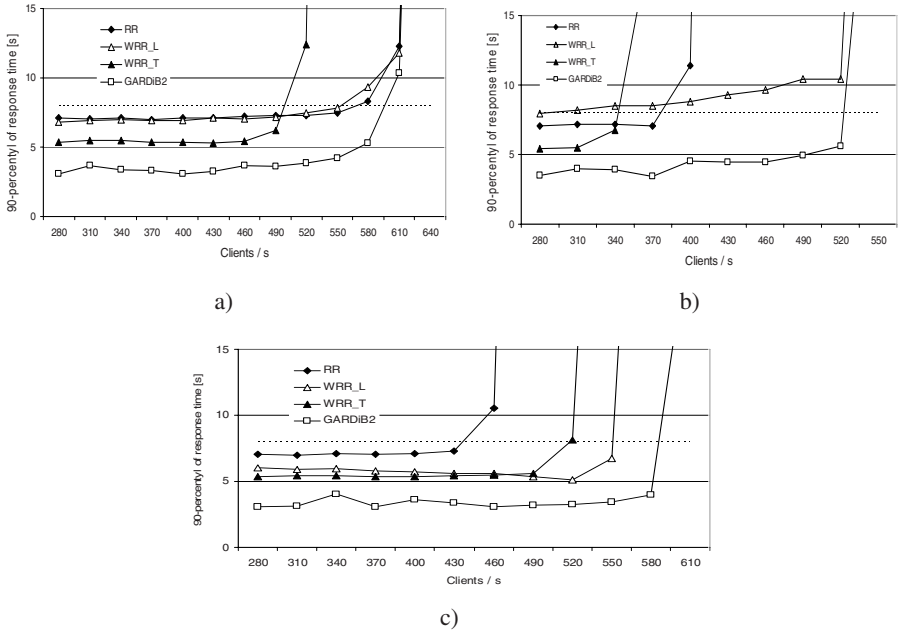


Fig. 6. 90-percentile of page response time vs. number of clients per second: a) NL/25; USA/25; AU/25; B/25, b) NL/40; USA/20; AU/10; B/30, c) NL/10; USA/25; AU/35; B/30.

memory size used for caching was 85 MB. The data transmission time for the request and the response was calculated as the sum of the Round-Trip Time (RTT) and (requested_object_size+290)/Throughput [10] where 290 stands for usual HTTP header size. The throughput was determined as the effective number of bytes that could be transferred in per second. The distribution of RTT was modeled on the basis of the real-life end-to-end network measurements performed by the authors targeting Web servers in the Netherlands (NL), Australia (AU) and USA. We compare our algorithms with two commonly used global dispatching policies: *Round Robin* (RR) and *Weighted Round Robin* (WRR). RR distributes the requests evenly among LSs regardless of their loads. WRR_L takes into account the site load measured by the number of request serviced at the moment whereas WRR_T considers data transmission time.

We assumed a workload scenario consisting of 20% of *dynamic* requests for *low*, *medium* and *high* intensive workload sizes. The rest are *static* requests. Static requests are serviced directly by the Web server whereas in the dynamic ones the objects are generated by the back-end servers. We simulated three homogeneous LSs, each equipped with three Web and three database servers. Three load scenarios were considered – the numbers after "/"show how many requests (in %) have been sent to specific LS directly (this is a background load) or via the broker (B) (these requests are under control of global distribution). The page response time is the sum of object's downloading times, assuming that the service of every object takes the same time. The results of the simulations are presented in Fig. 6. The dotted line is the threshold of 8 seconds – the time the user may meekly wait for the respond.

The results show that GARDiB2 outperforms their rivals for the whole range of the load size and for all scenarios. The response times for GARDiB2 are in the range of 3 – 4 seconds – such times are not viable for either of the competitors.

4 Conclusion

We presented a fuzzy-neural HTTP global distribution policy called GARDiB2. It achieved very good performance for the whole load range. The simulations showed that our design is valuable in the content-aware distributed request dispatching. Our policy outperformed very popular content-blind RR and WRR algorithms.

Acknowledgements

This work was supported by the Polish Ministry of Science and Higher Education under Grant No. N516 032 31/3359 (2006-2009).

References

1. Borzemski, L., Nowak, Z.: Using the Geographic Distance for Selecting the Nearest Agent in Intermediary-Based Access to Internet Resources. In: Khosla, R., Howlett, R.J., Jain, L.C. (eds.) KES 2005. LNCS (LNAI), vol. 3683, pp. 261–267. Springer, Heidelberg (2005)
2. Borzemski, L., Zatwarnicki, K.: A Fuzzy Adaptive Request Distribution Algorithm for Cluster-Based Web Systems. In: Proc. of 11th PDP Conf., pp. 119–126. IEEE Press, Los Alamitos (2003)
3. Borzemski, L., Zatwarnicki, K.: Using Adaptive Fuzzy-Neural Control to Minimize Response Time in Cluster-Based Web Systems. In: Szczepaniak, P.S., Kacprzyk, J., Niewiadomski, A. (eds.) AWIC 2005. LNCS (LNAI), vol. 3528, pp. 63–68. Springer, Heidelberg (2005)
4. Borzemski, L., Zatwarnicki, K.: Fuzzy-Neural Web Switch Supporting Differentiated Service. In: Gabrys, B., Howlett, R.J., Jain, L.C. (eds.) KES 2006. LNCS (LNAI), vol. 4252, pp. 195–203. Springer, Heidelberg (2006)
5. Cardellini, V., Casalicchio, E., Colajanni, M., Mambelli, M.: Web Switch Support for Differentiated Services. ACM Perf. Eval. Rev. 29(2), 14–19 (2001)
6. Cardellini, V., Casalicchio, E., Colajanni, M., Yu, P.S.: The State of the Art in Locally Distributed Web-Server Systems. ACM Comp. Surv. 34(2), 263–311 (2002)
7. Casalicchio, E., Colajanni, M.: A Client-Aware Dispatching Algorithm for Web Clusters Providing Multiple Services. In: Proc. WWW10, pp. 535–544 (2001)
8. Lee, K-M., Kwak, D-H., Leekwang, H.: Tuning of fuzzy models by fuzzy neural networks. Fuzzy Sets and Systems 76(1), 47–61 (1995)
9. Mamdami, E.H.: Application of Fuzzy Logic to Approximate Reasoning Using Linguistic Synthesis. IEEE Trans. on Computers, 26(12), 1182–1191 (1977)
10. Menasce, D., Almeida, V.: Capacity Planning for Web Performance. Metrics, Models, and Methods. Prentice Hall, New York (1998)
11. Pai, V.S., Aron, M., Banga, G., Svendsen, M., Druschel, P., Zwaenpoel, W., Nahum, E.: Locality-Aware Request Distribution in Cluster-Based Network Servers. SIGOPS Oper. Syst. Rev. 32(5), 205–216 (1998)

Global and Local Approach to Complex Systems Modeling Using Dynamic Neural Networks– Analogy with Multiagent Systems

Jarosław Drapała and Jerzy Świątek

Institute of Information Science and Engineering, Wrocław University of Technology
jaroslaw.drapala@pwr.wroc.pl, jerzy.swiatek@pwr.wroc.pl

Abstract. In this paper, the global and local approach to dynamic input-output complex systems modeling is presented. Local and global modeling problems are formulated. Dynamic neural network's learning algorithm for global modeling of complex system has been constructed. Analogy between models of input-output complex systems and multiagent systems is discussed.

1 Introduction

There are many ways of understanding term "complex system". We can take into consideration interactions between components of investigated system, high dimensionality, sophisticated behavior, a large number of interacting entities, multi-level organization of the system, etc. We are surrounded by many complex systems of a different nature (biomedical, economical, technical), we are part of many overlapped complex systems and what's more, we are itself quite efficient example of such a system.

In order to control and understand complex systems, one should construct it's accurate model. Depending on the nature of the problem, modeling techniques of different kind can be applied. This paper deals with modeling of systems, which complexity comes from cooperation of interconnected components. It is assumed that each of system's component can operate separately transforming input data into output signal. Connections between inputs and outputs of each of them give us a complex system. Modeling of such a system involves modeling of each process separately, taking into account connection between them. This leads to the concept of local and global modeling of complex systems, [11]. The so-called locally optimal model cames into existence when we construct a collection of models, optimal with respect to some criteria, for each element of the complex system separately, and then join them together. Important feature of locally optimal model is that despite it's optimality for each single element, it is not optimal - in general case - for a complex system as a whole. Unlike locally optimal model, the so-called globally optimal model is the best for the system as a whole, but not necessarily for each single element. It is also possible to construct a model, which keeps compromise between quality of models of single

B. Apolloni et al. (Eds.): KES 2007/ WIRN 2007, Part II, LNAI 4693, pp. 279–286, 2007.
© Springer-Verlag Berlin Heidelberg 2007

elements and quality of the complex system's model, the part of which they are. Similar problems occur when analyzing a two-level (generally multi-level) management process. Dependent on the level, we expected a model of different quality. At the lower level, it is better to prepare locally optimal model, but at upper level, globally optimal description is preferred.

Ideas described above had been firstly formulated as a part of Systems Identification Theory in work [2], then developed, among other things, in works [7,9,10,12]. Global and local approach to complex systems modeling proved to be useful in practical applications. Some biomedical (modeling of hemodynamics of cardiovascular system) and technical applications are presented in work [9], modeling of ammonium nitrate production process in [4], modeling of copper ore's enrichment process in [7] and modeling of human's elbow neuromuscular system in [1]. However, input-output complex systems modeling methods are not restricted to biomedical or technical applications. This paper aims to work out an analogy between the process of constructing globally optimal model of complex interconnected system and the process of gaining knowledge by the system of agents.

Let us note that general description of complex systems given at the beginning of this section resembles the one of multiagent systems, which are composed of multiple interacting computing elements, known as agents [13]. Agents are capable of performing autonomous action satisfying their design objective and interacting with other agents. But design objectives of each single agent are not the most important one's, because they all are supposed to satisfy the purpose of multiagent system as a whole. In work [14] this problem is referred as global and local problem solving by multiagent system. This very general analogy encouraged us to look closer at possibility of applying concepts from identification of complex systems on the field of multiagents systems. Section 2 gives a brief introduction on complex systems description based on systems identification. We have constructed an algorithm for determining parameters of dynamic neural network which acts as a model of dynamic complex system and described it in section 3. In section 4 we discuss the analogy between the process of creating a globally optimal model and the process of gaining knowledge by agents.

2 Description of Dynamic Complex System

Let us consider an input-output dynamic complex system containing M subsystems $O_1, O_2, \ldots, O_M$. Convenient model for a single input-output system is difference equation. Let:

$$\overline{y}_m(n{+}1) = \Phi_m(\overline{y}_m(n), \ldots, \overline{y}_m(n{-}d_{ym}{-}1), u_m(n), \ldots, u_m(n{-}d_{um}{-}1), a_m) \quad (1)$$

denotes model of m-th element of the complex system, where $u_m(n)$ and $y_m(n)$ are element's input and output, $\overline{y}_m(n)$ is the model output, n – discrete time step, d_{um}, d_{ym} – numbers of previous time steps taken into account by the model for input and output signals respectively, Φ_m – is some, generally nonlinear, function

(model) with parameters a_m. Input, output and model parameters are vectors from respective spaces, i.e.:

$$u_m(n) \in U \subseteq \Re^{S_m}; y_m(n), \overline{y}_m(n) \in Y \subseteq \Re^{L_m}; a_m \in A \subseteq \Re^{R_m}.$$

The structure of complex system is given by connections between inputs and outputs of each element. It is necessary to define which input of a particular element is output of the other element. Let us note that some system inputs are not the outputs of any other element, see Fig. 1. Those inputs are called external inputs – $x(n)$. Furthermore, let us distinguish some global plant outputs $v(n)$, which are important for model application, for example outputs which are used to formulate control problem on the upper level. By joining models Φ_m together according to the system's structure description (which example is showed on Fig. 1) and time dependencies, we obtain the model $\overline{\Phi}$ of complex system as a whole.

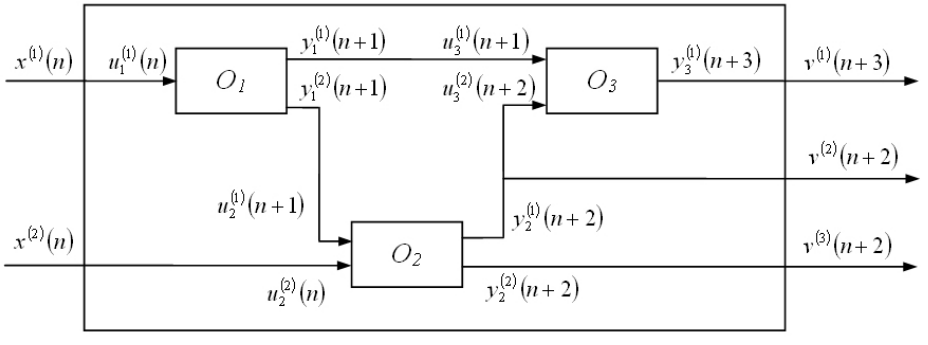


Fig. 1. Example of input-output dynamic complex system

3 Global and Local Modeling Using Neural Networks

Nowadays, scientist focus attention mainly on nonlinear systems, [8] or systems for which relation between their inputs and outputs is hardly sophisticated. Most of classical identification methods cannot handle this case of modeling tasks, so artificial intelligence methods, such as dynamic neural networks, are used instead. In order to determine the optimal parameters of neural network, measurements of input time series and corresponding output time series (selected examples of behavior of the real system) for each element of the systems are collected. Performance index is used as measure of the difference between model outputs time series and respective measurements. Because of complexity of the system, there are different possibilities to define performance index. On the one hand, it is possible to determine an optimal network model for each system element separately (locally optimal model), and on the other hand we can define performance index for the system as a whole (globally optimal model).

Generally speaking, the idea of local approach is to construct separate models for each element of complex system not taking into account connections between them. As we can see, the local approach doesn't need any innovative methods. Problem was precisely formulated in preceding paper [11]. Algorithms useful for local modeling can be found also in work [6]. Another view on the system, i.e. comparison only selected plant output time series with corresponding global network model output time series, defines the globally optimal model. In this case global identification performance index has the form:

$$Q(a) = \sum_{n=1}^{N} q(v(n), \overline{v}(n)), \tag{2}$$

where: $v(n)$ – is a vector of measurements taken from global outputs of complex system at n-th time step, $\overline{v}(n)$ – is corresponding vector worked out from the model, dependent on the parameters $a = [a_1, a_2, \ldots, a_M]$, q – is a measure of difference between $v(n)$ and $\overline{v}(n)$. Globally optimal parameters $\widetilde{a}$ of the complex system's network model are obtained by minimization of the performance index (2) with respect a, i.e.:

$$\widetilde{a} \to Q(\widetilde{a}) = \max_{a} Q(a). \tag{3}$$

Equations (1) for $m = 1, 2, \ldots, M$ with parameters $\widetilde{a}$ are globally optimal network models of the separate elements. After composition in accordance to the system's structure we obtain globally optimal network model of the system as a whole (Fig. 2). Naturally, there are many possible structures of complex systems. Now we focus attention on a very important case of a complex system, namely a system of cascade structure, see Fig. 3. It's significance results from the fact, that to the mentioned structure can be transferred all complex systems without feedbacks [12]. Let us also assume for simplicity, that all elements that are part of the system, are single input and single output plants. As a model of m-th element let's take for illustration simple dynamic linear neuron depicted in Fig. 4 and described by following differential equation:

$$\overline{y}_m(n+1) = \Phi_m(\overline{y}_m(n), u_m(n), a_m) = \varphi(a_m^{(2)}\overline{y}_m(n) + a_m^{(1)}u_m(n)), \tag{4}$$

where for linear neuron $\varphi(\gamma) = \gamma$. In order to construct model of a plant with more inputs and outputs or a plant of higher order, one should connect appropriate number of such a neurons. Quality of neural network used to global modeling of the complex system can be calculated from the specific case of global performance index (2):

$$Q(a) = \frac{1}{2} \sum_{m=1}^{M} \beta_m \sum_{n=m}^{N+m-1} (y_m(n) - \overline{y}_m(n))^2, \tag{5}$$

where β_m, $m = 1, 2, \ldots, M$, are coefficients taken from interval $[0, 1]$ and their task is to control local models quality. As can be seen, neural network learning algorithm can be regarded as an optimization process. It's goal is to find such

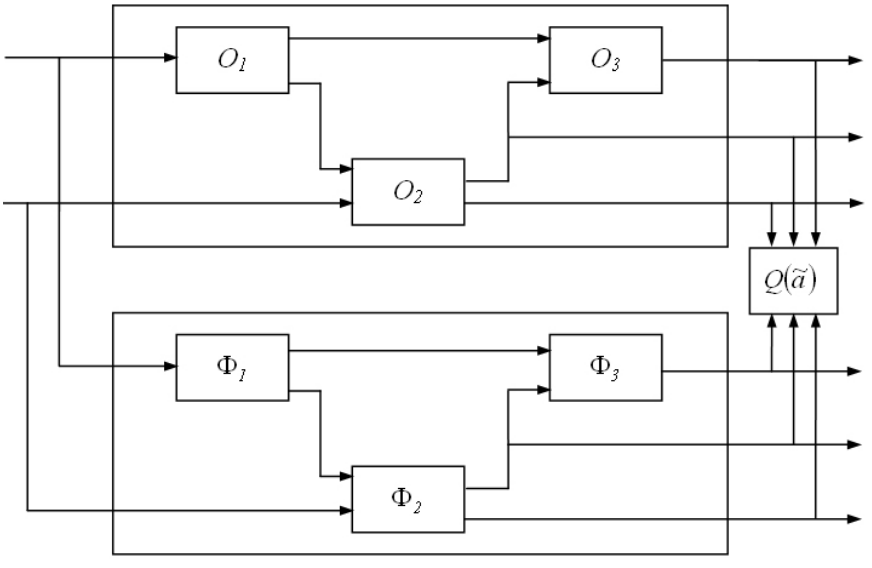


Fig. 2. Complex system with global model

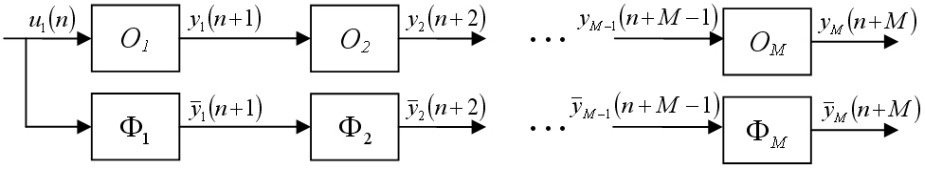


Fig. 3. Complex system with cascade structure and appropriate network model

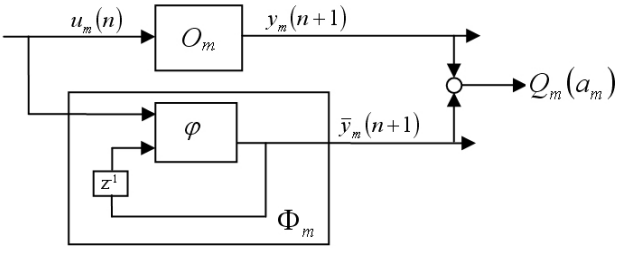


Fig. 4. Dynamic neuron as a model of first-order SISO element

a values of neural network parameters (so called weights), for which a given performance index - (5) in this case - is optimal. Because we assumed to deal with nonlinear systems, only numerical optimization methods can be considered.

One of the most popular one is the steepest gradient descent [3]. According to this method, globally optimal parameters of dynamic neural network can be found by performing iterations described by the formula:

$$a(i+1) = a(i) - \eta \nabla_a Q(a)|_{a=a(i)}, \tag{6}$$

where: i – denotes iteration number, η – is a correction coefficient. Initial solution $a(0)$ is chosen arbitrary. Iterations are performed until assumed convergence criteria is satisfied, [3].

For complex system with cascade structure which modeling quality is calculated from global performance index (5), and for linear neurons, recursive formulas giving gradient of (5) in each iteration step are as follows:

For M-th neuron:

$$\frac{\partial Q(a)}{\partial a_M^{(1)}} = -\beta_M \sum_{n=M}^{N+M-1} (y_M(n) - \overline{y}_M(n)) D_{u_M}(n), \tag{7}$$

where

$$\begin{cases} D_{u_M}(n) = \overline{y}_{M-1}(n-1) + a_M^{(2)} D_{u_M}(n-1) \\ D_{u_M}(M) = \overline{y}_{M-1}(M-1) \end{cases},$$

$$\frac{\partial Q(a)}{\partial a_M^{(2)}} = -\beta_M \sum_{n=M}^{N+M-1} (y_M(n) - \overline{y}_M(n)) D_{y_M}(n), \tag{8}$$

where

$$\begin{cases} D_{u_M}(n) = \overline{y}_M(n-1) + a_M^{(2)} D_{y_M}(n-1) \\ D_{y_M}(M) = 0 \end{cases}.$$

For m-th neuron ($m = 1, 2, \ldots, M-1$):

$$\frac{\partial Q(a)}{\partial a_m^{(1)}} = -\beta_m \sum_{n=m}^{N+m-1} (y_m(n) - \overline{y}_m(n)) D_{u_m}(n) \quad +$$

$$- \sum_{j=m+1}^{M} \beta_j \sum_{n=j}^{N+j-1} (y_j(n) - \overline{y}_j(n)) \left[\prod_{v=m+1}^{j} a_v^{(1)} \right] D_{u_m}(n-j+1), \tag{9}$$

where

$$\begin{cases} D_{u_m}(n) = \overline{y}_{m-1}(n-1) + a_m^{(2)} D_{u_m}(n-1) \\ D_{u_m}(m) = \overline{y}_{m-1}(m-1) \end{cases},$$

$$\frac{\partial Q(a)}{\partial a_m^{(2)}} = -\beta_m \sum_{n=m}^{N+m-1} (y_m(n) - \overline{y}_m(n)) D_{y_m}(n) \quad +$$

$$- \sum_{j=m+1}^{M} \beta_j \sum_{n=j}^{N+j-1} (y_j(n) - \overline{y}_j(n)) \left[\prod_{\upsilon=m+1}^{j} a_{\upsilon}^{(1)} \right] D_{y_m}(n-j+1), \tag{10}$$

where

$$\begin{cases} D_{y_m}(n) = \overline{y}_m(n-1) + a_m^{(2)} D_{y_m}(n-1) \\ D_{y_m}(m) = 0 \end{cases}.$$

4 Final Remarks

As mentioned before, models of complex input-output systems and multiagent systems share very important general features, it seems. Firstly, both are composed of many entities which can operate separately. Secondly, all these subelements can interact, resulting in the structure of higher order of complexity. Thirdly, agents gain knowledge about the environment inside which they act int the form of internal parameters called state [14], models Φ_m capture knowledge about real objects in form of parameters a_m during so-called learning process (which is in fact an iterative optimization algorithm).

Fourthly, agents are capable of exchanging knowledge, so as neural networks Φ_m. The question arises, how it is performed by neural networks connected in the complex system's model? Algorithm described by formulas (8)-(11) can be regarded as an equivalent of backpropagation algorithm for complex systems (but it is'not the same). It's important quality is that calculations can be distributed spatially in a very characteristic manner. Some neural networks Φ_m are able to calculate error they make as a models (M-th neural network in the example from section 3), which results from possibility of comparing directly the desired outputs with the one's generated as a response for the same input signals. The same neural networks are also able to work out and send back the error signal for networks, which outputs are their own inputs, and so on. Algorithm organized in such a way, can work properly even if models are spatially distributed. Each model can make calculations useful for other models independently, and send them to the destination place. Error signals are used to adjust values of neural network's parameters a_m. From this point of view, neural networks Φ_m act as agents gaining knowledge about environment and exchanging useful information.

There is, however, one important difference: connections between agents changes in the sense of information exchange, while in this work connection between element of complex systems were assumed to be fixed. But general idea of local and global modeling still holds. Analogy between multiagent systems and models of complex system is a very general concept, but it seems that it's worth to look closer to the details in the future works.

References

1. Brzostowski, K., Świątek, J.: Adaptive Control Algorithm with Two-Stage Identification. In: Proc. of 18th International Conference on Systems Engineering, pp. 55–58. Coventry University, England (2006)

2. Bubnicki, Z.: Identification of Control Plants. Elsevier, Oxford, New York (1980)
3. Chong, K.P., Zak, S.H.: An Introduction to Optimization. Series in Discrete Mathematics and Optimization. John Wiley & Sons, Inc., West Sussex, England (2001)
4. Drałus, G., Świątek, J.: Global Modeling of Complex Systems by Neural Networks. In: Proc. of 7th International Symposium on Artificial Life and Robotics, Oita, Japan, pp. 618–621 (2002)
5. Drapała, J., Świątek, J.: Modeling of Dynamic Complex Systems by Neural Networks. In: Proc. of 18th International Conference on Systems Engineering, pp. 109–112. Coventry University, England (2006)
6. Gupta, M.M., Jin, L., Homma, N.: Static and Dynamic Neural Networks - From Fundamentals to Advanced Theory. IEEE Press, Wiley- Interscience, A John Wiley & Sons, Inc. (2003)
7. Hasiewicz, Z.: Identification of Control System with Composite Structure. Scientific Papers of the Institute of Control and systems Engineering of the Wrocaw Technical University (1993)
8. Nelles, O.: Nonlinear System Identification - From Classical Approaches to Neural Networks and Fuzzy Models. Springer, Heidelberg (2001)
9. Świątek, J.: Two stage identification and its technical and biomedical applications. Scientific Papers of the Institute of Control and systems Engineering of the Wrocaw Technical University (1987)
10. Świątek, J., Drałus, G.: Global Network Modelling of Complex Systems with Respect to Local Models Quality. In: Świątek, J. (ed.) Proc. of 15th International Conference on Systems Engineering, Las Vegas, Nevada, pp. 218–226 (2002)
11. Świątek, J.: Global and Local Modelling of Complex Input-Output Systems. In: Proc. of 16th International Conference on Systems Engineering, pp. 669–671. Coventry University, England (2003)
12. Świątek, J.: Global Identification of Complex Systems with Cascade Structure. In: Rutkowski, L., Siekmann, J.H., Tadeusiewicz, R., Zadeh, L.A. (eds.) ICAISC 2004. LNCS (LNAI), vol. 3070, pp. 990–995. Springer, Heidelberg (2004)
13. Wooldridge, M.: An Introduction to Multiagent Systems. John Wiley & Sons, Ltd., West Sussex, England (2002)
14. Yoshida, T.: Cooperation Learning in Multi-Agent Systems with Annotation and Reward. International Journal of Knowledge-based and Intelligent Engineering Systems 11(1), 19–34 (2007)

A Content Dependent Visualization System for Symbolic Representation of Piano Stream

Alexander Adli[1], Zensho Nakao[2], and Yasunori Nagata[2]

[1] Graduate School of Engineering & Science, University of the Ryukyus, Okinawa 903-0213, Japan
[2] Department of Electrical & Electronics Engineering, University of the Ryukyus, Okinawa 903-0213, Japan
{alex,nakao}@augusta.eee.u-ryukyu.ac.jp

Abstract. This paper provides an overview on the advances of music information retrieval in symbolic representation of music. Such musical aspects as key, tonality, bass, melody, dynamics, rhythm and patterns are considered as foundation for visualizing systems for the piano stream in classical music. The paper then describes the proposed visualizing system and Malinowski's music animation machine. It lays light on the challenges facing creating contemporary visualizing systems. It is supplied with a related references list for further study on the issue.

Keywords: Music content analysis, Visualization, MIDI, piano stream, MIR.

1 Introduction

Most contemporary popular music player software (Windows Media Player, Winamp, iTune, etc.) contain a visualization module which provides a real time graphical complement for the performed digital music. These visualization modules rely mainly on amplitude and frequency information available in the audio files and use it as source information for the graphical module which usually performs complicated actions and creates complex patterns. In spite of this graphical complexity, it does not reflect the semantic of the music, and does not consider such musical concepts like tonality, bass and melody. A music information retrieval (MIR) system in audio files for those aspects is computationally expensive and still far from being able to satisfy the users (listeners/watchers).

The MIR community applies a lot of effort on analysis the musical contents both in digital audio and in symbolic representations. The symbolic representation like the MIDI files differs from the former in terms that it explicitly provides the information about the pedals, notes, and their timing. Different musical instruments are also divided into channels or tracks [9].

These characteristics of MIDI files makes them more suitable domain for content analysis than digital audio files and as a result visualizing systems for MIDI at the moment can be potentially more related to the musical content than their digital audio analogues [1]. Although MIDI files are popular, they are mostly used for composing

B. Apolloni et al. (Eds.): KES 2007/ WIRN 2007, Part II, LNAI 4693, pp. 287–294, 2007.
© Springer-Verlag Berlin Heidelberg 2007

aims or in musical databases but rarely among end-users for listening purposes. This is because MIDI requires high quality synthesizers which are too expensive or simply not good enough as their acoustic instruments equivalent.

As the quality of software synthesizers and the computing power of computers are rapidly improving, MIDI music listener auditory is expected to increase. A visualizing module for MIDI files based on musical content analysis might become am interesting contribution to the music industry.

In this paper the authors shed light on some of the music content analysis methods and techniques for MIDI; we will discuss such musical concepts such as key, tonality, melody and dynamics and their retrieval from MIDI data. In latter sections authors show how information about musical content can benefit visualizing systems in two examples: the authors' MIDI visualizer, and Malinowski Music Animation Machine (MAM). This is followed by conclusions and further research discussion, acknowledgment, and references to related works to both of visualizing music and content analysis.

2 Music Content Analysis in MIDI Files

The MIDI representation of music contains symbolic information of performance events: pedals and notes (onset time, pitch, duration). In this sense it can be compared to musical score and hence, it can be analyzed by tools and rules of theory of music. There are however many differences between the musical score and MIDI; the key is not necessarily known in MIDI representation, also notes like C# and D are indistinguishable. MIDI is a performance descriptor and not musical score for the musicians; this is why it is potential musical source for the end-user. We mentioned previously in the introduction the reasons why it is still not popular for this purpose.

A visualization system which can reflect graphically much of the musical contents in a way which makes sense for the listener/watcher must be able to extract musical aspects like key, tonality, bass, melody line, dynamics, rhythm, patterns, form, genre, emotions, etc. Behind every aspect there is at least few a algorithms and competitive methods for better extracting.

The authors will discuss the first seven aspects mentioned here and then give examples of implementing them into a visualizing system. They are discussed from the point of view of classical genre and single instrument – the piano.

2.1 Key and Tonality

Almost all classical tonal pieces are composed using one of the existed 24 tonalities of the diatonic music. This tonality is called the key. Tonality is one of the main musical attributes along with the meter, melody and harmony. The theory of music defines tonality as the hierarchy of importance that exists between tones in musical piece [12].

Tonality dynamically changes via progressions and modulations while the key rarely changes. The key mode (major or minor) reflects the overall "mood" of the piece. Parncutt [16] created a method for key detection in a sequence of chords. Later many techniques appeared for key detection by analyzing the composition beforehand

and determining the key as the most frequent tonality with its tonic sound frequently appearing. The authors of this paper suggested a method for tonality recognition dynamically. This method is designed for monophonic piano stream in MIDI file. It divides the stream basing mainly on pitch salience values into segments where every segment represents a single tonality [8].

MIDI files contain a meta event command where the key information (meta command 59) can be stored. In case this information is available, visualizer can use it, and apply the tonality methods to detect tonality progressions and modulations. If not, then it has to apply some actions to determine the key. Fig.1 shows the tonalities in the famous Beethoven's "For Elise".
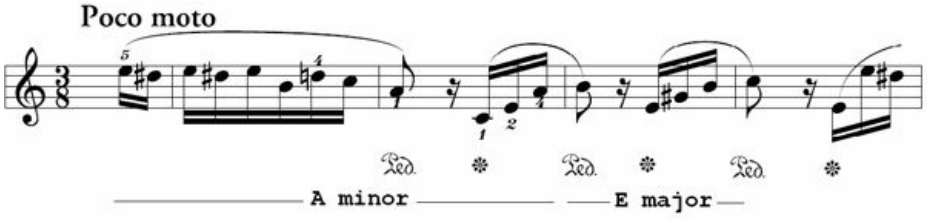


Fig. 1. Beethoven: For Elise. The piece is composed in A minor (the key). In this fragment the composition starts with its key tonality A minor, and then progresses to E major.

2.2 Bass and Melody Line

Basses are the foundation of the harmony, and usually are the lowest pitch in the harmonic line. Melody is the main part of the music, and it is the part people would sing if they were asked to express the music vocally. Most bass recognition algorithms are based on the assumption that the bass note is lowest than its neighbors. The authors in their work about tonality [8] also suggested an algorithm for bass recognition in piano stream based on the same principle.

Melody extraction is a difficult task. Uitdenbogerd and Zobel proposed the skyline technique [13]; it selects the highest pitch line of the note sequence into melody. However, even with later revisions of the technique such like in [14], it still fails to recognize complicated structures of melody or simply when the melody is in the lowest or middle pitches. Chopin's fragment in Fig. 2 shows the melody line where notes are circled. Skyline algorithm successfully classifies the higher notes as melodic. The lower notes "E" and "B" shown by arrows are basses and each one represents a harmony of E major and B major tonalities respectively.

Another techniques are based on choosing the best MIDI channel which contains the melody [2][13]. Usually the melody channel has louder level than the average of all channels, less average distance between pitches, less rhythmic and so on. However, this technique is not much useful in classical piano music, as it is seldom in such music to divide into melody and harmony channels especially when both are played on the piano.

Recently a psychoacoustics models for melody extraction are in discussion.
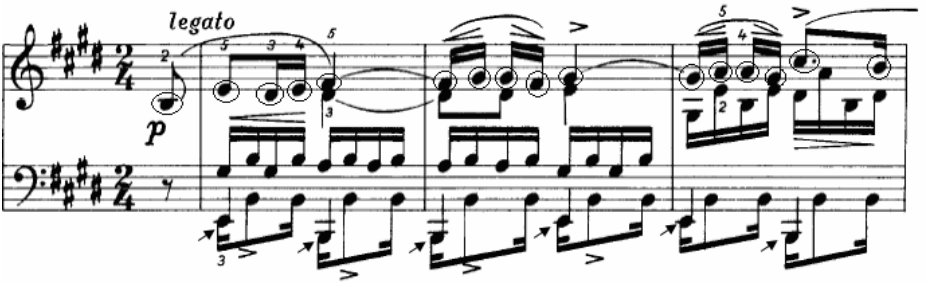


Fig. 2. Chopin: Etude No 3 in E major. The melody notes are circled. Arrowed notes are basses.

2.3 Dynamics, Rhythm and Patterns

Although MIDI files are considered as a relatively convenient domain for the musical contents analysis, the loudness aspects of music or dynamics are relatively difficult to extract. In digital sound there are good deterministic methods for loudness extraction from sound wave amplitude [15]. In MIDI files, the loudness is reflected in the velocity parameter of the note which is related to the speed of the piano key pressing. The authors studied the factors affecting the loudness in piano [7][8]. The loudness can be estimated from notes velocities, pitch and piano amplitude envelope. The loudness then reflects the dynamics levels of the music (forte, piano, crescendo and diminuendo).

Rhythm is the variation of length of the notes. There are two more relative aspects here: The meter, and the tempo. The rhythm is important as it reflects such features of music like accelerando and ritenuta. In 1983 Lerdahl and Jackendoff published their text, A Generative Theory of Tonal Music (GTTM). Since that time many methods based of music segmentation appeared for rhythm extraction.

The music segmentation however, depends on another feature of the music – the patterns. The pattern recognition in music leads to segmentations. In Fig. 3 the musical fragment can be classified into three parts with single pattern. The first part consists of 9 notes in C major starting in C which is the bass note here. The second and third repetitions repeat the note intervals of the first part but in different tonalities (G major and F major). Algorithms based on segmentation division by note interval repetition can extract the rhythm of the fragment in Fig. 3 successfully. However, in classical music segmentation and pattern recognition is more complicated process. Robert Rowe in his book machine musicianship [17] describes advanced methods for music segmentation based on neural networks techniques.
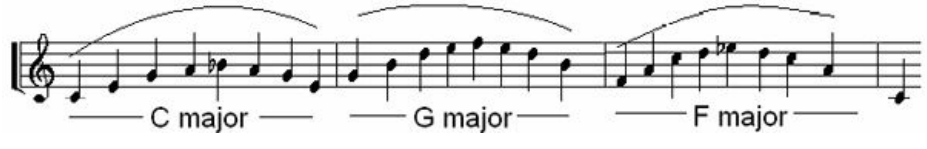


Fig. 3. Musical pattern been repeated in different tonalities

3 MIDI Visualizing Systems

Visualizing music is a creative process. There might be infinite ideas of representing the contents of music via graphics. As we mentioned before, most of the popular contemporary players depends on the sound wave but not on the musical context. In this section we will describe two visualizers for MIDI files based on musical content (The authors visualizer and Malinowski Visualizer). We hope that describing the visualizing ideas here might inspire many graphic and music researchers and programmers to create their own original creative visualizing systems.

3.1 Author's MIDI Visualizer

This visualizing system was created cooperatively by Adli and Ermakov. The system can receive MIDI messages from MIDI files as well as directly from MIDI device like digital piano.

The visualizing idea is a space flight in real time with the music. Every note causes a small star to appear (fig. 4.) exactly in the onset time of the note. The small star starts move toward the user (in a 3D space) which makes the user feels as if he/she is flying forward. Bass notes cause bigger rotating stars which move exactly like the previous ones. The color of the small stars reflects the current tonality.

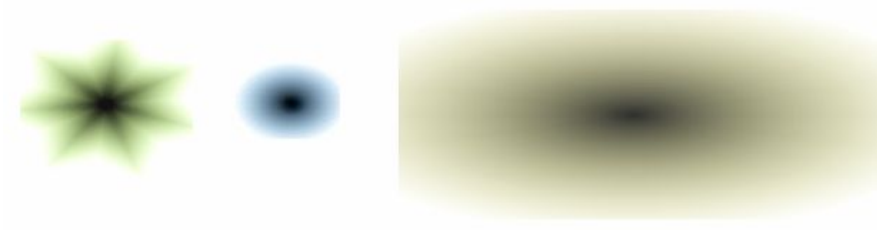


Fig. 4. The shapes of MIDI visualizer. Bass star (Left), small star (middle), explosion (right). As the original background is black, all colors are inverted.

Minor tonalities are associated with dark and gloomy colors while the major ones are associated with bright and cheerful colors. The color and rotation speed of the bass stars depends on the pitch of the bass. The lower the bass the slower the rotation speed is. Chords when happen cause an explosion. The color of this explosion depends on tonality while the size depends on the overall loudness of all the notes of the chord. The dynamics is represented by the size of the bass stars; for quite music bass stars are small and vice versa. Rhythm is reflected by the speed of the flight. And when the system knows the key tonality it adds a trembling feature for all the objects whenever the dominant tonality appears to reflect the "anxious" feeling frequently provoked by this tonality.

This visualizer is a real time one. And although it uses simplified algorithms for the content analysis, non-expert users noticed that the visualization performed by our visualizer were "closer" to the musical context than other popular visualization

modules which were basing on frequency and amplitude extraction. The graphical and musical qualities however, were not necessarily better.

3.2 Malinowski Music Animation Machine (MAM)

Stephen Malinowski during many years created his own algorithms and own visulalizer which he called MAM [4]. MAM handles MIDI messages from file or directly via MIDI port. As for 2007 this MAM can visualize MIDI in 12 different ways. Below we will describe one of them.

Every note is represented by a circle (fig. 5). The "now line" divides the notes into two groups: the notes in the past are to the left of the line, and the future notes are to the right. The now line advances and the notes are played once the now line crosses the center of their circles. As X axis is the time axis, Y axis reflects the note number. The lower notes are located at the bottom of the screen at vice versa. The color of the circles reflects the pitch class of the note while the size reflects its duration. Once the now line "hits" the circle, its inner part starts following the line and decreases in size gradually until it disappears (both graphically and as a sound). The melody line and harmony line are also separated.

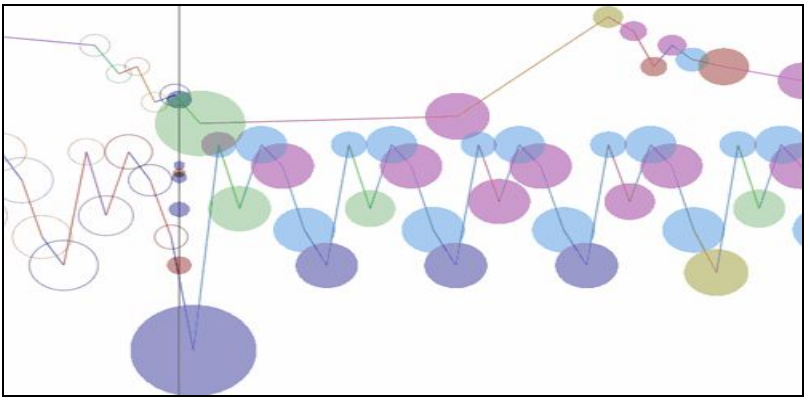


Fig. 5. The so called "part motion" visualization of Malinowski's MAM. It might be difficult to distinguish the different colors of the circles in grayscale.

3.3 Challenges in Visualizing Systems

Real time music information retrieval system in MIDI is frequently late in detecting musical aspects exactly in time of their occurrence. However in most cases music is beforehand stored in a MIDI file, and the visualizing system can first analyze the music in advance.

As the MIDI files are not digital audio files, they rarely can produce high quality sounds on average platforms. There are however, many hybrid musical standards where music is stored in both audio and symbolic way. Such files can serve as good domain where the music is performed from the audio data, and the visualization uses the symbolic data.

Visualizing system can be used for other purposes like visual music representation, pattern recognition for performers or genres. For example the performance worm by Goebl et. al. [6] draws different shape "worms" in tempo-loudness space which helps to distinguish performers.

Visualizing systems are encouraged to create dynamically their own graphical patterns in real time relevant to the performed music.

4 Conclusions and Further Work

MIDI files are expected to become more popular among the end users due to the improvements of both software and hardware synthesizers. Symbolic analysis of music in MIDI can provide more accurate information about the musical aspects like tonality, melody line and rhythm with less computational requirements if compared with the analysis in audio files.

Music visualizing systems are also becoming popular. The created visualization might be an aim by itself, or it can be used for different music classification applications. Visualizations in audio files are limited to amplitude and frequency information. There are however, many graphically advanced visualizators such as some Winamp plug-ins. Visualizations based on MIDI files can create visual patterns closely related to musical context as the musical information can be explicitly or implicitly obtained. The "MIDI visualizer" and "Music Animation Machine" are examples of such visualizing systems.

The authors wanted to invoke the MIR community creativity toward the potential visualizing possibility of music based on its contents.

Acknowledgments. The authors of this paper would like to express their gratitude to Ermakov A. who made a lot of efforts to make our MIDI visualizer look more beautiful.

References

1. Isaacson, E.: What You See Is What You Get: On Visualizing Music. In: Proceedings of International Symposium in Music Information Retrieval, London (2005)
2. Ozcan, G., Isikhan, C., Alpkocak, A.: Melody Extraction on MIDI music Files. In: Proceedings of the seventh IEEE International Symposium on Multimedia (2005)
3. Li, T., Ogihara, M.: Detecting Emotions in Music. In: Proceedings of International Symposium in Music Information Retrieval (2003)
4. Malinowski, S.: Music Animation Machine (2005), http://www.well.com/user/smalin/mam.html
5. Adli, A., Nakao, Z., Nagata, Y.: Calculating the expected sound intensity level for solo piano sound in MIDI files. In: Proceedings of the SCIS & ISIS, Tokyo Institute of Technology, Tokyo (September 2006)
6. Langner, J., Goebl, W.: Visualizing Expressive Performance in Tempo-Loudness Space. Computer Music Journal 27(4), 69–83 (2003)

7. Adli, A., Nakao, Z., Yokoda, T., Nagata, Y.: Piano Sound Characteristics: A Study On Some Factors Affecting Loudnes. In: Digital And Acoustic Pianos, ICICIC, Japan (2007) (in print)
8. Adli, A., Nakao, Z., Nagata, Y.: A Score Based Algorithm for Tonality Recognition in Monophonic Piano Streams (submitted)
9. MIDI Manufacturers Association: http://www.midi.org
10. Martin, K.D., Scheirer, E.D., Vercoe, B.L.: Music Content Analysis Through Models of Audition. In: Proceedings of the ACM Multimedia Workshop on Content Processing of Music for Multimedia Applications, Bristol, UK, ACM Press, New York (1998)
11. Rizo, D., de Le'on, P.P., P'erez-Sancho, C., Pertusa, A., Inesta, J.M.: A Pattern Recognition Approach for Melody Track Selection in MIDI Files. In: Proceedings of International Symposium in Music Information Retrieval (2006)
12. Levitin, D.J.: This is your brain on Music, Dutton (2006)
13. Uitdenbogerd, A., Zobel, J.: Manipulation of Music for Melody Matching. In: Proceedings of ACM Internaional Multimedia Conference, ACM Press, New York (1998)
14. Chai, W.: Melody Retrieval on the web, MS Thesis, Massachusetts Institute of Technology, Boston (2000)
15. Fastl, H., Zwicker, E.: Psychoacoustics: Facts and Models, Springer, 3rd edn. Springer, Heidelberg (2006)
16. Parncutt, R.: Revision of Terhardt 's psychoacoustical model for the roots of a musical chord. Music perception 6, 65–94 (1988)
17. Rowe, R.: Machine Musicianship. The Mit Press, Cambridge (2001)

Audio Watermarking Based on the Psychoacoustic Model and Modulated Complex Lapped Transform

Hanane H. Mirza, Alexander Adli, Hien D. Thai, Yasunori Nagata, and Zensho Nakao

Department of Electrical and Electronics Engineering, University of the Ryukyus, Okinawa 903-0213, Japan
{hanane,alex,tdhien,nakao}@augusta.eee.u-ryukyu.ac.jp

Abstract. The music industry suffers a multi-billion dollar annual revenue loss due to piracy. Audio watermarking technology is an important technique that is gaining widespread use. However, one of the challenges facing watermarkers is the high sensitivity of the Human Auditory System (HAS). In this paper, we present a novel audio spread spectrum watermarking scheme where the watermark is extracted from the original file.By using the psychoacoustic model this technique efficiently takes advantages of masking phenomena in HAS, in order to embed an important amount of watermark data below the masking threshold of the original input signal. The watermarking is performed in the Modified Complex Lapped Transform (MCLT) domain, which presents a good loss-less reconstruction for audio reconstruction. The computer simulation shows a good imperceptibility and a high robustness to a wide range of unintended and intended attacks.

Keywords: Audio watermarking, Psychoacoustic system, MCLT.

1 Introduction

Today's digital media have opened the door to an information marketplace where the true value of the product (digital content) is dissociated from any particular physical medium. It enables a greater degree of flexibility in its distribution and at a lower cost. Also, digital data can be duplicated and re-distributed at virtually no cost. One possible solution to this problem is the audio watermarking technique. This involves the addition of an imperceptible and statistically undetectable signature to multimedia content. From the principle of audio watermarking it is apparent that the embedded watermarking should be maximally resistant to common methods of signal processing and, at the same time, it should not change the quality of the original audio file. Also the amount of time necessary for embedding and extracting the watermark should be as short as possible [8]. A number of the presented schemes rely on the imperfections of the

B. Apolloni et al. (Eds.): KES 2007/ WIRN 2007, Part II, LNAI 4693, pp. 295–302, 2007.
© Springer-Verlag Berlin Heidelberg 2007

Human Auditory System (HAS). Zezula et al[6] exploits the magnitude coefficients of the MCLT domain as previously exhibited by Malvar [1].The MCLT is considered as the appropriate transform for audio signals as it presents a very low loss in the reconstruction phase. Our new robust method for digital audio watermarking is based on the psychoacoustic Model(PAFM)[7][9][5][4], which is used twice in the algorithm; First, the PAFM is applied directly to the audio signal in order to determine the threshold of tonal and non-tonal frequencies. By adding the noise-like (non-tonal) components we generate wave of noise that will serve as watermark data. After generating the watermark data, the input signal is transformed to the MCLT domain and the PAFM is used to determine the masking capacity for each MCLT subband [2]. As for detection process, we chose to pre-correlate the algorithm using a whitening function [3][8]. In order to test the imperceptibility of the algorithm and its robustness, the Signal to Noise Ratio (SNR) of the watermarked signal was calculated and a number of attacks were applied to the marked file.

2 Background

2.1 Modulated Complex Lapped Transform

The modulated complex lapped transform(MCLT) has been applied to noise reduction and echo cancelation with promising results[1]; it has been proposed as a very simple extension to the MLT that includes in addition, a sine modulated function, which form a transform with explicit phase information. The MCLT transform coefficients $X(k)$ are computed from the input signal block $x(n)$ with length $2M$ by:

$$X(k) = \sum_{n=0}^{2M-1} x(n)p_a(n,k) \tag{1}$$

For every M real valued input sample MCLT computes M complex frequency components, and the basis functions for analysis are introduced by:

$$p_a(n,k) = p_a^c(n,k) + jp_a^s(n,k)$$

with $j \equiv \sqrt{-1}$ and

$$p_a^c(n,k) = h_a(n)\sqrt{\frac{2}{M}}\cos[(n+\frac{M+1}{2})(k+\frac{1}{2})\frac{\pi}{M}]$$

$$p_a^s(n,k) = h_a(n)\sqrt{\frac{2}{M}}\sin[(n+\frac{M+1}{2})(k+\frac{1}{2})\frac{\pi}{M}]$$

From the above analysis we can see that the MCLT in fact is similar to a 2x oversampled DFT filter bank(using a doubly odd DFT[10]instead of the traditional DFT). The MCLT basis functions for the synthesis are computed as:

$$y_c(n) = \sum_{k=0}^{M-1} Y(k)p_s(n,k) \tag{2}$$

where:

$$p_s(n,k) = \frac{1}{2}[p_s^c(n,k) - jp_s^s(n,k)]$$

and:

$$p_s^c(n,k) = h_s(n)\sqrt{\frac{2}{M}}\cos[(n+\frac{M+1}{2})(k+\frac{1}{2})\frac{\pi}{M}]$$

$$p_s^s(n,k) = h_s(n)\sqrt{\frac{2}{M}}\sin[(n+\frac{M+1}{2})(k+\frac{1}{2})\frac{\pi}{M}]$$

with:

$$h_a(n) = h_s(n) = -\sin[(n+\frac{1}{2})\frac{\pi}{2M}]$$

The fact that the MCLT coefficient carry at the same time the real and imaginary parts which make it a good transform for audio theoretically since there is no information loss in the audio files reconstruction.

2.2 Psychoacoustic Frequency Model(PAFM)

The human ear acts as a frequency analyzer and can detect sounds with frequencies varying from 10 to 20kHz. The HAS can be modeled by a set of 26 band pass-filters called the critical bands. These are defined around a center frequency in which the noise band increases until there is a just noticeable difference in the tone at the center frequency. Thus if a faint tone lies in the critical band of a louder tone, the faint tone will not be perceptible. The psychoacoustic model analyzes the audio signal in order to account for the amount of noise masking available depending on its frequency and loudness. The psychoacoustic model used in this paper is defined in ISO-MPEG Audio psychoacoustic Model 2, and it operates through 4 steps:

Step 1: Align audio data in time, segment it into audio blocks, and calculate the spectrum value (Fig.1) for each block, with length N=1024 samples weighted by Hann window, $h(n)$:

$$h(n) = \frac{\sqrt{\frac{8}{3}}}{2}[1-\cos(\frac{2\pi n}{N})]$$

The power spectrum of the signal is calculated as:

$$S(k) = 10\log_{10}[\frac{1}{N}\|\sum_{n=0}^{N-1} s(n)h(n)exp(-j2\pi n\frac{k}{N})\|^2]$$

By processing the spectral values in groupings related to critical band widths to simplify calculation, the spectral values are perceptually grouped according to the bandwidth.

Step 2: Compute the tonality index as a function of frequency. This index gives a measure of whether the component is more tone-like or noise-like. This helps

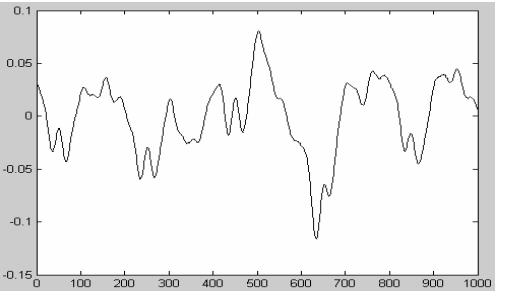


Fig. 1. Spectrum value

to interpolate between pure tone-masking-noise and noise-masking-tone values since the masking ability is different for each.

Step 3: Determine the masking threshold. First, spread the masking ability around critical bands and determine a lower bound for the threshold values in the quiet. This threshold is the lower bound of the audibility of sound. Then find the masking threshold for each subband. The PAFM selects the minimum of the masking thresholds covered by the subband only where the band is wide relative to the critical band in that frequency region.

Step 4: Calculate the signal to mask ratio (SNR). It computes the signal-to-mask ratio as the ratio of the signal energy within the subband to the minimum masking threshold for that subband.

3 Proposed Algorithm

3.1 Watermark Generation

1- Apply PAFM (Step1 through Step4) to the original signal in order to generate the appropriate watermark noise signal.

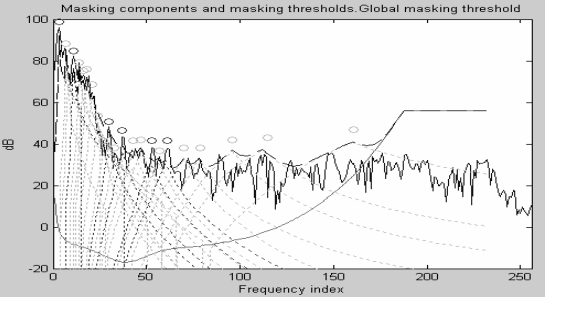


Fig. 2. Psychoacoustic Frequency Model

2- Generate the masking threshold for each audio block, and determine the noise-like components according to the representation in the original signal $x(n)$, and shape it into a noise-like wave signal w, where $w_i \in \{\pm 2.5\} dB$ with same length as the original signal.

3.2 Embedding Process

The embedding process as shown in Fig. 3 consists of the following steps:

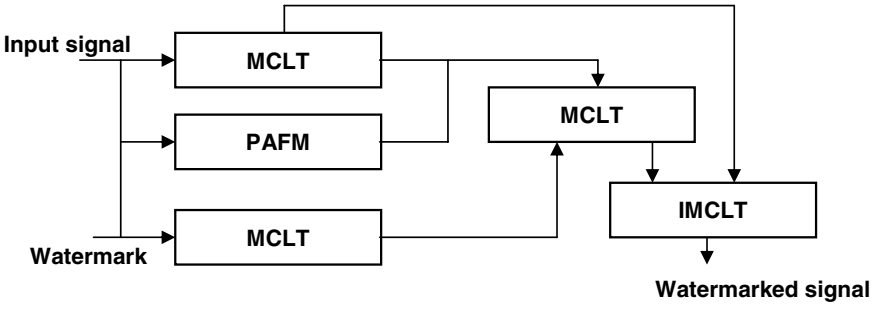


Fig. 3. Embedding process

1- The generated watermark is cut to subbands of length N=4096 and each frame is transformed to MCLT domain (Eq. 1).
2- The original signal into frames of length N=4096. And apply MCLT transform for each frame.
3- Use the PAFM individual masking threshold for each block . The general masking threshold is shown in (Fig 2) as tonal and non tonal component. This allows us to conclude the value of α_i for each block.
4- The addition of both MCLT coefficients(Fig.2, bottom) is happening in the magnitude domain by the formula:$y_i = x_i + \alpha_i w_i$ where $\alpha \in 0.5, 1.5$
5- The recovered signal (Fig.5) is obtained by the Inverse MCLT (Eq. 2).
6- In order to test the imperceptibility of the algorithm we calculate the signal to noise ratio:

$$SNR = 10log \frac{\sum_{i=1}^{n} X^2(i)}{\sum_{i=1}^{n} [X(i) - Y(i)]^2} \tag{3}$$

3.3 Detection Process

In the watermarking detection module it is necessary to consider the desired performance and robustness. The watermarking scheme proposed in this paper (Fig. 4) uses a blind detector and is done by manipulating the following functions for an under test signal Y^*:

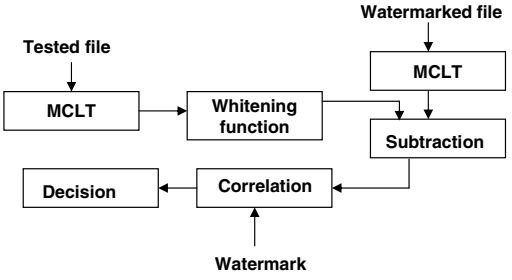


Fig. 4. Detection process

1. Whitening function: The tested audio file is transformed to MCLT domain, and because the MCLT frames are short and the watermarks are short, a sequence in consecutive chips could possibly cause false alarm problems and because our watermark is highly correlated, thus, a whitening function is required to solve this problem. To accomplish this task we choose to work with a Mean filter, which allows us to keep a low computing complexity and gain satisfactory results with correlation procedure. The whitening filter was applied to each MCLT frame.
2. Get the error E between the filtered frames and the original watermarked MCLT frames, $E[y^*.y]$ where $E[y^*.y] = 0$ if the watermarked frame wasn't attacked or modified and $E[y^*.y] = e$, otherwise.
3. Compute the correlation C between the error and the watermark w : Lets L be the length of the watermark w, then $C = \frac{1}{L}\sum_{i=1}^{L} e_i w_i$ with threshold $T = 0.39$.

4 Computer Simulation

The computer evaluation, five different music files (classic , pop, country, blues and folk)of wave format of 20 second each all sampled at $44.1KHZ$ stereo, were marked with the generated $w_i \in \{\pm 2.5\}dB$ noise-like signal (as explained in section 3.1 and 3.2). Figure 5 illustrates the results for classic music file. As for the detection process (section 3.3), The detector response is shown in Fig. 6, the threshold was set at 0.39, and all files passed the threshold to prove the presence of the watermark. Before applying the common signal distortion to the five audio files the watermark was extracted between 0.99 and 1.00 accuracy. In order to check the robustness of our algorithm, the watermarked signals were subjected to some of the common signal processing especially: Mp3 compression, resampling, filtering, noise and echo addition, the SNR results and the watermark extraction accuracy is given in Table1. The SNR results are higher for the blues and the classic music files, due to the file's structure that maintains the power of the noise-like introduced signal at a very low level.

Table 2 shows a comparison in term of imperceptibility calculated by SNR, between the technique introduced in this paper and the previously proposed techniques for audio watermarking.where :

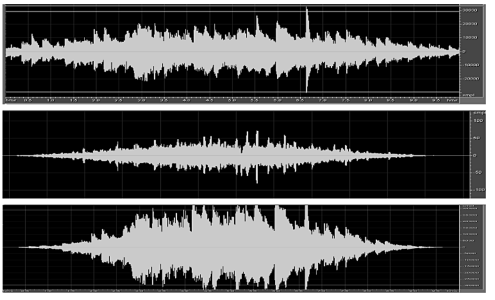


Fig. 5. Original signal(top), generated watermark (middle) and watermarked signal(bottom) for classic wave

Table 1. Simulation results

Attacks/file	Classic	Pop	Blues	Country	Folk
SNR	32.1	29.5	31.1	29.2	26.8
Noise addition (10 sec. of white noise)	0.94	0.82	0.71	0.70	0.68
Echo addition (100ms with 40% decay)	0.85	0.75	0.65	0.68	0.61
MP3	0.58	0.42	0.48	0.50	0.41
Band-pass filter (cutoff 100HZ and 6kHZ frequencies)	0.91	0.88	0.71	0.75	0.68
Resampling (44.1kHZ-22.05kHZ-44.1kHZ)	0.58	0.51	0.45	0.40	0.51

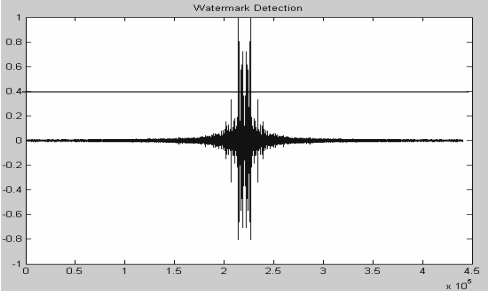


Fig. 6. Detector response

Table 2. SNR compariason with different watermarking algorithms

File	Method A	Method B	Method C	Method D
Classic	32.1	21.0	18.0	14.5
Pop	29.5	22.0	16.0	12.25
Blues	31.1	23.5	19.1	13.05
Country	29.2	21.1	17.2	12.47
Folk	26.8	23.0	17.1	12.47

Method A: Watermarking based on MCLT and Psychoacoustic Model.(Current technique)
Method B: Spread spectrum without use of MCLT.
Method C: Audio watermarking Using PAFM without MCLT.
Method D: MCLT with Aid of 2D pattern (No use of PAFM).

5 Conclusion

A combination between MCLT and PAFM audio watermarking algorithm was presented. Quality evaluation for five different files was made and the results showed that the watermarked signal is quite close to the original signal in terms of quality. PAFM was useful to generate the appropriate watermark and also to determine in which MCLT coefficients we can embed using a right value of the robustness factor. The algorithm showed the robustness to several well-known attacks. However, Mp3 compression still poses a challenge. A comparison with several previous techniques led to the conclusion that the proposed technique gives the highest imperceptibility value of watermarked files.

References

1. Malvar, H.: A modulated Complex Lapped Transform And Its Applications to Audio Processing IEEE. In: ICASSP, Phoenix (March 1999)
2. Kirovski, D., Malvar, H.: Spread-Spectrum Watermarking of Audio Signals. IEEE Transactions on signal Processing (Special Issue on Data Hiding, Phoenix) (2002)
3. Garcia Hernandez, J.J, Miyatake, M.N., Meana, H.P.: Real-Time MCLT Audio Watermarking and Comparison of Several Whitening Methods in Receptor Side. In: Proceeding of the Eighth IEEE International Symposium on Multimedia, IEEE Computer Society Press, Los Alamitos (2006)
4. Davies, M.E, Daudet, L.: Sparse audio representations using the MCLT. Elsevier, Amsterdam (2005)
5. Seok, J., Hong, J., Kim, J.: A Novel Audio Watermarking Algorithm for Copyright Protection of Digital Audio. ETRI Journal 24 (June 2002)
6. Zezula, R., Misurec, J.: Audio signal Watrmarking in MCLT domain with the Aid of 2D Pattern. IEEE, NJ, New York (2006)
7. Swanson, M.D., Zhu, B., Tewfik, A.H., Boney, L.: Robust audio watermarking using perceptual masking. Signal processing. Elsevier, Amsterdam (1998)
8. Gordy, J.D., Bruton, L.T.: Performance Evaluation of Digital Audio Watermarking Algorithms. Masterthesis, University of Calgary, Canada, pp. 337–355 (2000)
9. Cvejic, N., Seppanen, T.: Improving Audio Watermarking Scheme Using Psychoacoustic Watermark. Mediateam, Finaland (2004)
10. Malvar, H.: Signal processing with lapped transforms. Artech house, Norwood, MA (1992)

A New Optimization Algorithm of Kinoforms Based on Simulated Annealing

Shinya Nozaki[1], Yen-Wei Chen[2], and Zensho Nakao[3]

[1] Department of Information and Communication Systems Engineering,
Okinawa National College of Technology, 905 Henoko, Nago, Okinawa 905-2192, Japan
[2] Ritsumeikan University, Kusatsu, Shiga 525-8577, Japan
[3] Department of Electrical and Electronics Engineering,
University of the Ryukyus, Okinawa 903-0213 Japan
nozaki@okinawa-ct.ac.jp

Abstract. A new optimization scheme based on Simulated Annealing (SA) is proposed for synthesizing the kinoforms. An initial kinoform is obtained by the Fourier transform of the input image with a random phase. The random phase distribution is then optimized by simulated annealing in order to reduce the variance of the spectrum. Compared with conventional SA-based method, in which optimization is performed to the kinoform, the proposed method can significantly improve the convergence and reduce the computation cost.

Keywords: Computer Generated Holography, Kinoform, Optimization, Random Phase, Simulated Annealing.

1 Introduction

Computer Generated Holography (CGH)[1,2] is a technique which generates a hologram by using computers. The CGH can be used in many applications, such as optical interconnection, optical information, optical elements and so on. A great advantage of the CGH is that it can generate virtual object images.

The kinoform[3,4], one of the CGHs, is an attractive device that can manage wave fronts arbitrary with only the phase information of the complex amplitude. The kinoform is designed in such a manner that the maximum phase modulation of the incident light is 2π. Compared with other types of holograms, its diffraction efficiency is much higher. Because the kinoform is a phase-only optical element designed by a computer, the amplitude of the transfer function of the kinoform is assumed to be a constant. Therefore, the kinoform may include there construction noise caused by the amplitude negligence and the phase quantization. Because the reconstruction noise is a serious problem in many applications, it is necessary to optimize the kinoform in order to reduce the reconstruction noise.

The reconstruction noise due to the amplitude negligence can be reduced if the amplitude is controlled to be uniform for all coefficients. Various schemes have been proposed and used for synthesizing the kinoform. Random phase coding[5] causes a considerable reduction in the dynamic range of the spectrum, but sustained variance

B. Apolloni et al. (Eds.): KES 2007/ WIRN 2007, Part II, LNAI 4693, pp. 303–310, 2007.
© Springer-Verlag Berlin Heidelberg 2007

of the spectrum remains. The iterative methods, such as direct binary search (DBS)[6], simulated annealing[7-10], genetic algorithm[11], have also been proposed and used to optimize the kinoform, in which the mean square error between the input image and the reconstructed image is minimized. These methods are comparatively successful but they require high computation cost and the convergence is limited for complicated input images such as gray images. In this paper, we propose a new optimization scheme for synthesizing the kinoforms. An initial kinoform is obtained by the Fourier transform of the input image with a random phase. The random phase distribution is then optimized by the simulated annealing in order to reduce the variance of the spectrum.

In section 2, the conventional SA-based kinoform synthesis method is given; in section 3, the proposed new optimization scheme is discussed; in section 4, computer simulation results are presented to demonstrate the applicability of the proposed method.

2 The Conventional SA-Based Kinoform Synthesis Method

Consider a Fourier-transformed kinoform as shown schematically in Fig. 1, in which the coordinates of the image plane are denoted by (x,y) and those of the kinoform plane are denoted by (u,v). The kinoform is illustrated by a plane wave. The reconstructed image is derived in the focal plane of the Fourier lens.

When the kinoform is illuminated by a plane wave, the wave front U(u,v) from the kinoform is given by

$$U(u,v) = \exp[i\theta(u,v)], \tag{1}$$

where $\theta(u,v)$ is the phase distribution of the kinoform designed to have a value between 0 and 2π.

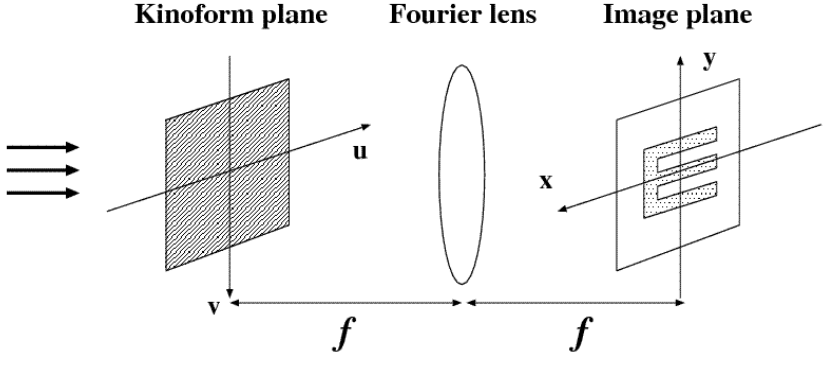


Fig. 1. Schematic illustration of the optical reconstruction setup used for the synthetic Fourier-transform kinoform

It can be obtained by Fourier transform of an image which we design on the computers. This equation tells that the kinoform has only phase information in Fourier domain. In most cases the phase θ can be quantized because of the phase-modulation property of the recording device, and θ can be described by a step function expressed as

$$\theta(u,v) = \frac{2n\pi}{L}, \quad n = 0,1,\ldots,L-1 \tag{2}$$

where *L* is the number of the quantization level The kinoform is then represented by a discrete phase distribution, and the reconstructed image *g(x,y)* from the kinoform is given by

$$g(x,y) = F\left[U(u,v)\right]^{-1}, \tag{3}$$

$$F\left[U(u,v)\right]^{-1} = \sum_{u=-\frac{M}{2}}^{\frac{M}{2}-1} \sum_{u=-\frac{N}{2}}^{\frac{N}{2}-1} U(u,v)\exp\left[i2\pi\left(\frac{x}{M}u + \frac{y}{N}v\right)\right], \tag{4}$$

where $F\left[U(u,v)\right]^{-1}$ is the inverse Fourier transform of $U(u,v)$, and the kinoform consists of an M x N array of rectangular cells. The intensity $I(x,y)$ of the reconstructed image can be expressed as

$$I(x,y) = \left|g(x,y)\right|^2. \tag{5}$$

If we use only the phase information obtained from the Fourier transform of the input image, the reconstructed image may contain the noise because of the amplitude negligence. The optimization techniques using a genetic algorithm or simulated annealing have been proposed[11]. A typical flow chart of the simulated annealing[7] for kinoform optimization is shown in Fig. 2.

3 A New Optimization Scheme

Although SA is a powerful method for the optimization of kinoforms, the limitations of SA or other iterative methods are their high computation cost and their poor convergence for complicated input images such as gray images, which will be shown in the next section. In this paper, we propose a new approach based on SA to improve the convergence and reduce the high computation cost. In the conventional SA-based method(Fig. 2), the initial kinoform is obtained by Fourier transform of the input image with zero phase, while in the proposed method, the initial kinoform distribution

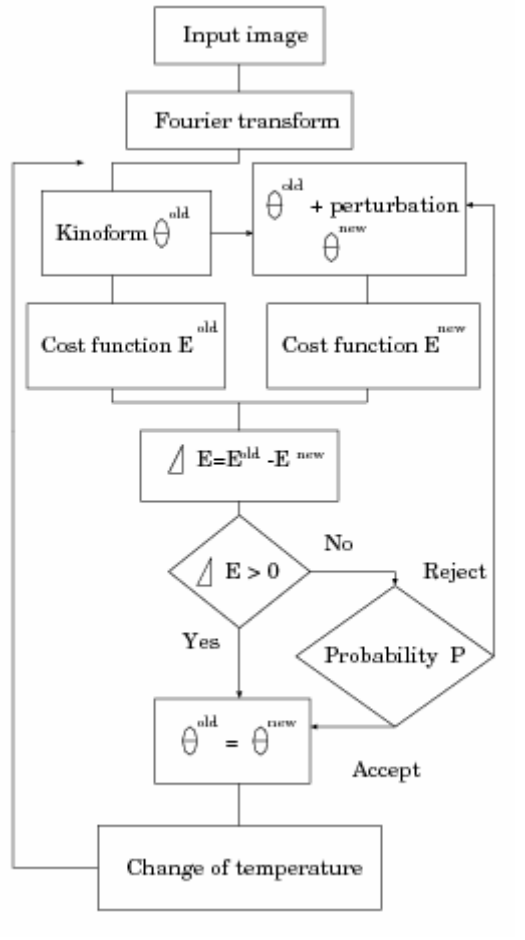


Fig. 2. A flow chart of SA for phase optimization of the kinoform

is obtained by Fourier transform of the input image with a random phase. By adding a random phase to the input image, the spectrum of its Fourier transform gets closer to a constant. Thus, we can start with an initial distribution near an optimum distribution, where the complex amplitude is a constant. It can increase the ability of the convergence. Furthermore, in the conventional SA-based image, the kinoform is optimized. The perturbation of the SA should be done for the entire pixels of the kinoform. In the proposed method, we propose to do the optimization for the random phase added to the input image. Since the phase of the pixels with zero intensity is zero, which does not require optimization, the optimization should be done to only the pixels with non-zero intensity. The search space can be reduced by a factor of N/N_s; where N_s is the total number of the pixels and is the number of non-zero intensity pixels. Thus in the proposed method, the searching ability of SA and the computation cost can be significantly enhanced.

As shown in the previous section, the phase of the input image's Fourier transform is used as a kinoform and the amplitude is assumed to be a constant. We use the variance of the spectrum as a cost function for simplicity. Thus, smaller is the variance of the spectrum, better is the kinoform. The flowchart of the proposed method is shown in Fig. 3.

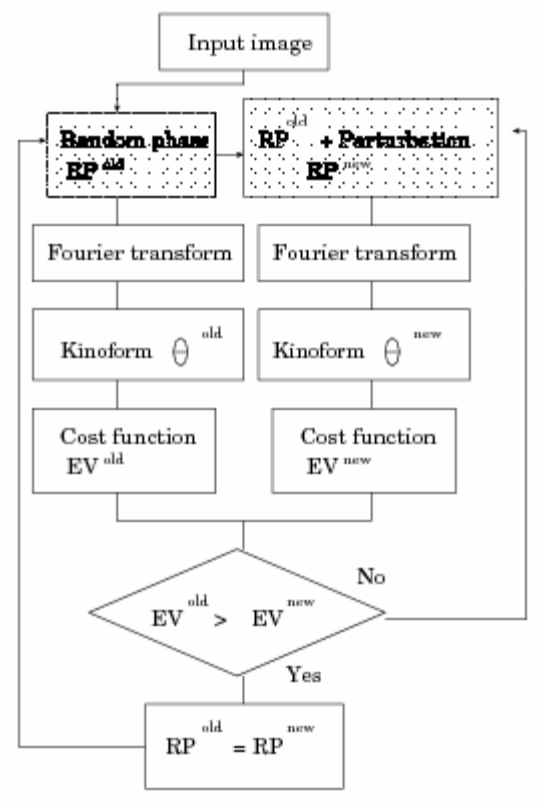


Fig. 3. A flow chart of the proposed method

4 Digital Simulation

In order to validate the capability of the proposed method, a typical example is used in simulation, which is a gray image "Lena"' as shown in Fig. 4(a). The number of pixel is 32 x 32. The reconstructed images by conventional method and proposed method in each annealing cycle are shown in Fig. 4(b) and (c), respectively. Although these show that these converge to input image, the proposed method performes better convergence than the conventional method. In particular, the initial reconstructed images (n=0), the reconstructed images by the proposed method is much better and we can see again that the initial random phase is powerful to optimize kinoform.

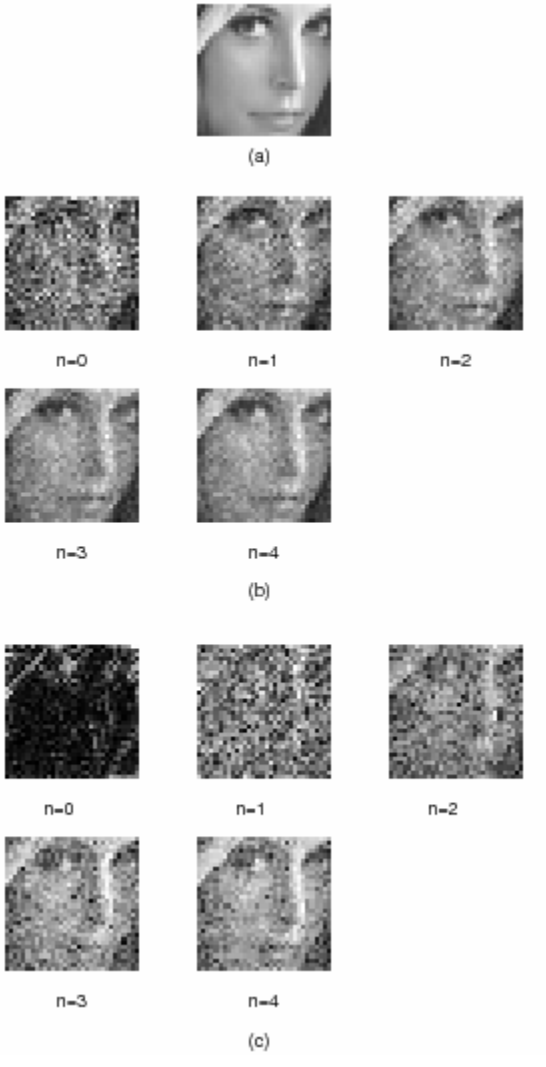


Fig. 4. Comparison of the reconstructed images by the conventional methodsand proposed method (a): the input image, (b): proposed method, (c): conventional method

The reconstructed images by conventional method and proposed method in each annealing cycle are shown in Fig. 4(b) and (c), respectively. Although these show that these converge to input image, the proposed method performes better convergence than the conventional method. In particular, the initial reconstructed images (n=0), the reconstructed images by the proposed method is much better and we can see again that the initial random phase is powerful to optimize kinoform.

Convergence history of the mean square error (MSE) is shown in Fig. 5. Since these costs can not compare between the conventional method and the proposed method, we used the mean square error the input image and the reconstructed image

for the comparison in the simulation. When the number of the iteration is zero, the error by the proposed method is much better due to the initial random phase. When the number of the iteration by the conventional method is 2, it nearly corresponds that initial error by the proposed method. As increased the number of iterations, the error by proposed method is reduced and is converged. That is, the proposed method shows the efficiency of the initial random phase as well as the good convergence. From the figure, it can be seen that the proposed method is better than conventional method.

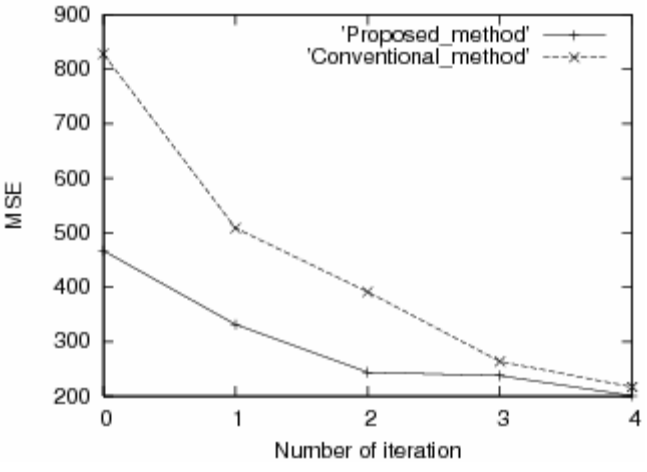


Fig. 5. Comparison of convergence between the proposed method and conventional method

5 Conclusion

We proposed a new approach for the optimization of kinoforms. In the proposed method, the initial kinoform was obtained by Fourier transform of the input image with a random phase and the random phase distribution was optimized by the simulated annealing. The simulation results show that the convergence is improved and the computation cost can be reduced significantly by the proposed method.

References

1. Lee, W.H.: Sampled Fourier-transform hologram generated by computer. Appl. Opt., 9, 639–643 (1970)
2. Lohmann, A.W., Paris, D.P.: Binary Fraunhofer Holograms. Generated by Computer. Opt., 6, 1739–1747 (1967)
3. Lesem, L.P., Hirsch, P.M., Jordan Jr., J.A.: The kinoform: a new wave front reconstruction device. IBM J. Res. Dev. 13, 150–155 (1969)
4. Sasagawa, T.: A multiple imaging system using holographic optical elements for laser machining. IEICE Technical Report 97(90), 1–6 (in Japanese, 1997)
5. Takeda, Y., Oshida, Y., Miyaura, Y.: Random Phase Shifters for Fourier Transformed Holograms. Appl. Opt., 11(4), 819–822 (1972)

6. Seldowitz, M.A., Allebach, J.P., Sweeney, D.W.: Synthesis of digital holograms by direct binary search. Appl. Opt. 26, 2778–2798 (1987)
7. Yoshikawa, N., Yatagai, T.: Phase optimization of two-dimensional Fourier kinoform by simulated annealing. Appl. Opt. 33, 863–858 (1994)
8. Nozaki, S., Chen, Y.W., Nakao, Z.: An Adaptive Simulated Annealing Applied to Opimization of Phase Distribution of Kinoform. In: Methodlogies for the Conception, Design and Application of Soft Computing (Proceedings of IIZUKA '98), pp. 979–982 (1998)
9. Chen, Y.W., Yamauchi, S., Wang, N., Nakao, Z.: A fast kinoform optimization algorithm based on simulated annealing. IEICE Trans. Fundamentals E83-A, 774–776 (2000)
10. Nozaki, S., Chen, Y.W., Nakao, Z.: A New Approach Based on Simulated Annealing to Kinoform Optimization. In: Proc. of SPIE, vol. 3956A-16 (2000)
11. Yoshikawa, N., Yatagai, T.: Quantized phase optimization of two-dimensional Fourier kinoforms by a genetic algorithm. Opt. Lett. 20, 752–754 (1995)

Curvelet-Domain Image Watermarking Based on Edge-Embedding

Thai Duy Hien[1], Ikeda Kei[1], Hanane Harak[1], Yen Wei Chen[2], Yasunori Nagata[1], and Zensho Nakao[1]

[1] Faculty of Engineering, University of the Ryukyus, Okinawa 903-0213, Japan
[2] College of Information, Science and Engineering, Ritsumeikan University, Shiga 525-8577, Japan
{tdhien,nakao}@augusta.eee.u-ryukyu.ac.jp
chen@is.ritsumei.ac.jp

Abstract. A new curvelet-based watermarking technique is presented in this paper, in which watermark signals are selected to be a gray-scale logo image. The curvelet transform was developed in order to represent edges along curves much more efficiently than the traditional transforms. We apply the transform to watermarking and evaluate the effectiveness of the method. Our watermarking algorithm embeds a watermark in curvelet coefficients which are selected by a criterion whether they contain as much edge information as possible. We evaluated the effectiveness of the method against some watermark attacks. Experiment results show that our new method yields quite good visual quality in watermarked images, and is robust to typical signal processing attacks such as compression, cropping, adding noise and filtering.

1 Introduction

Digital watermarking is an emerging field in computer science, digital signal processing and communication. And it is intended by its developers as the solution to the problems of data copyright, content protection and ownership proof, digital watermarking is technically the process of inserting pieces of information into digital data (audio, video, or still images), which can be detected or extracted later to make an assertion about the data.

Recently, several algorithms have used image representations where most information is concentrated in small number of coefficients. In this fashion, watermarking methods in the discrete cosine transform (DCT) domain [4], in the wavelet domain [5], and in the independent component analysis domain and in the principle component analysis (PCA) domain [6, 11, 12] have been proposed. A novel approach is used in this work where we propose a watermarking method based on the use of the curvelet transform domain. The curvelet transform is a relatively new technique and is motivated by the needs of image analysis. The transform is designed to present the edges much more efficiently than the traditional transforms. The proposed watermarking algorithm embeds a watermark in curvelet coefficients which are

B. Apolloni et al. (Eds.): KES 2007/ WIRN 2007, Part II, LNAI 4693, pp. 311–317, 2007.
© Springer-Verlag Berlin Heidelberg 2007

selected by a criterion whether they contain as much edge information as possible. In the following sections, we present an outline of the curvelet technique and explain the proposed method.

2 Curvelet Transform

Recently, Candes and Donoho [1] developed a new multiscale transform which they called the curvelet transform. The transform was designed to represent edges and other singularities along curves much more efficiently than the traditional transforms, i.e. using much fewer coefficients for a given accuracy of reconstruction. Thus, already it has been applied to processing of edge properties e.g., noise removal [3] and contrast enhancement [10].

The curvelet transform executes some processing in order to attain those properties. Figure 1 displays the procedure for the curvelet transform. The idea is to first decompose the image into a set of wavelet bands, and to analyze each band by a local ridgelet transform. The block size can be changed at each scale level. Roughly speaking, different levels of the multiscale ridgelet pyramid are used to represent different subbands of a filter bank output. At the same time, this subband decomposition imposes a relationship between the width and length of the important frame elements so that they are anisotropic and obey $width \approx length^2$.

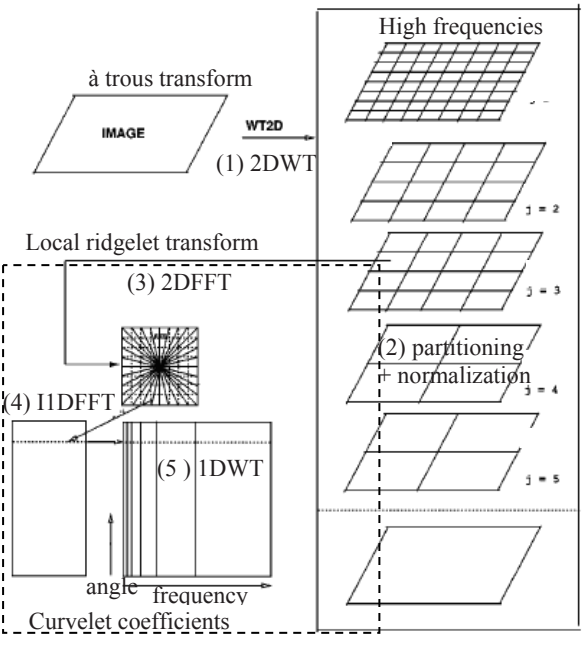


Fig. 1. Curvelet transform

The curvelet decomposition is the sequence of the following steps [1, 7, and 8]:

1. Subband Decomposition: The object is decomposed into sub-bands by "à trous transform" algorithm.
2. Smooth Partitioning: Each sub-band is smoothly windowed into "squares" of an appropriate scale (of side length $\approx 2^{-s}$).
3. Ridgelet Analysis: Each square is analyzed via the discrete ridgelet transform [9].

In this definition, the two dyadic sub-bands $[2^{2s}; 2^{2s+1}]$ and $[2^{2s+1}; 2^{2s+2}]$ are merged before applying the ridgelet transform. In the result, we achieve curvelet coefficients. with involvement of edge information.

Figure 1 shows the detailed process of curvelet transform of an image.

3 Watermark Embedding/Detection

In our watermarking embedding system, watermarking gives priority to embedding of the edges of the given image. Given the original image, we take the curvelet transform and get the curvelet coefficients. It is noted that in the same finite frequency scale, the edges have more important information. Since the coefficient where absolute value is large gives close agreement with the edge information of the image. Specifically, the watermarking algorithm embeds a watermark in curvelet coefficients which are selected by a criterion whether they contain edge information. Selection of the coefficients to which a watermark is embedded is based on a pre-defined threshold and the watermark is cast into coefficients whose absolute values are larger than the threshold.

3.1 Embedding Algorithm

The proposed algorithm is based on three input arguments: "*scale,*" "*pctg,*" and "*alpha*" to embed a watermark. "*scale*" is to select watermark embedding scale coefficients, e.g., if we are selecting input scale = [4, 5], then the watermark embed to from scale 4 coefficients to scale 5 coefficients until it is all embedded. Then in each selected scales, the coefficients are sorted and find the criterion for choosing watermark embed position by "*pctg*" parameter. It were the number of index of sorted coefficients, e.g., "*pctg*" = 0.8, the length (sorted coefficients) = 1000, then we determine that watermark embed threshold is the value of coefficient of index number = 0.8*1000 = 800. "*alpha*" is a strength parameter. Finally, the watermark is embedded to coefficients to satisfy the conditions: "larger than threshold." The architecture of our programs is shown in the Figure 2.

3.2 Watermarks

In this algorithm, the watermark is a logo image whose 32x32 pixels and marked a kanji character, which is the first character of place name 'Okinawa'. In embedding process, the logo changes the form into binary bit stream (the length is 1024 bit), W =

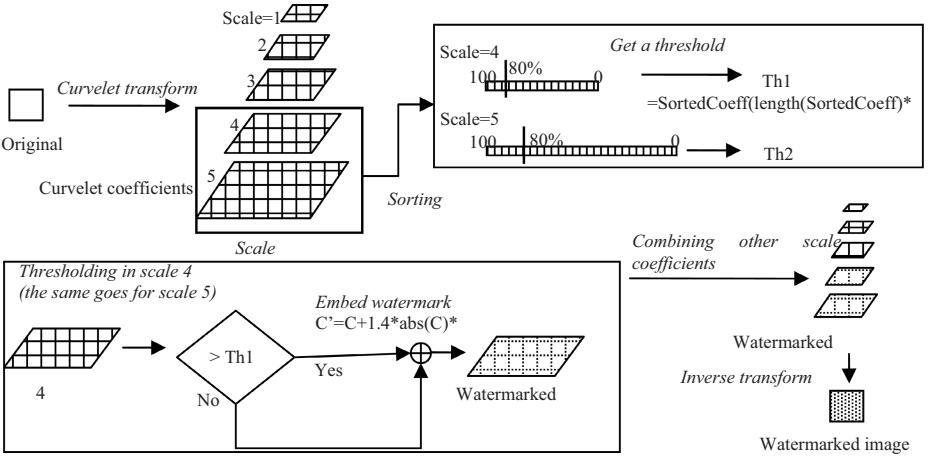


Fig. 2. The embedding algorithm (*scale*=[4,5], *pctg*=0.8, *alpha*=1.4)

w1, w2, w, ... ,w1024, and then cast to the curvelet coefficients. The embedded coefficients were modified by the following equation:

$$C' = C + alpha \cdot w_i \tag{1}$$

where *C* is a curvelet coefficient to satisfy the conditions, *alpha* is a strength parameter (this parameter is the same as the input argument *alpha*).

3.3 Watermark Detection

A logo is detected from the altered curvelet coefficients, and binarized with specific threshold.

We used the Normalization Correlation (NC) to evaluate the quality of detected logo watermark. NC is calculated using the following equation:

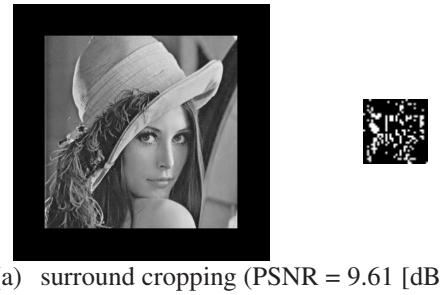
$$NC = \frac{\sum_{m=1}^{N/2}\sum_{n=1}^{N/2}\left[W_k(m,n)\cdot W_k^{'}(m,n)\right]}{\sum_{m=1}^{N/2}\sum_{n=1}^{N/2}\left[W_k(m,n)\right]^2} \tag{2}$$

where W_k is original logo and W_k' is detected logo. The value of NC has the unit range [0, 1], and if we get at higher NC values, the embedded watermark is more similar to the detected one.

4 Experimental Results and Conclusion

We use standard 512x512 pixel image 'Lena' for evaluation of our proposed method. In our program, the image is transformed through fast discrete curvelet transform via

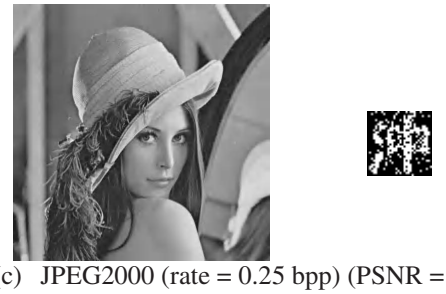
wrapping [2]. We conducted experiments logo watermarking with the noted above parameters is scale=[3], pctg=0.31, alpha=50.0. For watermarking evaluation we used the Peak Signal to Noise Ratio (PSNR) and Mean Square Error (MSE). We apply the proposed method to the image 'Lena,' the obtained watermarked image (Figure 4) shows that there is no noticeable difference between the watermarked and the original image, which confirms the invisibility requirement in our watermarking method. Its PSNR is 48.73 [dB] and MSE is 0.87 and The NC value of detected logo is 0.9063. From Figure 4, we can confirm that the recognition of the watermark in the experimental images is giving satisfactory results. In addition, we perform further experiments to evaluate for the robustness against more commonly used image processing attacks such as JPEG compression, surrounding cropping, adding noise, image resizing and so on. We apply the attacks to the watermarked image and then detect the watermark. Variation in the detector response depending on a parameter changing of various attacks is displayed in Figure 3.



(a) surround cropping (PSNR = 9.61 [dB] / NC = 0.504)



(b) JPEG (Quality factor = 15%) (PNSR = 37.1 [dB] / NC = 0.826)



(c) JPEG2000 (rate = 0.25 bpp) (PSNR = 33.3 [dB] / NC = 0.672)

Fig. 3. Detector response against various attacks



(d) Adding noise (PSNR = 22.4 [dB] / NC = 0.52)



(e) Lowpass filtering (PSNR = 39.1 [dB] / NC = 0.754)



(f) Median filtering(PSNR = 40.4 [dB] / NC = 0.857)

Fig. 3. (*continued*)



Fig. 4. Watermarked image (PSNR = 48.73 [dB] and MSE = 0.87), and detected logo 'Oki' (NC = 0.9063)

In this paper we proposed a novel use of the curvelet transform in watermarking digital images. Our preliminary results show that the method is robust against some important image processing attacks.

References

1. Donoho, D.L., Duncan, M.R.: Digital curvelet transform: Strategy, implementation an experiments. In: Proc. SPIE, vol. 4056, pp. 12–29 (2000)
2. Candes, E.J., Demanet, L., Donoho, D.L., Ying, L.: Fast Discrete Curvelet Transforms, Tech. Report, Calif. Inst. of Tech. (2005)
3. Starck, J.L., Candes, E., Donoho, D.L.: The Curvelet Transform for Image Denoising. IEEE Trans. Image Proc. 11(6), 670–684 (2002)
4. Starck, J.L., Murtagh, F., et al.: A DCT domain system for robust image watermarking, Signal Proc., pp. 357–372 (1998) (Special Issue in Copyright Protection and Access Control for Multimedia Services)
5. Dugad, R., Ratakonda, K., et al.: A new wavelet-based scheme for watermarking images. In: ICIP 98. Proc., IEEE Inter., Conf., on Image Proc, Chicago, IL, USA, pp. 419–423. IEEE Computer Society Press, Los Alamitos (October 1998)
6. Hien, T.D., Nakao, Z., Chen, Y.-W.: Robust digital watermarking based on principal component analysis, Internat. Internat. J. Comput. Intell, Appl. 4(2), 183–192 (2004)
7. Cand‘es, E.J., Donoho, D.L.: New tight frames of curvelets and optimal representations of objects with piecewise-C2 singularities. Comm. on Pure and Appl. Math. 57, 219–266 (2004)
8. Starck, J.L., Donoho, D.L., Candes, E.J.: Astronomical Image Representation by the Curvelet Transform. Astronomy and astrophysics(Berlin. Print) 398, 785–800 (2003)
9. Campisi, P., Kundur, D., Neri, A.: Robust Digital Watermarking in the Ridgelet Domain. IEEE Signal Processing Letters 11(10), 826–830 (2004)
10. Starck, J.-L., Murtagh, F., Candes, E., Donoho, D.L.: Gray and Color Image Contrast Enhancement by the Curvelet Transform. IEEE Transaction on Image Processing 12(6), 706–717 (2003)
11. Hien, T.D., Nakao, Z., Chen, Y.-W.: Robust multi-logo watermarking by RDWT and ICA. Signal Processing archive 86, 2981–2993 (2005)
12. Hien, T.D., Nakao, Z., Chen, Y.-W.: RDWT/ICA for Image Authentication. In: The 5th IEEE International Symposium on Signal Processing and Information Technology, Athens

Linking KANSAI and Image Features by Multi-layer Neural Networks

Xinyin Huang[1], Shouta Sobue[2], Tomoki Kanda[2], and Yen-Wei Chen[2]

[1] School of Education, Soochow University, Suzhou, Jiangsu 215006, China
[2] College of Information Science and Eng., Ritsumeikan Univ., Shiga, 525-8577, Japan

Abstract. KANSEI is a Japanese term which means psychological feeling or image of a product. KANSEI engineering refers to the translation of consumers' psychological feeling about a product into perceptual design elements. Recently several researches have been done for image indexing or image retrieval based on KANSEI factors. In this paper, we report a quantitative study on relationship between image color features and human KNASEI factors. We use the semantic differential (SD) method to extract the KANSEI factors (impressions) such as bright, warm from 4 group subjects (Children, students, adults, elderly person) while they viewing an image (painting). A neural network is used to learn the mapping functions (relationships) from the image feature space to human KANSEI factor space (psychological space).

Keyword: KANSEI, impression, image, color, neural network, SD method.

1 Introduction

Recently a growing interests has been seen in KANSEI engineering or emotion design. KANSEI is an individual subjective impression from a certain artifact, environment or situation using all the senses of sight, hearing, feeling, smell, taste, recognition and balance [1]. KANSEI engineering refers to the translation of consumers' psychological feeling about a product into perceptual design elements. KANSEI can "measure" the feelings and shows the relationship to certain product properties. In consequence, products can be designed to bring forward the intended feeling. KANSEI is the integrated functions of the mind, and various functions exist in during receiving and sending. Filtering, acquiring information, estimating, recognizing, modeling, making relationship, producing, giving information, presenting and etc. are the contents of KANSEI. A person's KANSEI will be expressed through physiological functions. There are 4 ways of measuring the KANSEI: (1) words; (2) physiological response (heart rate, EEG); (3) people's behaviors and actions; (4) Facial and body expressions [2]. The most common way of measuring the KANSEI is through the words. Recently several researches have been done for image indexing or image retrieval based on KANSEI factors (impression words) [3-5]. In this paper, we propose a method to study the relationship between image color features and human KNASEI factors. We use the semantic differential (SD) method to extract the KANSEI factors (impressions) such as bright, warm from 4 group subjects (Children,

B. Apolloni et al. (Eds.): KES 2007/ WIRN 2007, Part II, LNAI 4693, pp. 318–325, 2007.
© Springer-Verlag Berlin Heidelberg 2007

students, adults, elderly person) while they viewing an image (painting). A neural network is used to learn the mapping functions from the image feature space to human KANSEI factor space (psychological space).

2 Image Feature Space and KANSEI Factor Space

In order to make a quantitative study on the relationship between the image features and KANSEI factors, we construct an image feature space and a KANSEI factor space (psychological space) as shown in Fig.1. One image (painting) has one point in the image feature space and has a corresponding point in the KANSEI factor space, which is an impression (KANSEI) of the subject to the image, just like a projection of the image feature to the psychological space. The relationship between the image features and the KANSEI can be described by the mapping function from the image feature space to the psychological space. The input of the mapping function is the image features and the output of the function is the KANSEI factors (impressions). Since different person, especially the person at the different age will have a different KANSEI or impression for a given image (painting), we have to use a proper mapping function for KANSEI-based applications such as impression words based image indexing or retrieval. The mapping functions (relationships) can be learned by finding the corresponding points in the image feature space and psychological space.

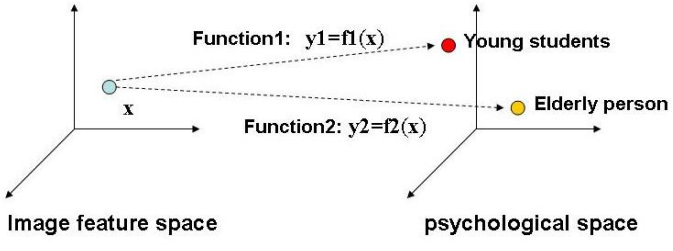


Fig. 1. Mapping Functions from image feature space into psychological space

2.1 KANSEI Factors (Psychological Features)

In order to find a corresponding point in psychological space, we use the semantic differential (SD) method to extract the KANSEI factors (impressions) such as bright, warm from 4 group subjects (Children, students, adults, elderly person) while they viewing an image (painting). In this research, we use words (adjective pairs) to measure the KANSEI. By careful selections, we chose 12 pairs of adjectives, which are shown in Table 1, as measures of KANSEI.

We chose 15 images (oil painting) shown in Fig.2 for learning of mapping functions and validations. We displayed the images to 4 group subjects. (group 1: children at the age of about 10; group 2: college students at the age of about 20; group 3: adults at the age of 40-50; group 4: elderly person at the age of 60-70). Each group consists of 30 subjects. We asked the subjects to rate the impression with 12 pairs of

adjectives in 5-step scales (-2, -1, 0, 1, 2). The average values over 30 subjects are used as represent values of each group. The averaged standard deviations are 1.1, 0.6, 0.3, 0.4 for group 1, group 2, group 3 and group 4, respectively. The children group has a relatively large variation. The typical impression data for image 1 is shown in Fig.3. The number of horizontal axis is the No. of adjective pair in Table 1. It can be seen that there are significant differences among groups.

Table 1. 12 pairs of adjectives

1. bright - dark
2. vivid - subdued
3. clear - fainted
4. gaudy - plain
5. hard - soft
6. warm - cool
7. pure - impure
8. simple - complex
9. jaunty - placid
10. wide - narrow
11. dry - wet
12. like - dislike

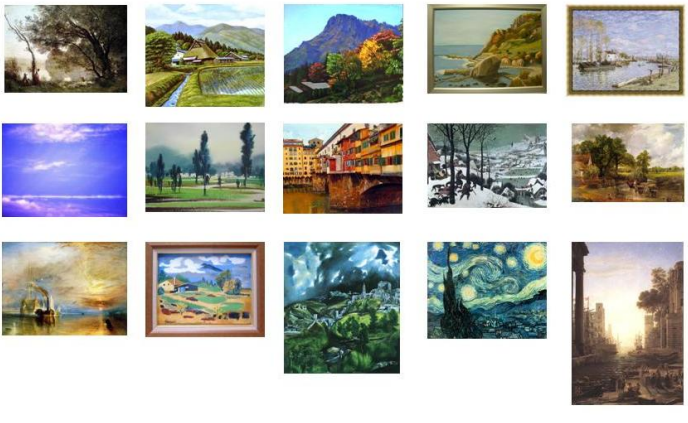


Fig. 2. 15 images (oil painting) used for training

In order to remove the redundancy and reduce the dimensionality of the impression feature vector, the impression data are analyzed by Principal Component Analysis (PCA) [6] for each group. PCA is to linearly transform the data such that the outputs are uncorrelated. The impression data is expressed by a 12-dimensional vector $\mathbf{z} = (z_1, z_2, \cdots z_{12})^T$. The linear image analysis process is to find a matrix **W**,

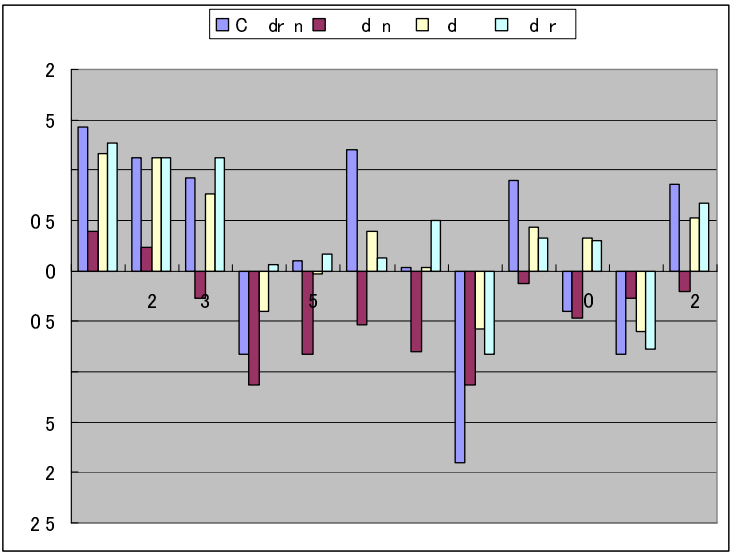


Fig. 3. The typical impression data for image 1

so that the resulting vector:

$$\mathbf{y} = \mathbf{W}^T \mathbf{z} \qquad (1)$$

has uncorrelated components. $\mathbf{W}$ is constructed in such a way that $\mathbf{W}\mathbf{W}^T = \mathbf{I}$. Actually, the columns of $\mathbf{W}$ are the eigenvectors of the covariance matrix of $\mathbf{z}$. The accumulated cover rate is shown in Fig.4. It can be seen that the 12-dimensional adjective feature space can be approximated with top 4 factors, while maintaining more than 90% of information. Thus we chose the top 4 factors as features of KANSEI and constructed a 4-dimensional psychological space. The KANSEI feature vector $\mathbf{y}$ is given by

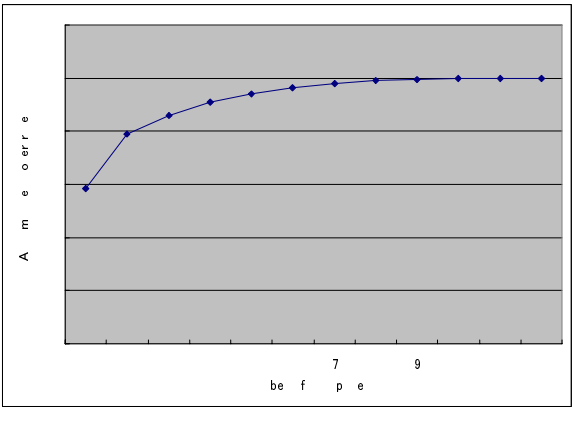


Fig. 4. Accumulated cover rate

$$\mathbf{y} = \begin{pmatrix} y_1 \\ y_2 \\ y_3 \\ y_4 \end{pmatrix} = \mathbf{W}^T \mathbf{z} = \begin{pmatrix} \mathbf{u}_1^T \mathbf{z} \\ \mathbf{u}_2^T \mathbf{z} \\ \mathbf{u}_3^T \mathbf{z} \\ \mathbf{u}_4^T \mathbf{z} \end{pmatrix} \qquad (2)$$

where $\mathbf{u}_1$, $\mathbf{u}_2$, $\mathbf{u}_3$, $\mathbf{u}_4$ are top 4 eigenvectors of the covariance matrix of $\mathbf{z}$, respectively.

2.2 Color Image Features

There are many low-level visual features such as color, texture and shape information in an image. Among them, color is perhaps the most dominant and distinguishing visual feature. In this paper, we focused our researches on the relationship between the color features and the KANSEI factors. Several color features have been proposed to represent color compositions of images [7]. Color histograms are widely used to capture the color information in an image [8]. They are easy to compute and tend to be robust against small changes of camera viewpoints. An image is given in some color space (e.g. red, green, blue). The color channels are quantized into a coarser space with k bins for red, m bins for green and l bins for blue. Therefore the color histogram is a vector $\mathbf{h} = (h_1, h_2, \cdots h_n)$, where $n = k \times m \times l$. In this paper, k=m=n=8 and n=512. Each element h_j represents the number of pixels of the discretized color in the image and we normalize histogram elements as

$$h_j' = h_j / \sum_{j=1}^{n} h_j \tag{3}$$

The normalized histogram vector is used as image features. The typical images and their color histograms are shown in Fig.5(a) and 5(b), respectively.

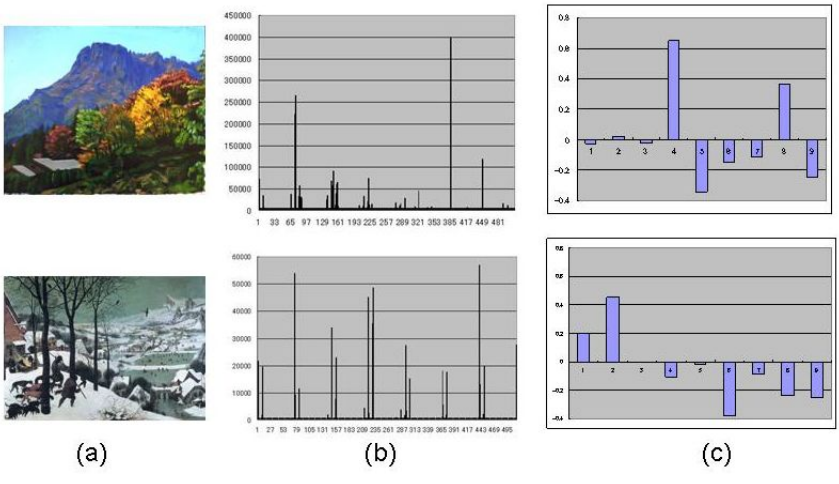


Fig. 5. Typical images (a); their 512-D color histograms (b); their 9-D PCA color features (c)

The disadvantage of color histogram is its large dimension, such as 512-dimension. Several methods have been proposed to transform the high dimensional data to low dimensional features. In this paper, we use Principal Component Analysis (PCA), which has been described in Sec.2.1, to reduce the large dimension of color histograms[9]. By using PCA, we can project the 512-dimensional feature space into a

9-dimensional subspace, while maintaining 85% of information. Thus, we select the top 9 components as image features and construct a 9-dimensional image feature space. The examples of PCA color features (9-dimension) are shown in Fig.5(c). It is clear that different color images have different features.

3 Learning and Validations

As we described in the previous section, by constructing both image feature space (9-D) and psychological space (4-D), we can easily find a corresponding points for each input images. We use a three-layer neural network to learn the mapping function (linking) between the two spaces. The points (9-dimensional image feature vector) are used as input of the neural network. Thus the number of neurons in the input layer is 9. The output is the point (4-dimensional KANSEI factors) in the psychological space and the number of neuron in the output layer is 4. The number of neurons in the middle layer is 32. The 15 images shown in Fig.2 are used for learning and validations (testing). One image is selected as testing image and remaining 14 images are used as training samples and the corresponding impression data obtained by SD method as described in 2.1 section are used as teaching signals. The learning process is shown in Fig.6. A backpropagation algorithm is used as a learning algorithm. The learning is done for each group. Finally, we got 4 trained neural networks for each group.

Leave-one-out method is used for validations. No.1 image is used as the testing image for validations. The remaining 14 images (except the testing image) are used for learning. We first calculate the image feature vector (the point in image feature space) and then the corresponding point (KANSEI feature vector) $\mathbf{y}=(y_1, y_2, y_3, y_4)^T$ in psychological space is calculated by using the trained neural networks (mapping functions), which is the output of the neural network. The KANSEI features $\mathbf{y}$ are then transformed into impression data (12 pairs of adjectives) using inverse PCA as shown in Eq.(4). The validation process is shown in Fig.7.

$$\hat{\mathbf{z}} = \mathbf{W}\mathbf{y} = y_1\mathbf{u}_1 + y_2\mathbf{u}_2 + y_3\mathbf{u}_3 + y_4\mathbf{u}_4 \tag{4}$$

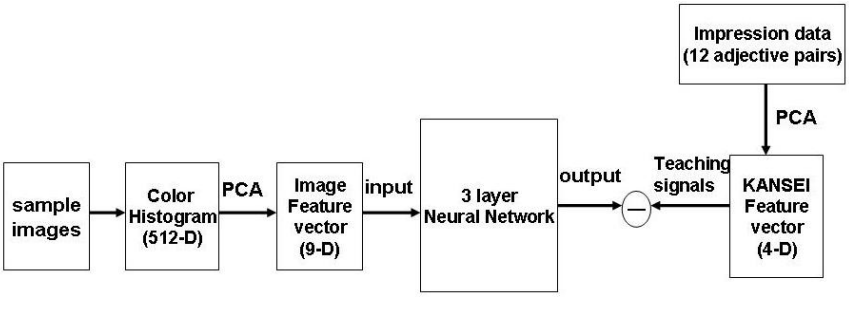


Fig. 6. Learning process

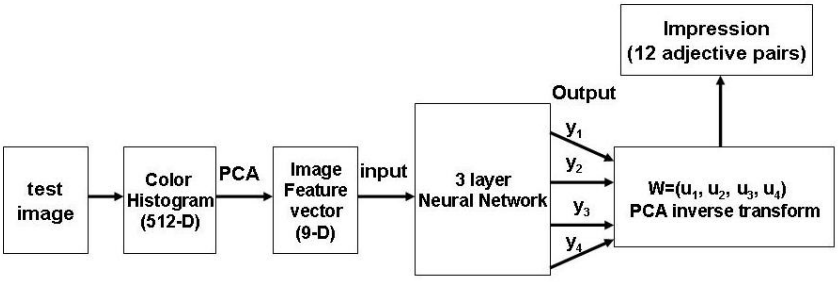


Fig. 7. Validation process

The estimated impressions (estimations) by using the trained neural networks (mapping functions) for each group are shown in Fig.8 and in order to make a comparison, the real impressions obtained by using SD method are also shown in Fig.8. The root mean square errors (RMSE) between the real impressions and the estimated impressions are summarized in Table 2. It can be seen that good agreements between the real impression and estimated impression for college students, adults and elderly person have been achieved. The estimation to children is not so accurate. The main reason is that the impression data of children group (group 1) used for learning have relatively large variations as shown in Sec.2.1 (the standard deviation of group 1 is 1.1, while other groups are about 0.5). It is difficult to make an accurate estimation based on leaning from the samples which have large variations.

Table 2. Root mean square errors (RMSE) between real impressions and estimations

	Children	Students	Adults	Elderly
RMSE	1.09	0.55	0.35	0.31

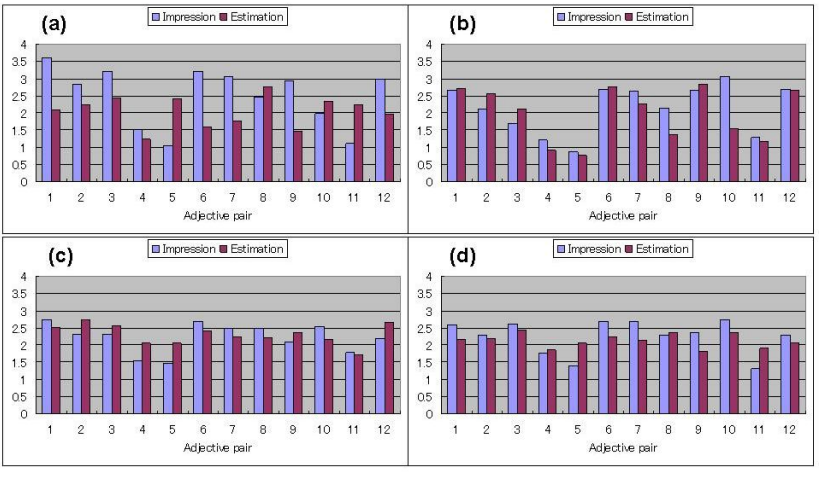


Fig. 8. Comparison between estimated results with real impressions. (a) Children; (b) Students; (c) Adults and (d) Elderly person.

4 Conclusions

In this paper, a quantitative study has been done for evaluation of relationship between image features and human KNASEI factors. We used the semantic differential (SD) method to extract the KANSEI factors (impressions) such as bright, warm from 4 group subjects (Children, college students, adults, and elderly person) while they viewing an image (painting). A neural network was used to learn the mapping functions from the image feature space to human KANSEI factor space (psychological space). Leave-one-out method has been used for testing. The testing results showed that it is possible to obtain a proper mapping function from the image feature space to the psychological space by using our proposed method. It is also important to use a proper mapping function for KANSEI based applications.

This work is supported in part by the "Open Research Center" Project for Private Universities: matching fund subsidy from Japanese Ministry of Education, Culture, Sports, Science, and Technology.

References

1. Schwte, S., Eklund, J., Axelsson, J.R.C., Nagamachi, M.: Concepts, methods and tools in Kansei Engineering. Theoretical Issues in Ergonomics Science 5, 5214–5232 (2004)
2. Grimsæth, K.: Linking emotions and product features. KANSEI Engineering , 1–45 (2005)
3. Black, J., Kahol, K., Tripathi, P., Panchanathan, S.: Indexing natural images for retrieval based on Kansei. In: Proc. of Human Vision and Electronic Imaging conference (2004)
4. Takagi, H., Noda, T.: Media Converter with Impression Preservation Using a Neuro-Genetic Approach International. J. of Hybrid Intelligent Systems 1, 49–56 (2004)
5. Takagi, H., Noda, T., Cho, S.: Psychological Space to Hold Impression Among Media in Common for Media Database Retrieval System. In: SMC'99. Proc. of IEEE International Conference on System, Man, and Cybernetics, Tokyo, Japan, vol. VI, pp. 263–268. IEEE Computer Society Press, Los Alamitos (1999)
6. Gianfelici, F., Biagetti, G., Crippa, P., e Turchetti, C.: A Novel KLT Algorithm Optimized for Small Signal Sets. In: IEEE ICASSP 2005. Proceedings of 2005 IEEE International Conference of Acoustics, Speech and Signal Processing, pp. 18–23. IEEE Computer Society Press, Los Alamitos (2005)
7. Sticker, M.A., Orengo, M.: Similarity of color images. In: Proc. SPIE Storage Retrieval Still Image Video Database, pp. 381–392 (1996)
8. Deng, Y., Manjunath, B.S.: An efficient color representation for image retrieval. IEEE Trans. Image Processing 10, 140–147 (2001)
9. Zeng, X.Y., Chen, Y.W., Nakao, Z.: Independent Component Analysis for Color Indexing. IEICE Trans. Information & Systems E87-D, 997–1003 (2004)

Pattern Analysis and Prediction of *O*-Linked Glycosylation Sites in Protein by Principal Component Subspace Analysis

Yen-Wei Chen[1,2], Xuemei Yang[2,3], Masahiro Ito[2], and Ikuko Nishikawa[2]

[1] Elect & Information Eng. School, Central South Univ. of Forestry and Technology, Changsha 410004, China
[2] College of Information Science and Eng., Ritsumeikan Univ., Shiga, 525-8577, Japan
[3] Department of Mathematics, Xianyang Normal Univ., Xianyang 712000, China

Abstract. Glycosylation is one of the most important post-translation modifications steps in the synthesis of membrane and secreted proteins and more than half of all proteins are glycosylated. In this paper, we propose a principal component analysis (PCA) based subspace approach for pattern analysis and prediction of *O*-glycosylation sites in protein. PCA is used to find principal components and subspaces of glycosylated residues and nonglycoslylated residues, respectively. From the calculated principal compoents, we found that the glycosylted proteins all have a high serine, threonine and proline content. The prediction can be viewed as a 4-classes classification problem or 2-classes classification problems. We project the protein sequence (test vector) into each subspace and calculate the distance between the test vector and its projection on the subspace. The protein sequence can be classified into the "nearest" class. The prediction accuracy for nonglycosylated sites (negative sites) is about 70%-90%, and the accuracy for *O*-glycosylated sites (positive sites) is about 70%-100%.

Keywords: Protein *O*-linked glycosylation, principal component analysis, pattern analysis, prediction.

1 Introduction

Glycosylation is one of the most important post-translation modification steps in the synthesis of membrane and secreted proteins and more than half of all proteins are glycosylated [1]. Two types of glycosylation exist: *N*-linked glycosylation to the amide nitrogen of asparagines side chains and *O*-linked glycosylation to the hydroxyl of serine and threonine side chains. It has been known that *N*-linked glycosylated asparagine residues practically always occur in the sequence Asn-Xaa-(Ser/Thr)-Yaa, where Xaa and Yaa can be any amino acid except proline [2, 3]. On the other hand, though *O*-linked glycosylation is known to be serine or threonine specific, not all serine or threonine residues are glycosylated and no consensus sequence for *O*-linked glycosylation has been found.

Many efforts have been made for finding some rules or consensus sequence for *O*-linked glycosylation. Wilson et al. found that proline occurs at increased frequency at positions -1 and +3 relative to the glycosylation site [4]. Elhammer et al. [5] extended

B. Apolloni et al. (Eds.): KES 2007/ WIRN 2007, Part II, LNAI 4693, pp. 326–334, 2007.
© Springer-Verlag Berlin Heidelberg 2007

this by *in vitro* experiments using bovine colostrum transferase and matrix statistics on a data set of 196 glycosylation sites. They found that a high abundance of not only proline, but also serine and threonine, at all positions from -4 to +4 favoured glycosylation.

In recent years, many computational methods based on weight matrix [6,7], artificial neural networks [6,8,9] and support vector machines [9,10] have also been developed for prediction of *O*-glycosylation sites. The prediction accuracy can be achieved more than 70%.

In this paper, we propose a new approach based on principal component analysis (PCA) for simultaneous pattern analysis and prediction of *O*-glycosylation sites in protein. PCA is used to find principal components and subspaces of glycosylated residues and nonglycoslylated residues. From the calculated principal compoents, we found that the *O*-glycosylted proteins all have a high serine, threonine and proline content. The prediction is viewed as a 4-class classification problem or 2-class classification problems. We project the protein sequence (test vector) into each subspace and calculate the distance between the test vector and its projection on the subspace. The protein sequence can be classified into the "nearest" class. The prediction accuracy for nonglycosylated sites (negative sites) is about 70%-90%, and the accuracy for *O*-glycosylated sites (positive sites) is about 70%-100%.

2 Principal Component Subspace Analysis

Principal Component Analysis (PCA) is to linearly transform the data such that the outputs are uncorrelated [11]. In this paper, PCA is used to find the principal components and subspace of glycosylated sites and nonglycosylated sites from the training data set. Let the N training vectors of a particular class be $\{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N\}$. The covariance matrix is defined as

$$\mathbf{S} = \frac{1}{N}\sum_{i=1}^{N}(\mathbf{x}_i - \mathbf{m})(\mathbf{x}_i - \mathbf{m})^T \tag{1}$$

where $\mathbf{m}$ is the mean vector of the training set and subscript T is the transpose.

For finding the principal components, we solve the eigenvalue equation as

$$\lambda\mathbf{v} = \mathbf{S}\mathbf{v} \tag{2}$$

In this fashion, the basis vectors for each class i are computed as a set of N eigenvectors $\mathbf{v}_j^i$, $j = 1, \cdots, N$. Each eigenvector is normalized so that the basis is orthogonal. The test vector $\mathbf{x}_{\text{test}}$ is projected to each class subspace which is defined by its basis vectors (eigenvectors). The distance of the test vector to the subspace of class i is defined as the distance between the $\mathbf{x}_{\text{test}}$ and its projection on the subspace $\hat{\mathbf{x}}_{test}^i$ as Eq.(3)

$$\left\|\mathbf{x}_{test} - \hat{\mathbf{x}}_{test}^i\right\| \tag{3}$$

The projection of $\mathbf{x}_{\text{test}}$ into the basis vector $\mathbf{v}^{\text{i}}_{\text{j}}$ is given by

$$y_j^i = \mathbf{v}_j^{i\,T} \mathbf{x}_{test} \tag{4}$$

Since the basis vector (eigenvectors) are orthogonal, the projection (or reconstruction) of $\mathbf{x}_{\text{test}}$ into each class subspace is given by

$$\hat{\mathbf{x}}_{test}^i = \sum_{j=1}^{M} y_j^i \mathbf{v}_j^i = \sum_{j=1}^{M} \mathbf{v}_j^{i\,T} \mathbf{x}_{test} \mathbf{v}_j^i \tag{5}$$

where M is the number of components or the dimension of the subspace. Smaller is the dimension of the subspace, larger is the distance. In order to make a robust classification, we use an averaged distances over M as a measure, which is shown in Eq.(6).

$$D^i = \frac{1}{MM} \sum_{M=1}^{MM} \left\| \mathbf{x}_{test} - \sum_{j=1}^{M} \mathbf{v}_j^{i\,T} \mathbf{x}_{test} \mathbf{v}_j^i \right\| \tag{6}$$

where MM is the dimension of feature vector (maximum dimension of subspace). Finally, the test vector is assigned to the "nearest" class as Eq.(7).

$$C = \arg\min_i D^i = \arg\min_i \sum_{M=1}^{MM} \left\| \mathbf{x}_{test} - \sum_{j=1}^{M} \mathbf{v}_j^{i\,T} \mathbf{x}_{test} \mathbf{v}_j^i \right\| \tag{7}$$

3 Protein Sequence Data and Encoding

The protein sequence data used in this research are searched from glycosylation database Uniprot (v8.0) [12]. We selected only 99 mammalian protein entries (protein sequence) with annotation of *O*-glycosylation for analysis. The entries, which are "potentially" or "probably" annotated, were excluded. Since *O*-glycosylation is known to be serine (S) or threonine (T) specific, we focus on these two kinds of amid acid residues. Each selected protein sequnce is then truncated by a window (window size: *N*) into several subsequences with S or T residues at the center. These protein subsequences are then divided into four classes. The first and second classes are for serine and threonine residue sites, which are annotated experimentally as being glycosylated, respectively. The third and fourth classes are for serine and threonine residue sites, which have no such annotations. The four classes are called here as positive S sites, positive T sites, negative S sites and negative T sites, respectively, for convenience. One example of protein sequence and its protein subsequences (N=7) are shown in Fig.1. The number of samples for each class is summarized in Table 1. Since the number of negative sites are much larger than that of positive sites. We randomly choose 200 samples from each negative site for training and randomly choose 10 samples from each class for testing. The statistical analysis (PCA) is done for each class separately.

Table 1. Number of samples

	Total Number	Number for training	Number for testing
Positive S	174	164	10
Positive T	292	282	10
Negative S	693	200	10
Negative T	874	200	10

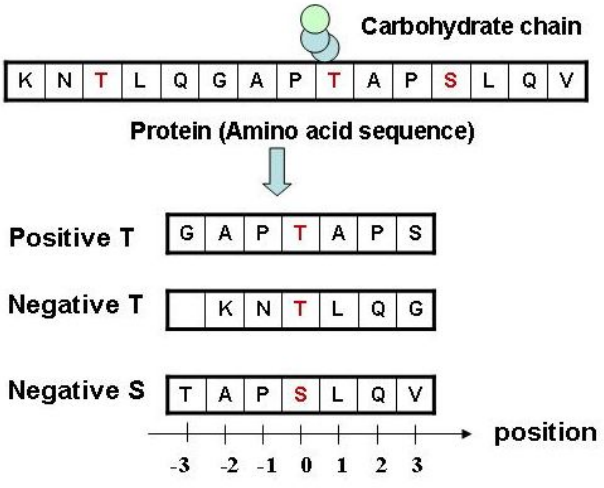


Fig. 1. Subsequences (*N*=7) with S or T at the center used for statistical analysis and prediction

The protein sequences (exclude S or T at the center) with a length of *N*-1 are used for analysis. There are many methods have been proposed for protein sequence coding, such as sparse coding [9,10], 5-letter coding [9], hydropathy coding [9] and physical properties based coding [10]. In this paper, we use the sparse coding scheme for representation of the protein sequence. In sparse coding, 21-binary sequence is used to code one site of amino acid or vacancy. Thus the total length of coded sequence or dimension of feature vector (**x**) is $(N-1)\times 21$. Though the window size (*N*) is an important parameter, in this paper we only focus on *N*=7 as a typical window size because of limited space. The detailed discussion on window size will be presented in other paper.

4 Positional Distribution and Principal Components

By taking a histogram of amino acid at each position or by calculating the averaged protein sequence, we can obtain a positional probability function (PPF). The calculated PPFs from the samples are shown in Fig.2.

It can be seen that the PPF of positive S site is similar to that of positive T, while the PPF of negative S is similar to that of negative T. Compared to negative sites,

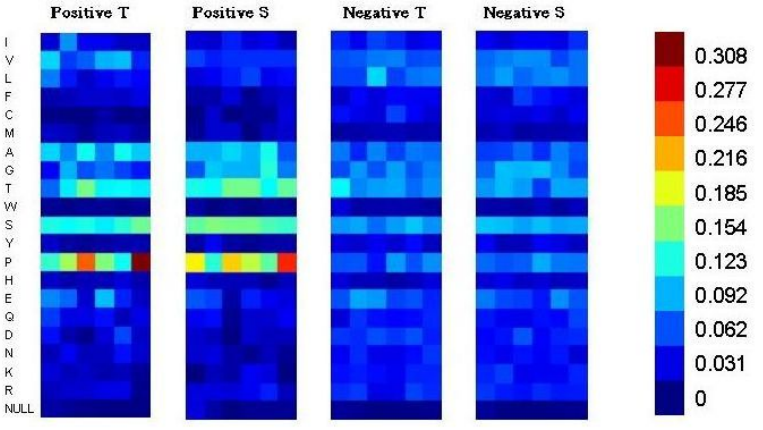


Fig. 2. Positional Probability Functions (PPFs)

both positive S and positive T have a high abundance of not only proline, but also serine, threonine and Alanine, at all positions. The detailed PPFs of proline, serine, threonine and alanine are shown in Figs.3-6, respectively. We can see that the abundance of proline is much higher especially at positions -1 and +3 relative to the glycosylation site, which agrees well with Wilson's results [4].

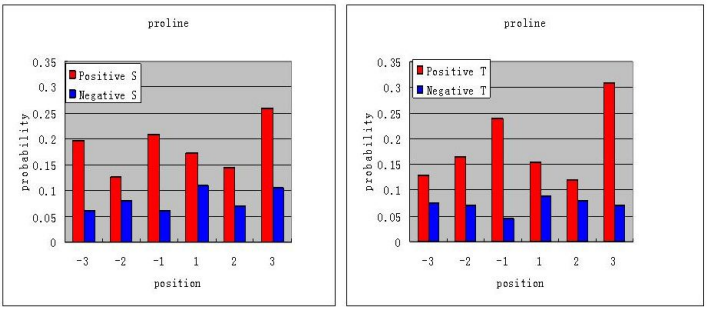


Fig. 3. PPFs of proline

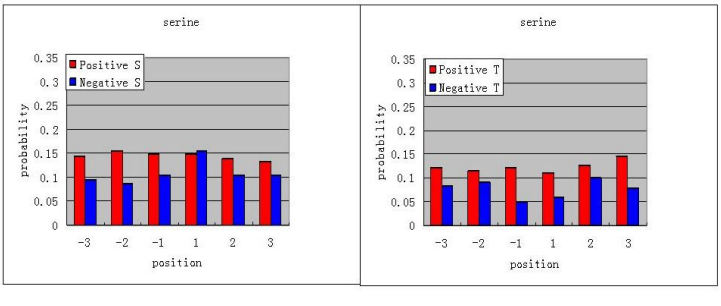


Fig. 4. PPFs of serine

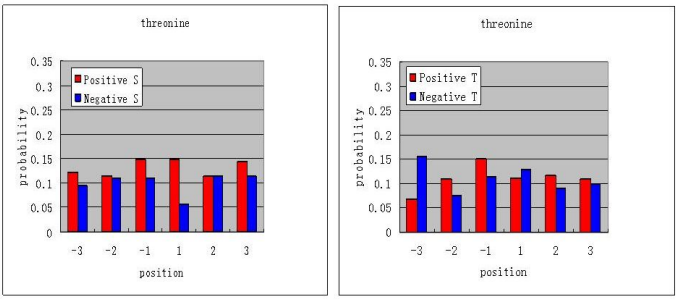


Fig. 5. PPFs of threonine

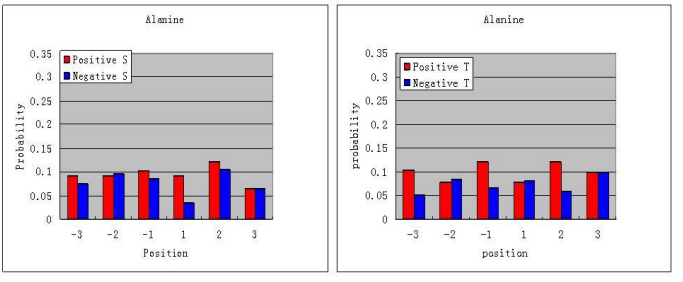


Fig. 6. PPFs of alanine

The top four principal components (the basis of subspace) are shown in Figs.7-10. It can also be seen that the glycosylated proteins all have a high proline content.

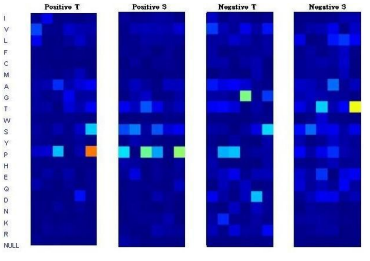


Fig. 7. 1st principal components

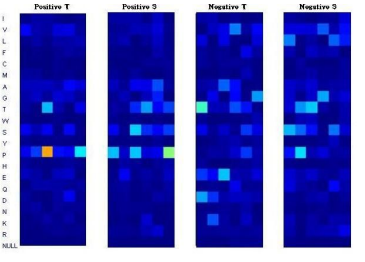


Fig. 8. 2nd principal components

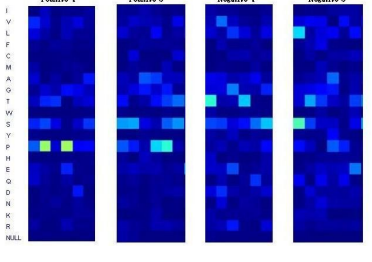


Fig. 9. 3rd principal components

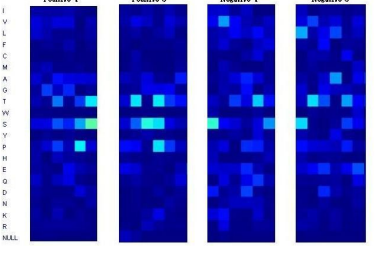


Fig. 10. 4th principal components

5 Predictions and Validations

As described in Sec.2, classification of an input protein sequence can be done by calculating the distance between the input vector and its projection into each subspace. We use the cross validation method to validate the prediction accuracy. We randomly selected 10 samples from each class respectively, which are not included in the training samples. We projected each sample into each subspace and then calculated the averaged distances using Eq.(6). One typical example of negative T is shown in Fig.11. Figure 11(a) shows the distances as a function of number of components (dimension of subspace). The calculation of Eq.(6) is corresponding to calculations of the area under the curve. The results for each subspace (curve) are shown in Fig.11(b).

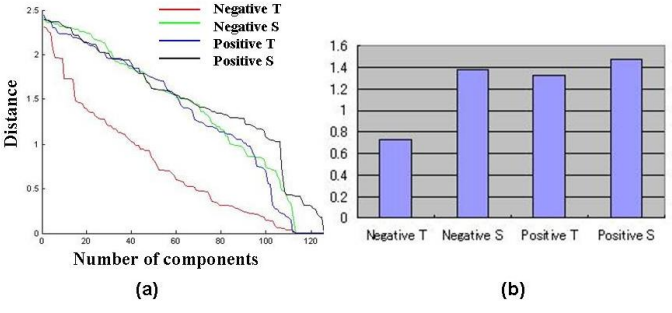


Fig. 11. Example of negative T. (a) Distance vs. number of components; (b) averaged distance.

The prediction can be viewed as a 4-class classification problem. The input protein sequence can be classified into the "nearest" class. The prediction accuracies for each class are shown in Fig.12 as results of method 1. Furthermore, the prediction can be simplified as two 2-class classification problems. The first one is to classify the test vector into either of negative class or positive class. The negative T and negative S are combined as a negative class, while positive T and positive S are combined as a positive class. The prediction accuracies are also shown in Fig.12 as results of method 2. The second 2-class classification method is to classify the test sequence into either of negative S or positive S if the residue at the center is S, otherwise it will be classified into either of negative T or positive T. The prediction results are also shown in Fig.12 as results of method 3. It can be seen that the predication

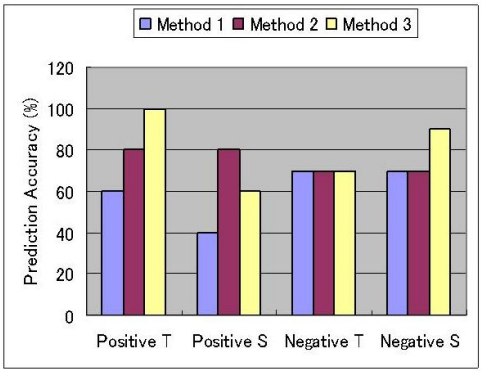


Fig. 12. Prediction accuracy for each class

accuracy for negative sites is about 70%-80%, while the prediction accuracy for positive sites is about 70%-100%.

6 Conclusions

We proposed a new approach based on principal component analysis (PCA) for pattern analysis and prediction of *O*-glycosylation sites in protein. PCA was used to find principal components and subspaces of glycosylated residues and nonglycoslylated residues, respectively. From the calculated principal compoents, we found that the *O*-glycosylted proteins have a high abundance of proline especially at positions -1 and +3. The results also show that the *O*-glycosylated proteins have a higher abundance of not only proline, but also serine, threonine and alanine than non-*O*-glycosylated sites. Futhermore, *O*-glycosylation site was predicted by PCA subspace analysis as a classification problem. We projected the protein sequence (test vector) into each subspace and calculated the distance between the test vector and its projection on the subspace. The protein sequence was classified into the "nearest" class. The prediction accuracy for non-*O*-glycosylated sites (negative sites) was about 70%-90%, and the accuracy for *O*-glycosylated sites (positive sites) was about 70%-100%.

Acknowledgments

The authors would like to thank Ms I.Nouno and Mr H.Sakamoto for their preparation of protein data. This work was supported in part by the Grant-in Aid for Scientific Research from the Japanese Ministry for Education, Science, Culture and Sports under the Grant No. 19300080.

References

1. Hart, G.W.: Glycosylation, Curr. Opin. Cell Biol. 4, 1017–1023 (1992)
2. Bause, A.: Biochem. J. 209, 331–336 (1983)
3. Gavel, Y., von Heijne, G.: vol. 3, pp. 433–442 (1990)
4. Wilson, I.B.H., Gavel, Y., Heijne, G.: Amino acid distributions around O-linked glycosylation sites. Biochem. J. 275, 529–534 (1991)
5. Elhammer, A.P., Poorman, R.A., Brown, E., Maggiora, L.L., Hoogerheide, J.G., Kezdy, F.J.: The specificity of UDP-GalNAc: polypeptide N-acetylgalactosaminyltransferase as inferred from a database of in vivo substrates and from the in vitro glycosylation of proteins and peptides. J. Biol. Chem. 268, 10029–10038 (1993)
6. Hansen, J.E., Lund, O., Engelbrecht, J., Bohr, H., Nielsen, J.O., Hansen, J.S., Brubak, S.: Prediction of O-glycosylation of mammalian proteins: specificity patterns of UDP-GaINAc: polypeptide N-acetylgalactosaminyltransferase. Biochem. J. 308, 801–813 (1995)
7. Eisenhaber, B.B., Eisenhaber, F.: Prediction of potential GPI-modification sites in proprotein sequences. J. Mol. Biol. 292, 741–758 (1999)

8. Julenius, K., Molgaard, A., Gupta, R., Brunak, S.: Prediction, conservation analysis and structural characterization of mammalian mucin-type O-glycosylation sites. Glycobiology 15, 153–164 (2004)
9. Nishikawa, I., Sakamoto, H., Nouno, I., Iritani, T., Sakakibara, K., Ito, M.: Prediction of the O-glycosylation sites in protein by layered neural networks and support vector machines. In: Gabrys, B., Howlett, R.J., Jain, L.C. (eds.) KES 2006. LNCS (LNAI), vol. 4252, pp. 953–960. Springer, Heidelberg (2006)
10. Li, S., Liu, B., Zeng, R., Cai, Y., Li, Y.: Predicting O-glycosylation sites in mammalian proteins by using SVMs. Computational Biology and Chemistry 30, 203–208 (2006)
11. Bishop, C.M.: Neural Network for Pattern Recognition. Oxford University Press, Oxford (1995)
12. http://www.ebi.uniprot.org/

Prediction of the *O*-Glycosylation with Secondary Structure Information by Support Vector Machines

Ikuko Nishikawa, Hirotaka Sakamoto, Ikue Nouno, Kazutoshi Sakakibara, and Masahiro Ito

College of Information Science and Engineering,
Ritsumeikan University, Kusatsu, Shiga 525-8577, Japan
nishi@ci.ritsumei.ac.jp

Abstract. Mucin-type *O*-glycosylation is one of the main types of the mammalian protein glycosylation. It is serine (Ser) or threonine (Thr) specific, though any consensus sequence is still unknown. In this report, support vector machines (SVM) are used for the prediction of *O*-glycosylation for each Ser or Thr site in the protein sequences. 29 mammalian protein sequences are selected from UniProt8.0, and its structure information is obtained from Protein Data Bank (PDB). A protein subsequence with a prediction target of Ser or Thr site at the center is used as input to SVM, and its amino acid sequence information, and the secondary structure or accessibility, which are calculated by DSSP from PDB data, are encoded as an input data. The results of the preliminary experiments show the effectiveness of the local structure information added to the sequence information.

Keywords: Prediction, SVM, *O*-glycosylation, Protein, Secondary Structure.

1 Introduction

Investigation of the protein glycosylation and its function is one of the main topics in understanding the life systems[1]. The carbohydrate chains play an important role in various life phenomena by combining to the protein or lipid. Especially, more than a half of the protein is glycosylated to acquire the function, biological diversity, structural stability and so on. It is also known to affect the onset of serious diseases such as Alzheimer's disease, cancer or respiratory illness. *O*-glycosylation and *N*-glycosylation are two main types in the mammalian protein glycosylation. *N*-glycosylation is a binding process to an asparagine (Asn) residue in a sequence of Asn-X-Ser/Thr, and its mechanism is clarified including the synthetic pathway of the glycoprotein. On the other hand, *O*-glycosylation is known to serine (Ser) or threonine (Thr) specific, though any consensus sequence nor the mechanism is still unknown.

The experimental findings in bioinformatics are rapidly growing recently. UniProt[2] is the largest database of the protein sequence information, which has been build by unifying the former leading databases, SwissProt and PIR. The data in UniProt are updated according to the latest experimental reports, and opened to the public through the Internet. Protein Data Bank (PDB)[3] is the largest database of the protein structure information, which contains the three dimensional structure data of the composite elements

B. Apolloni et al. (Eds.): KES 2007/ WIRN 2007, Part II, LNAI 4693, pp. 335–343, 2007.
© Springer-Verlag Berlin Heidelberg 2007

observed by the X-ray diffraction or NMR study, and it is also opened to the public. In the present research, we use the latest mammalian protein sequence data obtained from UniProt8.0, and the protein structure data from PDB to calculate the secondary structure information and the accessibility using DSSP[4], and predict the *O*-glycosylation sites by support vector machines (SVM)[5]. Similar approach was reported by Julenius et al.[6], who had constructed the original *O*-glycosylation database O-Glycbase[7], and used the layered neural network for the prediction of the glycosylation sites. Li et al.[8] used SVM for the prediction on the protein data obtained from UniProt6.1, and the approach is closer to the authors'[9,10]. In this report, the protein sequence data and its structural information are the input to SVM, and the output is the prediction on each target site of Ser or Thr in the sequence whether it is *O*-glycosylated under an appropriate condition. The effective input information for the prediction is elucidated by adding the structural information such as the secondary structure information and accessibility, changing the encoding method of the amino acid sequence, and the window size around a target site in the prediction experiments.

The paper is organized as follows. Section 2 describes the protein data used in the present experiments, and Section 3 shows the encoding method as a preprocessing of the input data. Section 4 explains the prediction by SVM and the evaluation method, and then the experimental results are shown in Section 5. Section 6 is for the present summary and future works.

2 Protein Data

2.1 Protein Sequence Data from UniProt and Structure Data from PDB

99 protein sequences of mammalian protein are selected from UniProt8.0[2], which have annotation as *O*-glycosylated from the experimental findings. Among 99 sequences, 41 proteins are found to be registered in PDB[3] with three dimensional structure data of each amino acid residue. Each data in PDB is a protein data observed in a specified experimental condition, and it is often the case that one protein is registered as multiple data in PDB with different experimental conditions and with different informations on the structure. As the result, there are 242 data files of the three dimensional structures in PDB for the above 41 proteins. Each line which describes a mutual reference information between PDB and UniProt is checked in detailed for 242 files, and the reference pairs with perfect consistency are selected. As the results, 29 proteins are finally chosen to be used in the present prediction experiment.

Among 29 proteins, there are 90 sites of Ser or Thr with the annotation (which are called positive sites in the following), while the rest 2077 sites of Ser or Thr are without annotation (which are called negative sites).

2.2 Secondary Structure and Accessibility Estimated by DSSP from PDB Data

Next, the three dimensional structure data of each amino acid residue obtained from PDB is input to DSSP[4]. DSSP is a software available on the Internet, which is able to estimate a solvent accessibility and a secondary structure of each amino acid residue, both of which are the information of a local protein structure, obtained from the three

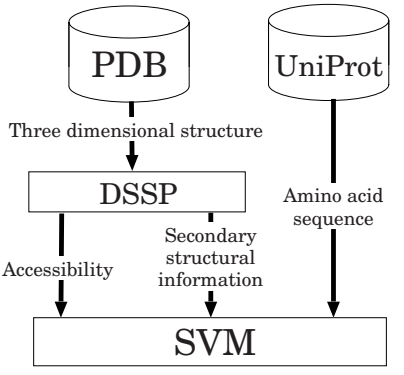


Fig. 1. Protein data obtained from UniProt, PDB and estimated by DSSP

dimensional structure data in PDB. Fig. 1 shows the flow to acquire the protein data used in the prediction.

The secondary structure is a regular structure observed in a relatively local range of a protein sequence, which is distinguished as a helix, a sheet, a turn and so on. DSSP outputs a secondary structure of each residue as one of the following seven classes, that is, H = alpha helix, B = beta-bridge, E = beta ladder, G = 3/10 helix, I = pi helix, T = hydrogen bonded turn and S = bend.

The solvent accessibility is an index of the exposure of each amino acid residue to the solvent, and given by a non-negative integer. If a residue is more exposed to have a higher accessibility, then a carbohydrate chain or enzymes may have more chance to access to the reside, therefore may affect the glycosylation process. DSSP outputs an accessibility of each residue.

3 Preprocessing for Input Data

3.1 Subsequence of Amino Acid Residues

A subsequence of window size W_s is cut out from the original protein sequence with a prediction target site at the center. For example, a target site and the first and second nearest neighbor amino acid residues on both sides compose a subsequence of $W_s = 5$. Each amino acid residue is encoded by the following three ways, to be a binary input vector to SVM. If a target site is near a terminal of a protein sequence and closer to the terminal than $(W_s - 1)/2$ residues, then the outside of the terminal is expressed by null characters.

An input subsequence of amino acid residues is expressed by a binary sequence using the following three encoding methods.

Sparse encoding: 20 kinds of amino acid residues and a null character are expressed by 21 bits, only one bit of which takes 1, in sparse encoding. Because a target site is limited to be Ser or Thr, it is expressed by 2 bits.

5-letter encoding: Amino acid residues are classified into five groups of aliphatic, charged, polarized, cyclic and the other depending on the biochemical characteristics according to [6], then 5-letter encoding expresses each site by 6 bits, including a null.A target site is expressed by 2 bits, here again.

Multiple encoding: Multiple encoding is a combination of the sparse encoding and the 5-letter encoding. That is, a length $W_c (< W_s)$ neighborhood of a target site is encoded by the former method, while the rest outer part is encoded by the latter.

3.2 Secondary Structure and Accessibility

Secondary Structure: Seven classes of the output from DSSP are unified into three main classes. Namely, H and E are remained as distinguished classes of alpha helix and beta strand, respectively, and the other B, G, I, T and S are all unified into C, which stands for coil. Three class of each amino acid residue is coded by 3 bits, only one bit of which takes 1. A residue for which DSSP fails to output the estimation, or for which there is no three dimensional information in PDB, is coded by 3bits of zeros.

Accessibility: Accessibility takes a non-negative integer. Therefore, the value is normalized and expressed by a real number in $[0, 1]$.

3.3 Numbers of Positive and Negative Data

The present goal is to predict whether each target site will be *O*-glycosylated (positive) or not (negative). However, the number of Ser or Thr sites which are experimentally verified as positive still remain rather small compared with the negative, which are defined as not verified as positive up to now. Therefore in the prediction experiment in the next section, all available 90 positive sites on 29 protein sequences are used, while only a part of 2077 negative sites are randomly chosen to be used. The twice number of positive sites, that is, 180 negative sites are chosen to compose 270 sites in total.

4 Prediction by SVM

4.1 SVM

SVM is used for the prediction. One of the advantages to use SVM in the present study is that the increase of the dimension of the input vector does not lead to the explosion of the calculation time during the learning. Moreover, for a reasonable learning of the layered neural network, for example, the number of the adjustable parameters should be kept limited owing to the limited number of the available learning data, therefore W_s should be limited, while SVM gets rid of this kind of limitation.

Radial Basis Function given by (1) is used as a kernel;

$$K(\boldsymbol{x}, \boldsymbol{x}') \propto \exp(-\gamma ||\boldsymbol{x} - \boldsymbol{x}'||^2) \ , \tag{1}$$

where γ is a kernel parameter. Another parameter is a margin size denoted by C. A famous open software package SVM-light[11] is used in the following experiment.

4.2 Parameter Setting

W_s is set to 3, 5, 7, 9, 11, 21 or 31 for each of the three encoding methods. The margin size parameter C of SVM is varied from 0.1 to 100, and kernel parameter γ is varied from 1.0×10^{-4} to 1.0×10^{0}.

4.3 Cross Validation

3-fold cross validation is used to fully utilize the limited number of the data. The 29 protein sequences are divided into 3 sections, and the positive and negative Ser or Thr sites on the proteins contained in each section make up one dataset. The number of positive sites in each dataset is 29, 32 and 29. The sizes of 3 datasets are set equal by adjusting the numbers of randomly chosen negative sites. Therefore, the number of negative sites in each dataset is 61, 58 and 61, respectively.

4.4 Evaluation Methods

The following two kinds of quantities are used for the quantitative evaluation of the classification into the positive and negative classes.

Matthews Correlation Coefficient: The sizes of two classes are not equal as the number of the positive sites are limited, therefore we use Matthews correlation coefficient r, which is defined by;

$$r = \frac{t_{\mathrm{p}} t_{\mathrm{n}} - f_{\mathrm{p}} f_{\mathrm{n}}}{\sqrt{(t_{\mathrm{n}} + f_{\mathrm{n}})(t_{\mathrm{n}} + f_{\mathrm{p}})(t_{\mathrm{p}} + f_{\mathrm{n}})(t_{\mathrm{p}} + f_{\mathrm{p}})}} \ , \tag{2}$$

where t_{p} is the number of correctly classified as positive sites, and f_{p} is the number of falsely classified as positive sites. Similarly, t_{n} and f_{n} are the numbers of truly negative and falsely negative sites, respectively. r takes a real value in $[-1.0, 1.0]$. $r = 1.0$ if all the data are correctly classified, while $r = -1.0$ if all are falsely classified. If all the data are classified into one class, which means no correlation between the prediction and the correct class, then r is 0.0.

Accuracy, Sensitivity and Specificity: As the evaluation indices of the correct classification, accuracy A_{c}, sensitivity S_{n} and specificity S_{p} are used, which are defined by the following (3), (4) and (5), respectively;

$$A_{\mathrm{c}} = \frac{t_{\mathrm{p}} + t_{\mathrm{n}}}{t_{\mathrm{p}} + t_{\mathrm{n}} + f_{\mathrm{p}} + f_{\mathrm{n}}} \ , \tag{3}$$

$$S_{\mathrm{n}} = \frac{t_{\mathrm{p}}}{t_{\mathrm{p}} + f_{\mathrm{n}}} \ , \tag{4}$$

$$S_{\mathrm{p}} = \frac{t_{n}}{t_{\mathrm{n}} + f_{\mathrm{p}}} \ . \tag{5}$$

These are the ratios of the correct classification of all, positive and negative sites, respectively.

5 Results and Discussion

5.1 Dependency on the SVM Parameters

First, the results obtained by various values of SVM parameters are compared by Matthews correlation coefficients r. Margin size parameter C ranges from 0.1 to 100, RBF kernel parameter γ from 1.0×10^{-4} to 1.0×10^{0}, and window size is $W_s = 3, 5, 7, 9, 11, 21$ or 31. The sparse encoding is used for a sequence information, and 3bits of secondary structure class information or a normalized accessibility are added to compose an input vector.

Fig. 2 (left) is the C, γ-dependency of the correlation coefficient r for the input with an accessibility at $W_s = 7$. r drops for $C < 1.0$, while there is little dependency on C for $C > 1.0$. On the other hand for γ dependency, large r is attained by small γ, especially around $0 < \gamma < 0.3$, and the maximum r is obtained at $\gamma = 0.0025$ with $C = 25$. This is found to be the case for the input without accessibility, and with a secondary structure.

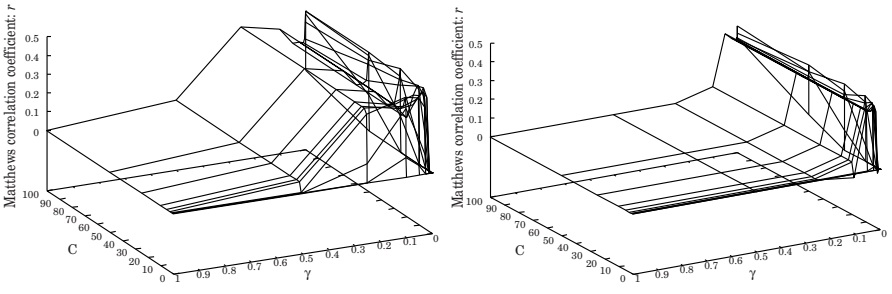


Fig. 2. SVM parameters C, γ-dependency of Matthews coefficients at $W_s = 7$ (left) and 21 (right). Input is a sequence information and an accessibility.

Fig. 2 (right) is the C, γ-dependency of r for the input with an accessibility at $W_s = 21$. In this figure, a rather sharp peak is found in γ dependency. This is also found to be the case for the input without accessibility, and with a secondary structure. C, γ-dependency are similar to these figures in the present range of W_s, and therefore we can chose appropriate values for C and γ within this range.

5.2 Comparison of the Results With and Without Structure Information

Then, the prediction results with and without a secondary structure class or an accessibility are compared by Matthews correlation coefficient r. The followings are the results obtained by the sparse, 5-letter, and multiple encodings with window size $W_s = 3, 5, 7, 9, 11, 21$ and 31. For the multiple encoding, $W_c = 3, 5, 7, 9$ and 11. The best C and γ are used for each setting.

Figs. 3 and 4 summarizes the results for the sparse, and the 5-letter and multiple encodings, respectively. For the multiple encoding the results with $W_c = 7$ are shown.

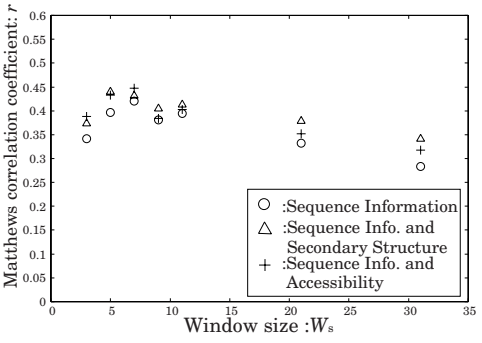


Fig. 3. Matthews coefficients obtained by SVM with/without the structure information using the sparse encoding

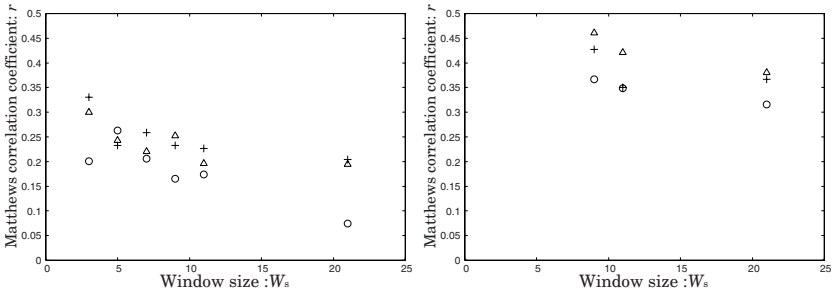


Fig. 4. Matthews coefficients obtained by SVM with/without the structure information using the 5-letter encoding (left) and the multiple encoding with $W_c = 7$ (right)

By the sparse encoding shown in Fig. 3, the highest value $r = 0.477$ is obtained with a sequence information with an accessibility at $W_s = 7$, with $C = 25$ and $\gamma = 0.0025$. r tends to decrease for larger window size $W_s > 11$ for all three kinds of input information. This may be caused by the small number of data, because it is observed that when all the data of 99 proteins are used for the input without the structure information, then r never decrease even if W_s is extended to 51. The results obtained by the inputs with either structure information, a secondary structure class or an accessibility, attain higher r at all window sizes. In many cases, a secondary structure gives better performance than an accessibility. The similar tendencies are shown in Fig. 4 for the 5-letter and multiple encodings, though the performance is rather worse in the former.

Table1 shows the accuracy A_c, sensitivity S_n and specificity S_p obtained by the sparse encoding at $W_s = 5, 7$ and 21 , together with the results obtained by the 5-letter and multiple ($W_c = 7$) encodings at $W_s = 9$. The results by the sparse encoding indicate that the better performance with a structure information is mainly given by the better specificity for smaller W_s, while by the better sensitivity for larger W_s. Overall values are not satisfactory, especially by the 5-letter encoding, and for the sensitivity S_n, that is a correct classification ratio of the positive sites. The worse results for the positive

Table 1. Accuracy, sensitivity and specificity by the sparse encoding at $W_s = 5$, 7 and 21 (left), by the 5-letter encoding at $W_s = 9$, and the multiple encoding with $W_c = 7$ at $W_s = 9$ (right)

	A_c	S_n	S_p	A_c	S_n	S_p
	Sparse, Ws=5			5-letter, Ws=9		
Sequence information	0.733	0.600	0.800	0.663	0.289	0.850
Secondary Structure	0.759	0.522	0.877	0.689	0.378	0.844
Accessibility	0.755	0.522	0.872	0.678	0.389	0.822
	Sparse, Ws=7			Multiple(Wc=7), Ws=9		
Sequence information	0.755	0.533	0.866	0.722	0.578	0.795
Secondary Structure	0.759	0.477	0.900	0.763	0.600	0.844
Accessibility	0.751	0.622	0.816	0.759	0.522	0.878
	Sparse, Ws=21					
Sequence information	0.696	0.555	0.767			
Secondary Structure	0.707	0.644	0.739			
Accessibility	0.688	0.644	0.711			

sites may be caused by the difference in the numbers of the positive and negative sites in the learning data. Even by the 5-letter encoding, the learning data are mostly correctly classified to attain the accuracy A_c about 0.9 for any input information. However, more intensive investigation is necessary for any conclusion.

The present experiments are preliminary, but the results show the effectiveness of adding the local structure information, which is expressed by a secondary structure and an accessibility, for the prediction of *O*-glycosylation.

6 Summary and Future Works

The present report on the prediction of *O*-glycosylation by SVM shows the effectiveness of the structure data, additional to a sequence information. More intensive investigation with different ratios of the negative and positive sites data with more number is necessary for further discussion. The future works include a study of new encodings and other kernel functions to elucidate the effective information for *O*-glycosylation and thereby to clarify the mechanism.

Acknowledgment

This work has been supported in part by the Japanese Ministry of Education, Culture, Sports, Science and Technology, under a Grant-in-Aid for Scientific Research on Priority Areas “Emergence of Adaptive Motor Function through Interaction between Body, Brain and Environment”, and under a Grant-in-Aid (B) Project No.19300080.

References

1. Taylor, M.E., Drickamer, K.: Introduction to Glycobiology. Oxford Univ. Press, Oxford (2003)
2. `http://www.ebi.uniprot.org`

3. http://www.rcsb.org/pdb/home/home.do
4. http://swift.cmbi.kun.nl/swift/dssp
5. Cristianini, N., Shawe-Taylor, J.S.-: An Introduction to Support Vector Machines and other kernel-based learning methods. Cambridge University Press, Cambridge (2000)
6. Julenius, K., Molgaard, A., Gupta, R., Brunak, S.: Prediction, conservation analysis, and structural characterization of mammalian mucin-type *O*-glycosylation sites. Glycobiology 15(2), 153–164 (2004)
7. http://www.cbs.dtu.dk/databases/oglycbase/
8. Li, S., Liu, B., Zeng, R., Cai, Y., Li, Y.: Predicting *O*-glycosylation sites in mammalian proteins by using SVMs. Computational Biology and Chemistry 30, 203–208 (2006)
9. Nishikawa, I., et al.: Prediction of the *O*-glycosylation sites in protein by layered neural networks and support vector machines. In: Gabrys, B., Howlett, R.J., Jain, L.C. (eds.) KES 2006. LNCS (LNAI), vol. 4252, pp. 953–960. Springer, Heidelberg (2006)
10. Nouno, I., et al.: Prediction of mucin-type *O*-glycosylation by layered neural networks and support vector machines. In: Proceedings of the 17th Int. Conference on Genome Informatics (December 2006)
11. http://svmlight.joachims.org

A Simple Model for Human-Robot Emotional Interaction

Isabella Cattinelli and N. Alberto Borghese

Dept. Computer Science, Università degli Studi di Milano
Via Comelico 39/41, 20135 Milan, Italy
{cattinelli,borghese}@dsi.unimi.it

Abstract. This paper describes a model of human-robot interaction. It computes the robot's simulated emotional state as a result of the emotional input received from the user and the robot's internal emotion dynamics. The latter is determined by the key elements of personality and attitude. The proposed model is a probabilistic finite state automaton, where the probabilistic transition function is called personality and is modified during the interaction according to the history of the interaction itself. The criterion which drives this personality change is called attitude. Such a model may be used in affective computing applications in order to design emotional agents and robots.

Keywords: Affective Computing, Probabilistic Finite State automata, Human-Robot Interaction.

1 Introduction

Affective computing [1] is an emerging branch of Computer Science that is aimed at merging psychology, cognitive science and technology in order to design devices able to understand human emotions and to react accordingly, even displaying their own (simulated) emotions. The final goal is to realize machines endowed with *emotional intelligence*. Such devices can communicate with users not only by a limited set of standard commands, but also in simpler and more natural ways, for instance through body language [2]. This results in a simpler and more useful technology, since people can approach a machine like they would do with another person. Also, adding an emotional component to a machine, especially a robot, may help social acceptance of this technology.

Applications include social robots [3,4], emotion detection using non verbal communication (posture [5], gestures [6], facial expressions [7], etc.), systems for health care [8], computational humor [9], and many others. In the following, we describe a novel model of human-robot interaction, which computes the simulated emotional states produced by the robot in response to the user's input, according to the robot's internal emotion dynamics.

B. Apolloni et al. (Eds.): KES 2007/ WIRN 2007, Part II, LNAI 4693, pp. 344–352, 2007.
© Springer-Verlag Berlin Heidelberg 2007

2 Interaction Model

In our model, the user and the robot communicate by exchanging emotions: the user displays her current emotional state, e.g. by a specific facial expression. This input, along with a set of parameters, triggers a state transition inside the robot, which will then express its new emotional state through an appropriate behavior (emission of sounds, lights, motions, etc.).

User's emotion detection is quite a difficult task, since an emotion is a complex phenomenon which involves cognitive, neurophysiological and motor expression events. Devices have been created to record physiological parameters [10], such as electromyogram or blood volume pressure, that can reveal modifications in a person's state when subject to specific emotions. A less invasive approach consists in analyzing exterior displays of emotion, via non verbal communication: gestures, postures, voice pitch can all be used to convey emotions. However, the most important channel of body language, when it comes to expressing emotions, is face; facial expressions have been extensively analysed as a mean for extracting information about human emotions. A widespread system for cataloging facial expression is FACS, *Facial Action Coding System* [11], which codes basic face local movements, named *Action Units* (AU). Different sets of AUs can be associated to different emotions. Therefore, by recognizing the AUs on the user's face and linking them to their associated emotion expression, it is possible to infer the user's emotional state (or, at least, the emotional state she wants to communicate, since facial expression can be controlled and faked).

The detected emotions constitute the inputs for our interaction model. In such a context, a natural formalism to be adopted is that of finite state automata [12], also proposed in [13] and [14]. In particular, our model is a probabilistic finite state automaton (PFSA) $\langle Q, U, P, q_0 \rangle$, where:

- Q is the (finite) set of emotional states (e.g. happy, sad, angry, etc.);
- U is the (finite) set of input, that is user's emotions;
- $P = \{P_0, P_1, \ldots\}$ is the set of probabilistic transition functions, such that $P_t : Q \times U \times Q \to [0, 1]$ for $t = 0, 1, \ldots$;
- q_0 is the initial state.

The set of states and the set of inputs can be defined according to the specific context where the interaction takes place. According to Ekman's theory [15], we can isolate six basic emotions (joy, sadness, surprise, anger, fear and disgust) which may constitute our set of states; we can also expand this set by considering different intensity levels for each basic emotion. The initial state may be a neutral condition, suggesting an initial absence of emotion. Similarly, the set of inputs will coincide with the set of emotional states that the interaction partner can experience. So, the set of inputs U and the set of states Q may coincide, meaning that both partners in interaction can feel and express to each other the same emotions.

The transition function at time 0, P_0, is called *personality*. We can fill different personality files containing the initial probability values for state transitions. Each file will include different values, reflecting a specific personality the robot

can assume. For instance, the friendly personality file will be characterized, in general, by high probabilities of entering the joy state. Setting P_0 results therefore in setting the personality of the robot during the interaction.

Interactions are not static phenomena. It is likely that, during an interaction, one person's behavior is going to change, according to the overall trend of the interaction. This idea of evolution in interaction can be captured by modifying the transition probabilities according to the interaction history and to the *attitude*, that is the criterion which regulates the update. This can be achieved by measuring the quality of inputs received so far and modifying the stored probability values so that:

- the robot's behavior will follow the one shown by the user so far (imitative attitude); or
- the robot's behavior will diverge from that of the user (compensatory attitude).

Let us group the inputs in categories (nice inputs, bad inputs, sad inputs, etc.); for instance, the user's expression of joy or surprise can constitute the nice input category, anger and disgust can be associated to the bad input category, and fear and sadness to the sad input group. Let us also define a counter for each category, which is updated every time the robot receives an input belonging to that category. We then update the transition probability values (given by P_t) when at least one of the counters exceeds the update threshold. The result will be new probability values for all transitions, represented by P_{t+1}. The update consists in increasing the probability of entering some *target states*, which are determined, for each input category, by the selected attitude. For instance, assuming that the counter associated to the nice input category went over threshold at time t, and that the selected attitude is imitative, we increase all the probabilities of entering the JOYFUL state (that is, the target state associated, in imitative attitude, to the nice input group) and, of course, we decrease every other value so that $\sum_{q' \in Q} P_{t+1}(q, u, q') = 1 \ \forall q \in Q, u \in U$. After updating, the counter is reset to zero.

Therefore, at each time, the emotional state of the robot depends not only on the input received from the user, but also on the internal dynamics, represented by the concepts of personality and attitude. By setting these parameters and providing adequate inputs, the user can lead the interaction toward specific results (i.e. frequently obtaining joyful responses). The probabilistic nature of the model allows for rich interactions and reflects the unpredictability of human relationships.

3 Extended Model

We presented, so far, a model where inputs come from a user and a robot enters different emotional states. It is possible to generalize our model if we assume that the interaction partners are two generic agents, each of whom represented by a finite state automaton. These two automata are going to interact in the

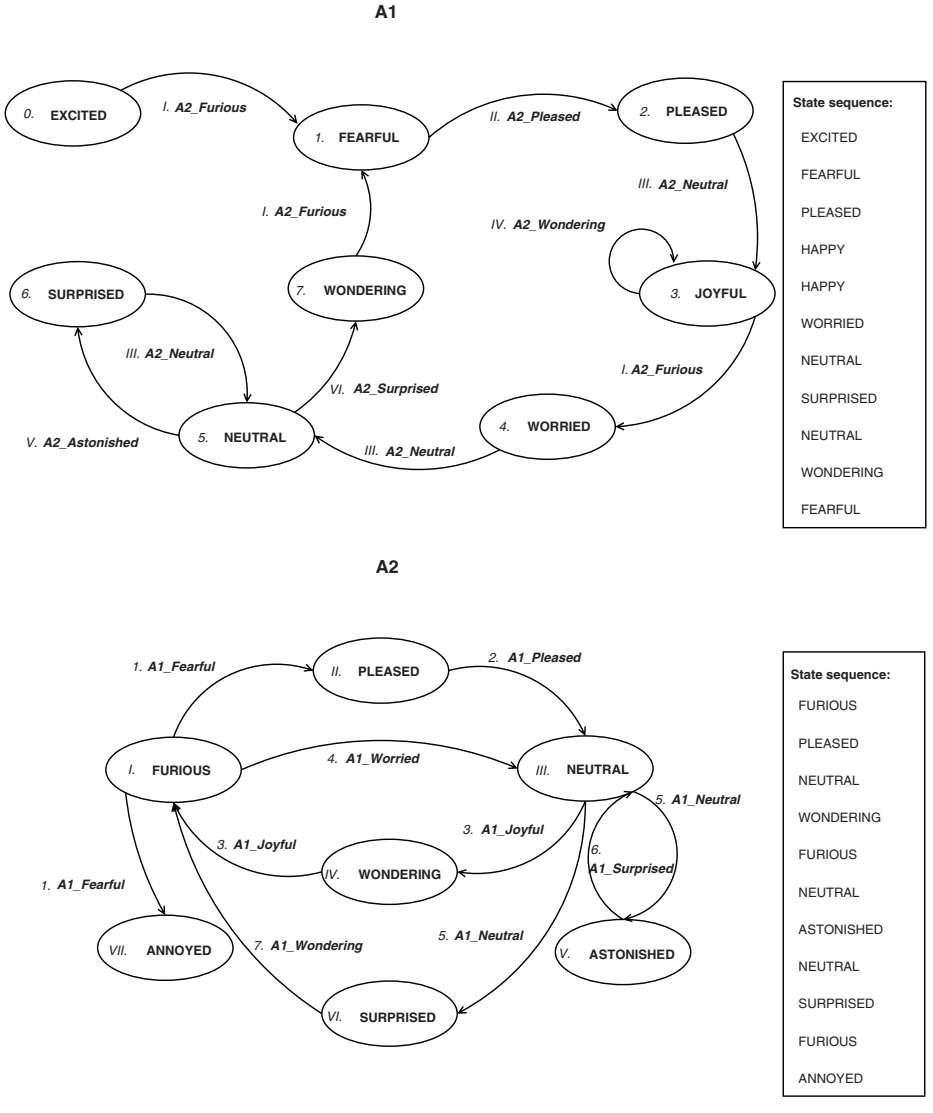


Fig. 1. Example of interaction on the two-automata model. We label with arabic numbers the states in A^1, corresponding to the inputs to A^2; with roman numbers we show the states in A^2 and, consequently, the inputs in A^1.

sense that one's state becomes the other's input, and vice versa. The model now includes two automata, $A^1 = \langle Q, U, P^1, q_0 \rangle$ and $A^2 = \langle U, Q, P^2, u_0 \rangle$, where the set of states of A^1 coincides with the set of inputs of A^2, and vice versa. By means of this model, it is possible to observe a wide variety of interactions, obtained by simply modifying the personalities and attitudes of the two agents. Moreover, we extended the set of emotional states to include different levels of

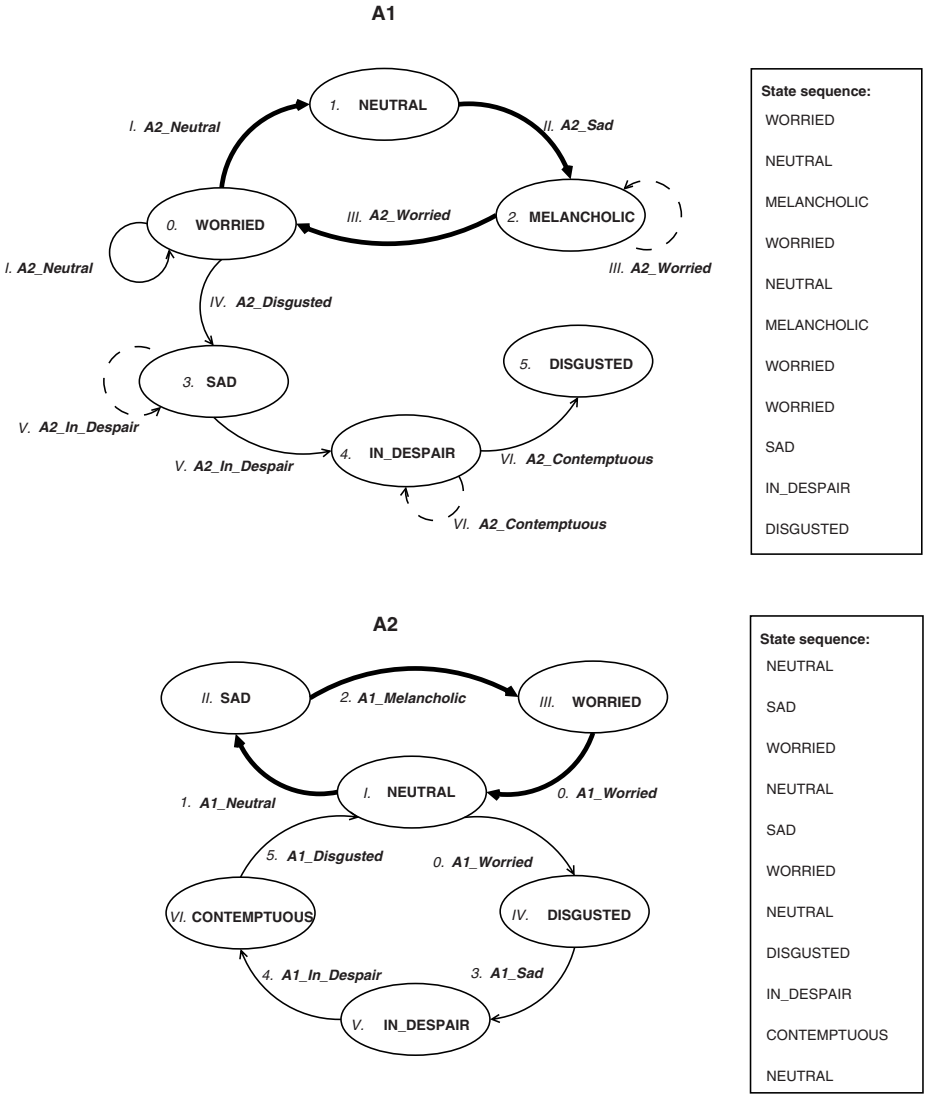


Fig. 2. Example of interaction: the cycle in bold was predicted as a high probability cycle

emotions, so that, for instance, the basic emotion of joy corresponds now to three different states: PLEASED (low intensity joy), JOY (medium level) and EXCITED (high level). In Figure 1 an example of interaction on such a model is shown: ten steps of interaction are shown through the state transitions occurred in each of the two automata. Numbers are used to show correspondence between states in one automaton and inputs in the other one.

Such a model can be used as a tool for studying human-robot or human-human interactions; in particular, high probability cycles of interaction can be extracted and interpreted as behavioral patterns. We define *cycle* a sequence of states where the first and final states are the same; we are interested in extracting all cycles of length n on one of the two automata (e. g. A^1), whose probability to occur is higher than a given threshold. To do so, let us assume that the transition function remains unchanged during the interaction (stationary problem), and that A^2 is deterministic, that is $A^2 = \langle U, Q, \delta, u_0 \rangle$ where $\delta : Q \times U \rightarrow Q$ returns univocally the next state, given the current state and input. In such a context, we can explore the tree of all possible state sequences on A^1, starting from q_0, computing the probability of each sequence and pruning every branch whose probability is less than the given threshold. Since our model is Markovian (i. e. the probability of a transition does not depend on previous transitions), the probability of a sequence is given by the product of the probabilities of each transition composing the sequence. Adding a new transition to the current sequence, therefore, cannot increase the overall probability of that sequence, since this requires a multiplication by a number in $[0, 1]$. This allows for an early pruning of possible solutions, because all sequences under construction whose probability goes under the threshold can immediately be discarded.

All high probability cycles in A^1 will be obtained; these cycles can help to predict the overall trend in the interaction and extract specific interaction patterns. Changing the interacting personalities will result in emerging of new interaction patterns. In Figure 2 we show another example of interaction: we highlighted a high probability cycle (its probability is 0.24) that in the current instance actually took place. In dashed arrow we show high probability transitions that, however, in the considered example, did not occur.

4 Discussion and Conclusion

The main result of this paper is the design of an interaction model, which can easily be extended and personalized. One can quickly add new states, inputs, personalities and attitudes. Such a model can be employed in many different fields, as a basis for the creation of rich emotional agents; the definition of a personality file allows to easily manipulate the agent's behavior at run-time. The model can also be used to endow robots of an emotional component, in order to make them more useful and integrated tools in the society. The interaction model can also be used to investigate behavioral dynamics, for instance by extracting high probability patterns of interaction or estimating the frequency of a particular emotional state, etc.

We should stress that the aim of this work was to provide a simple framework for human-machine emotional *interaction*. In the proposed model, emotional states are monolithic, discrete and finite states of an automaton. This is necessarily a simplification of what an emotion truly is: emotion theorists largely agree that emotions are composed of several different features, such as the motivational or the motor expression component. Scherer [16] proposes that emotion

is a process made up of at least three aspects: evaluation of the relevance for the subject of some (possibly external) input, preparation of an action in response to the stimulus, and communication of the emotional state to the social environment. In particular, the first of these steps requires subsequent checks on different features of the stimulus, for instance its novelty, or its inherent pleasantness; the emotional state of the organism is determined by the outcome of this hierarchical sequence of checks. A direct consequence of this theory is that emotional states continuously change in response to these cognitive appraisals. Such a continuous process is, in our model, discretized: evaluation of the external input (the user's emotion) prompts the entering of the machine's emotional state, and no other state is entered until a new input is provided. We can imagine that the one state produced by the robot at each interaction step is just the final product of the dynamic emotional process started by the user's input; in other words, the user is only able to see the final emotional state entered by the robot, not all the intermediate steps (which might, even in human-human interactions, be concealed). Currently, our aim consists in modeling emotional interaction, rather than focusing on emotion production and display: discretized emotional states can be a simple yet reasonable way to reach our goal. However, future work on the model could investigate more deeply the opportunity to introduce a temporal dynamic to emotional state transitions.

While the extended model (see Section 3) considers different intensity levels for the basic emotions, allowing us to represent part of the variability inherent in the emotion space (e.g. one can be happy in very different ways), it is still true that emotions have a number of different gradations which cannot be entirely captured by our model. The same is true for mixed emotions, too. Gradations could be better described in dimensional approaches, where each emotional state is represented by a point in the emotion space, organized by affective dimensions (such as positive/negative). However, this kind of granularity is not required in an interaction framework: we can see that our interlocutor is mildly angry, or very angry, but we cannot be certain that the emotion displayed is, say, 30% high-intensity anger mixed with 70% very low-intensity fear. While interacting with another individual, we roughly categorize the displayed emotional states: the number and nature of these categories may vary (and they can be easily modified in the model), but, even though we can, in principle, perceive a high number of emotion gradations, the tendency will be that of clustering them into wider categories. Of course, this may be appropriate while modeling emotional interaction, but may not when focusing on emotion production and display: in such cases, a fuzzy approach would be more suitable.

We are now experimenting the proposed model in a real human-robot interaction, employing the Sony AIBO [17] robot. We implemented a simple module for facial expression recognition based on FACS, using basic image processing techniques (color segmentation, edge extraction, etc.) in order to get a real-time response (less than 1 s of processing time on AIBO's on board processor). Figure 3 shows a preview of this application: the user's emotional expression is processed by AIBO to determine, along with its personality and attitude, its next state.



Fig. 3. A step of interaction with AIBO: the user shows an expression of joy (left image), and AIBO enters the JOYFUL state (right image). Its new state is displayed by motions, sounds and lights.

Acknowledgments. We are grateful to Prof. Maria Rita Ciceri for helpful discussions and for providing us with useful references in the psychology field.

References

1. Picard, R.W.: Affective Computing. MIT Press, Cambridge (1997)
2. Argyle, M.: Bodily Communication. International Universities Press, Inc. (1975)
3. Kismet: (1998), Web site
   `http://www.ai.mit.edu/projects/humanoid-robotics-group/kismet/kismet.html`
4. Leonardo: (2001), Web site
   `http://robotic.media.mit.edu/projects/Leonardo/Leo-intro.html`
5. Silva, P.R.D., Bianchi-Berthouze, N.: Measuring posture features saliency in expressing affective states. In: Proceedings of the 2004 IEEE International Conference on Intelligent Robotics and Systems (IROS'04), pp. 2003–2008 (2004)
6. Silva, P.R.D., Osano, M., Marasinghe, A., Madurapperuma, A.P.: Towards recognizing emotion with affective dimensions through body gestures. In: FGR '06: Proceedings of the 7th International Conference on Automatic Face and Gesture Recognition (FGR06), pp. 269–274 (2006)
7. Anderson, K., McOwan, P.: A real-time automated system for the recognition of human facial expressions. IEEE Transactions on Systems, Man and Cybernetics, Part B 36(1), 96–105 (2006)
8. Lisetti, C., LeRouge, C.: Affective computing in tele-home health. In: HICSS '04: Proceedings of the Proceedings of the 37th Annual Hawaii International Conference on System Sciences (HICSS'04) - Track 6 (2004)
9. Binsted, K., Bergen, B., Coulson, S., Nijholt, A., Stock, O., Strapparava, C., Ritchie, G., Manurung, R., Pain, H., Waller, A., O'Mara, D.: Computational humor. IEEE Intelligent Systems 21(2), 59–69 (2006)
10. Picard, R.W., Vyzas, E., Healey, J.: Toward machine emotional intelligence: Analysis of affective physiological state. IEEE Transactions on Pattern Analysis and Machine Intelligence 23(10), 1175–1191 (2001)

11. Ekman, P., Friesen, W.V.: Manual for the Facial Action Coding System. Consulting Psychologists Press, Inc., Palo Alto (1978)
12. Hopcroft, J.E., Ullman, J.D.: Introduction to automata theory, languages and computation. Addison-Wesley, Reading (1979)
13. Kopecek, I.: Constructing personality model from observed communication. In: Proc. 9th International Conference on User Modeling; Assessing and Adapting to User Attitides and Effect: Why, When and How?, pp. 28–30 (2003)
14. Gmytrasiewicz, P.J., Lisetti, C.L.: Emotions and personality in agent design. In: AAMAS '02: Proceedings of the First International Joint Conference on Autonomous Agents & Multiagent Systems, July 15-19, 2002, Bologna, Italy, pp. 360–361 (2002)
15. Ekman, P.: An argument for basic emotions. Cognition and Emotion 6(3-4), 169–200 (1992)
16. Scherer, K.R.: On the nature and function of emotion: A component process approach. In: Approaches to emotion, pp. 293–317. Lawrence Erlbaum Associates, Mahwah (1984)
17. AIBO: Web site (1999), `http://www.sony.net/Products/aibo`

Applying Trust, Reputation and Intuition Aspects to Support Virtual Communities of Practice

Juan Pablo Soto, Aurora Vizcaíno, Javier Portillo-Rodríguez, and Mario Piattini

Alarcos Research Group, Information Systems and Technologies Department, Indra-UCLM Research and Development Institute, University of Castilla-La Mancha
Ciudad Real, Spain
jpsoto@proyectos.inf-cr.uclm.es,
{aurora.vizcaino,mario.piattini}@uclm.es,
javier.portillo@alu.uclm.es

Abstract. Communities of Practice (CoPs) are places in which the flow of knowledge is generated as a result of the confidence that members have in each other. The main goal of this research is to emulate these communities' behavior in knowledge management systems. As a consequence of this, an increase in the exchange information between the members of a virtual community is expected. However, upon starting this work, we encountered several problems. For instance, how could we implement or rate social aspects such as trust and reputation?. This paper describes our advances in resolving these problems.

Keywords: Trust, Reputation, Communities of Practice.

1 Introduction

In recent years Knowledge Management (KM) has become an important success factor for companies. The purpose of knowledge management is to help companies to create, share and use knowledge more effectively [5]. Information technologies play a key role in achieving these goals but are only a small component in an overall system that must integrate the supporting technology with people-based business processes. KM is not a technology solution but is rather primarily a people oriented process, dealing with, for example, leadership, culture, expertise and learning, with technology playing a supporting role. Using this idea as a base we have studied how people obtain and increase their knowledge in their daily work. From this study we realised that employees frequently exchange knowledge with people who work on similar topics as them and consequently, either formally or informally, communities are created which can be called "communities of practice", by which we mean groups of people with a common interest where each member contributes knowledge about a common domain [17].

Communities of practice (CoPs) enable their members to benefit from each other's knowledge. This knowledge resides not only in people's minds but also in the interaction between people and documents. CoPs share values, beliefs, languages, and ways of doing things. Many companies report that such communities help reduce problems due to lack of communication, and save time by "working smarter" [18]. An

B. Apolloni et al. (Eds.): KES 2007/ WIRN 2007, Part II, LNAI 4693, pp. 353–360, 2007.
© Springer-Verlag Berlin Heidelberg 2007

interesting fact is that members of a community are frequently more likely to use knowledge built by their community team members than that created by members outside their group [6]. This factor occurs because people trust in the information offered by a member of their community more than in that which is supplied by a person who does not belongs to that community. Of course, the fact of belonging to the same community of practice already implies that the members have similar interests and perhaps the same level of knowledge about a topic. Consequently, the level of trust within a community is often higher than that which exists outside the community. Because of this, as is claimed in [6], knowledge reuse tends to be restricted within groups. Therefore, people, in real life in general and in companies in particular, prefer to exchange knowledge with "trustworthy people" by which we mean people they trust. It is for these reasons that we consider the implementation of a mechanism in charge of measuring and controlling the confidence level in a community where the members share information to be important.

Bearing in mind that people exchange information with "trustworthy knowledge sources", we have designed a prototype in which software agents try to emulate the human process of evaluating knowledge sources with the goal of fostering the use of knowledge bases in companies where agents provide the employees with "trustworthy knowledge".

The remainder of this work is organized as follows. The next section defines two important concepts that take place in the process of obtaining information (trust and reputation). Then, in Section 3 our proposal to rate reputation is described. In Section 4 we illustrate how the formulas defined for calculating reputation in virtual communities has been implemented in a prototype which detects and suggests trustworthy documents for members of CoPs. Finally, in Section 5, conclusions are described.

2 Trust and Reputation

Trust and reputation mechanisms have been proposed for use in various domains such as e-commerce [20], peer-to-peer computing [15], recommender systems [14], etc. However, there is no universal agreement on the definition of trust and reputation. Because the main goal of our work is to rate the credibility of information sources and of knowledge we first need to define these two important concepts.

2.1 Trust

Trust is a complex notion whose study is usually narrowly scoped. This has given rise to an evident lack of coherence amongst researchers in the definition of trust. For instance in [3] the authors define trust as confidence in the ability and intention of an information source to deliver correct information. Wang and Vassileva in [15] define trust as a peer's belief in another peer's capabilities, honesty and reability based on his/her own direct experiences. In [12] trust is defined as a subjective expectation an agent has about another's future behavior based on the history of their encounters.

Social scientists have collectively identified three types of trust, which are:

- *Interpersonal trust* which is the trust one agent has in another agent directly. This trust is agent and context specific [10].

- *System trust* or *impersonal trust* refers to trust that is not based on any property or state of trustee but rather on the perceived properties or reliance on the system or institution within which that trust exists. For instance, inherited experiences of an organization.
- *Dispositional trust*, or *Basic trust*, describes the general trusting attitude of the truster. This is "a sense of basic trust, which is a pervasive attitude toward oneself and the world" [10].

Experience and knowledge form the basis for trust in future familiar situations [9]. For this reason, the frequency and intensity of interactions between people provide an increased level of habituation which reinforces trust between the parties. Another concept highly related to trust is reputation which is described in the following section.

2.2 Reputation

Several definitions of 'reputation' can be found in literature. For instance, Mui et al in [12] define reputation as a perception that one agent has of another's intentions and norms. Barber and Kim define this concept as the amount of trust an agent has in an information source, created through interactions with information sources [3]. This definition is the most appropriate for our research since the level of confidence in a source is based on previous experience of it. The reputation of an information source not only serves as a means of belief revision in a situation of uncertainty, but also serves as a social law that obliges us to remain trustworthy to other people. Additionally, a good reputation may also be used to a person's advantage, as reputation is also considered as a form of social capital, especially in commerce [4]. Therefore, people, in real life in general and in companies in particular, prefer to exchange knowledge with "trustworthy people" by which we mean people they trust. People with a consistently low reputation will eventually be isolated from the community since others will rarely accept their justifications or arguments and will limit their interaction with them. It is for this reason that the remainder of this paper deals mainly with reputation. However, if we attempt to imitate the behavior of the employees in a company when they are exchanging and obtaining information we observe that apart from the concept of reputation other factors also influence. For this reason, in this paper we argue that reputation is not a single notion but one of multiple parts. These are:

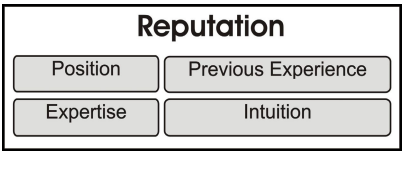


Fig. 1. Reputation factors

- *Position:* employees often consider information that comes from a boss as being more reliable than that which comes from another employee in the same (or a lower) position as him/her [16]. However, this is not a universal truth and

depends on the situation. For instance in a collaborative learning setting collaboration is more likely to occur between people of a similar status than between a boss and his/her employee or between a teacher and pupils [7]. In an enterprise this position can be established in different ways, for instance by using an organizational diagram or classifying the employees according to the knowledge that a person has, as can be seen in Allen's proposal in [2], which distinguishes between:

- Technological gatekeepers, defined as those actors who have a high level of knowledge interconnectedness with other local firms and also with extra-community sources of knowledge. These act basically by channeling new knowledge into the community and diffusing it locally.
- External stars, that are highly interconnected with external sources of knowledge but have hardly any interaction with other local firms.

Such different positions inevitably influence the way in which knowledge is acquired, diffused and eventually transformed in the local area. Because of this, as will later be explained, this factor will be calculated in our research by taking into account a weight that can strengthen this factor to a greater or to a lesser degree.

- *Expertise*: this term can be briefly defined as the skill or knowledge of a person who knows a great deal about a specific thing. This is an important factor since people often trust in experts more than in novice employees. In addition, "individual" level knowledge is embedded in the skills and competencies of the researchers, experts, and professionals working in the organization [13].
- *Previous experience*: People have greater trust in those sources from which they have previously obtained more "valuable information". Therefore, a factor that influences the increasing or decreasing reputation of a source is "previous experience" and this factor can help us to detect trustworthy sources or knowledge.
- *Intuition*: When people do not have any previous experience they often use their "intuition" to decide whether or not they are going to trust something. Other authors have called this issue "indirect reputation or prior-derived reputation" [11]. In human societies, each of us probably has different prior beliefs about the trustworthiness of strangers we meet. Sexual or racial discrimination might be a consequence of such prior belief [11]. We have tried to model intuition according to the similarity between the user profiles: the greater the similarity between one agent and another, the greater the level of intuition.

Taking all these factors into account, we have defined our own "concept of reputation", which is summarized in Figure 1.

3 Our Proposal

In order to implement this concept we have developed a reputation model. The main goal of this model is to rate the level of confidence in an information source or in a provider of knowledge. As the model is going to be used in virtual communities where people are usually distributed in different locations we have implemented a multi-agent architecture where each software agent acts on behalf of a person.

We have chosen the agent paradigm because it constitutes a natural metaphor for systems with purposeful interacting agents, and this abstraction is close to the human way of thinking about their own activities [19]. This foundation has led to an increasing interest in social aspects such as motivation, leadership, culture or trust [8]. Our research is related to this last concept of "trust" since artificial agents can be made more robust, resilient and effective by providing them with trust reasoning capabilities.

Considering that the reputation concept described in Section 2 is composed of position, expertise, previous experiences and intuition, we propose a formula to measure the level of reputation in agents' communities.

R_{ij} is the value of reputation of j in the eyes of i. This value is computed as follows:

$$R_{ij} = w_e * E_j + w_p * P_j + w_i * I_{ij} + (\sum_{j=1}^{n} QC_{ij})/n \quad (1)$$

where E_j is the value of expertise which is calculated according to the degree of experience that the person upon whose behalf the agent acts has in a domain.

P_j is the value assigned to a person's position.

I_{ij} is the value assigned to intuition which is calculated by comparing each user's profile.

In addition, previous experience should also be calculated. We suppose that when an Agent *i* consults information from another Agent *j*, the Agent *i* should evaluate how useful this information was. This value is called QC_{ij} (Quality of *j*'s Contribution according to the Agent *i*). To attain the average value of an agent's contribution, we calculate the sum of all the values assigned to these contributions by the Agent i, for instance n and we divide it by the number of evaluations (n).

Finally, w_e, w_p and w_i are weights with which the reputation value can be adjusted to the needs of the organizations or communities. These weights represent different values depending on the category of each employee. For instance, if an enterprise considers that all its employees have the same category, then $w_p = 0$. The same could occur when the organization does not take its employee's intuitions or expertise into account.

In this way, an agent can obtain a value related to the reputation of another agent and decide to what degree it is going to consider the information obtained from this agent.

Moreover, when a user wants to join to a community in which no member knows anything about him/her, the reputation value assigned to the user in the new community is calculated on the basis of the reputation assigned from other communities where the user is or was a member. For instance, a User Agent called j, will ask each community manager where s/he was previously a member to consult each agent which knows him/her with the goal of calculating the average value of his/her reputation (R_{Aj}). This is calculated as:

$$R_{Aj} = (\sum_{i=1}^{n} R_{ij})/n \quad (2)$$

where n is the number agents who know j and R_{ij} is the value of j's reputation in the eyes of i. In the case of being known in several communities the average of the values R_{Aj} will be calculated. Then, the User Agent j presents this reputation value (similar to when a person presents his/her curriculum vitae when s/he wishes to join a company) to the Manager Agent of the community to which it is "applying". This mechanism is similar to the "word-of-mouth" propagation of information for a human [1].

In the case of the user being new in the system then this user is assigned a "new" label in order for the situation to be identified.

Once the Community Manager has obtained a Reputation value for Agent j it is added to the community member list.

4 Prototype

In order to test our proposal we have developed a prototype system into which people can introduce documents and where these documents can also be consulted by other people. The goal of this prototype is to allow software agents to help employees to discover information that may be useful to them thus decreasing the overload of information that employees often have and strengthening the use of knowledge bases in enterprises. In addition, we try to avoid the situation of employees storing valueless information in the knowledge base.

When a person searches for knowledge in a community, and after having used the knowledge obtained, that person then has to evaluate the knowledge in order to indicate whether the knowledge was useful and how it was related to the topic of the search (for instance a lot, not too much, not at all).

To design this prototype we have implemented a *User Agent* and a *Manager Agent*. The former is used to represent each person that may consult or introduce knowledge in a knowledge base. Therefore, the *User Agent* can assume three types of behavior or roles similar to the tasks that a person can carry out in a knowledge base. The User Agent plays one role or another depending upon whether the person that it represents carries out one of the following actions:

- The person contributes new knowledge to the communities in which s/he is registered. In this case the User Agent plays the role of **Provider (Pr)**.
- The person uses knowledge previously stored in the community. Then, the User Agent will be considered as a **Consumer (Co)**.
- The person helps other users to achieve their goals, for instance by giving an evaluation of certain knowledge. In this case the role is of a **Partner (Pa)**.

The second type of agent within a community is called the *Manager Agent* (which must manage and control its community.

The prototype provides the options of using community documents and updating reputation values, proposing new topics in the community, etc. Due to space limitations, we shall now describe only one situation:

Using community documents and updating reputation values. People can search for documents in every community in which they are registered. When a person searches for a document relating to a topic his/her User Agent consults the Manager Agent about which documents are related to their search. Then, the Manager Agent answers with a list of documents. The User Agent sorts this list according to the reputation

value of the authors, which is to say that the contributions with the best reputations for this Agent are listed first. On the other hand, when the user does not know the contributor then the User Agent consults the Manager Agent about which members of the community know the contributors. Thus, the User Agent can consult the opinions that other agents have about these contributors, taking advantage of other agents' experience. To do this the Manager consults its interaction table and responds with a list of the members who know the User Agent. Then, this User Agent contacts each of them which answer the R_{sj} value, obtained from formula 1 explained in the previous section. If nobody knows the contributors then the information is listed, taking their authors' expertise and positions into account. In this way the User Agent can detect how worthy a document is, thus saving employees' time, since they do not need to review all the documents related to a topic but only those considered most relevant by the members of the community or by him/her according to his/her previous experience with the document or its authors.

Once the person has chosen a document, his/her User Agent adds this document to its own document list (list of consulted documents), and if the author of the document is not known by the person because it is the first time that s/he has worked with him/her, then the Community Manager adds this relation to the interaction table. This step is very important since when the person evaluates the document consulted s/he provides the QC of the contribution and his/her User Agent will be able to assign a trustworthy value for this document, and will thus gradually obtain the reputation values of all the members with which it has interacted.

Therefore, this is a prototype where agents are continuously learning by their own experience and thank to the exchange of messages between agents they can also learn from the other's experience, increasing in this way the knowledge of all the community.

5 Conclusions

The main contributions of this paper are the reputation model and our proposal to give support to virtual communities by using software agents since according to literature the exchange of knowledge is likely to take place in these communities thanks to the trust that members have in each other. Therefore, the goals of this research are:

- To assist employees in identifying trustworthy entities and to foster the idea of sharing experiences.
- To help groups work effectively together.
- To encourage knowledge exchange between community members.
- To identify individuals with the right knowledge: expertise locator.

All these advantages provide organizations with a better control over their knowledge, which will be more trustworthy, and consequently it is expected that employees will feel more willing to use it.

Acknowledgments. This work is partially supported by the ENIGMAS (PIB-05-058) and MECENAS (PBI06-0024) project. It is also supported by the ESFINGE project (TIN2006-15175-C05-05) Ministerio de Educación y Ciencia (Dirección General de Investigación)/ Fondos Europeos de Desarrollo Regional (FEDER) in Spain and CONACYT (México) under grant of the scholarship 206147 provided to the first author.

References

1. Abdul-Rahman, A., Hailes, S.: Supporting Trust in Virtual Communities. In: 33rd Hawaii International Conference on Systems Sciences (HICSS'00) (2000)
2. Allen, T.: Managing the Flow of Technology: Technology Transfer and the Dissemination of Technological Information within the R&D Organization. MIT Press, Cambridge, MA (1984)
3. Barber, K., Kim, J.: Belief Revision Process Based on Trust: Simulation Experiments. In: 4th Workshop on Deception, Fraud and Trust in Agent Societies, Montreal Canada (2004)
4. Dasgupta, P.: Trust as a Commodity. In: Gambetta, D. (ed.) Trust: Making and Breaking Cooperative Relations, pp. 49–72. Basil Blackwell, Oxford (2000)
5. Davenport, T., Prusak, L.: Working Knowledge: How Organizations Manage What They Know, Project Management Institute. Harvard Business School Press, Boston, Massachusetts (1997)
6. Desouza, K., Awazu, Y., Baloh, P.: Managing Knowledge in Global Software Development Efforts: Issues and Practices. IEEE Software, 30–37 (2006)
7. Dillenbourg, P.: Introduction: What Do You Mean By Collaborative Learning? In: Dillenbourg (ed.) Collaborative Learning Cognitive and Computational Approaches, Elsevier Science, Amsterdam (1999)
8. Fuentes, R., Gómez-Sanz, J., Pavón, J.: A Social Framework for Multi-agent Systems Validation and Verification. In: Wang, S., Tanaka, K., Zhou, S., Ling, T.-W., Guan, J., Yang, D.-q., Grandi, F., Mangina, E.E., Song, I.-Y., Mayr, H.C. (eds.) Conceptual Modeling for Advanced Application Domains. LNCS, vol. 3289, pp. 458–469. Springer, Heidelberg (2004)
9. Luhmann, N.: Trust and Power. Wiley, Chichester (1979)
10. McKnight, D., Chervany, N.: The Meanings of Trust. In: Technical Report 94-04, Carlons School of Management, University of Minnesota (1996)
11. Mui, L., Halberstadt, A., Mohtashemi, M.: Notions of Reputation in Multi-Agents Systems: A Review. In: Alonso, E., Kudenko, D., Kazakov, D. (eds.) Adaptive Agents and Multi-Agent Systems. LNCS (LNAI), vol. 2636, pp. 280–287. Springer, Heidelberg (2003)
12. Mui, L., Mohtashemi, M., Ang, C., Szolovits, P., Halberstadt, A.: Ratings in Distributed Systems Systems: A Bayesian Approach. In: 11th Workshop on Information Technologies and Systems (WITS), New Orleans (2001)
13. Nonaka, I., Takeuchi, H.: The Knowledge Creation Company: How Japanese Companies Create the Dynamics of Innovation. Oxford University Press, Oxford (1995)
14. Schafer, B.J., Konstan, A.J., Riedl, J.: Recommender Systems in E-Commerce. In: 1st ACM Conference on Electronic Conference (EC-99), pp. 158–166 (1999)
15. Wang, Y., Vassileva, J.: Trust and Reputation Model in Peer-to-Peer Networks. In: Proceedings of IEEE Conference on P2P Computing, IEEE Computer Society Press, Los Alamitos (2003)
16. Wasserman, S., Glaskiewics, J.: Advances in Social Networks Analysis. Sage Publications, Thousand Oaks (1994)
17. Wenger, E.: Communities of Practice: Learning Meaning, and Identity. Cambridge University Press, Cambridge (1998)
18. Wenger, E., McDermott, R., Snyder, W.: Cultivating Communities of Practice. Harvard Business School Press, Boston (2002)
19. Wooldridge, M., Ciancarini, P.: Agent-Oriented Software Engineering: The State of the Art. In: Okamoto, E., Pieprzyk, J.P., Seberry, J. (eds.) ISW 2000. LNCS, vol. 1975, Springer, Heidelberg (2000)

Evaluation of Trivial Dialogue Phrase Databases Through Practical Application to User-Computer Conversation - Case Study: English - Spanish

Calkin S. Montero and Kenji Araki

Graduate School of Information Science and Technology, Hokkaido University,
Kita 14-jo Nishi 9-chome, Kita-ku, Sapporo, 060-0814 Japan
{calkin,araki}@media.eng.hokudai.ac.jp

Abstract. This paper presents the results obtained during the application of automatically generated open-domain trivial dialogue phrase databases to conversational partners. The databases were generated by a language independent algorithm. Evaluative experiments were carried out for English and Spanish. Results show the applicability of the analyzed phrase databases.

Keywords: User-computer conversation, trivial dialogue phrases, dialogue evaluation.

1 Introduction

The challenging goal of implementing computer applications capable to sustain a conversation with humans has long been of particular interest to NLP-researchers. The significant advances in the core fields of human language technology, i.e. computational linguistics, speech technology, etc., have led to the recent development of dialogue systems applications. These applications range from goal specific spoken dialogue systems, e.g., Jupiter, providing a telephone-based conversational interface for international weather [1], airline travel information systems [2], restaurant guides [3], telephone interfaces to emails or calendars [4], car navigation systems [5], and so forth, to attempts to simulate human trivial dialogue - chat - e.g. ELIZA [6].

In this respect, natural language generation (NLG) as a sub-field of computational linguistics and language-oriented Artificial Intelligence (AI) has devoted itself to researching and simulating the production of written or spoken discourse. The applications and limitations of language generation have been widely studied. In this regard well-known approaches are the *canned text* approach, in which the computer prints out a text given by a programmer, and the *template filling* approach, in which predetermined templates are filled up to produce a desired output. Other approaches include the *phrase based* approach, which can be conceptualized as generalized templates that recursively match parts of the input, and the *feature based* approach, in which the generation can be conceptualized as a collection of features appropriate for each part of the input [7].

B. Apolloni et al. (Eds.): KES 2007/ WIRN 2007, Part II, LNAI 4693, pp. 361–368, 2007.
© Springer-Verlag Berlin Heidelberg 2007

The template filling approach has been applied to the simulation of human chat as a language generation technique. One of the most famous examples of conversational systems, ELIZA [6], uses the template filling approach to generate the system's response to a user input, however due to the templates limitations the dialogue tends to become repetitive and monotonous, hardly ever satisfying the user's expectations.

In more recent research Inui et al. [8] have used a corpus based approach to language generation for dialogue system. In their approach, Inui et al. (2003), applied keyword matching in order to extract sample dialogues from a *dialogue corpus*, i.e., utterance-response pairs. However the overall evaluation of the system is not given, making it hard to compare the system's performance. Other conversational systems, e.g. ALICE by Wallace [9] and Jabberwacky by Carpenter [10], apply as well a corpus based approach to NLG in order to retrieve system's responses to a user inputs being evaluated most of the time by on-line users. A disadvantage of corpus based conversational system is that they suffer from the tedious labor of handcrafting their dialogue corpus[1]. In previous research we have proposed a language independent approach for the automatic generation of trivial phrases database, arguing to reduce the load of work and the amount of time needed in order to create a chatbot dialogue corpus [11].

However, it is difficult to objectively evaluate the performance of a conversational partner - chatbot - system. One reason is that there are not fixed tasks to accomplish throughout the dialogue. Another reason is that topical coherence must prevail in spite of the topic-shifts that occur during the conversation [12]. Therefore the evaluation of a chatbot is in general solely based on the user's satisfaction. This can be observed during chatbot evaluative contests, e.g., the Loebner Prize competition [13], where the system is evaluated based on the judges opinion on how well the system responds to a set of fixed questions.

In this paper it is described the evaluation of the applicability of automatically generated trivial dialogue phrases[2] databases to chatbot programs. We apply the semantic differential (SD) method in order to identify users attitude towards the chatbots *overall performance* during the dialogue. Results show the applicability of the analyzed phrase databases.

2 System Overview

In this system the automatically generated trivial dialogue phrases database is used as part of a chatbot database. The system general overview is shown in Fig.1. A brief explanation of the creation of the Phrases DB is given in Sect.2.1.

[1] ALICE chatbot database (ALICE brain) took over 10 years work to accomplish. The Jabberwacky database is being developed since 1988
http://www.alicebot.org/superbot.html
http://alicebot.org/articles/wallace/dont.html
http://feeling.jabberwacky.com:8081/j2about

[2] A trivial dialogue phrase is referred as an expression used by a chatbot program as the answer of a user input.

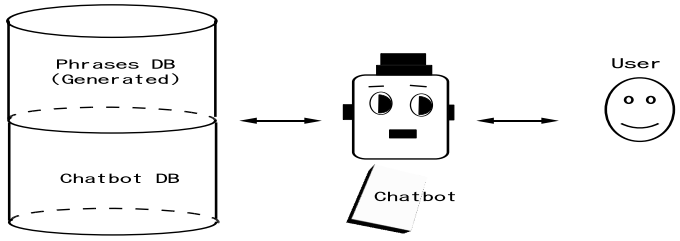


Fig. 1. System Overview

2.1 Generated Trivial Phrases DB

The generation of the trivial Phrases DB is performed using a GA-like transfer approach, details given in [14]. Fig.2 shows the generation process. A brief description of each step is given hereunder.

Initial Population. In the population selection process a small population of phrases are selected randomly from a set of phrases given to the system beforehand. For the experiments, this set of phrases consisted of 30 phrases for English and 20 phrases for Spanish.

Crossover. In order to ensure a language independent method the algorithm does not use syntactic information (part of speech tagging). Therefore, in order to avoid the distortion of the selected phrase during the initial population process due to completely blind crossover, the crossover rate was set to 0%.

Mutation. During the mutation process, each one of the seed phrases selected in the initial population is mutated at a rate of $1/N$, where N is the total number of words in the phrase. The mutation is performed through a transfer process, using features randomly selected, i.e., nouns, adjectives, adverbs, given beforehand.

Evaluation. In order to evaluate the newly generated expression, the WWW is used as knowledge base. Since matching an entire phrase on the Web might result in very low retrieval, or in some cases non retrieval at all, we apply the sectioning of the phrase into its respective n-grams. First, each n-gram for a given phrase is evaluated according to its frequency, i.e, number of hits, on the WWW. Then the system evaluates the phrase as "good" if all of its n-grams were labeled "accepted" by the system, that is to say, all of the n-grams are above the given threshold. If for a given phrase there is at most *one rejected n-gram* or *one weakly accepted n-gram*, the n-gram is below a given threshold, and the phrase is evaluated as "usable", otherwise the phrase is "rejected".

In order to verify its portability, the algorithm was applied to the generation of English and Spanish trivial dialogue phrases. According to a user evaluation, 243 out of the 1,767 generated phrases for Spanish were *semantically useful* for dialogue. In the case of English, 1,030 out of 2,512 phrases generated were

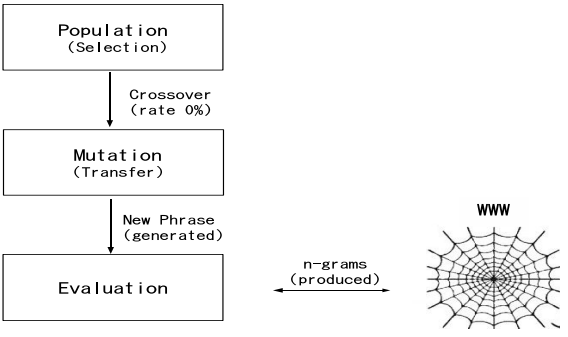


Fig. 2. Phrases DB Generation Process

semantically useful[3]. In this paper, these English and Spanish semantically useful phrases were applied as part of a chatbot database as explained in Sect.3.

3 Design Experiments: Databases Evaluation and Results

As stated before, it is difficult to empirically evaluate a conversational partner system. For the experiments a comparative evaluation between a plain system (baseline) and a system using the generated database was performed. Phrases from both English and Spanish databases were evaluated within real dialogue context. The experiment combined aspect for the evaluation of the chatbots performance based on the *semantic differential (SD) method* measures, as well as other criteria such as user satisfaction and acceptance.

The SD method was developed by the psychologist and communication scholar Charles E. Osgood [15]. This method connects scaled measurement of attitudes with the *connotative meaning* of words, i.e., subjective cultural or emotional association with a word. Osgood worked with the semantics of words and ideas involved in scaling opinions and created the SD method to plot a psychological distance between words by mapping a subject's connotations of the words. During the experiment this method was applied to determine the user's attitude towards the chatbots performance.

In the SD method there are three major factors or dimensions of judgment:

- Evaluative factor (good - bad)
- Potency factor (strong - weak)
- Activity factor (active - passive)

For the experiments, 15 adjective-pairs were used within a 7 points ranking scale where "7" was the strongest, positive meaning, and "1" was the weakest,

[3] The term *"semantically useful"* refers here to the ability of the phrase to convey information within a given context. This implies that a given phrase does not necessarily have to be grammatically correct in order to be understood and used.

negative meaning. Fig.3 shows the selected adjective-pairs[4]. Here the *evaluative factor* is represented by the pairs {a-e, g, i, k, l, n}, the *potency factor* is represented by {j}, and the *activity factor* is represented by {f, h, m, o}.

a. Fun	7	6	5	4	3	2	1	boring
b. Interesting	7	6	5	4	3	2	1	uninteresting
c. Human-like	7	6	5	4	3	2	1	un-natural
d. Good	7	6	5	4	3	2	1	bad
e. Entertaining	7	6	5	4	3	2	1	unentertaining
f. Communicative	7	6	5	4	3	2	1	unexpressive
g. Understandable	7	6	5	4	3	2	1	not-understood
h. Innovative	7	6	5	4	3	2	1	monotonous
i. Intelligent	7	6	5	4	3	2	1	dumb
j. Easy	7	6	5	4	3	2	1	difficult
k. Enjoyable	7	6	5	4	3	2	1	unpleasant
l. Kind	7	6	5	4	3	2	1	unkind
m. Relaxing	7	6	5	4	3	2	1	stressful
n. Like	7	6	5	4	3	2	1	dislike
o. Positive impression	7	6	5	4	3	2	1	negative imp.

Fig. 3. Adjective-pairs Seven-points Scale Ranking

The performed experiments were carried out as follows. The baseline system, ALICE[5] [9], uses Artificial Intelligence Markup Language (AIML), a version of XML, to represent patterns (user's input) and templates (chatbot's response). In a preliminary stage, the databases (English and Spanish correspondingly) were converted into AIML files of general patterns to be used externally as part of the chatbot inner database. However, users feedback (3 non native English speakers, 2 of them native Spanish speakers) showed that when the chatbot used the databases, its performance (in terms of user satisfaction) was much lower than the chatbot performance without the databases. A fundamental cause is considered to be the lack of correlation between the user input and the context for using the phrases from the databases as a response. Therefore, the phrases from the generated databases (English and Spanish) were inserted manually into the chatbots inner databases. In this way it is possible to evaluate the quality of the phrases within the dialogue as well as how influential these phrases were to the chatbots overall performance. Three native Spanish speakers and three non native English speakers (university students, 4 males, 2 females) were prompted to have a conversation with the baseline chatbot and the chatbot using the augmented databases for both Spanish and English. The results of the averaged evaluation of the chatbots performance according to the SD method are shown in Fig. 4 for Spanish and Fig. 5 for English.

[4] Adjectives were selected according to their appearance in the Web after searching for *"chatbot is"*.

[5] Public version of 2001.

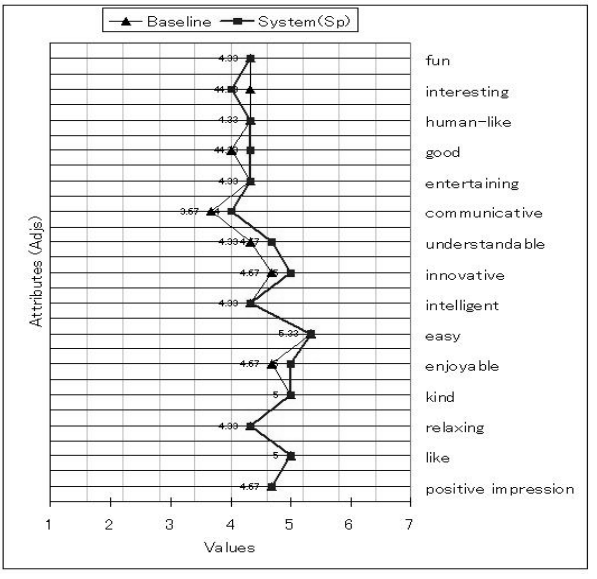


Fig. 4. Average Evaluation for Spanish System

In Fig.4 it is shown that the baseline system was evaluated as more "interesting" than the system using the generated Spanish database (System(Sp)). However, looking at the evaluation for "good" the System(Sp) is evaluated higher that the baseline system. This apparent contradiction could be explained regarding to the users *novelty feeling* towards the chatbots. The users were prompted to have a chat with the baseline system *first*. When they were prompted to have a chat with the System(Sp), the novelty feeling seemed to decrease since they had already used a similar system. However, observing the performance itself, the System(Sp) scored higher than the baseline system, showing improvement.

In Fig.5 it is shown that the baseline system was evaluated as more "understandable" than the system using the generated English database (System(Eng)). This might be considered caused by the used of *semantically usable phrases*, which does not imply that the phrases are always *completely grammatically correct*. However, with the use of phrases from the automatically generated database the System(Eng) outperformed the baseline system. The results shown in these figures reveal the validity of the application of the automatically generated databases to conversational systems.

3.1 Discussion

When observing the experimental results, the English database appears to work better than the Spanish one. These results are presumably caused by the difference in size of the databases. However, the general tendency of the systems using the databases is to outperform the baseline ones.

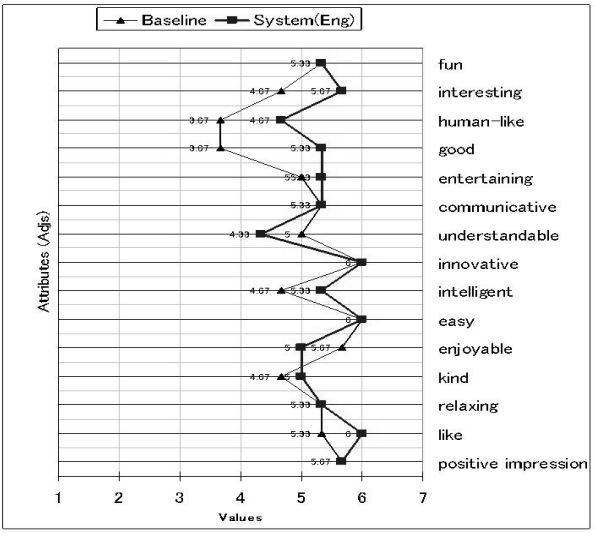


Fig. 5. Average Evaluation for English System

The results also suggest that the placement of the new databases was an important factor that showed to influence the performance of the chatbots. When the phrases of the databases were used to match general patterns, the performance of the systems was considerably lower than the performance of the systems without the phrases. On the other hand, when the databases were inserted to match specific patterns the systems performance showed improvement. Determining the appropriate time within a given context when a phrase might be used it is still a topic that requires deep research.

4 Conclusion and Future Work

In this paper the practical evaluation of automatically generated trivial dialogue phrase databases was carried out. The databases were evaluated for Spanish and English in conversational context using the SD method. Results showed a betterment in the system performance when the databases were applied, indicating the validity of the databases for conversational systems. As an ongoing work there is room for researching on the applicability of the conversational systems to interactive question-answering (QA).

Acknowledgment

This work has been partially supported by a grant of the Ministry of Internal Affairs and Communications Strategic Information and Communications R&D Promotion Program (SCOPE).

References

1. Zue, V., Seneff, S., Glass, J., Polifroni, J., Pao, C., Hansen, T., Hetherington, L.: Jupiter: A telephone-based conversational interface for weather information. IEEE Transactions on Speech and Audio Processing 8, 85–96 (2000)
2. Bratt, H., Dowding, J., Hunicke-Smith, K.: The SRI telephone-based ATIS system. In: Proceedings of the ARPA Spoken Language System Technology Workshop, pp. 22–25 (1995)
3. Lau, R., Flammia, G., Pao, C., Zue, V.: Webgalaxy: Beyond point and click - a conversational interface to a browser. In: Proceedings of the 6th International WWW Conference, pp. 119–128 (1997)
4. University of Texas at Austin: Smartvoice, `http://www.utexas.edu/its/smartvoice/`
5. Pellom, B., Ward, W., Hansen, J., Cole, R., Hacioglu, K., Zhang, J., Yu, X., Pradhan, S.: University of Colorado dialog systems for travel and navigation. In: Human Language Technology Conference (2001)
6. Weizenbaum, J.: Eliza a computer program for the study of natural language communication between man and machine. Commun. ACM 9(1), 36–45 (1966)
7. Web file: Natural language generation, `http://www.dfki.de/fluids/Natural_Language_Generation.html`
8. Inui, N., Koiso, T., Nakamura, J., Kotani, Y.: Fully corpus-based natural language dialogue system. In: Natural Language Generation in Spoken and Written Dialogue, AAAI Spring Symposium (2003)
9. Wallace, R.: A.L.I.C.E. artificial intelligence foundation (2005), `http://www.alicebot.org`
10. Carpenter, R.: Jabberwacky. learning artificial intelligence, `http://www.jabberwacky.com` `http://www.icogno.com/`
11. Montero, C.S., Araki, K.: Is it correct? - towards web-based evaluation of automatic natural language phrase generation. In: Proceedings of the COLING/ACL 2006, Interactive Presentation Sessions, pp. 5–8 (2006)
12. Herring, S.C.: Interactional coherence in CMC. Journal of Computer Mediated Communication 4(1) (1999)
13. Web file: Loebner Prize in artificial intelligence, `http://www.loebner.net/Prizef/loebner-prize.html`
14. Montero, C.S., Araki, K.: Unsupervised language independent genetic algorithm approach to trivial dialogue phrase generation and evaluation. In: NLDB07 - 12th International Conference on Applications of Natural Language to Information Systems (2007)
15. Osgood, C., Suci, G., Tannenbaum, P.: The Measurement of Meaning. University of Illinois Press, Urbana (1957)

Formalizing Recognition of Sketching Styles in Human Centered Systems

Danilo Avola, Fernando Ferri, and Patrizia Grifoni

Istituto di Ricerca sulla Popolazione e le Politiche Sociali - Consiglio Nazionale delle Ricerche,
Via Nizza 128, 00198 Rome, Italy
{danilo.avola,fernando.ferri,patrizia.grifoni}@irpps.cnr.it

Abstract. Sketch-based interaction enables users' simple communication and it is used to represent concepts and commands in human centered systems. This communication approach can be used in different contexts with different devices. The ink style through which the user performs a sketch is a critical component in the recognition and interpretation processes. In particular, the different users' styles adopted to perform the sketch can introduce over-tracing and/or cross-hatching phenomena that are respectively represented in the sketch like bold style or dashed style. The paper provides an approach to formalize the recognition of the different ink styles (bold, solid and dashed) performed by the user during his/her sketch activity.

Keywords: Human centered systems, sketch-based interaction, ink styles, stroke style, sketch style, solid style, bold style, dashed style.

1 Introduction

People use the free hand drawing to express concepts and to represent ideas in an immediate and intuitive simple way. Therefore, the use of sketch-based interfaces to interact with different human centered systems provides a natural convenient way to carry out concepts and commands. The user interaction is influenced by the individual drawing style. Sometimes, tools used to interact with the different devices (such as: smart phone, PDA, palmtop, tablet-PC, and so on) and/or specific contexts can influence the individual ink style too. Different users can draw the same object/concept using different styles, this can happen without specific reasons because a sketch action is intrinsically an informal and messy human expression. Therefore, the over-tracing and/or cross-tracing phenomena can frequently occur. Figure 1 shows the same object/concept (circle) drawn by different sketches obtained using the three main drawing styles (solid, bold, and dashed). These different drawing styles make the interpretation of the sketch very hard. The purpose of this paper is to propose and discuss an approach to formalize the recognition of the different main styles performed by the user during her/his sketch activity. More specifically, an approach to formalize the recognition of bold, solid and dashed style is presented.

There is an extensive body of related work on sketch style recognition in human centered systems.

B. Apolloni et al. (Eds.): KES 2007/ WIRN 2007, Part II, LNAI 4693, pp. 369–376, 2007.
© Springer-Verlag Berlin Heidelberg 2007

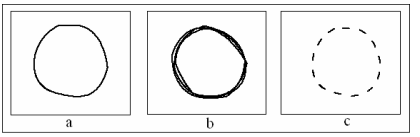


Fig. 1. A circle drawn in: (a) solid style, (b) bold style, (c) dashed style

In [1][2][3] authors suggest a fast simple approach to recognize sketches drawn with a stylus on a digitizing tablet. The approach enables the system to identify shapes (such as: triangles, lines, rectangle, and so on) of different size and with different orientation in the space, drawn with continuous, dashed-tracing or over-tracing strokes. This approach is based on the computation and ratio of particular geometrical features of shapes in the sketch (such as: convex hull, the largest triangle and the largest quadrilateral inscribed in the convex hull, the smallest enclosing rectangle of the convex hull). The over-tracing action, considered as a user imprecision, is discussed in [7]. The system generates geometric approximations for single stroke shapes that are over-traced. According to this approach the system has to match input strokes with geometric primitives producing fits for several shapes (such as: lines, arcs and circles) by computing model's parameters that minimize the least squares fitting error. Also in [9] the problems that involve the interpretation of over-tracing freehand sketch are faced. In this approach over-tracing strokes are interpreted according to the following steps: 1) the strokes are first classified into lines and curves by a linearity test, 2) after this a strokes grouping process that handles lines and curves separately is performed. Afterwards, the grouped strokes are fitted with 2D geometry and further tidied-up with endpoint clustering and parallelism correction. Finally, the in-context interpretation is applied to detect incorrect stroke interpretation based on geometry constraints and to suggest the most probable correction based on the overall sketch context. In [5] the authors present an approach that, unlike the previous two, deals with both the over-tracing and hatching sketch on the 3D sketch. With this approach the strokes are divided into core strokes (strokes that touch the characteristic curves of the object), and hatching strokes (strokes that are mapped to the faces of the object). This approach allows the interpretation of the user's sketch that presents over-tracing strokes, by grouping strokes into bundles. Another approach to recognize the free hand sketches drawn using different styles (with over-traced, dashed or continuous strokes) is given in [8]. The key advance of this approach, is an integrated sketch parsing and recognition model designed to enable natural and pen-based computer interaction. With this approach, the stream of pen strokes is firstly examined to identify delimiter patterns called "markers". These then anchor a spatial analysis, which groups the remaining strokes into distinct clusters, each representing a single visual object. Finally, a shape recognizer is used to find the best interpretations of the clusters. Indeed there are several approaches that face the over-tracing or (less frequently) cross-hatching phenomena. But there are not many works that face the problems from the stroke style point of view. This paper proposes an approach to formalize the recognition of the different drawing styles. The approach does not need to know the features of the geometric shapes in the sketch. It works only on each single stroke. The paper is organized as follows. Section 2 describes both the approach and its formalization to recognize the solid and bold ink

styles, Section 3 describes the approach to recognize the dashed ink style. Finally, Section 4 concludes the paper.

2 Recognition and Formalization of Solid and Bold Ink Styles

In this paper a stroke is considered as a drawing action defined by the sequence: pen down, pen movement, pen up. As shown in Figure 2-a, the stroke is spatially characterised by the coordinates (x,y) and temporally characterised by the temporal interval (Δt) used by the user to perform the sketch.

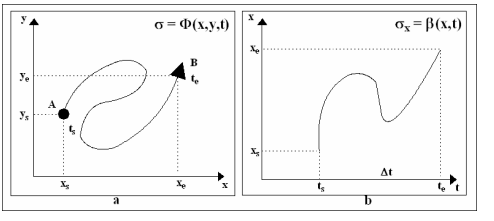


Fig. 2. (a) The stroke σ, (b) The projection, on x, of the stroke σ

In Figure 2-a, the label A shows the start point of the stroke characterized by the start coordinates (x_s,y_s) at the time of pen down t_s. While the label B shows the end point of the stroke characterized by the end coordinates (x_e,y_e) at the time of pen up t_e. For this reason the drawing of a stroke σ can be described as function Φ of both spatial coordinates (x,y) and temporal coordinate (t):

$$\sigma = \Phi(x,y,t) \ . \tag{1}$$

The first step in detecting stroke style consists of the analysis of the spatial and temporal sequence of the drawn stroke. The aim of analysis is to detect the possible direction change conveyed by the user during stroke drawing. As shown in Figure 2, this analysis can be performed by studying the behavior of a single space component, of the stroke, in function of the time. Considering only one of the two spatial coordinates of each pixel (for instance x) and the time (t) in which the pixels have been drawn, it is possible to obtain a reliable transposition (respect the x axes) of the stroke drawn by the user, that is:

$$\forall t \in \Delta t \ \exists x \in \mathbb{N} : \ \sigma = \Phi(x,y,t) \Rightarrow \sigma_x = \beta(x,t); \text{ where } \Delta t = t_e - t_s \ . \tag{2}$$

In this way, as shown in Figure 2-b, a function (σ_x) depending on both the chosen axis (x) and the time (t) can be considered. This function represents the projection on x axes of the stroke (σ). The Δt represents the drawing stroke time interval.

The projection function (σ_x) allows highlighting the direction change performed by the user during stroke drawing (σ). In fact, as shown in Figure 3, these changes (on stroke σ) are conveyed in proximity to the relative and/or absolute maximum and minimum points belonging to the projection function (σ_x).

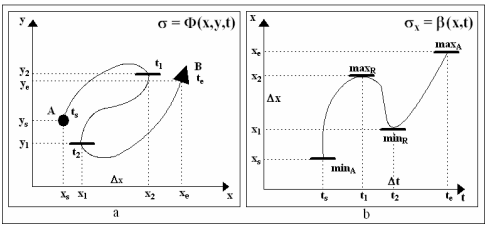


Fig. 3. (a) Stroke with two direction changes, (b) Relative and absolute maximum and minimum points

In this context, the direction changes of the stroke (σ), is defined as the transition of x values from non-increasing monotone to non-decreasing monotone (and vice versa). In Figure 3-b the projection function (σ_x) clearly shows four maximum and minimum points, given below:

$$\begin{aligned}
&\text{absolute minimum point } \min_A \text{ of coordinates } (x_s, t_s) \text{ since: } \forall t \in \Delta t,\ \exists x \in \Delta x \Rightarrow \beta(x,t) \geq \beta(x_s, t_s)\,. \\
&\text{relative minimum point } \min_R \text{ of coordinates } (x_1, t_2) \text{ since: } \exists CR_t : \forall t \in CR_t \cap \Delta t,\ \exists x \in \Delta x \Rightarrow \beta(x,t) \geq \beta(x_1, t_2)\,. \\
&\text{relative maximum point } \max_R \text{ of coordinates } (x_2, t_1) \text{ since: } \exists CR_t : \forall t \in CR_t \cap \Delta t,\ \exists x \in \Delta x \Rightarrow \beta(x,t) \leq \beta(x_2, t_1)\,. \\
&\text{absolute maximum point } \max_A \text{ of coordinates } (x_e, t_e) \text{ since: } \forall t \in \Delta t,\ \exists x \in \Delta x \Rightarrow \beta(x,t) \leq \beta(x_e, t_e)\,.
\end{aligned} \tag{3}$$

Where CR_t represents a circular round of t, and Δx represent the variability range of the x values during stroke drawing. In particular the points max_R and min_R identify the direction changes (respect x axis) of the stroke (σ), while the other two points represent the start (min_A) and the end (max_A) points of the stroke (σ). These four points univocally fix the related coordinates of the stroke (σ). As shown in Figure 3-b, the points min_A, max_R, min_R and max_A are respectively related to the points of coordinates (x_s, y_s, t_s), (x_2, y_2, t_1), (x_1, y_1, t_2), and (x_e, y_e, t_e) of the Figure 3-a. The direction changes on stroke (σ) are thus recognized. In Figure 3-a the stroke (σ) shows two direction changes according to the points (x_2, y_2, t_1) and (x_1, y_1, t_2). These changes identify, on the stroke (σ), three different sub-strokes, more exactly:

$$\begin{aligned}
&\text{first sub-stroke fixed by the function : } \sigma = \Phi(x,y,t),\ t \in [t_s, t_1] \subset \Delta t\,. \\
&\text{second sub-stroke fixed by the function : } \sigma = \Phi(x,y,t),\ t \in [t_1, t_2] \subset \Delta t\,. \\
&\text{third sub-stroke fixed by the function : } \sigma = \Phi(x,y,t),\ t \in [t_2, t_e] \subset \Delta t\,.
\end{aligned} \tag{4}$$

The stroke (σ) is subdivided in several strokes, where every identified sub-stroke has a unique direction. The identification of the single sub-stroke is independent from the speed with which the user has drawn it. That is, if the user draws the stroke in Figure 3-a two times with a different speed, the approach will identify the same sub-strokes, as shown in Figure 4.

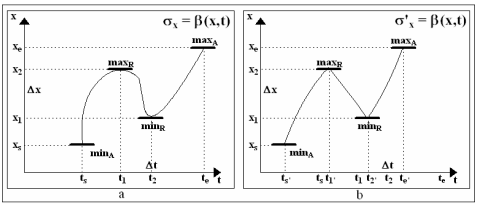


Fig. 4. (a) Transposition of the stroke σ, (b) Transposition of the stroke σ drawn with a different temporal sequences

The user has drawn the stroke which transposition is shown in Figure 4-b faster and keeping a more constant speed than the stroke which transposition is shown in Figure 4-a. Despite of this, the approach has detected (spatially on x axis) the same coordinates for the four maximum and minimum points. Therefore the approach will consider the same three sub-strokes founded for the stroke in Figure 3-a. During sketch activity a stroke can cross itself one or more time, to accomplish the proposed approach each over-traced pixel time has to be considered. When all the sub-strokes have been detected the second step of the approach can be performed.

In the second step, for every established sub-stroke an enclosing rectangle has to be considered, as shown in Figure 5-d, 5-e and 5-f. To perform this result, a reliable transposition, this time respect the y axes, of the stroke has been considered:

$$\forall t \in \Delta t \; \exists y \in \mathbb{N} : \sigma = \Phi(x,y,t) \Rightarrow \sigma_y = \alpha(y,t) . \tag{5}$$

Through the projection function (σ_y), as shown in Figure 5-a, 5-b and 5-c, for each sub-stroke can be identified two points: the absolute maximum ((y_1,t_b), (y_3,t_d), (y_5,t_f)) and the absolute minimum ((y_2,t_a), (y_4,t_c), (y_6,t_e)) respect the y axes.

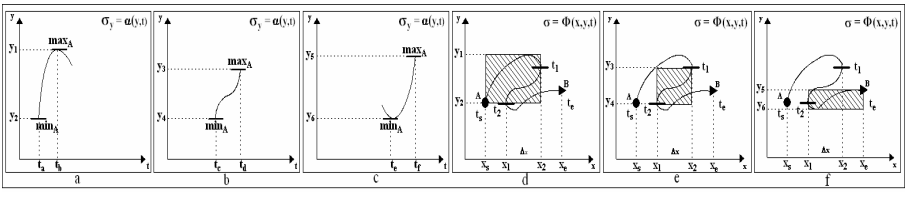


Fig. 5. Absolute maximum and minimum points respect y axes on: (a) first, (b) second, (c) and third sub-stroke. Enclosing rectangles on: (d) first, (e) second, (f) and third sub-stroke.

As shown in Figure 5-d, 5-e and 5-f, for each stroke (first, second and third) the absolute maximum and minimum points respect both x and y axis can be used to identify the connected enclosing rectangles as follow:

$$\begin{aligned}
ER_1 &= \int_{y_2}^{y_1}\int_{x_s}^{x_2} \sigma \, dy dx = \int_{y_2}^{y_1}\int_{x_s}^{x_2} \Phi(x,y,t) \, dy dx \;\; \text{where } t \in [t_s,t_1] \subset \Delta t . \\
ER_2 &= \int_{y_4}^{y_3}\int_{x_1}^{x_2} \sigma \, dy dx = \int_{y_4}^{y_3}\int_{x_1}^{x_2} \Phi(x,y,t) \, dy dx \;\; \text{where } t \in [t_1,t_2] \subset \Delta t . \\
ER_3 &= \int_{y_6}^{y_5}\int_{x_1}^{x_e} \sigma \, dy dx = \int_{y_6}^{y_5}\int_{x_1}^{x_e} \Phi(x,y,t) \, dy dx \;\; \text{where } t \in [t_2,t_e] \subset \Delta t .
\end{aligned} \tag{6}$$

The last step of the approach to discriminate the stroke style (bold or solid) is to analyze the relationship among the associated areas to the enclosing rectangles (Ar_1, Ar_2 and Ar_3 as shown in Figure 6-a and 6-c). A stroke drawn in a solid style usually has all its associated areas to the enclosing rectangles weakly overlapped, as shown in Figure 6-a and 6-b, while, a stroke drawn in a bold style usually has all its enclosing rectangles strongly overlapped, as shown in Figure 6-c and 6-d, formally:

$$A_{rel} = A_1 \cup A_2 = (A_{tot}) - ((Ar_1 \cap Ar_2) \cup (Ar_1 \cap Ar_3) \cup (Ar_2 \cap Ar_3) \cup (Ar_1 \cap Ar_2 \cap Ar_3)) . \quad (7)$$

Where $A_{tot} = (Ar_1 \cup Ar_2 \cup Ar_3)$ and A_{rel} represents the relationship among associated areas. As it is possible to observe, in Figure 6-b, there are two wide areas (A_1 and A_2) that are covered only by two different associated areas to the enclosing rectangles. Also in Figure 6-d, it is possible to observe two areas (A_1 and A_2) that are covered only by two different associated areas to the enclosing rectangles. The area A_{rel} related to the bold stroke (Figure 6-d) is smaller than one related to the solid stroke (Figure 6-b).

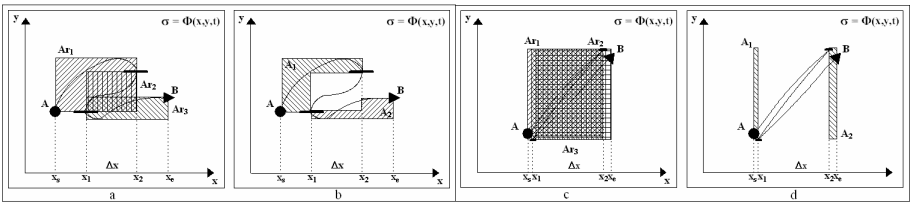


Fig. 6. Enclosing rectangle on: (a)(b) solid stroke, (c)(d) bold stroke

Our training data set were composed by 200 shapes (100 drawn in solid style and 100 drawn in bold style). Each shape has been drawn by a stroke. The experimental results have shown that in a solid stroke the area A_{rel} represents usually a large portion (among the 46% to 100%) of the related area A_{tot}. While, in a bold stroke the area A_{rel} represents usually a small portion (among the 0% to 37%) of the related area A_{tot}. Obviously, it is always possible to draw a stroke where the areas (A_{rel} and A_{tot}) have an ambiguous relationship (among 38% to 45%), it depends on the personal concept of bold and solid style which fixes the ratio of the areas. In table 1 are shown some meaningful examples of shapes style detection.

Table 1. Some visual experimental results of the proposed approach

	SOLID					BOLD				
Single Stroke					..					..
N° Rectangles	5	4	4	3	..	4	8	3	6	..
Ratio(A_{rel},A_{tot})	81%	93%	89%	78%	..	16%	14%	7%	22%	..

The computational complexity required by proposed approach for all basic operations (such as: direction change detection, building rectangles, and so on) can be considered constant. For this reason the whole complexity of the related algorithm can be considered linear in the size of the pixels that make up the stroke σ.

3 Recognition of Dashed Ink Style

The dashed style is composed from more solid and/or bold strokes. As shown in Figure 7-a and 7-b for each one of the five strokes (three solids, two bolds) an enclosing rectangle is considered. Afterwards, the barycentre of each enclosing rectangle is determined and, as shown in Figure 7-c and 7-d, a neighbourhood relationship has to be identified among the rectangles' barycentres. The neighbourhood relationship is based on simple distance measure (fixed threshold, thr) among rectangles' barycentres.

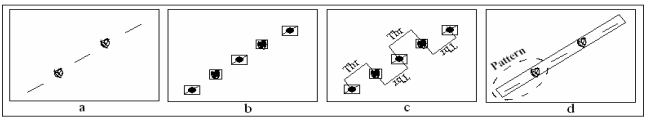


Fig. 7. (a) Five solid/bold strokes, (b) enclosing rectangle and barycentre, (c) threshold, (d) a dashed stroke

Obviously, when two rectangles' barycentres are too far (respect the thr) more than one dashed stroke is considered. Introducing similarity measures (such as the entropy, that measures the disorder degree of a statistical distribution) into the enclosing rectangle could be possible to recognize the specific pattern that made up the dashed stroke. In this case the simple pattern is represented by a single solid stroke and a single bold stroke (as shown in Figure 7-d) but a generic dashed stroke can have a more complex pattern. Finally, as added value, taking into account the temporal information on the strokes that composes a pattern (that is the temporal appearance of a stroke respects the others) it is also possible to identify the verse of a pattern. Combining the shapes into the previous data set has been possible to test the dashed ink style recognition. Obviously, the effectiveness of this approach step depends on the fixed thr, thus if the strokes that compose the pattern are "homogenous" (that is: the enclosing rectangles that compose each stroke of a pattern have a comparable size) than it will always possible to identify the connected pattern.

4 Conclusions

The spread of sketch based interfaces to support the different human centered systems is mainly due to the fact that they make the interaction in different contexts and on different devices intuitive and spontaneous. The ink style through which the user performs a sketch is a critical aspect, therefore recognizing and interpreting steps are

hard tasks. The selection of a specific ink style is driven by a personal ink style. Sometimes, it also depends from the interaction tools and/or the specific contexts.

This paper proposes both a strategy and its formalization to detect the stroke ink style (solid, bold and dashed) for sketch based interfaces on human centered systems. The shown approach takes into account only the single stroke and it does not need to know the shapes features contained in the sketch. In this way it is possible to recognize also the different ink styles that can compose the different sides of the same object/shape. The proposed approach discussed starting from the x axis can be extended to the other spatial coordinates.

References

1. Fonseca, M.J., Jorge, J.A.: Experimental Evaluation of an on-line Scribble. Pattern Recognition Letters. 22(12), 1311–1319 (2001)
2. Fonseca, M.J., Jorge, J.A.: Using Fuzzy Logic to Recognize Geometric Shapes Interactively. In: Fuzz IEEE, the 9th IEEE International Conference on Fuzzy System (FUZZIEEE), San Antonio, TX USA, vol. 1, pp. 291–296 (2000)
3. Fonseca, M.J., Jorge, J.A.: A simple approach to recognize geometric shapes interactively. In: Chhabra, A.K., Dori, D. (eds.) GREC 1999. LNCS, vol. 1941, pp. 266–276. Springer, Heidelberg (2000)
4. Fonseca, M.J.: Sketch-Based Retrieval in Large Sets of Drawings. Ph.D. Thesis Department of Information Systems and Computer Engineering. Technical University of Lisbon Superior Technical Institute (July 2004)
5. Mitani, J., Suzuki, H., Rimura, F.: 3D sketch: sketch-based model reconstruction and rendering, pp. 85–98. Kluwer Academic Publishers, Dordrecht (2002)
6. Pratt, W.K.: Digital Image Processing, 3rd edn. John Wiley & Sons, Chichester (2001)
7. Sezgin, T.M., Davis, R.: Handling Overtraced Strokes in Hand-Drawn Sketches. In: Proceedings of the AAAI Spring Symposium Series: Making Pen-Based Interaction Intelligent and Natural, Washington, D.C. (October 21-24, 2004)
8. Kara, L.B.: Automatic parsing and recognition of hand-drawn sketches or pen-based computer interfaces. Ph.D. Thesis Mechanical Engineering Department Carnegie Mellon University (September 2005)
9. Ku, D.C., Qin, S.F., Wright, D.K.: Interpretation of Overtracing Freehand Sketching for Geometric Shapes. In: The 14th International Conference in Central Europe on Computer Graphics, Visualization and Computer Vision (WSCG ' 2006) (January 30 - February 3, 2006)

Fostering Multi-Modal Summarization for Trend Information

Tsuneaki Kato[1], Mitsunori Matsushita[2], and Noriko Kando[3]

[1] The University of Tokyo
3-8-1 Komaba, Meguro-ku, Tokyo 153-8902, Japan
[2] NTT Communication Science Laboratories, NTT Corp.
2-4 Hikaridai, Seika-cho, Soraku-gun, Kyoto 619-0237, Japan
[3] National Institute of Informatics
2-1-2 Hitotsubashi, Chiyoda-ku, Tokyo 101-8430, Japan

Abstract. It is more important than ever to be able to extract relevant information and to understand or review its content quickly. Spotting trends is one such way of understanding or reviewing. We propose multi-modal summarization for trend information as a new technology to help this activity. We also describe the design of a data set that will foster research in the field, and introduce an ongoing unique workshop using this data set.

1 Introduction

People are inundated by a huge amount of information from several media, making it more important than ever to be able to extract relevant information and to understand or review its content quickly. Spotting trends is one such way of understanding or reviewing.

A trend, a general tendency in the way a situation is changing or developing, is based on temporal statistical data, and can be obtained by synthetically summarizing it, but not by simple enumeration. Changes in a product's price and sales, and the financial status of a company are examples of trend information. Trends sometimes must be described by multidimensional information: the trend in the market share of a given product should be discussed with reference to competitive companies, and trends in land prices should be depicted as a function of time and location. Understanding and reviewing a huge amount of information requires the assistance of a mechanism that can properly answer questions on trends, such as how the price of gasoline changed during the year, what the situation of the domestic personal computer market has been, and how severe the typhoons were.

In considering compilation of trend information, we notice that its input consists of not only non-linguistic information, that is, temporal data spanning a given period, but also linguistic information consisting of explanations and evaluations of that data, which can be more important than the data itself. The output consists of both plain and concise text and non-linguistic information,

B. Apolloni et al. (Eds.): KES 2007/ WIRN 2007, Part II, LNAI 4693, pp. 377–386, 2007.
© Springer-Verlag Berlin Heidelberg 2007

such as information graphics, that are used to help users reach a more effective understanding. With these considerations in mind, we propose *multi-modal summarization* that uniformly summarizes linguistic and non-linguistic information and compiles it into multimedia presentations in which text and visual information work cooperatively[1].

Multi-modal summarization needs the creative fusion of researches in several fields. Such fields include multimedia presentation generation, text summarization, information visualization, information extraction, and question answering, among some of which there has been little interaction so far. We proposed and have been coordinating MuST, a workshop on multi-modal summarization for trend information, in order to encourage the fusion and progress of researches in those fields and to foster researches on multi-modal summarization for trend information. The major characteristic of the MuST workshop is the use of a shared resource, the MuST data set. In this paper, we introduce the specification of the MuST data set and demonstrate its importance by explaining its role in the MuST workshop. The major contributions of this paper are the proposal of a novel research field called multi-modal summarization and the design of research resources and a workshop that will foster the research.

This paper is organized as follows: Section 2 explains what multi-modal summarization technology is and demonstrates that it can be used effectively to compile trend information. In Section 3, the data set, whose core is made up of annotated newspaper articles, is described and the specification of the annotation schema is explained in detail. Section 4 presents a workshop on multi-modal summarization for trend information using the data set as a motivating force for interested researchers. This section includes the introductions of current efforts in the workshop. Section 5 discusses possible future directions, and concludes the paper.

2 Multi-Modal Summarization for Trend Information

2.1 Multi-Modal Summarization

Text summarization extracts important content from a huge amount of text and presents it in condensed form. Information visualization generates visual representation of abstract data in order to support understanding and access. The two technologies have the common objective of supporting quick understanding or review of large amounts of information. Text summarization can be divided into extraction- and abstraction-based types. Abstraction-based summarization first builds semantic representations of given texts, then processes them by selection, aggregation, and generalization, finally generating a concise text from them [1]. The first two steps can be viewed as a kind of information extraction in its wider sense. The last step is similar to report generation. Information visualization transforms raw data in some idiosyncratic format into data tables,

[1] We avoid the term, "multimedia summarization", because it is commonly used to summarize audio and visual information, and use "multi-modal summarization" instead.

relational representations of data with explicit semantics, then maps it into visual structures, and creates those views by specifying parameters [2].

Thus, in addition to their common objectives, there are parallels between text summarization, especially the abstraction-based kind, and information visualization. Since intermediate representation is independent of modes, information built or extracted from texts could be mapped to visual structures and vise versa. Semantic representations with different origins could be merged, re-categorized and then assigned to different media. This presents possibilities of a productive fusion of text summarization and information visualization. Multi-modal summarization pursues those possibilities.

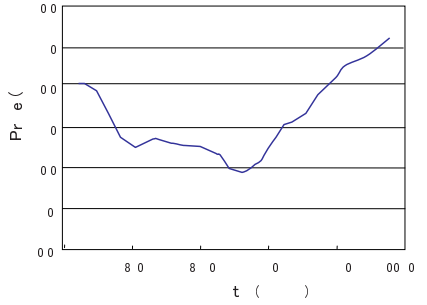
In the real world, information is not usually provided in an exclusively linguistic or visual form. Rather, it is generally used in a complementary form, such as a textual report with numerical charts or information graphics with annotations in the text. Multimedia presentation generation aims to generate such multimedia presentations from media-independent semantic representation taking the user's interests and intentions into account [3][4]. Current information technologies, however, should fulfill the great need to manipulate the currently available floods of unstructured and uncoordinated information. Since robust technologies with broad coverage are expected, text summarization and information visualization rather than report generation and multimedia presentation generation are required.

Combined with the importance of handling information in several modes, multi-modal summarization is expected, which has the same relationship to multimedia presentation generation as the relationship that text summarization has to report generation. In other words, multimedia presentation generation, which presumes the existence of a well-formed semantic representation, should be developed into multi-modal summarization, which can compile different kinds of real world information into multimedia presentations.

2.2 Trend Information

Figure 1 is a line chart with a short text, both representing changes in oil prices in 1998 and 1999. It is easier for a user to grasp the overall picture of change and see a trend as a general tendency by looking at a line chart than it is by looking at tables or numbers. On the other hand, the text, which mentions each turning point, may be preferable to the chart, in which it is obscure where the turning points are. This shows that it is important that linguistic and visual information complement each other for effective presentation of trend information. Specifically, textual descriptions supported by charts or charts with textual annotations are the most effective format in promoting understanding of trends.

In addition, trend information comes in both linguistic and numerical forms and usually from multiple sources. For example, data in a white paper allows us to draw charts like that in Fig. 1, but it does not indicate where the turning points are and how those changes should be interpreted. In contrast, newspaper articles published during such turning points do interpret the changes. These situations arouse crucial needs of multi-modal summarization for compiling trend information.



Oil prices declined throughout 1998, falling to $10 something per barrel by the end of the year, dipping below $10 for a short time in February 1999. Then they advanced over 50% during April and May, reaching around $15. They continued to climb, passing the $20 mark at $20.20 per barrel by late August. They have hit a new high since February 1998. They climbed to between $21 and $22 in late September.

Fig. 1. Changes of the oil price

3 Data Set and Research Paradigm

3.1 A Possible Framework

Let us think of a possible framework for compiling and providing trend information. A user's query on oil industry trends in 1999, for example, is processed as follows.

(i) Using some ontology, which may be constructed by human efforts or automatically, the system determines that statistics such as oil prices and retail gasoline prices are needed to describe those trends and also understands the relationships among the statistics.

(ii) Related articles and/or data on those statistics are retrieved from a background document set, such as newspaper articles, and/or a background numerical data set, such as white papers.

(iii) With newspaper articles, since only certain portions of each retrieved article are likely to be relevant, the system must extract those portions and categorize them according to their relevance. The portions to be extracted include those mentioning each specific related statistic, such as "the oil price advanced over 50% since the end of last year" and "the retail price of gasoline is declining slowly." Also to be extracted are general comments related to the trends in the question itself; such as "Overproduction of gasoline and excessive rivalry among gas stations cause this peculiar situation."

(iv) From the sets of extracted portions, the system generates summaries and/or builds semantic representations for summaries. These are conducted for each statistic and for the queried trend itself. The output may be text, tables, or graphics such as charts visually representing the change in given statistics. Then, the numerical data collected is referred to and incorporated into the output.

(v) Finally, it is integrated into one report, a multimedia presentation that considers the relationships between statistics, such as "Based on market principles, the cost ought to affect the retail price."

3.2 MuST Data Set

We designed and constructed the Multi-modal Summarization for Trend information (MuST) data set with the objective of fostering development of the systems discussed above and research on their component technologies.

The MuST data set consists of data used to compile trend information, intermediate results of the compilation process, and examples/references of components of multimedia presentations to be obtained on several topics, each of which corresponds to a users' query or information request. There is a wide range of topics: social and economic domains or activities, such as the oil industry and the personal computer market; groups of incidents such as earthquakes and typhoons; and organizations such as Sony corp. Those trends can be compiled from linguistic descriptions of statistical values and/or reports on events. For example, trends in the personal computer industry include statistics on shipment volume, shipment value, and market share of major makers. Typhoon trends consist of a review of typhoon-related events, such as their development and landfalls, and damage statistics. In the MuST data set, these descriptions of statistics and events are considered units of information. In addition to those primitives, situation descriptions and evaluations found in, say, newspaper articles and numerical data of those statistics can be materials for compiling trend information.

The current MuST data set makes much use of linguistic information and uses as its document set the Japanese Mainichi newspapers from 1998 and 1999. We selected 27 topics, and for each topic, selected about three statistics and/or event types from which its trend is compiled, 77 in total. These correspond to input examples and results of the first step of the framework discussed. For each topic, we collected and constructed the following data:

- A list of newspaper articles (582 articles in total) concerning the topic,
- Annotations of those articles
- Examples of textual summaries on the selected statistics and events
- Tables or charts on those statistics

The list of articles provides articles obtained through the second step for information collection in the framework. *The textual summaries* are examples of the output of the fourth step for summarizing each statistic, *The tables and charts* provides data collected for summarization and examples of parts of summarization results.

The annotated articles is the most important data in the MuST set, which corresponds to the intermediate results of the fourth step for summarization. The annotation schema is compliant with the XML format. Examples in which it was applied to English texts are shown in Fig. 2, instead of examples of real data, which is in Japanese.

<unit stat="Dubai oil price"><name>The oil price (Dubai Oil)</name> has kept dropping since its <rel type="ord">peak</rel> <date gra="month" abs="199710">last October</date>, of <val>around $20</val> <name part="foot">per barrel</name>, and fell to <val>$12.50</val>. in <date gra="ten-days" abs="19980121">late January</date></unit>. <unit stat="Dubai oil price">After <date gra="ten-days" abs="19980121">that</date>, ø<ins type="name">Dubai oil price</ins> rose temporarily, <del type="rsn">because of the tension of the Iraq situation,</del> but has been struggling recently at <val>around $10</val><del type="other">, being "on the oversupplied side"</del></unit>.

<unit event="typhoon landfall">Medium-strength <par>typhoon No. 10</par> struck <par>Makuraszaki-shi, Kagoshima</par>, at <date gra="hour" abs="19980917">about 4:30 pm on the 17th</date>, and will strike in <par>the vicinity of Shukumo-shi, Kochi</par> <date gra="hour" abs="19980917">the same night</date></unit>.

Fig. 2. Examples of annotations on English text

In this annotation schema are 13 kinds of elements. First, sentences mentioning selected statistics or events are annotated as `unit` elements, each of which has a `stat/event` attribute that shows which statistic or event it is related to. Most annotations are conducted on sentences selected as `unit` elements. To make them understandable after extraction from the surrounding context, elliptical phrases (zero anaphora) are supplied and marked as `ins` elements. `Ins` elements have an attribute that indicates what elements such as the name of the statistic and the dates of the report are supplied. Some clauses in `unit` elements are marked as `del` elements when their parts do not relate directly to the statistics or events concerned. They include the sources of information and the impact of the value change or event, which are categorized and described as an attribute of the `del` element.

From the text of a `unit` element, phrases mentioning the name of the statistic (`name` element), the value of the statistic (`val` element), the relative values, which relate to the statistic but are not the value itself (`rel` element), dates (`date` element), and other parameters (`par` element) are determined and marked as the respective element (shown in parenthesis). `Name` elements have an attribute that shows which statistic that element denotes, since there is sometimes more than one statistic mentioned per `unit` element. In addition, since the name of a statistic can be expressed in a variety of ways and can be scattered in multiple non-consecutive phrases in a sentence, some `name` elements consist of just part of a statistic name. To cope with this situation, `name` elements have an attribute that indicates whether they are complete names, main portions of a name, or complementary portions. `Val` and `rel` elements are annotated to make their relationship to a particular statistic clear; `rel` elements are subcategorized into differences, proportions, and orders, which categorizations are indicated as attributes. `Date` elements have attributes that indicate their granularity and absolute dates when using relative expressions. For example, "Yesterday" is annotated with `gra="day", abs="19990203"`. `Par` elements correspond to arguments

of the statistic or event, such as companies for market share and the name of a typhoon and a place for typhoon landfalls. Dates are handled separately because of their special position, though they could be a kind of arguments. Anaphoric expressions are marked as such, and constitute `ref` elements with an attribute that shows their antecedent. Again, dates are special in this treatment as well.

It could be understood from the specification above that the annotation of the MuST data set represents the intermediate results of semantic and pragmatic analysis specialized for statistical and/or event information. In the flow of summarization in general, extraction and analysis of important sentences is followed by rephrasing and sentence generation to eliminate redundancy and maintain consistency. The annotation corresponds to the output of the former, extraction and analysis, and the input of the latter, rephrasing and sentence generation. Using terms of information extraction, named entity extraction and temporal expression analysis [5] are completed by this annotation. Therefore, for researchers interested in important sentence extraction or language processing, such as named entity extraction and temporal expression analysis, the annotation can be referred to as the correct result of their process and used as training data in their machine learning approach. For researchers interested in rephrasing, sentence generation, and information visualization, it can be used as input data in which several fundamental analyses are already completed. In this sense, the annotated articles behave as a hub of a multi-modal summarization.

Data sets with these characteristics differ from test collections used in the information retrieval field [6]. Neither are they pure linguistic resources. The MuST data set has a wider range in its use, though it shares the characteristic of application dependency with QuestionBank [7].

4 MuST Workshop

The MuST workshop[2] was designed to encourage cooperative and competitive studies on multi-modal summarization for trend information [8]. It is a pilot workshop of the NTCIR workshop[3], which is a series of evaluation workshops designed to enhance research in information access technologies including information retrieval, question answering, text summarization, extraction, and so on. One purpose of the MuST workshop is to promote discussion and community conformity. Another is to construct and accumulate resources such as tools and corpora. Those would be encouraged by sharing the same research resource and by conducting research concerning common themes. An objective and quantitative evaluation using a test set is outside the scope of the present workshop, which distinguishes this workshop from other NTCIR tasks.

The MuST data set, the shared resource, motivates participation in the workshop and substitutes for objective evaluation using a common test set in evaluation workshops. The resource shared includes not only the materials to be processed, but also the intermediate results behaving as a hub, and references for the output

[2] `http://must.c.u-tokyo.ac.jp` (In Japanese).
[3] `http://research.nii.ac.jp/ntcir/index-en.html`.

design. Although multi-modal summarization requires several component technologies that are widespread in many research fields, which makes it difficult to construct a total system, using this data set makes it possible for the participants to address their own subjects, that is, parts of the whole framework. Participants can discuss their interests with each other using the data set as common ground, and can see how their studies or their modules fit into the framework. This is especially important for those studying elemental technologies.

MuST was announced in November 2004. At the beginning of 2005, a mailing list of interested persons was developed, and a call for participants started. At the same time, we started to deliver the data set to the participants. In this first cycle, MuST had fifteen participants. A closed workshop was held in March 2006 as the finale of the first cycle, where all participants gave a presentation on their research progress and held an intensive discussion. Now MuST has eighteen participants and is in the second cycle, which will close at the workshop in March 2007. This cycle, just as the first did, welcomes all research using the MuST data set to promote active discussion and research: system development, studies on elemental technologies, and data analysis. In addition, we started resource accumulation, in which the data and tools made in the first cycle are shared by the participants.

4.1 Current Efforts

A major effort conducted in the workshop is information extraction for visualizing trend information. In order to extract more points of data to be plotted on a graph, some participants exploit not only direct expressions like "It rose to $20.20 per barrel," but also comparative expressions like "Oil prices as of this February have dropped 40% since last October." Another group emphasizes the importance of changes of data during some time period rather than data points, as the unit of information extracted and represented. Their proposal includes a widely applicable information extraction method that regards the changes of time-series information as information primitives and a visualization method of qualitative and quantitative characteristics of changes using schematic arrow shapes plotted with textually represented comments.

There are several possibilities for fusion of linguistic and non-linguistic information. One of those is information graphics of numerical data with linguistic annotations. The annotated contents can be qualitative descriptions and explanations of the cause or impact of that value. Examples proposed in the workshop are a chart depicting public approval rating of the cabinet annotated with automatically extracted possible causes affecting the rating at the time, and changes of stock prices with linguistic summaries, which change their summarization ratios based on the granularity of the chart.

The MuST data set is used in several ways in the workshop besides in extraction using annotated information. Some participants are studying summarization and information extraction using the article sets in the data set without using annotation information. It can be said that they are using the sets as the input of multi-document summarization or to try a kind of relation discovery [9] by finding important statistics in the articles.

Other participants, employing an XML parser and simple heuristics, avoided the difficulties of information extraction, and emphasized visualization. Visualization of trends related to spatial statistical information such as land prices and earthquakes are being actively studied. Such visualization often involves designing an interface for information access, which allows users to grasp an overall picture through a visual representation of numerical data, and through that representation to access more detailed linguistic information of interest.

Generating linguistic expression that explains time series data and their changes is another challenge, for which summaries in the MuST data set may be referenced and compared to the output. Those researches sometimes use numerical data from other resources, since the current MuST data set is relatively poor in this aspect. There is a study on simulating the annotations of the MuST data set. While it is relatively easy to assign date and value elements, automatic annotation of names of statistics is still challenging.

All titles and abstracts of the works conducted in the MuST workshop can be found on the MuST home page for those interested.

5 Conclusion

Multi-modal summarization was proposed as a technology for confronting the huge amounts of information currently available by coordinating text summarization and information visualization. Trend information is important because it summarizes huge amounts of linguistic and non-linguistic information expressed in several modes. Multi-modal summarization for trend information, therefore, is worth studying. This paper describes how the MuST data set was designed to foster multi-modal summarization and presents some ongoing efforts that demonstrate the potential of the field and the importance of the data set.

As pressing future work, we have to draw a road map of the technology, since multi-modal summarization of trend information is relevant to several research fields. In addition to those mentioned in section 2.1 and illustrated in section 4.1, generating linguistic summaries of information graphics [10] and of time series data [11] [12] may be component technologies. Some systems were proposed already such as [13]. As mentioned briefly in section 4.1, multi-modal summarization could support users' access to original information, in addition to supporting review of it. Interactive rather than final and static presentations as interfaces for information access should be examined.

Compiling trend information is not the only application of multi-modal summarization. Fashion and public opinions formed by reading personal evaluations and opinions are other fields to which multi-modal summarization could be applied. Such applications should be examined.

Acknowledgments

The participants of the MuST workshop inspired many ideas and insights. We appreciate their contributions. The MuST workshop and related research are

supported in part by grants from a joint research project of Nippon Telegraph and Telephone Corp. and the University of Tokyo, and a joint research project of the National Institute of Informatics and the University of Tokyo.

References

1. Mani, I.: Automatic Summarization. John Benjamins Publishing Co., Amsterdam (2001)
2. Card, S.K., Mackinlay, J.D., Sheiderman, B.: Information visualization. In: Readings in Information Visualization Using Vision to Think, pp. 1–34. Morgan Kaufman Publishers, Inc., Seattle (1999)
3. Fasciano, M., Lapalme, G.: Postgraphe: A system for the generation of statistical graphics and text. In: Proceedings of 8th International Workshop on Natural Language Generation, pp. 51–60 (1996)
4. Roth, S.F., Mattis, J.: Data characterization for intelligent graphic presentation. In: Proceedings of Conference on Human Factors in Computing Systems, pp. 193–200 (1990)
5. Mani, I., Wilson, G.: Robust temporal processing of news. In: Proceedings of ACL-2000, pp. 69–76 (2000)
6. Voorhees, E.M., Harman, D.K. (eds.): TREC: Experiment and Evaluation in Information Retrieval. The MIT Press, Cambridge (2005)
7. Judge, J., Cahill, A., van Genabith, J.: Questionbank: Creating a corpus of parse-annotated questions. In: Proceedings of COLING/ACL 2006, pp. 497–504 (2006)
8. Kato, T., Matsushita, M., Kando, N.: MuST: A workshop on multimodal summarization for trend information. In: Proceedings of the Fifth NTCIR Workshop Meeting on Evaluation of Information Access Technologies, pp. 556–563 (2005)
9. Shinyama, Y., Sekine, S.: Preemptive information extraction using unrestricted relation discovery. In: Proceesings of HLT-NAACL 2006, pp. 320–327 (2006)
10. Carberry, S., Elzer, S., Green, N., McCoy, K., Chester, D.: Extending document summarization to information graphics. In: Proceedings of the ACL-04 Workshop: Text Summarization Branches Out, pp. 3–9 (2004)
11. Boyd, S.: Trend: A system for generating intelligent descriptions of time series data. In: Proceedings of the IEEE International Conference on Intelligent Processing Systems (CIPS-1998) (1998)
12. Sripada, S.G., Reiter, E., Hunter, J., Yu, J., Davy, I.P.: Modeling the task of summarizing time series data using ka techniques. In: Macintosh, A., Moulton, M., Preece, A. (eds.) Applications and Innovations in Intelligent Systems IX, Proceedings of ES2001, pp. 183–916. Springer, Heidelberg (2001)
13. Ahmad, S., de Oliveria, P.C.F., Ahmad, K.: Summarization of multimodal information. In: Proceedings of LREC-2004 (2004)

Interactive Virtual Oriental Brush with Pen-Tablet System

Won-Du Chang and Jungpil Shin

Department of Information System, The University of Aizu, Fukushima, Japan
12cross@gmail.com, jpshin@u-aizu.ac.jp

Abstract. In spite of the importance of natural drawing with input device, the interface between the device and virtual brush's status has been dealt less in literatures. Through this paper, we present a natural interface between pen-tablet device and virtual brush. Proposed system uses pressure instead of z-coordinate of pen to allow drawing with the similar feeling to the real brush. The contribution of this paper is as follows: first, a brush model interacting with pressure is generated. Second, 3D brush model is allowed to be folded back decreasing droplet size intensively according to pen direction changes.

Keywords: Oriental brush implementation, Black-ink painting, Electronic pen device, Human-computer interaction.

1 Introduction

Pen-tablet system is one of the most popular input devices for computer in these days, except keyboard and mouse. Compared to mouse, it provides more familiar environment to its users allowing elaborate touch, and also provides additional information such as pen pressure and directional data. It is often used to design models or draw pictures with graphical software as much image contents are required in digital format in these days [1].

There have been several researches on simulating painting systems in oriental and western styles [2-18]. General painting system can be divided into four parts through the researches in the human interactive system's point of view. The first part is where users input to the system. Users send signals using a certain device at this step. The second part is modeling brush. According to the input data, the brush is stated to a certain shapes. Typically, in the results of this part, 2D droplet is generated on the paper and a stroke is generated connecting two droplets. Conventional method to generate a droplet or stroke is controlling the droplet shape directly on 2D [2-10], and relatively small number of system use a 3D brush model and controls droplet shape with it [11-13]. In spite of its relatively heavy computational requirement, 3D system allows more freedom to its users [12]. The third part is modeling paper. The structure of paper is pre-determined and not related to the inputs directly, which is highly related to the ink or color's activities on the paper. Papers are modeled with pipes between cells [14], filters between cells [3], and layers at each cell [10]. The fourth

B. Apolloni et al. (Eds.): KES 2007/ WIRN 2007, Part II, LNAI 4693, pp. 387–394, 2007.
© Springer-Verlag Berlin Heidelberg 2007

part is modeling the activities of ink or colors. The activities inside of brush [14], inside of paper [3,10,14], and on the paper [17] are modeled in the previous researches.

Meanwhile, the most of the researches mentioned above focus on the processes after the data is inputted and only two research groups have dealt with the first part of entire system. Chu and Tai used a real oriental brush as an input device with ultrasound emitter/receivers and gyroscopes [12]. Users can watch the virtual brush on the monitor, where the brush is moving according to the real brush. Using the real brush, this system provides very natural environment for the conventional painters. However, the virtual brush may act differently to the real brush because there is no sensor at the bristles and the direction of the bristles is basically determined according to the pen direction. They additionally refer a system using a pen-tablet device in the same paper, where the pressure is simply converted to the height of the brush. This simple converting can decrease the naturalness of painting because the required pressure to lower the brush is different according to the angle and the height of the brush.

Baxter *et al.* proposed a system using SensAble Technologies' PHANTOM which provides the z coordinates of the brush [17]. Similar to the Chun and Tai's system, they used z axis as the height of the brush on the paper. One of the drawbacks in this system is that it is impossible to have feeling of touching paper when users paint with the system, and the bristles' direction is basically determined according to the pen direction also.

In this paper, we propose an interface between pen-tablet device and virtual 3D brush, where their correspondence is more natural compared to the conventional methods in a certain points. The proposed interface is mainly described in the section 2. Section 3 discusses the generated results from the proposed method, and the conclusions and future works are followed at the last section.

2 Brush Model

2.1 General Design

In this section, we present a model of an oriental brush only, but the proposed model could be easily applied to other brushes modifying the parameters. Fig. 1. shows the oriental brush used as a sample. Because oriental brush uses a number of different bristles for its hairs, it shows more elastic energy when it is bent. The structure of general oriental brush can be found in [12].

We model virtual brush so that its horizontal cross section becomes circle (see fig. 2). This model is designed inspired by [12] but different in many points. Main part of the model is the core line from its root to the tip. The core is able to be bent, where a disk crossing the core is orthogonal. The radiuses of the discs according to their positions are measured from the real brush. When the disk touches paper, it is distorted to prohibit the disk from be lowered under the paper preserving its dimensions. The contact area to the paper is painted with ink, and a droplet is



(a) Entire shapes (b) Magnified bristles' region

Fig. 1. Real oriental brush

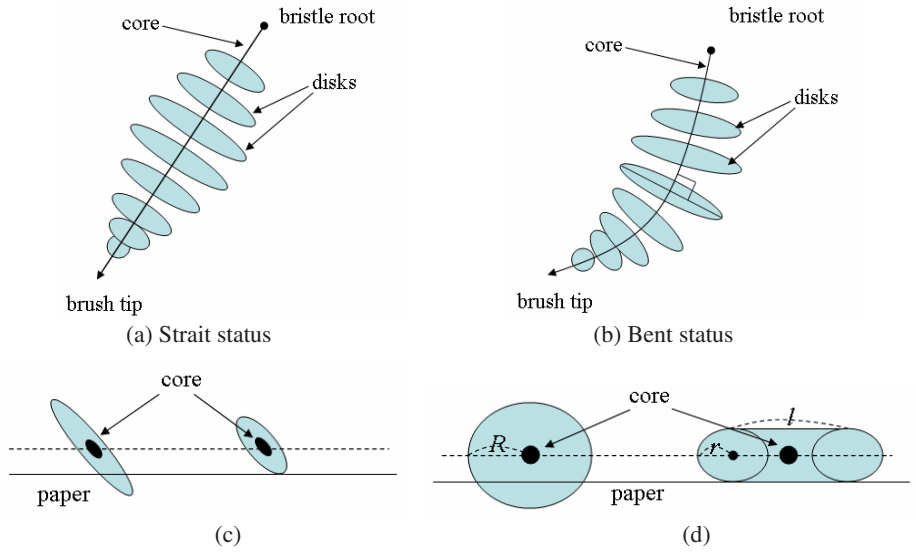
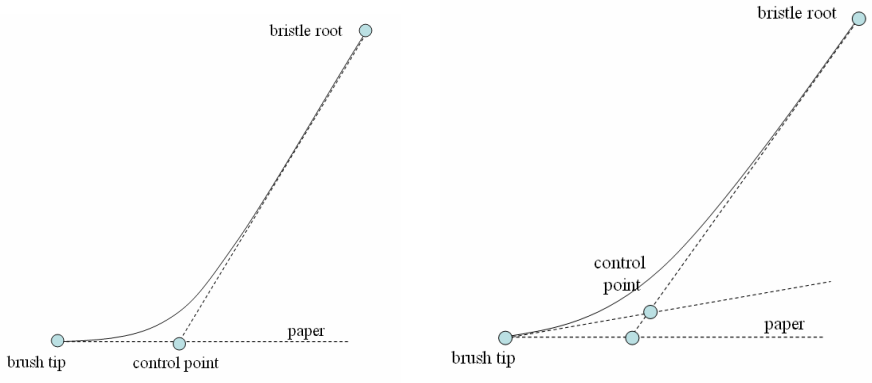


Fig. 2. Brush (bristles) model. (a) and (b) shows the status of bristles when they are strait and bent respectively. (c) and (d) show the shape change of a disk when it touches paper. (c) is The side-view and (d) is front-view of the disk. The left images of (c) and (d) represent the disk before changed, and the right images show the changed shape. The size of the disk section is equal, so that $\pi \cdot R^2 = \pi \cdot r^2 + 2rl$ where R is radius of a disk before changed and r is distance between the core at the disk and paper. When r is smaller than a threshold, it is set to the threshold value.

generated on the paper when the contact area by all the disks is painted. Additionally, to prevent the height of distorted area from being too small and the contact length to paper from being too long, a minimum value is adopted for the height and the disk is not allowed to have shorter height than the minimum.

The curve of the core representing a bent bristles is defined with Bezier curve, which generate similar curve to the real curve of bent bristles with small number of control points. Bristle root, brush tip, and an additional point are used for control points (see fig.3a). In this model, however, the amount of disk distortion increases too high at the positions on the core near the tip, which is not realistic by itself and also generating strange droplet. Hence, we set the control point slightly higher and slant the vector from the tip to the control point as in fig 3(b).



(a) Normal Bezier curve (b) Curve with slant control vector

Fig. 3. The core line's curve for bent bristles

2.2 Brush Dynamics

In this section, we define the correlation between the input data from pen-tablet and the 3D brush model. The data from the pen-tablet are pen pressure and directions (altitude and azimuth), and the required data to determine the shape of 3D brush are the positions of the control point and the bristle's root when the length of bristles' core and the disk radiuses are pre-defined.

First, we define the angle between the vector from the control point to bristle's root and the paper as the altitude of the input data since it is natural. Because the bristles' length and the slant angle of the control point is constant, only required variables to determine 3D brush is the length between the pen tip and the control point.

Here, at a certain status of 3D brush, we define three types of forces - the force to endure elastic force, the force to bend the bristles to the current status, and the given force by a user. Basically the first force can be thought as an integral of the second force according to time, and its size is equal to the third force's size.

Let the brush tip point as O, the control point as V, bristle's root point as R, the point where the vector RV meets the paper as S, the angle between vector VR and OS as α, and the angle between vector OV and OS as θ. We adopt the elastic force equation from [12]. The size of the force to straighten bristles up, $\left\|F^e\right\|$, is defined as follows:

$$\left\|F^e\right\| = \kappa\left|\alpha-\theta\right|^m, \tag{1}$$

where κ and m are control parameters for elastic force.

Meanwhile, the size of the force to bend bristles to the status is proportional to $\left\|F^e\right\|$ if α and θ are not changed when they are bent. With the assumption, the size of total force $\left\|F^T\right\|$, to bend the bristles into a certain status, can be written as follows:

$$\left\|F^T\right\| = \left\|F^e\right\| \cdot l, \tag{2}$$

where l is the length between the pen tip and the control point.



(a) Normal status (b) Reversed status

Fig. 4. Two status of brush in writing

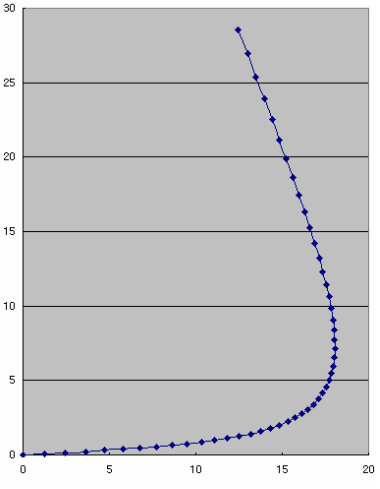


Fig. 5. Modeled brush core when the pen is reversed slightly

$\|F^T\|$ is needed to be normalized to be compared with the pressure of pen-tablet device, so that both $\|F^T\|$ and pressure are normalized within [0 1]. Here, if the maximum value of α and l is set up, κ can be easily calculated. In this paper, $\alpha^{\max}$ and $l^{\max}$ are $3\pi/4$ and $\tau \cdot L$ respectively, where L is the length of bristles and τ is control parameter. For the experiment in the paper, m is set to 0.8 and τ is set to 0.6 experimentally.

2.3 Reversed Brush

It is often in oriental painting that drawing with brush angle α in section 2.2 bigger than 90 degrees. Fig. 4 shows the difference of two statuses: normal and reversed. Not only the status of the brush is completely different, but also the shape of the bristles and their size show the big differences. It causes inconvenience and limits the degree of freedom if the tip direction of the brush is basically set according to the input

direction from pen-tablet and not concerning the reversed brush. In our system, the altitude and azimuth information is used to detect the reversing of input pen. We basically decide the reverse if the difference of brush's azimuth to the pervious value is bigger than a certain threshold. However, the previous value of azimuth is only updated when the altitude difference is bigger than another threshold. The 3D brush model's core generated by the proposed system is shown in fig. 5. The equation is as follows:

for i^{th} input data,

$$\begin{aligned} &if\,(\psi_l(i) < \theta_l(i))\ \text{then},\ prev = i-1, \\ &if\,(|\psi_z(i) - \psi_z(prev)| > \theta_z) \\ &\quad \text{then, change the brush status (normal} \leftrightarrow \text{reversed),} \end{aligned} \tag{3}$$

where ψ_l is an altitude input value, $prev$ is a previous data index to compare, and ψ_z is an azimuth.

3 Results and Discussions

The described painting model is developed and we are able to paint oriental black-ink painting. Because our subject is the interaction between pen-tablet and virtual brush status, the droplet shapes according to the pen pressure and altitude are displayed in fig. 5. In the results, it is very clear that the effect of pressure and altitude is different. The width of droplets increases more as the pressure increases, and the height of droplet are controlled intensively by the altitude. But they can be effect strongly even when it is applied separately. For example, whether the pressure is in a low or high status, droplet could be wide highly if the altitude is extreme low. The dashed line divides the droplets generated by normal and reversed brushes. Though their actual altitudes are getting lower for the droplets below the dashed line, the size of the brushes are getting smaller, and it shows proposed algorithm is applied successfully.

4 Conclusions and Future Works

In this paper, we first summarized the previous works on the computer-aid painting in the human interactive system's viewpoint, and also we have presented an interactive correlation between pen-tablet and virtual 3D brush. 3D brush is modeled in accordance with pen pressure and directions, and droplets are generated with the brush model on paper. In addition, we define a reversed status for the virtual brush using pen directional information and showed it is successfully with proposed system.

Compare to the conventional method, the proposed brush model reacts according to the pressure of pen, allowing users writing with their feeling as well as their sight, and the proposed definition of reversed status is simple and expected to be easily applied to the other systems, while it increases the degree of freedom much.

In our future works, we are going to try to apply the more complicated physical formulas. Though we did not focus on the problem because of its complexity and

computational workload, it is expected to be an interesting challenge to overcome the difficulty of the problem.

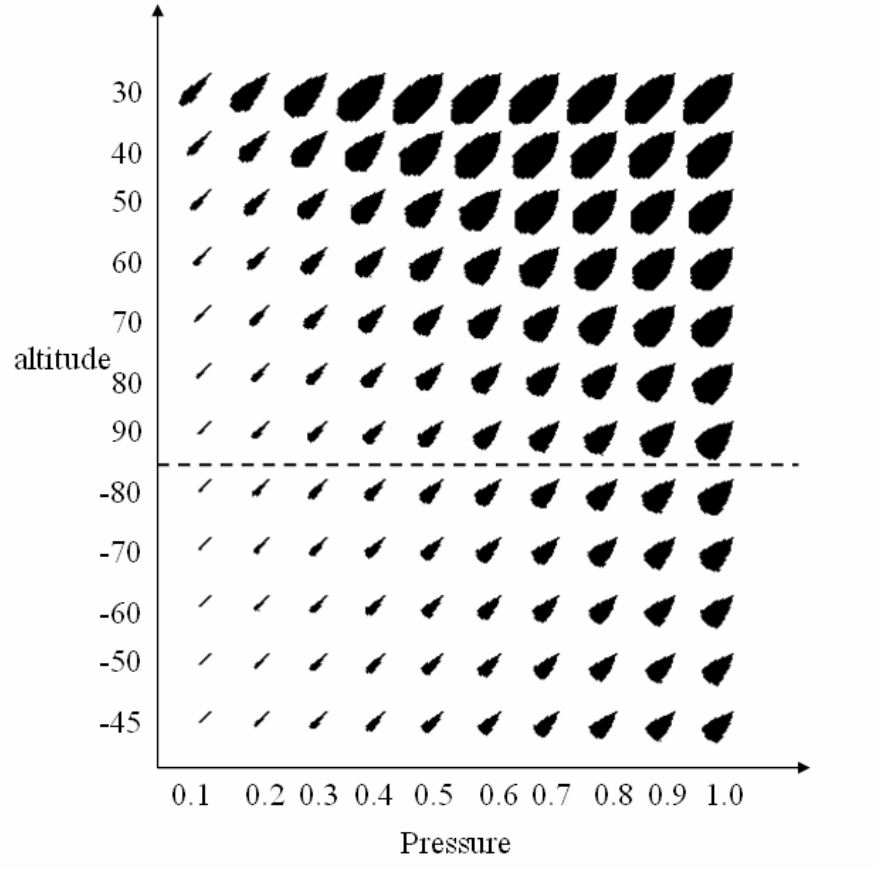


Fig. 6. Generated droplets using proposed system

References

1. Interview with Animators and a Painter (in Korean), http://www.wacomdigital.co.kr
2. Xu, S., Lau, F.C.M., Cheung, W.K., Pan, Y.: Automatic Generation of Artistic Chinese Calligraphy. IEEE Intelligent Systems 20(3), 32–39 (2005)
3. Huang, S.-W., Way, D.-L., Shih, Z.-C.: Physical-based Model of Ink Diffusion in Chinese Ink Paintings. Journal of WSCG 11(3), 520–527 (2003)
4. Strassmann, S.: Hairy Brushes. In: Proceedings of the 13th Annual Conference on Computer Graphics and Interactive Techniques, vol. 20(4), pp. 225–232 (1986)
5. Hertzmann, A.: Painterly Rendering with Curved Brush Strokes of Multiple Sizes. In: Proceedings of the 25th Annual Conference on Computer Graphics and Interactive Techniques, pp. 453–460 (1998)

6. Hsu, S.C., Lee, I.H.H.: Drawing and Animation Using Skeletal Strokes. In: Proceedings of the 21st Annual Conference on Computer Graphics and Interactive Techniques, pp. 109–118 (1994)
7. Yu, J., Peng, Q.: Realistic Synthesis of Cao Shu of Chinese Calligraphy. Computers and Graphics 29, 145–153 (2005)
8. Mi, X., Xu, J., Tang, M., Dong, J.: The Droplet Virtual Brush for Chinese Calligraphic Character Modeling. In: Procedings of the 6th IEEE Workshop on Applications of Computer Vision, p. 330 (2002)
9. Ishigame, M., Miura, N., Hosaka, A., Fukushima, T.: Creation of Brush-Written Characters using Electronic Writing Brush Model. Human Interface Research Report in Japan (in Japanese) 71(7), 45–51 (1997)
10. Yu, Y.J., Lee, D.H., Lee, Y.B., Cho, H.G.: Interactive Rendering Technique for Realistic Oriental Painting. Journal of WSCG, 11(1) (2003)
11. Baxter, W.V., Lin, M.C.: A Versatile Interactive 3D Brush Model. In: Proceedings of the 12th Pacific Conference on Computer Graphics and Applications, pp. 319–328 (2004)
12. Chu, N.S.H., Tai, C.-L.: Real-Time Painting with an Expressive Virtual Chinese Brush. IEEE Computer Graphics and Applications 24(5), 76–85 (2004)
13. Lee, J.: Simulating Oriental Black-Ink Painting. IEEE Computer Graphics and Applications 19(3), 74–81 (1999)
14. Zhang, Q., Sato, Y., Takahashi, J., Muraoka, K., Chiba, N.: Simple Cellular Automaton-based Simulation of Ink Behaviour and Its Application to Suibokuga-like 3D Rendering of Trees. The Journal of Visualization and Computer Animation 10, 27–37 (1999)
15. Baxter, W., Wendt, J., Lin, M.C.: IMPaSto: A Realistic, Interactive Model for Paint. In: Proceedings of the 3rd International Symposium on Non-photorealistic Animation and Rendering, pp. 45–148 (2004)
16. Curtis, C.J., Anderson, S.E., Seims, J.E., Fleischer, K.W., Salesin, D.H.: Computer-Generated Watercolor. In: Proceedings of the 24st Annual Conference on Computer Graphics and Interactive Techniques, pp. 421–430 (1997)
17. Baxter, B., Scheib, V., Lin, M.C., Manocha, D.: DAB: Interactive Haptic Painting with 3D Virtual Brushes. In: Proceedings of the 28st Annual Conference on Computer Graphics and Interactive Techniques, pp. 461–468 (2001)
18. Baxter, W., Liu, Y., Lin, M.C.: A Viscous Paint Model for Interactive Applications. In: ACM SIGGRAPH 2004 Sketches SIGGRAPH '04, p. 3 (2004)

Multidimensional Emotional Appraisal Semantic Space (MEAS): Evaluating HM Affective Interactions

Maria Rita Ciceri and Stefania Balzarotti

Laboratory of Communication Psychology,
University Cattolica del Sacro Cuore,
Largo Gemelli 1, 20123 Milano, Italy
{maria.ciceri,stefania.balzarotti}@unicatt.it

Abstract. In this paper we'll present a dimensional semantic model of emotion locating the former in a four axes space in the attempt to quantify emotion considered as constantly changing with time rather than to label it. The Multidimensional Emotional Appraisal Semantic Space (MEAS) is a method according to which the most relevant features to measure emotions are: a dimensional approach to emotion, the multimodality of expressive signals and the embodied nature of the emotional interaction. Four dimensions have been selected: novelty, pleasantness, coping and arousal. The componential organization all these different levels gives birth to multiple and fuzzy emotional experiences that are in part signalled through patterns of non verbal expression and that can change their intensity in relation to the relevance of the event for the individual and in relation to the arousal given by physiological signals.

Keyword: Emotion.

1 Introduction

The increasing use of computers and machines that support the human user in any kind of task has nowadays transformed them into important and constitutive members of the physical and social environment with which people interact. For this reason, researchers from several disciplines have turned their attention to emotion and its role in Human Machine interaction [1][2][3][4][5]. Emotional models are needed to implement new generation interfaces as emotions are thought to constitute key components to achieve more effective Human Computer Interaction. However, modelling emotions has revealed a complex challenge and different computational models of the user's emotion have been developed to support the complexity of human emotion and the multimodal richness of face-to-face communication [6][7][8][3]. Research goal in this area has been twofold: on one side, theoretical knowledge on emotion has been applied towards emotion simulation implementing artificial human-like autonomous agents (e.g. avatars, robots, ECAs, etc.) able to interact with the user reproducing human facial and/or vocal emotional expressions [9][10][11] [6]; (2) a second research field has concerned emotional decoding through

B. Apolloni et al. (Eds.): KES 2007/ WIRN 2007, Part II, LNAI 4693, pp. 395–402, 2007.
© Springer-Verlag Berlin Heidelberg 2007

the implementing and design of interfaces able to recognize the user's emotional responses from real-time capturing and processing of multiple – physiological, facial and vocal – signals [4]. Whatever the research goal is – simulation or recognition – researchers have to face the complexity of the emotion: in the last years a widespread consensus has grown in psychological literature on the *componential* nature of emotion [12] [13] [14] [15] and thus theorists have stressed the need to analyze multiple response systems – physiological, behavioural and experiential. A critical issue in the implementation of interfaces able to manage affective interactions with the user concerns the semantic dimension, i.e. the criteria or rules to attribute an emotional meaning to patterns of signals. To date, the main approach has been the fixed and stable attribution of an emotional label to a specific configuration of signals. For example, the well-known Facial Action Coding System (FACS) [16]– or better the EMFACS - involves a stable link between configurations of Action Units and six discrete basic emotions. However, such an approach seems not able to account for the multiplicity of emotional signals that characterize human interactions and seems limited to the decode and interpret stereotyped sets of few basic emotions.

In this paper we'll present a dimensional semantic model of emotion locating the former in a four axes space in the attempt to quantify emotion considered as constantly changing rather than to label it. The Multidimensional Emotional Appraisal Semantic Space (MEAS) is a method according to which the most relevant features to measure emotions are: a dimensional approach to emotion, the multimodality of expressive signals and the embodied nature of the emotional interaction.

2 Evaluating Affective Interactions: Dimensional Approach, Multimodality and Embodiment

The ability to «measure» the user emotional state seems to be still the hardest challenge today both when attempting to decode emotions and to simulate them in a believable manner within the interaction. Three characteristics of the emotional interaction seem to emerge as relevant: the multimodality of expressive signals, the dimensional approach, its embodied nature.

2.1 Emotional Appraisal and Dimensional Approach

In recent years, appraisal emotion theories have suggested that the elicitation and differentiation of emotions is based on a process of appraisal of the situations that affect an individual's needs and goals [15]. This means that emotions may be differentiated on the basis of a few appraisal dimensions, according to which individuals evaluate situations and events. The component-process model [14] describes emotion as the dynamic and constantly changing affective tuning of organisms as based on the continuous evaluative monitoring of their social and physical environment. In particular, it has been defined a fixed sequence of event evaluation checks: novelty and expectancy, hedonic valence, relevance and goal conduciveness, agency and responsibility (SECs, [17]). Thus, emotion is considered as *a process, not a state*: it is defined as an episode of temporary synchronization of all major subsystems of the organism functioning represented by five components (cognition; physiological regulation; motivation; motor expression, and monitoring/feeling) [18].

2.2 Multimodality

Researchers have based their empirical investigations including the combination of multiple modalities combining more classical methods with the collection of body measures that may capture aspects of an affective state [2] [4], trying to reproduce the multimodal richness of the emotional process and of human communication.. In this perspective, combining all the features extracted might provide consistent information in order to increase the accuracy, reducing the error related to the use of single parameters [19] [20].



Fig. 1. Multimodal signals

Research on multimodal communication have suggested that human expressive systems are characterized by the semantic interdependence, in which each signal in relation to the other participates in the meaning construction. The coordinated use of signals belonging to different expressive systems enables humans to soften, to stress, to modify and emotional states and adapt communicative intentions. For these reasons, it becomes relevant to focus on the role of different signal systems in emotional expression: a first distinction concerns two macro categories, involuntary physiological signals on one side and communicative signals such as facial expression on the other. In our model, physiological signals act as a generic and transverse layer besides the specific channels used to realize the communication and the interaction between Human and Machine. As a matter of fact, the principal use we may do in our framework of the physiological signals is as detectors of the beginning of emotional arousal in the human [21]. Then, the progressive analysis of the nonverbal signals allows us to identify successive levels of the emotional appraisal.

A multimodal system aimed at coding and decoding emotion should require: 1) the synchronized acquisition of multimodal signals; 2) relevant features extraction and 3)

the transmission to a system aimed at the semantic attribution [22]. The method presented here tries to focus on this last point.

2.3 Embodied Nature of Emotional Interaction

Our model is not aimed at recognizing and labeling emotions, but rather it tries to accompany the appraisal process along the temporal line of action in order to define rules that support an efficient behavior selection and enhance the attuning of the HM system. Emotional interactions are embedded in a physical context, i.e. a) they are *situated interactions*, where agents share a common intention (we-intention), which may be different with respect to the context: for example, have fun, give/receive information, learn within an e-learning context, etc.; b) they occur through and within joint actions. In particular, agents need to coordinate their action and tune their emotional reactions accordingly.

As a consequence, in the definition of possible rules of semantic attribution, our model tries to take into account two different contextual factors: the event or situation (that is the stimulus/action of agent) and the user action (performance).

3 A Multidimensional Approach

One of the most challenging problems today seems to concern the selection of relevant features and the process of semantic attribution. According to theories and approaches previously presented, we can argue that a system based on linear emotional coding-decoding and functioning through the attribution of a fixed number of emotional labels to a fixed number of expressive indexes seems not much reliable and ecologically unlikely: first of all, continuous expressive modifications occur during the interaction; secondly complex or mixed emotions may be expressed.

According to appraisal theories [15], emotions are considered experiences that flow along a temporal line of action, elicited and differentiated on the basis of a person's subjective cognitive evaluation of the significance of a situation. The MEAS dimensional semantic model of emotion locates the former in a four axes space in the attempt to quantify a human attitude constantly changing with time rather than definitely label it. Four dimensions have been selected: the first three dimensions correspond to appraisal criteria derived from SECs (novelty, pleasantness and coping) while the last corresponds to the level of physiological arousal. The componential organization all these different levels gives birth to multiple and fuzzy emotional experiences that are in part signalled through patterns of non verbal expression and that can change their intensity in relation to the relevance of the event for the individual (graduation of the axes from -2 to +2) and in relation to the arousal given by physiological signals.

3.1 MEAS Procedures of Scoring

The MEAS procedure of scoring has as a starting point the analysis of the signals belonging to the different response systems taken into account (non verbal and physiological).

Regarding non verbal behavior, the first level of the method consists of the behavioral *frame by frame* micro-analysis (25ms). Two different indexes are extracted: behaviour rate (number of occurrences/min) and duration (% interval). This analysis is based on the previous elaboration of a coding grid and thus the selection of behavioural units to be extracted (BCS, Behaviour Coding system, [23]. Four macro-categories are considered. *Facial movements*: the fundamental muscle movements that comprise Facial Action Coding System [16] were selected. We considered action units relating to upper face and lower face (20 AUs and 10 ADs). For each unit intensity is rated (Low, Medium, High). *Gaze direction:* we considered if the subject looked at the task, at the keyboard, around. etc. *Posture*: behavioural units of approaching/withdrawing are considered. *Vocal behaviours:* it is recorded when the subject is speaking (verbal) or uses other kinds of vocalizations. Reliabilities between different coders need to be calculated. Moreover, THEME Software is used for the detection of multimodal recurrent pattern analysis (T-pattern). The number of different T-patterns detected in a behavioural stream, their average and/or maximum length (number of event types involved) and their average/maximum level (number of hierarchical levels in a pattern) may be used as measures of complexity or overall synchrony [24].

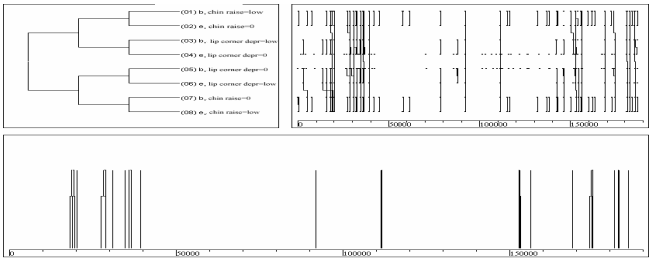


Fig. 2. The example shows one of the most frequent patterns from THEME 5.0 data [25]. Three diagrams are presented: 1) the pattern tree graph shows the event types composing the pattern (signalling begin or end) listed in the order in which they occur, while the pattern connections are on the left of the event list showing how many hierarchical levels are involved; 2) the connection graph deals with frequency and real-time distribution of events in the pattern: dots represent event occurrences and the lines connecting the dots represent pattern occurrences; 3) the instance graph provides information about the real-time structure of the pattern. This specific pattern combines the AU combination 15+17 onset and offset.

At a second level of analysis, starting from indications found in literature and from the results of the previously mentioned analysis of patterns and configuration of multimodal signals, the three axes derived from SECs (novelty, hedonic value, coping) are scored on a 5 point rating scale from -2 to +2 by a group of three judges to describe the user emotional state. Inter-judge agreement is calculated. Information on a fourth dimension concerning arousal are derived from the analysis of physiological data, i.e. we consider the possible physiological correlates rising from GSR and heart rate (HR) through the detection of different patterns, in order to identify possible

autonomous nervous system (ANS) arousal and modification of activity. More in detail we adopted the same descriptors used in previous research involved in the investigation and identification of emotion through bio-signals [2].

3.2 MEAS Application: Example of Qualitative Analysis

In this section we'll present a first and obviously limited attempt to investigate how multimodal sources of information can be combined in the evaluation of affective interactions: physiological signals (HR and EDA), non verbal behaviour (facial movements, head, posture. etc.), events of the ongoing task (context), the subject's performance (action) and finally the emotional axes which should represent the constantly changing emotional appraisal of the situation.

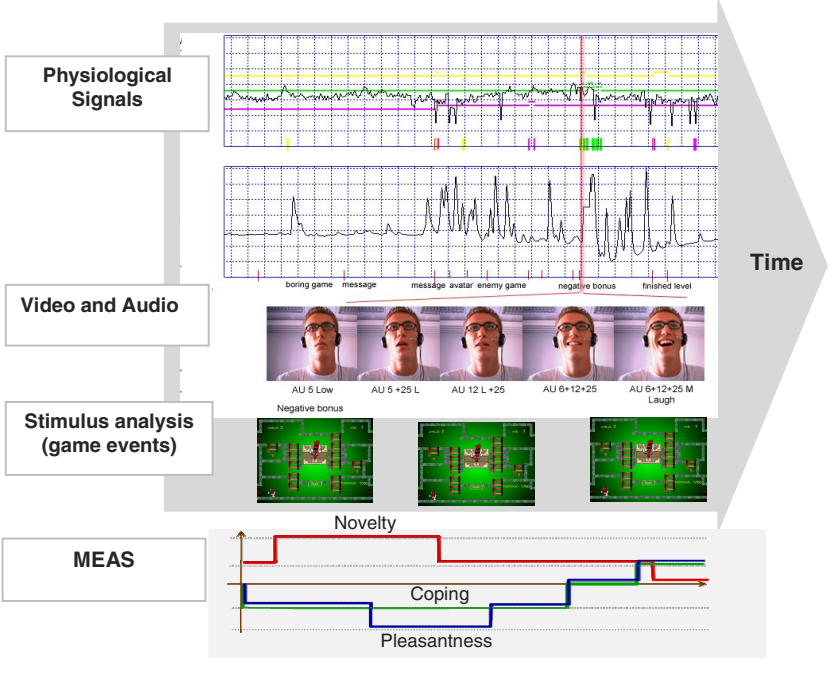


Fig. 3. Emotional attuning process. The figure shows: 1) the physiological signals (heart rate and EDA) corresponding to a temporal section of 16 minutes of videogame (avatar-boring game); 2) the simultaneous variations in the dimensional axes of novelty, pleasantness and coping on a rating scale ranging from -2 to +2; 3) one frame sequence extracted from video intervals of 3.18 sec showing the non verbal units exhibited in correspondence of 4) a specific event (negative bonus) and performance (dead).

From a qualitative analysis of this extracted data, it is possible to observe: (1) the need to rely on changes in multiple systems such as non verbal (facial) expression to enhance the interpretation level of the emotional experience with respect to semantic dimensions such as pleasantness or coping; (2) both variations in physiological and

non verbal signals are congruent to the task events; (3) non verbal behavioural units (especially facial movements) and physiological signals seem to require different temporal units of analysis: a 25 fps resolution allows to observe the onset, offset and intensity micro-variations of action units, while a longer temporal units is required by physiological variations.

Acknowledgments. The elaboration of the MEAS model is part of a wider research project and we would like to thank all the people who are involved and cooperated in its realization; a special thank to the scientific coordinator Prof. Bruno Apolloni of the DSI Department University Statale of Milan and to Eng. Luca Piccini, TBM Laboratory, Politecnico, Milan for the analysis of physiological signals.

References

1. Picard, R.W.: Affective Computing. MIT Press, Cambridge (1997)
2. Picard, R.W., Vyzas, E., Healey, J.: Toward Machine Emotional Intelligence: Analysis of Affective Physiological State, IEEE Transactions Pattern Analysis and Machine Intelligence, 23(10) (2001)
3. Cowie, R., Douglas-Cowie, E., Tsapatsoulis, N., Votsis, G., Kollias, S., Fellenz, W., Taylor, J.G.: Emotion recognition in human-computer interaction. IEEE Signal Process 18, 32–80 (2001)
4. Lisetti, C.L., Nasoz, F., LeRouge, C., Ozyer, O., Alverez, K.: Intelligent affective interfaces: a patient-modelling assessment for tele-home health care. International Journal of Human-Computer Studies 59, 245–255 (2003)
5. Lisetti, C.L., Brown, S.M., Alverez, K., Marpaung, A.H.: A social informatics approach to human robot interaction with a service social robot. IEEE Transactions on Systems, Man and Cybernetics, Special Issue on Human-Robot Interaction, 195–209 (2004)
6. Lisetti, C.L.: Personality, affect and emotion taxonomy for socially intelligent agents. In: Proceedings of the 15th International Florida Artificial Intelligence Research Society Conference (FLAIRS'02), Menlo Park, CA, AAAI Press, Pensacola, FL (2002)
7. Bianchi, N., Lisetti, C.L.: Modelling Multimodal Expression of User's Affective Subjective Experience. User Modelling and User Adapted Interaction 12(1), 49–84 (2001)
8. Kort, B., Reilly, R., Picard, R.W.: An Affective Model of Interplay Between Emotions and Learning: Reengineering Educational Pedagogy—Building a Learning Companion. In: International Conference on Advanced Learning Technologies (2001)
9. Breazeal, C.: Sociable Machines: Expressive Social Exchange Between Humans and Robots. Doctoral Dissertation. Department of Electrical Engineering and Computer Science. MIT, Cambridge (2000)
10. Bartneck, C.: How convincing id Mr Data's Smile: Affective expressions of Machines. User Modelling and User-Adapted Interaction 11, 279–295 (2001)
11. Gratch, J., Marsella, S.: Tears and Fears: Modelling Emotions and Emotional Behaviours in Synthetic Agents. Presented at AAMAS, Montreal, Canada (2001)
12. Lazarus, R.S.: Emotion and Adaptation. Oxford University Press, New York (1991)
13. Levenson, R.W.: Human emotions: A functional view. In: Ekman, P., Davidson, R.J. (eds.) The nature of emotion: Fundamental questions, pp. 123–126. Oxford University Press, New York (1994)

14. Scherer, K.R.: Emotions as episodes of subsystems synchronization driven by nonlinear appraisal processes. In: Lewis, M.D., Granic, I. (eds.) Cambridge studies in social and emotional development, Emotion, development, and self-organization: Dynamic systems approaches to emotional development, vol. 13, pp. 70–99. Cambridge University Press, New York (2000)
15. Scherer, K.R., Schorr, A., Johnstone, T. (eds.): Appraisal processes in emotion: Theory, methods, research. Series in affective science. Oxford University Press, London (2001)
16. Ekman, P., Friesen, W.V., Hager, J.: Facial Action Coding System. The Manual. Research Nexus Division of Network Information Research Corporation, Salt Lake City (2002)
17. Scherer, K.R.: Appraisal considered as a process of multilevel sequential checking. In: Scherer, K., Schorr, R., Johnstone, A. (eds.) Appraisal Processes in Emotion: Theory, Methods, Research. Series in affective science, Oxford University Press, London (2001)
18. Van Reekum, C.M., Johnstone, T., Banse, R., Etter, A., Wehrle, T., Scherer, K.R.: Psycho physiological responses to appraisal dimensions in a computer game. Cognition Emotion 18(5), 663–688 (2004)
19. Prendinger, H., Mayer, S., Mori, J., Ishizuka, M.: Using bio-signals to measure and reflect the impact of character-based interfaces. In: Proceedings Fourth International Working Conference on Intelligent Virtual Agents (2003)
20. Cacioppo, J.T., Tassinary, L.G., Berntson, G.G.: Handbook of psychophysiology, 2nd edn. Cambridge University Press, New York (2000)
21. Piccini, L.: Design and development of hardware and software architecture for the unobtrusive biosignals acquisition: applications in advanced human machine interfaces, PhD Thesis, Politecnico di Milano, Department of Bioengineering (2003)
22. Anisetti, M., Bellandi, V., Damiani, E., Beverina, F.: Facial identification problem: a tracking based approach. In: Proceedings of the First International Conference on Signal-Image Technology and Internet-Based Systems, IEEE SITIS, Yaoundé, Cameroon (2005)
23. Ciceri, R., Colombo, P., Balzarotti, S.: Analysis of the human physiological responses and multimodal emotional signals to an interactive computer. In: Proceedings of AISB, Agents that want and like: motivational and emotional roots of cognition and action (2005)
24. Magnusson, M.S.: Discovering hidden time patterns in behaviour: T-Patterns and their detection, behaviour research methods. Instruments and Computers 32(1), 93–110 (2001)

Multimodal Sentence Similarity in Human-Computer Interaction Systems

Fernando Ferri, Patrizia Grifoni, and Stefano Paolozzi

Istituto di Ricerca sulla Popolazione e le Politiche Sociali - Consiglio Nazionale delle Ricerche, Via Nizza 128, 00198 Rome, Italy
{fernando.ferri,patrizia.grifoni,stefano.paolozzi}@irpps.cnr.it

Abstract. Human-to-human conversation remain such a significant part of our working activities because its naturalness. Multimodal interaction systems combine visual information with voice, gestures and other modalities to provide flexible and powerful dialogue approaches. The use of integrated multiple input modes enables users to benefit from the natural approach used in human communication. However natural interaction approaches may introduce interpretation problems. This paper proposes a new approach to match a multimodal sentence with a template stored in a knowledge base to interpret the multimodal sentence and define the multimodal templates similarity. We have assumed to map each multimodal sentence to the natural language one. The system then provides the exact/approximate interpretation according to the template similarity level.

Keywords: Human-Computer interaction, multimodality, sentence similarity.

1 Introduction

Face-to-face conversation remains such a significant part of our working activities despite of the availability of a great number of communication technologies. Anyway, there is a great interest towards natural interaction approaches and great efforts in development of technologies to aid such type of approaches (see for example [1], [2], [3]) and in improvement the interpretation on the computer side.

From this perspective multimodality has the potential to greatly improve Human-Computer Interaction combining harmoniously different communication methods. A Users can use voice, handwriting, sketching and gesture to input information. On the other side the system can use icons, text, sound and voice (output)to present information.

This paper uses the concept of multimodal language defined as a set of multimodal sentences [4], by the extension of the definition of Visual Language given in [5]. A multimodal sentence contains atomic elements (glyphs/graphemes, phonemes and so on) that form the Characteristic Structure (CS). The CS is given by the elements that form functional or perceptual units for the user. A multimodal sentence is defined, similarly to [6], as a function of: 1) the multimodal message, 2) the multimodal description that assigns the meaning to the sentence, and 3) the interpretation function

B. Apolloni et al. (Eds.): KES 2007/ WIRN 2007, Part II, LNAI 4693, pp. 403–410, 2007.
© Springer-Verlag Berlin Heidelberg 2007

that maps the message with the description, and the materialization function that maps the description with the message.

The goal of this paper is to propose a new approach to understand how a multimodal input sentence is matched with a template considering how different modalities cooperate each other, taking into account the user's behavior that can alter the multimodal input recognized by the system. The resulting multimodal input sentence is matched with a template stored in a knowledge base to provide an interpretation of the sentence. The sentence can precisely match the template or approximates it.

We consider the speech modality as the prevalent one because, generally, users explain their intentions by speech and use other modalities to "support" the speech and eventually to resolve ambiguities.

In our system each multimodal sentence corresponds to a natural language one. The system returns the interpretation of the multimodal sentence if the template that exactly maps with the corresponding natural language sentence is available. When the user interacts with the system, the corresponding multimodal sentence must refer to a stored template in the knowledge base in order to be interpreted. If the corresponding sentence, expressed in natural language, doesn't match with any of the stored templates than the system can interpret those sentences that approximate the matching considering templates similar to the first one. Because of some multimodal sentences (with different templates) can have very close interpretations (some times they can have the same meaning), this paper proposes to calculate the templates' similarity starting from the semantic similarity of the natural language sentences corresponding to the Multimodal one, we also take into account the user's behavior that can alter the multimodal input recognized by the system. For this purpose it is possible the association of a sentence and its template with the most similar template computing semantic similarity between natural language sentences. In this way we can provide an interpretation of the multimodal sentence also in case of non-perfect matching.

The natural language sentence can be represented as a semantic network of objects and binary relations among them, where each object corresponds to one node.

The paper is structured as follows: Section 2 provides a running example, section 3 describes the proposed approach. section 4 reports the conclusion.

2 Running Example

In order to explain our approach we consider the following example: the user draws an Entity-Relationship scheme assigning a label to each construct. Without loss of generality we assume that the input is given by only two modalities: speech and sketch. This scenario is represented in Fig. 1.

Suppose that the user sketches the diagram as shown in Fig. 1a, and he/she speeches the sentences shown in Fig. 1b. The system has to interpret the multimodal sentence and then has to materialize it as shown Fig. 1c.

For the sake of simplicity we only consider the creation of the Teaching Relation.

The user says: "The rhombus is the relation Teaching", at the same time the user sketches the figure of a rhombus (in his/her intention).

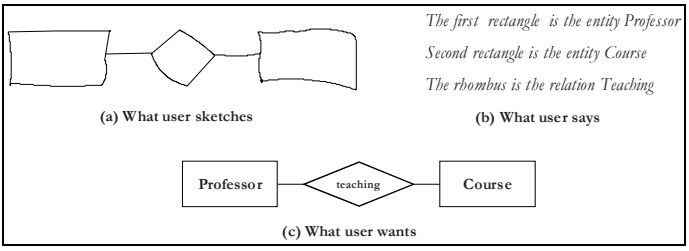


Fig. 1. An example of multimodal input

If the figure is properly interpreted by the sketch recognizer as a rhombus (Fig. 2a), we are in the typical situation of redundant information given by different modalities (i.e. the concept "rhombus" is expressed both by speech and sketch modalities) The Natural Language sentence associated to the multimodal one will be: "The rhombus is the relation Teaching" and it represents all the concepts expressed by speech and sketch.

Given such a sentence, a corresponding template to the sentence in the knowledge base, must be found.

The situation is more complex if we have some ambiguities problem. That is for example, if a user draws a figure that in his/her intention is a rhombus, recognized by the system as a different figure (Fig. 2b). So the system is unable to identify that the given input is redundant and can produce an incorrect multimodal sentence interpretation. In Fig. 2 both situations are illustrated with the proper timeline.

From the user point of view both situations must reproduce the same multimodal sentence, but due to ambiguities in sketch recognition the system does not produce the same sentences and they are not associated to the same template. In order to avoid these problems we propose an approach that take into account the user's behavior to reproduce multimodal sentences as near as possible to the user's input will.

3 Evaluating Multimodal Sentence Similarity

The first important problem to solve is to understand how different modalities cooperate and what is the template that the multimodal sentence matches according to the cooperation modality.

Six type of cooperation between modalities have been distinguished (see [7]): *Complementarity*, *Concurrency*, *Equivalence*, *Redundancy*, *Specialization* and *Transfer*.

Let us consider the interaction of two modalities (in particular we address speech and sketch modalities) M_1 and M_2 that transmit information in ΔT_1 and ΔT_2 time intervals respectively. The possible time intervals relationship between M_1 and M_2 are summarized in Fig. 3 and are:

- *Sequential*: The transmission of the second modality starts after the first one.
- *Disjoint*: The transmissions of the two modalities take place in two separated time intervals.

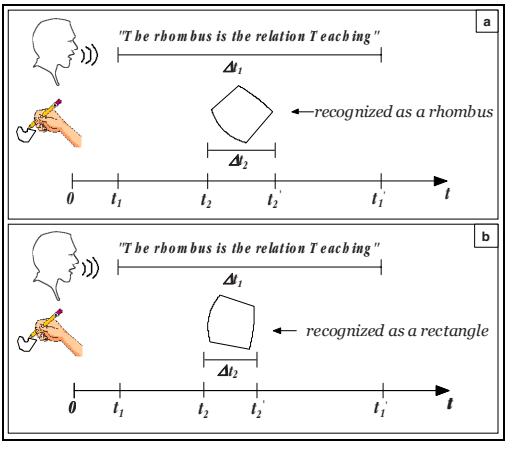


Fig. 2. Example of multimodal input (by speech and sketch) with timeline

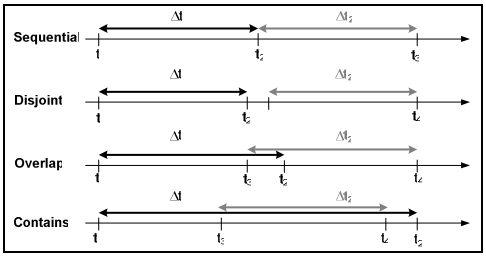


Fig. 3. Time intervals relationship in multimodal interaction

– *Overlap*: The transmission of a modality partially overlaps the transmission of the other one.
– *Contains*: The transmission of a modality is self contained in the transmission of the other one.

Evaluating time intervals for the involved transmission we analyze the possible combination of different input events. Multimodal input events can either be interpreted independently, or they can be merged.

Let us refer to the example given in the previous section. Firstly the system individually recognizes the concepts or the chunk of information for each modality involved in the input event. In this case the system must perform a speech and a sketch recognition in order to capture the initial information. Then each input modality is associated with its own time interval of transmission. The next step is to interpret these unimodal input on the base of the transmission time and the information recognized in order to extract a multimodal sentence representing the whole input event.

For our purposes, it is important to address two type of cooperation: Complementarity and Redundancy. Considering the aforementioned time intervals, Sequential, Overlap and Disjoint transmissions can denote Complementarity

interaction as Contains and Overlap transmissions can denote Redundancy interaction as previously stated in [7]. In Fig. 4 this relationships are presented.

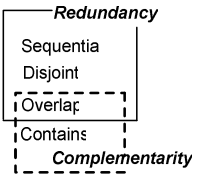


Fig. 4. Time intervals and type of interaction

In this paper we address redundancy, because it has been observed that a redundant multimodal input involving speech and sketch enables a more natural form of communication.

As stated before each modality transmission represents a finite number n of concepts.

Let us suppose that the modality M_1 transmits the concept C_1 in the time ΔT_1 and the modality M_2 transmits the concept C_2 in the time ΔT_2. ΔT_1 partially overlaps time ΔT_2. In our example M_1 is the speech modality and M_2 is the sketch modality. We have to identify the template for the multimodal sentence in order to correctly interpret it. That is, this "match" requires interaction between modalities has to be recognized. In redundant interaction C_1 and C_2 must be the same concept. However, the modality used to express one concept (for example concept C_1) can introduce some imprecision and approximations. This is the situation in which the user would draw a rhombus, but the system recognizes a rectangle due to the drawing imprecision. At the same time the concept expressed by speech mode is "the rhombus". In this case the two concepts C_1 and C_2 are similar concepts. The question is: what similarity measure has to be considered between two concepts in order to have a redundant interaction between modalities?

This measure car vary among different users. The system has to acquire knowledge on the user's behavior during the whole interaction process.

We have considered a sample of 20 different users, and we asked them to draw a rhombus, alternatively with other figures, using our sketch interface, for 20 times. This has permitted us to take into account the user's behavior, that is used to better calculate the similarity between concepts.

The sketch recognizer identifies, for each sample, a measure of similarity between user's sketch and the required figure (in this case a rhombus).

This measure represents a fundamental element for the computation of concept similarity. More formally, if C_i is the *i*-th sample drawn by the user, and n is the total number of the samples, the user behaviour approximation A_{ss} (for sketch-speech modalities interaction) is:

$$A_{SS} = \frac{1}{n}\sum_{i=1}^{n} C_i \,, \qquad A_{ss} \in (0,1) \,. \tag{1}$$

The concepts are stored in the concepts database as natural language terms. We adopt an extended words similarity algorithm in order to calculate semantic similarity between concepts using the WordNet lexical database [8].

3.1 Evaluating Concepts Similarity

The taxonomical structure of the WordNet knowledge base is important in determining the semantic distance between words. In WordNet, terms are organized into synonym sets (synsets), with semantics and relation pointers to other synsets.

One direct method for similarity computation is to find the minimum length of path connecting the two words [9].

However, this method may be not sufficiently accurate if it is applied to a large and general semantic net such as WordNet. To address this weakness, it is important to notice that concepts at upper layers of the WordNet's hierarchy have more general semantics and less similarity between them, while words that appear at lower layers have more concrete semantics and have a higher similarity. Therefore, also the depth of word in the hierarchy should be considered. In summary, we note that similarity between words is determined not only by path lengths but also by depth (level in the hierarchy). Moreover we take into account the user behaviour considering the user behaviour approximation Ass. The proposed algorithm is an extensions of the one proposed in [10] for words similarity.

Given two concepts, c1 and c2, we need to find the semantic similarity s(c1,c2).

Let be l the shortest path length between c1 and c2, and h the depth of subsumer in the hierarchical semantic nets, the semantic similarity can be written as:

$$s(w_1, w_2) = f_1(l) \cdot f_2(h) \cdot B_{ss} \,. \tag{2}$$

B_{ss} represents the user's behaviour contribution and its value depends on the value of the product between $f_1(l)$ and $f_2(h)$. If this value is higher than a threshold value (δ), B_{ss} is equal to $(A_{ss})^{-1}$, otherwise B_{ss} is equal to 0. More formally:

$$B_{ss} = \begin{cases} 0 & \text{if} \quad f_1(l) \cdot f_2(h) < \delta \\ \dfrac{1}{A_{SS}} & \text{if} \quad f_1(l) \cdot f_2(h) \geq \delta \end{cases} . \tag{3}$$

The path length between two concepts, c_1 and c_2, can be computed according to one of the following cases:

1. c_1 and c_2 belong to the same synset,
2. c_1 and c_2 do not belong to the same synset, but their synsets contains one or more common words,
3. c_1 and c_2 neither belong to the same synset nor their synsets contain any common word.

By the above considerations we can considered the function $f_1(l)$ to be a monotonically decreasing function of path l.

For depth contribution, function $f_2(h)$ should be a monotonically increasing function with respect to depth h.

The whole formula for words similarity computation is:

$$s(w_1, w_2) = e^{-\alpha l} \cdot \frac{e^{\beta h} - e^{-\beta h}}{e^{\beta h} + e^{-\beta h}} \cdot B_{ss} \,. \tag{4}$$

The values of α and β depend on the used knowledge base. For WordNet the optimal parameters are α=0.2 and β=0.45 as explained in [11].

3.2 Evaluating Multimodal Sentence Similarity

The interpretation of a multimodal sentence needs its corresponding sentence, expressed in natural language form, matches with one template contained in the knowledge base. An approximate interpretation, which uses a semantic approach, can be provided.

As discussed in the introduction, the need of interpreting sentences characterizing the multimodal dialog has led us to propose a process, devoted to obtain the exact interpretation or its approximation. The steps of our algorithm for evaluating multimodal sentence similarity can be summarized as follows:

- the user formulates his/her multimodal sentence,
- the resulting sentence is transformed in a Natural Language (NL) one and its template is compared with the templates sharing the same keywords (this reduce the number of the compared templates); the corresponding templates are selected,
- if there are no matches, we extract from each one of the selected templates the relative sentences that have been used to create (by example) it,
- then the system computes the semantic similarity between each of these sentences and the user's sentence. Indeed sentence similarity is calculated through computing similarity between templates of each sentence. After investigating a number of methods, we proposed a templates similarity measure following the approach reported in [10].
- the highest value of semantic similarity is used to choose the more similar template with the given sentence.

4 Conclusions

Multimodal human-computer interaction, in which the computer accepts input from multiple channels or modalities, is more flexible, natural, and powerful than unimodal interaction with input from a single modality. However naturalness of communication is directly proportional with the complexity of the interpretation of messages. In this paper is presented an approach to define how a multimodal sentence matches to a sentence's template according to: 1) the different type of cooperation between modalities; 2) the similarity approximations given by the user's behaviour in his/her multimodal input. We have implemented a multimodal system based on speech and sketch modalities that uses the aforementioned methodology to better understand user's input, also considering the user's way of inputting. Future work will address other input modalities, such as gesture, that could also help disambiguate the user's interaction behaviour.

References

1. Binot, J.L., Falzon, P., Perez, R., Peroche, B., Sheehy, N., Rouault, J., Wilson, M.D.: Architecture of a multimodal dialogue interface for knowledge-based systems. In: Proceedings of Esprit'90 Conference, pp. 412–433. Kluwer Academic Publishers, Dordrecht (1990)
2. Johnston, M., Cohen, P.R., McGee, D., Oviatt, S.L., Pittman, J.A.: Unification-based Multimodal Integration. In: Proceedings of the ACL, Madrid (1997)
3. Wahlster, W.: User and discourse models for multimodal communication. In: Sullivan, J., Tyler, S. (eds.). Intelligent User Interfaces. ACM Press, Addison-Wesley (1991)
4. Caschera, M.C., Ferri, F., Grifoni, P.: Multimodal interaction systems: information and time features. Journal of Web and Grid Services (to appear)
5. Bottoni, P., Costabile, M.F., Levialdi, S., Mussio, P.: Formalizing visual languages. VL 1995, 45–52 (1995)
6. Celentano, A., Fogli, D., Mussio, P., Pittarello, F.: Model-based Specification of Virtual Interaction Environments. In: Proceedings of the 2004 IEEE Symposium on Visual Languages - Human Centric Computing (VLHCC'04) - vol. 00, pp. 257–260 (2004)
7. Martin, J.C.: Toward intelligent cooperation between modalities: the example of a system enabling multimodal interaction with a map. In: Proceedings of Int. Conf. on Artificial Intelligence (IJCAI'97) Workshop on Intelligent Multimodal Systems, Nagoya, Japan (1997)
8. WordNet 2.1: A lexical database for the English language (2005), http://www.cogsci.princeton.edu/cgi-bin/webwn
9. Rada, R., Mili, H., Bichnell, E., Blettner, M.: Development and Application of a Metric on Semantic Nets. IEEE Trans. System, Man, and Cybernetics 9(1), 17–30 (1989)
10. Li, Y., McLean, D., Bandar, Z.A., O'Shea, J.D., Crockett, K.: Sentence Similarity Based on Semantic Nets and Corpus Statistics. IEEE Transactions on knowledge and data Engineering 18(8), 1138–1150 (2006)
11. Li, Y.H., Bandar, Z., McLean, D.: An Approach for Measuring Semantic Similarity Using Multiple Information Sources. IEEE Trans. Knowledge and Data Eng. 15(4), 871–882 (2003)

Reconsideration of the Effectiveness on Extracting Computer Diagnostic Rules by Automatically Defined Groups

Yoshiaki Kurosawa, Akira Hara, Kazuya Mera, and Takumi Ichimura

Graduate School of Information Sciences, Hiroshima City University,
3-4-1, Ozuka-higashi, Asaminami-ku, Hiroshima, Japan
{kurosawa,ahara,mera,ichimura}@its.hiroshima-cu.ac.jp

Abstract. Our aim is to manage computer systems without expert knowledge. We have proposed a method of diagnostic rule extraction from log files by using Automatically Defined Groups (ADG) based on Genetic Programming. However, this work less explained the effectiveness, especially, the characteristics of the acquired rules. Therefore, we re-evaluated the effectiveness by performing two experiments: the use of artificial log files and the use of real log files. As a result, we confirmed that ADG could acquire the rules composed of multiple terms. This characteristic is very important because we can judge the message that we must consider the co-occurrence of the words, i.e. 'Error' and 'not'. Thus, we conclude that the ADG is effective for the diagnosis of the systems.

Keywords: Genetic Programming, Rule Extraction, Data Mining.

1 Introduction

Recently, computer systems have increased in size. With this increase of them connecting with Local Area Network (LAN), many TCP/IP services such as http service have been offered for the system users. The administration of the systems, therefore, have become more complex and difficult to detect some errors.

Several studies have been made to automatically analyze system log files in order to decrease such administrators' duty (cf. [1,2,3]). However, the main purpose of these studies was to diagnose the problem of the systems, by focusing of only one service (http), file (/var/log/messages), or protocol (simple network management protocol), based on some expert knowledge that the researchers manually analyzed in detail. Certainly, this aim is useful, but the real purpose of the administrators is not to focus on only one service but to manage all their systems including many personal computers (PC), devices, and so on.

Based on the aim, we extended [4] an Automatically Defined Groups (ADG) based on Genetic Programming (GP) [5,6,7]. However, it was not clear whether this extended ADG could acquire rules composed of multiple terms. Therefore, we re-evaluated the effectiveness focusing on these points; 'Can acquire multiple rules?', 'Can the rules be ordered by the supported agents?', and 'Can the rules consist of multiple terms?'.

B. Apolloni et al. (Eds.): KES 2007/ WIRN 2007, Part II, LNAI 4693, pp. 411–418, 2007.
© Springer-Verlag Berlin Heidelberg 2007

```
Example 1   Oct 11 22:14:15 mymachine su: 'su root' failed for lonvick on /dev/pts/8
Example 2   "Use the BFG!"
Example 3   Aug 24 05:34:00 CST 1987 mymachine myproc[10]: %% It's time to make the do-nuts.
            %% Ingredients: Mix=OK, Jelly=OK # Devices: Mixer=OK, Jelly_Injector=OK, Frier=OK
            # Transport: Conveyer1=OK, Conveyer2=OK # %%
Example 4   1990 Oct 22 10:52:01 TZ-6 scapegoat.dmz.example.org 10.1.2.3 sched[0]: That's All Folks!
```

Fig. 1. Sample Messages in RFC3164

```
< Common Log Format >
    127.0.0.1 - frank [10/Oct/2000:13:55:36 -0700] "GET /apache_pb.gif HTTP/1.0" 200 2326

< Combined Log Format >
    127.0.0.1 - frank [10/Oct/2000:13:55:36 -0700] "GET /apache_pb.gif HTTP/1.0" 200 2326
    "http://www.example.com/start.html" "Mozilla/4.08 [en] (Win98; I ;Nav)"

< Extended Log File Format >
    00:34:23 GET /foo/bar.html
```

Fig. 2. Sample HTML Messages

2 Descriptions of System Log and Its Analysis

We do not simply deal with system log files because there are various formats. For example, RFC3164 proposed a syslog protocol [8]. According to the document, system log files describe messages generated by certain events on various types of sender such as computer. No concrete detail is described as this definition mentions, however. Fig.1 shows four valid sample logs described in RFC 3164. Moreover, we further exemplify the difference of the format; see Fig.2.

By comparing these samples, log files have no unified format to describe some events.We may automatically analyze even these types of format by using detailed and appropriate knowledge. However, such ambiguous format must make administrators difficult to maintain the systems in case of lack of appropriate knowledge. In next section, we will discuss this type of difficulty.

2.1 Problems as to Log Files When Managing Systems

According to the paper previosly mentioned [4], when administrators check their systems whether they work well or not, the administrators may be confronted with the following problems at least; 'lack of knowledge on system', 'enormous data size', and 'temporary lack of knowledge by change of some kind'. To solve them, simple translation method based on regular expressions was proposed.

In [4], some tags such as <DATE> and <TIME> were given to some fields because the fields can be easily understood. Furthermore, the other types of tags were given by using existing databases in the system. For example, tags such as <HOST> and <REMOTEHOST> relate to hostname. They can be distinguished each other by using appropriate commands related to the system database, e.g., "ifconfig." In the same way, by using the other command "id" or "groups," we can distinguish 'username' or 'groupname' with words appeared in the log messages. If the system can not understand the content of the fields, it gave <EXP> tag including explanation of some sort. Thus, their procedure is simple because of the realization without knowledge.

3 Extracting Rules Using Automatically Defined Groups

3.1 Concept of Automatically Defined Groups

In this section, we introduce the ADG method [4,5,6,7]. The method is based on evolutionary computing and can optimize both the grouping of agents and the program of each group in the process of evolution. By grouping multiple agents,

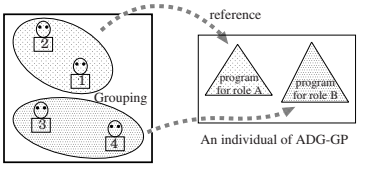


Fig. 3. Concept of ADG

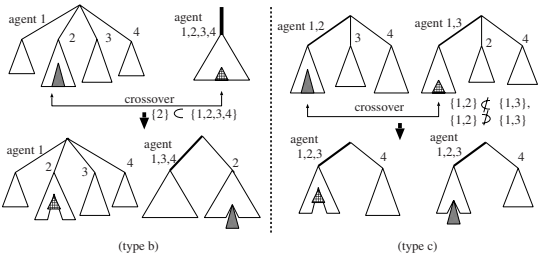


Fig. 4. Examples of crossover

we can easily analyze the behavior of agents. The acquired group structure is utilized for understanding how many roles are needed and which agents have the same role. That is, the following three points are automatically acquired by using ADG. (1)How many groups are required to solve the problem? , (2)Which group does each agent belong to? , and (3)What is the program of each group?

A team that consists of all agents is regarded as one GP individual. One GP individual maintains multiple trees, each of which functions as a specialized program for a distinct group as shown in Fig. 3. We define a group as the set of agents referring to the same tree for the determination of their actions. All agents belonging to the same group use the same program.

Generating an initial population, agents in each GP individual are divided into groups at random. Crossover operations are restricted to corresponding tree pairs. For example, a tree referred to by an agent 1 in a team breeds with a tree referred to by an agent 1 in another team. However, we consider the sets of agents that refer to the trees used for the crossover. The group structure is optimized by dividing or unifying the groups according to the relationship of the sets. The crossover operations are shown in Fig. 4. In addition, the (type a) is ommited because the structure of each individual is unchanged.

3.2 Extracting Multiple Rules Using ADG

The ADG method can be utilized as a rule extraction method for classifying positive and negative cases in database. Respective trees in an individual of ADG represent the logical expressions, and return true or false for each data. If one or more trees return true for an input data, the data is regarded as a positive case. The fitness f is calculated by the following equation (1).

$$f = \frac{H_{Abnormal}/L_{Abnormal}}{H_{Normal}/L_{Normal}} - \beta \frac{\sum_{N_{Normal}} fault_agent}{H_{Normal} \times N_{agent}} - \delta\ V_w \qquad (1)$$

In this equation, L means the number of line in log files. Thus, $L_{Abnormal}$ and L_{Normal} means the number of line in the abnormal/normal state respectively. H means that one or more trees in an individual return true for a log message. When the rule returns true for data in the normal state log file, $fault_agent$, that is, the number of agents who support the wrong rule, was counted. Thus,

the third term represents the average rate of agents who support the wrong rules when its misrecognition happens. Moreover, V_w is the variance of every agent's load, which is introduced from the viewpoint of load balancing among agents.

The concept of each agent's load arises from the viewpoint of cooperative problem solving by multiple agents. The load is calculated from the adopted frequency of each group's rule and the number of agents in each group. The adopted frequency of each rule is counted when the rule returns true to the messages in abnormal state log. If multiple trees return true for a message, the frequency of the tree with more agents is counted. When the agent a belongs to the group g, the load of the agent w_a is defined as follows:

$$w_a = f_g / n^g_{agent} \tag{2}$$

where n^g_{agent} represents the number of agents which belong the group g, and f_g represents the adopted frequency of g.

In addition, in order to inhibit the redundant division of groups, f is multiplied by γ^{G-1} according to the increase of the number of groups, G, in the individual. γ is a penalty coefficient on the number of groups.

Last, we explain their logical tree in their paper.

(`include <SORT> unexpected`) $\wedge$ (`include <EXP> warning`)

The second argument of each term such as <SORT> and <EXP> represents the attached tag. If the message enclosed in the tag includes the word specified by the third argument, this expression returns true. In addition, the logical expression has to return false for any logs in normal state.

3.3 System State Pattern for Supervised Information

In general, when machine learning method is performed, some sort of supervised information is needed. However, using 'System State Pattern (SSP)', we can deal with log data without supervised information [4]. The SSP means a kind of state transition of the system: from normal state to abnormal one, or from abnormal state to normal one. By comparing both states, we may find some differences. That is to say, this SSP approach is based on the heuristics of an idea that error messages will appear in the only abnormal log, and is adopt to extract rules without supervised information.

4 Experiments

We perform two experiments to achieve the aim mentioned above. One is an experiment (Experiment 1) on the basis of artificial log files, which is performed to check whether the ADG can extract rules composed of multiple terms. The other is an experiment (Experiment 2) by adopting log files stored from actual computers.

4.1 Experiment 1: Rule Extraction from Artificial Logs

In this paper, the artificial log files are made by the following procedure. At first, we assume the situation where we must manage five hosts having the same two log files. Each file includes 300 messages, that is, 300 lines. In addition, these 10 log files (two files by five hosts) are exactly the same. Each message consists of five kinds of tag-and-word set; <LOGNAME>, <TIME>, <HOST>, <REMHOST>, and <EXP>. The number of word list are 1, 4, 5, 5, and 17, and each word is arbitrarily selected from actual log files except for <HOST> and <REMHOST>.

Then, one of two files mentioned above is treated as normal, and the other is treated as abnormal. At this point, we change some tag-and-word sets only in the certain abnormal log files according to the following two steps. That is, this is the case that system errors occur on purpose.

Step 1. We arbitrarily select two hosts, i.e. host1 and host2. Then, we can change some words including a tag <REMHOST> only in the abnormal log files of the selected hosts. For example, we replace the word 'host1' and 'host2' with 'host5' including the tag. These replacements means that some errors occur when sending data from 'host1' to 'host5' or sending data from 'host2' to 'host5'. In addition, we change the tag once regarding 'host1', and three times regarding 'host2'. Thus, this means that the number of the error related to 'host1' is three times larger than the one related to 'host2'.

Step 2. We replace an arbitrary word in the <EXP> with the same word mentioned at (Step 1), 'host5', in both normal/abnormal files. In this paper, we perform this replacement five times. The replacement makes it difficult that we find errors by focusing on the word frequency because the files are almost equivalent.

Therefore, these log files are assumed that there are five hosts, and two of them encounter certain accidents.

We applied ADG to these artficial log message data. In this paper, GP functions and terminals are shown in Table 1.

Fig. 5 illustrates the change of the average number of groups. Because each group refer to one diagnostic rule, the number of groups corresponds to the number of the extracted rules. These agents belonging to approximately four groups mainly seek four rules for detecting system errors. Particularly in the best individual, two rules were extracted. The two extracted rules are shown in Fig. 7.

The first two lines means a rule to check whether <HOST> includes 'host1' and <REMHOST> includes 'host5', respectively. We clearly understand that this extracted IF-THEN rule can detect the error related to 'host5', made by the two steps mentioned above. In addition, these two terms are minimum, and this rule composed of the terms is the best one in order to detect the errors. If this type of error should not be detect by focusing on the frequency of the host 'host5' because this host name is even distribution in the log files. On the other hand, our proposed method can detect this type of error. Therefore, we

Table 1. GP Functions and Terminals

symbol	#args	functions
`and`	2	arg0 ∧ arg1
`include`	2	if arg0 (Tag) includes arg1 (Word), return T else return F
`<HOST>,<EXP>, ...`	0	Tag name
`0,... ,N-1`	0	selected number in word list

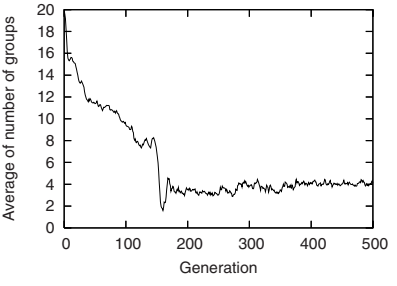


Fig. 5. Change of the average of the number of groups : Experiment 1

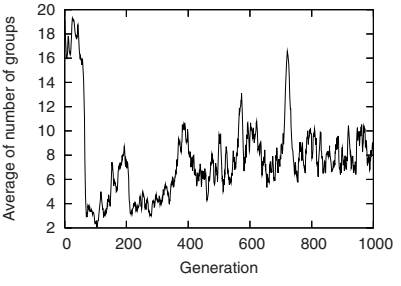


Fig. 6. Change of the average of the number of groups : Experiment 2

conclude that the method is effective to extract diagnostic rules from log files even including co-occurrence relations such as "error" and "not find."

Next, the second rule in Fig. 7 is complicated because the rule consists of many terms, and the same term exists in the rule: "(`include <HOST> host2`)." This rule, in a sense, is redundant because this type of error can be detect by only two terms as we previously mentioned. However, no misdetection could occur by using this rule. Therefore, all these terms work to identify the messages only in the abnormal message.

Last, we take the number of agents which support the rules into account. The support number of the first rule is 50, and the one of the second rule is 150. This means that all 200 agents is appropriately divided into two groups according to the error rate between 'host1' and 'host2' mentioned at Step 1.

4.2 Experiment 2: Rule Extraction from Actual System Logs

In Experiment 2, we apply ADG to the actual log message data. Our 322 target files, such as syslog files and html logs, were collected on a server. These files consisted of 90,762 normal state lines, and 88,336 abnormal state lines. The files differed from the ones collected in our previous work [4]. In order to describe these log files, we needed 25 tags as well as the previous work. However, maximum word list size N was 935, and the word list with max size belonged to the <EXP> tag. The respective weights in equation (1) and the other parameters were the same in Experiment 1.

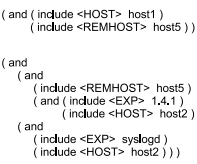


Fig. 7. Extracted Rules: Ex.1

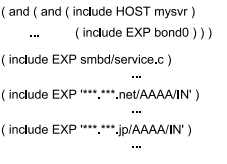


Fig. 8. Extracted Rules: Ex.2

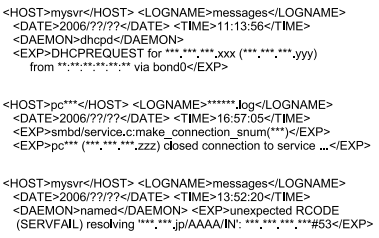


Fig. 9. Extracted Logs: Ex.2

Fig. 6 illustrates the change of the average number of groups[1]. These agents in each individual seem to be divided into approximately eight groups although these numbers fluctuate greatly. In the best individual, 200 agents were divided into 17 groups, that is, 17 rules were extracted. By using these rules, we tried to detect errors from the same log files, the abnormal state files, which were used when extracting the rules, that is, as closed testing procedure. As a result, we can detect 16,877 messages.

Fig. 8 shows a part of the acquired rules that correspond to the tree structured presentation in the best individual. These rules are arranged according to the number of agents. This figure illustrates the effectiveness of the ADG method because the first rule consists of multiple terms. In addition, this case has something to do with the 'DHCP' as shown in Fig. 9.

Both the third rule and fourth rule in Fig. 9 include "***.***.***/AAAA/IN," related to 'host name' in the tag <EXP>. This description is not to easy to understood at the first glance. However, by referring the description from the log files, we understand that the words have something to do with the DNS error because the message includes "unexpected RCODE(SERVFAIL)."

4.3 The Effectiveness of the ADG

We confirmed the following points from our experiments: (a)we could acquire multiple rules, (b)the rules could be ordered by the number of agents, and, (c)the rules could consist of multiple terms from both the artificial data and the real data.

These points are the ones which previous work less explained [7]. Especially, the third point is very important because the real log files include the message such as "Error was not found" and we can not detect whether the message is normal or abnormal without focusing on the co-occurrence of the word such as 'error' and 'not'. Therefore, the ADG method is effective in such case.

5 Conclusions and Future Work

We re-evaluated the ADG on the basis of evolutionary computation. As a result, we found 17 rules and one of them had two terms in Experiment 2. That is,

[1] We selected 1000 generations to unify the experimental condition in the paper [9].

we confirmed that we could extract multiple rules composed of multiple terms. Moreover, It is important that we could perform the extraction without any expert knowledge and supervised information compared with [1,2,3].

However, the optimization may not be sufficient in the real data as shown in Fig.6 because the number of tags (25) is relatively large and the number of terminals (935 words) is considerably large. Thus, the optimization such as the decrease of tags and terminals is planned for one of our future works.

In addition, we did not apply any natural language processing, e.g., morphological analysis, to log files. However, the files include various messages written by natural language, especially Japanese. Thus, we will adopt natural language processing techniques (i.e. [10,11]) as preprocessing, and apply to GP.

References

1. Andrews, J.H.: Theory and practice of log file analysis. Technical Report 524, Department of Computer Science, University of Western Ontario (1998)
2. Cooley, R., Mobasher, B., Srivastava, J.: Data preparation for mining world wide web browsing patterns. Knowledge and Information Systems 1(1), 5–32 (1999)
3. Büchner, A.G., Baumgarten, M., Anand, S.S., Mulvenna, M.D., Hughes, J.G.: Navigation pattern discovery from internet data. In: Proc. of the Web Usage Analysis and User Profiling Workshop, pp. 25–30 (1999)
4. Kurosawa, Y., Hara, A., Ichimura, T., Kawano, Y.: Extraction of Error Detection Rules without Supervised Information from Log Files Using Automatically Defined Groups. In: Proc. of the IEEE International Conference on Systems, Man and Cybernetics (SMC2006), pp. 5314–5319 (2006)
5. Hara, A., Ichimura, T., Yoshida, K.: Discovering Multiple Diagnostic Rules from Coronary Heart Disease Database Using Automatically Defined Groups. International Journal of Manufacturing 16(6), 645–661 (2005)
6. Hara, A., Ichimura, T., Takahama, T., Isomichi, Y.: Discovery of Cluster Structure and The Clustering Rules from Medical Database Using ADG; Automatically Defined Groups. In: Ichimura, T., Yoshida, K. (eds.) Knowledge-Based Intelligent Systems for Healthcare, pp. 51–86 (2004)
7. Hara, A., Nagao, T.: Construction and analysis of stock market model using ADG; Automatically Defined Groups. International Journal of Computational Intelligence and Applications (IJCIA) 2(4), 433–446 (2002)
8. Lonvick, C.: The BSD Syslog Protocol, RFC3164 (August 2001)
9. Kurosawa, Y., Hara, A., Ichimura, T.: Preprocessing techniques for extracting computer diagnostic rules by ADG. In: Proc. of the IEEE Three-Rivers Workshop on Soft Computing in Industrial Applications (SMCia2007), IEEE Computer Society Press, Los Alamitos (to appear, 2007)
10. Kurosawa, Y., et al.: A description method of syntactic rules on filmscripts. Journal of Natural Language Processing (in Japanese) 12(6), 25–62 (2005)
11. Mera, K., Kurosawa, Y., Ichimura, T.: Emotion Oriented Interaction system for Elderly People. In: Ichimura, T., Yoshida, K. (eds.) Knowledge Based Intelligent Systems for Health Care, Advanced Knowledge International (2004)

Adaptive Mechanisms for Classification Problems with Drifting Data

Zoheir Sahel[1], Abdelhamid Bouchachia[2], Bogdan Gabrys[1], and Paul Rogers[1]

[1] School of Design, Engineering and Computing, Bournemouth University, UK
{zsahel,bgabrys,progers}@bournemouth.ac.uk
[2] Dept. of Informatics, University of Klagenfurt, Austria
hamid@isys.uni-klu.ac.at

Abstract. Most work on supervised learning is undertaken on static problems. However, in many real world classification problems, the environment in which the classifiers operate is dynamic i.e. the descriptions of classes change with time. In this paper, the process of generating drifting data is introduced in order to assess two adaptive approaches that deal with dynamically changing data. These approaches are: retraining on evolving data set and incremental learning. The empirical evaluation has shown that both these approaches improve the performance compared to the non-adaptive mode though a number of outstanding research issues remain.

Keywords: Concept drift, adaptive classification, Incremental learning.

1 Introduction

Rapid development in computer and sensor technology has resulted in an increasing amount of data that is easily captured and stored. Examples of this data may include medical history records in health care, financial data in banking to list only a few. Inferring knowledge from such data using robust methods and systems became also possible with the increasing availability of computer power and storage medium at affordable prices. Pattern classification systems are amongst these methods receiving more research efforts. However, often the assumption made about the environment, is that it is static and the classification problem remains unchanged during time of operation. This fundamental assumption does not always hold, in fact it is often not the case when classifiers are operated on large amount of data produced potentially in real time. That is to say that during training, it is not always possible to have a representative training set of all possible classes. The result of not adapting a classifier model after deploying it is performance deterioration[1][2].

In order to address the problem of adaptation in classification systems, many adaptive methods have been proposed. One class of such methods is classifier retraining based on an evolving dataset which results in classifier forgetting old observations. Retraining classifiers in this way may result in maintaining their performance. In practice, it could be costly in terms of time which makes it unappropriate for applications where time is an issue. Furthermore, the evolving

B. Apolloni et al. (Eds.): KES 2007/ WIRN 2007, Part II, LNAI 4693, pp. 419–426, 2007.
© Springer-Verlag Berlin Heidelberg 2007

training set needs to be stored and the memory available varies from one application to another. Three primary strategies in evolving a training set can be distinguished: retraining using a fixed size window[3], using an adaptive window size[3] and density based training set[4].Having a large training set may result in a slow adaptive classifier not following the change of data whereas a small training set will result in a highly responsive classifier omitting past but important observations. Achieving a balance between these two extremes remains an open research question and it is referred to by the stability/plasticity dilemma [5],[6],[7]. Another adaptive mechanism, is online incremental learning[8],[9] whereby the classifier learns incrementally from data while in operation. The risk associated with updating the classifier in this manner is the continuous use of potentially miss-classified data in adjusting the model. In addition, incremental learning classifiers are sensitive to data ordering and may result in totally different models using the same data presented in different order[9].

It is also noticed in the literature that no comprehensive dynamically changing datasets are publicly available as a benchmark for the research community such as those static data available at the UCI repository[10]. Therefore the main goal of the research work reported in this paper was to provide a comparative analysis of the existing primary mechanisms for adaptation of classifiers within a controlled environment where the characteristics of the change i.e. data drift are known. The rest of this paper is organized as follow. In section 2, a systematic drift simulation is explained and applied to a selection of UCI static datasets. In section 3, we briefly describe selected incremental learning classifiers and standard classifiers made adaptive based on retraining using an evolving dataset. Section 4 covers the results of testing and comparison of the selected classifiers on the original and drifting data. Section 5 is a conclusion.

2 The Process of Generating Drifting Data

Although most real world data are subject to change over time, datasets simulating one aspect or more of the changing environment are not widely available as static data. No repository such as the UCI[10] is publicly available for the research community working on adaptive classifiers for algorithm comparison. Consequently, any work in this domain relies on privately gathered or simulated data. In this paper, we explain how we generated drifting data. Our aim is to create a controlled environment in order to tune the change and study its effect on classifiers performance. The reason for doing so and not using a real world dataset is because often the change is unknown in such data and sometimes it incorporates more than one type of change.

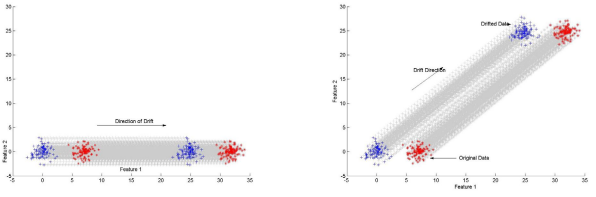
Having the original datasets, *Original*, the drifting data will be generated using eq. 1:

$$New(n,j) = Original(i,j) + \left(\frac{n}{O} * Rate(j)\right) + (0.1 * Noise(j)) \qquad (1)$$

where, $j = 1 \cdots F; n = 1 \cdots N$, F is the number of features, N is the number of new data points, O is the number of original data points (size of the static data

Table 1. Datasets used for dynamic data generation

Dataset	Features	Classes	Number of Instances
Artificial	2	2	200 [100 100]
Mixture of Gaussians	2	2	250 [125 125]
Iris	4	3	150 [50 50 50]
Wine	13	3	178 [59 71 48]



(a) Drift geared by 1 feature

(b) Drift geared by 2 features

Fig. 1. Artificial data drift

set), New is an $(N \times F)$ matrix of the new data points, $Original$ is an $(O \times F)$ matrix of the static data, $i = mod(n-1, O)+1$ is the index of the original point, $Rate$ is a $(1 \times F)$ vector representing the drift rate per feature and $Noise$ is a $(1 \times F)$ vector representing a random noise for each feature (generated from a uniform distribution), i.e., $Noise = rand(1, F)$.

Table 1 depicts the characteristics of the static datasets selected for our initial experiments. Figure 1 illustrates drifting data for artificial dataset.

3 On-Line Adaptive Mechanisms

In the literature, classifiers can be categorised as off-line or on-line classifiers. Off-line classifiers are trained using a set of representative data in an off-line manner. They update their model after visiting all data and often require to see the training data more than once. After that, they are deployed to classify unseen data with no further learning.

On-line incremental classifiers however, possess the ability to learn from each pattern as it is presented and do not necessarily require to revisit the data. Such incremental classifiers are trained in this manner but when deployed, they very often do not use explicitly any adaptive mechanism to update their model. In this paper, we use two adaptive strategies: (1) retraining using an evolving dataset which can be applied potentially to any classifier and (2) on-line incremental learning.

3.1 Re-training Classifiers Using Evolving Dataset

This strategy is based on retraining a classifier after processing each or a number of patterns on an evolved training set. In other words, the training set is updated by including recent patterns and excluding old and thus assumed not relevant

patterns. The new training set is used for retraining the classifier. There are three main approaches in evolving the training set: (1) Forgetting at a constant rate, (2) Forgetting at a variable rate, (3) Density based forgetting [4].

Forgetting at a constant rate involves obtaining the training data from a sliding window. After each new pattern presentation, this pattern is included in the training set for the next retraining step and the oldest observed pattern is excluded from the window. This method is usually too simple for handling the challenges presented by complex changing environment, however, it is appropriate when the change is systematic. We will verify the appropriateness of such forgetting at a constant rate for the drift scheme on which we focus in this paper.

Forgetting at a variable rate is also achieved by obtaining the training data from a type of moving window, but the difference from forgetting at a constant rate is the adaptability of the window size. It is expanded if the change is slow and shrinks if the change is faster. The challenge however is to obtain a highly adaptable classifier while maintaining its stability. This is still an open research problem with no obvious, uniformly good solution proposed.

An interesting alternative to the above moving window approaches is a density-adaptive reinforcement learning and forgetting algorithm proposed in [4]. The described method used hybrid decision trees for a variable resolution partitioning and labelling of the input space. The trees were generated on the basis of an evolving set of exemplars. What is interesting in this machine learning approach is the use of a density adaptive forgetting mechanism which deletes observations from the learning set depending on whether subsequent evidence is available in the local region of the parameter space. This allows for a regular adaptation of the regions where the data occurs frequently while not removing the data representing rare events.

Algorithm 1 describes the procedure of retraining an offline non-adaptive classifier using an evolving drifting dataset approach with a forgetting at a constant rate strategy. In the current study, we use a set of selected classifiers available within the PRTools package [11]. The algorithms are the(1) linear(LDC), (2) quadratic (QDC), (3) nearest neighbour (KNNC), (4) parzen (PARZENC), (5) feed forward neural network (BPXNC), (6) the support vector machine (SVC)

Algorithm 1. Testing classifiers based on re-training

Let

- O be the size of original dataset
- F be the number of features in the original dataset
- N be the number of new points to be generated
- M be the window size specified by the user
- $Original$ be an $O \times F$ matrix representing the original data points
- $Classed$ be an $(O + N) \times F$ matrix representing the drifting data as classified by a classifier $\mathcal{W}$
- New be an $(O + N) \times F$ matrix representing the generated data points

1. Train the classifier $\mathcal{W}$ on the original dataset $Original$ and calculate error e
2. Initialise classified data with the original data, i.e., $Classed = Original$.
3. Let $j = O + 1$
repeat
  4. Classify the data point $New(j)$ and add it to the classified data $Classed(j)$
  5. Compare the classified data C against their true labels in New and calculate errors e_j.
  6. Remove the oldest point (first point in the window) and append $Classed(j)$ to the window M.
  7. $j = j + 1$.
  8. Re-train the classifier $\mathcal{W}$ on the new window.
until j = n

and (7) radial basis function neural network (RBN). As mentioned previously, the aim of having a large spectrum of classifiers is to test them using drifting data in a non-adaptive mode and compare the results with testing them using the same data in an adaptive mode based on forgetting at constant rate.

3.2 Incremental/On-Line Learning

As explained in Sec. 3, incremental learning (IL) algorithms offer, in principle, a natural mechanism for adaptation to changing environments. The challenge with these algorithms lies in finding a balance between catastrophic interference (or forgetting) on one hand and the ability to incrementally and continually accommodate new data points whenever new data becomes available [12]. The known IL algorithms are often used in an offline manner and with static data. In the context of this study, we will investigate their appropriateness to handling a drifting data. The algorithms applied are of two types: (1) *hyperbox-based algorithms (HBAs)* and *point-based algorithms (PAs)*. The HBAs class includes ARTMAP (FAM) [13], nearest generalized exemplar (NGE) [14] and generalized fuzzy min-max neural networks (GFMMNN) [15]. In these algorithms, prototypes are modeled as hyperboxes defined by their upper and lower corners. They are updated using some operations such as shrinking and growing. On the other hand, the PAs class includes growing neural gas (GNG) [16] and incremental learning based on function decomposition (ILFD) [17]. While prototypes in GNG are nodes of a graph, in ILFD they are cluster centers in the sense of radial basis function neural networks and can take different geometrical shapes depending on the type of distance used. A more detailed comparative description of the characteristics of all these algorithms can be found in [9].

4 Experimental Evaluation

Table 1 shows the collection of data sets that will be used to evaluate the approach presented in this paper. After normalisation of the original data, the rate of change R was set to $0.2, 0.3$ and 0.4 for simulating slow, medium and fast drift respectively. With an exception of the artificial dataset which was drifted only along the first feature, these values have been applied to each of the features at the same time for all the other datasets. In effect, the main goal of the evaluation is to check the effect of drifting those data sets on the accuracy of the classification. Basically, two aspects are to be examined: (1) retraining and (2) on-line adaptive learning.

Applying the first option, retraining (Sec. 3.1), the selected classifiers produced the results shown in Fig. 2. In this experiment, we ignored the artificial data set since all classifiers have 0% error. Considering the mixture of gaussians data, all classifiers seem to have a similar trend. Indeed, the effect of drift appears to worsen the performance as the degree of drift rises from slow to fast. When tested on iris data, they achieve high accuracy regardless of the degree of drift except for SVC which provides the least accuracy. Similarly, all selected classifiers maintained a stable performance when tested on Wine data but they

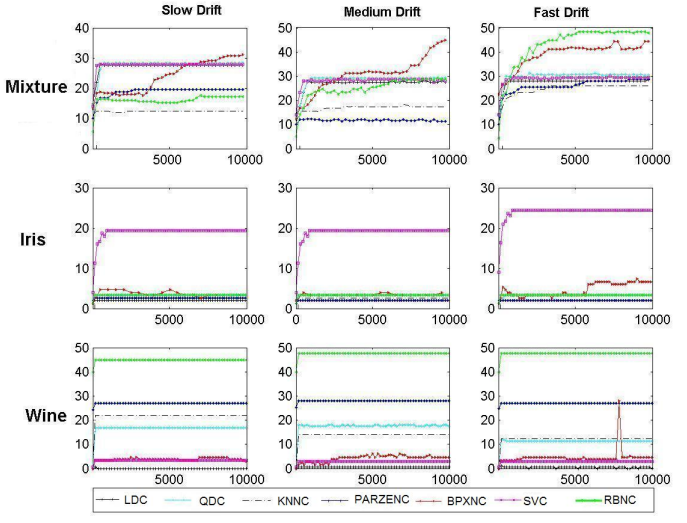


Fig. 2. Performance (errors in %) using the classifiers on the evolving dataset

differ in the level of accuracy. For instance, RBNC achieves the worst accuracy, in the contrast to KNNC which achieves the best accuracy result compared to the remaining classifiers.

On the other hand, the application of the the selected IL algorithms (Sec. 3.2) on each of the drifting data sets yields the results depicted in Fig. 3. As one can see the performance of the algorithms deteriorates as the degree of drift rises from slow to fast. This is clearly visible in most cases. Some exceptions have occurred. For instance, GNG achieved better results when the drift is significant as can be noticed in the case of Wine data. Similar remark holds for ILFD with the Iris data. It is however worth stressing that in the case of mixture data (overlap of classes), all algorithms behave in almost the same manner and tend to result in a 50% error which corresponds to a random guess accuracy. Moreover, the easiest data sets are the Artificial and the Wine data sets since at least one of the algorithms provides very high performance (error close to 0). The former case, ILFD achieves the best results, while in the latter case, GFMMNN outperforms the remaining algorithms. With the mixture of Gaussians data, FAM and GNG are slightly better than the others, while for the iris data, GFMMNN performs better. Nevertheless, in many cases the algorithms have the same tendency and clearly the original incremental learning rules do not cope very well with drifting data.

Comparing the outcome of retraining and incremental learning experiments (Fig. 2 and Fig. 3), the evolution trend of the performance indicates that the error in the first class of algorithms steadily increases as the the degree of drift increases when deployed in non-adaptive mode (see Tab. 2). However, the performance of the selected classifiers increases when they are adapted using either complete retraining or incremental learning. It is not entirely clear whether the degree of drift has similar effect on our selected adaptive algorithms as with the

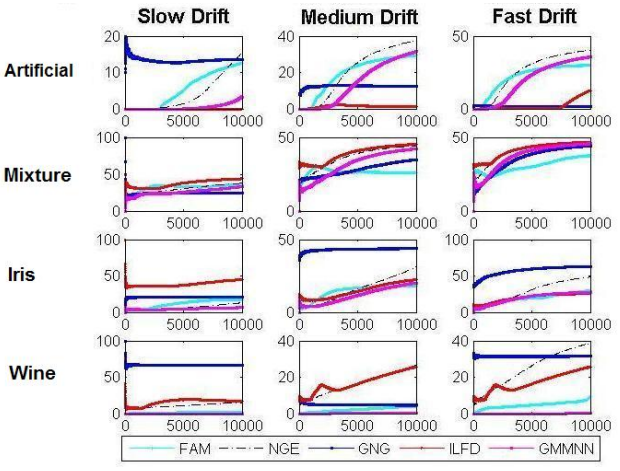


Fig. 3. Performance (errors in %) of the IL Classifiers on the drifting datasets

Table 2. Average classification error (%)

Alg. / Data	LDC	QDC	KNNC	PARZENC	BPXNC	SVC	RBNC
Artificial	0	0	0	0	0	0	0
Slow	15.00	15.00	14.81	14.92	09.72	14.86	21.30
Medium	37.61	37.65	37.42	60.07	35.57	37.46	59.74
Fast	40.42	40.43	40.37	57.54	38.88	40.36	56.98
Gaussians	14.40	13.60	10.80	10.00	09.60	14.00	04.80
Slow	28.79	28.31	29.24	29.41	31.46	27.96	35.67
Medium	40.44	41.14	40.97	40.87	55.90	40.04	54.39
Fast	44.79	52.23	44.90	44.89	53.92	44.56	52.14
Iris	02.00	02.00	02.00	0.67	03.33	04.00	0
Slow	08.99	09.04	10.18	08.89	13.89	18.65	13.53
Medium	22.32	34.98	28.39	25.45	25.09	27.05	45.76
Fast	39.63	50.31	47.03	45.38	29.10	31.82	55.71
Wine	0	0.56	0	25.28	0	0.56	03.93
Slow	02.76	04.84	0	24.75	03.77	02.21	54.64
Medium	23.09	28.47	20.00	24.58	17.88	19.55	57.38
Fast	43.95	42.68	63.00	24.36	28.60	42.11	58.23

non-adaptive mode. Depending on the window size, complete retraining builds a new model after each retraining iteration which is time consuming compared to incremental learning. IL algorithms adapt their existing model to include new observations whenever they appear. The drawback of both methods is the risk associated with using misclassified data to adapt the model which leads to performance deterioration. This calls for on-line performance monitoring to detect such continuous deterioration, however given that the true label of patterns is often unavailable, this issue remains a question for further research.

5 Conclusion

The paper evaluates two principle classes of approaches for handling the problem of data drift in the context of pattern classification. These approaches are: a) retraining classifiers based on moving data windows in order to capture change

and b) a range of incremental learning algorithms with their natural mechanisms to incrementally adapt the classification models without need for storing training data. While the empirical evaluation has shown that such approaches improve the classification performance on drifting data in comparison to non-adaptive classifiers, the accuracy of even the adaptable classifiers (as evaluated in this study) deteriorates with time irrespectively and much more robust approaches balancing the on-line learning and forgetting are needed if such classifiers are to be deployed with confidence in a dynamically changing environments.

References

1. Kelly, M.G., Hand, D.J., Adams, N.M.: The impact of changing populations on classifier performance. In: Proc of the fifth ACM SIGKDD international conference on Knowledge discovery and data mining, pp. 367–371 (1999)
2. Gabrys, B.: Learning hybrid neuro-fuzzy classifier models from data: to combine or not to combine? Fuzzy Sets and Systems 147(1), 39–56 (2004)
3. Widmer, G., Kubat, M.: Learning in the presence of concept drift and hidden contexts. Machine Learning 23(1), 69–101 (1996)
4. Salganicoff, M.: Density-adaptive learning and forgetting. In: Proceedings of the Tenth International Conference on Machine Learning, pp. 276–283 (1993)
5. Abraham, W.C., Robins, A.: Memory retention-the synaptic stability versus plasticity dilemma. Trends in Neurosciences 28(2), 73–78 (2005)
6. French, R.M.: Pseudo-recurrent connectionist networks: An approach to the Sensitivity-Stability dilemma. Connection Science 9(4), 353–380 (1997)
7. Lim, C.P., Harrison, R.F.: Online pattern classification with multiple neural network systems: an experimental study. Systems, Man and Cybernetics, Part C, IEEE Transactions on 33(2), 235–247 (2003)
8. Bouchachia, A., Mittermeir, R.: Towards incremental fuzzy classifiers. Soft Computing 11(2), 193–207 (2007)
9. Bouchachia, A., Gabrys, B., Sahel, Z.: Overview of some incremental learning algorithms (2007)
10. Hettich, S., Blake, C., Merz, C.: UCI repository of machine learning databases (1998), `www.ics.uci.edu/~mlearn/MLRepository.html`
11. Duin, R.P.W.: PRTools 3 Manual, provided with PRTools 3.1.6 Toolkit (2000)
12. Grossberg, S.: Nonlinear neural networks: principles, mechanism, and architectures. Neural Networks 1, 17–61 (1988)
13. Carpenter, G., Grossberg, S., Rosen, D.: Fuzzy ART: Fast stable learning and categorization of analog patterns by an adaptive resonance system. Neural Networks 4(6), 759–771 (1991)
14. Salzberg, S.: A nearest hyperrectangle learning method. Machine learning 6, 277–309 (1991)
15. Gabrys, B., Bargiela, A.: General fuzzy min-max neural network for clustering and classification. IEEE Transactions on Neural Networks 11(3), 769–783 (2000)
16. Fritzke, B.: A growing neural gas network learns topologies. In: Advances in neural information processing systems, pp. 625–632 (1995)
17. Bouchachia, A.: Incremental learning via function decomposition. In: Proceedings of the 5th ICMLA, pp. 63–68. IEEE Computer Society, Los Alamitos (2006)

Architecture of an Hybrid System for Experimentation on Web Information Retrieval Incorporating Clustering Techniques

Montserrat Mateos[1] and Carlos G. Figuerola[2]

[1] Universidad Pontificia de Salamanca, Spain
mmateossa@upsa.es
[2] Universidad de Salamanca, REINA Research Group, Spain
figue@usal.es
http://reina.usal.es

Abstract. The results retrieved from common search engines are frequently not acceptable because of the large amounts of Web pages that are returned and the ambiguity of the results. Several authors have suggested grouping the Web pages by topics using clustering techniques.

The application of these techniques involves different alternatives in different areas. Carrying out these research works could be easier and their cost lower if an experimentation hybrid system could be available that implemented the different alternatives or new ones could be added with a low cost.

This work proposes an architecture of a Web information retrieval experimentation hybrid system that incorporates clustering techniques. This hybrid system is composed of different replaceable components, so that we could test different alternatives with a low cost. Later, using the proposed and the implemented system an empirical study is exposed in order to evaluate what type of clustering techniques is the most appropriate.

Keywords: Web Information Retrieval, Clustering, Architecture, Hybrid Systems.

1 Introduction

This work shows the experimental results of applying clustering techniques to documents or Web pages retrieved from a search engine (Fig. 1).

The application of clustering techniques to the Web searches has been proposed in several works [5], [6], [7], [8], [11], [12], [15], but there are few empirical evaluations about the results of applying this type of techniques to the results retrieved from a search engine, so the application of this techniques is an important study and research area.

The most experimental and empirical works are based on computing the clustering solutions from a set of Web pages and then they determine and validate if the clustering solution is acceptable. In order to carry out these works the Web

B. Apolloni et al. (Eds.): KES 2007/ WIRN 2007, Part II, LNAI 4693, pp. 427–434, 2007.
© Springer-Verlag Berlin Heidelberg 2007

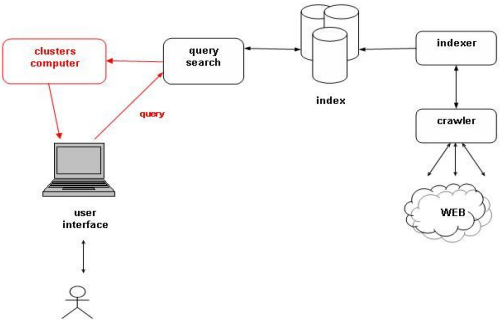


Fig. 1. Search engine architecture incorporating techniques of clustering

pages that answer to a query are retrieved, after the features to compute the clustering solutions are extracted, later the clustering solution is computed and finally the pages grouped by topics are displayed to the user; then the user can carry out the searches in this environment. There are a lot of different alternatives for each of above operations (information retrieval models [1], clustering techniques [2] or retrieval task from search engines) and the researchers choose those alternatives that they think are more adequate; so the cost of the definition and implementation of the different applications and tools to carry out the research is very high, moreover this approach duplicated most of the work and effort of development and implementation. For these reasons in our opinion developing an hybrid system that incorporates these techniques, models, etc. to carry out any experimentation with a low cost represents an important contribution.

In this work an architecture of an hybrid system of Web information retrieval based on search engine incorporating clustering techniques for experimentation is proposed and its application is proved with the evaluation of the performance of different algorithms of clustering (flat and hierarchical). Next the proposed architecture of this hybrid system is exposed, the study case used for the trial is explained and finally the conclusions are stated.

2 Architecture of a Hybrid System to Web Information Retrieval

The proposed and designed hybrid system (Fig. 2) has several modules. In this system each component or module is replaceable, so that different alternatives could be tested; In order to test a new alternative we only have to replace the component that we want to test.

The architecture of the proposed system has 5 components:

- ***Web Pages Retriever.*** This component retrieves the Web pages that answer the user's query. The input to this component is the user's query and the output is the Web pages retrieved from search engines.
- ***Documents Pre-Processor.*** With this component the *logical view* of each Web page is obtained. The input is the set of Web pages and the index

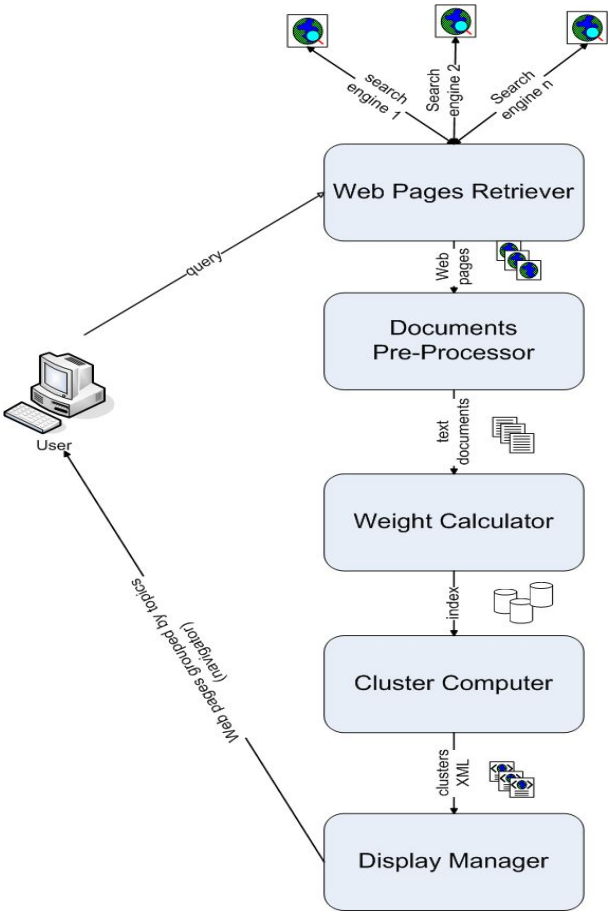


Fig. 2. Architecture proposed for a Web information retrieval experimentation hybrid system incorporating clustering techniques

terms of each page is the output. There are different alternatives to compute the logical view. The logical view could be obtained from different sources (snippet, n first words of the Web page, whole Web page, etc.), and on the other hand, the logical view could be obtained applying different operations (elimination of stop words, stemming, etc). In this component the documents have to be converted into plain text too because the Web documents can be in several formats (html, word, postscript, pdf,...)

- ***Weight Calculator.*** The logical view of the Web pages is the input to this component and the output is the weight of each of the index terms. Several schemas [13] have been implemented.
- ***Cluster Computer.*** The clustering solution using the chosen algorithm is computed. The algorithms used to compute the solutions of clustering

need as input the features/characteristics of the objects to cluster, and as output, they give the objects grouped. Usually the input/output information format is different for each algorithm. Therefore this component has also to transform the weight of index terms to the format demanded by the algorithm and after it transforms the output given by the algorithm to the format required by the next component. The format required by the next component has been designed to store any type of clustering solution (flat or hierarchical) just as any number of clusters. XML has been used to store this information.

- ***Display Manager.*** With this component the retrieved Web pages are displayed grouped by topics meaningful to the user. The input is the XML information created with the previous component.

3 Study Case

The designed system has been implemented to evaluate the performance of different algorithms of clustering when these algorithms are applied to the Web pages retrieved from search engines. To carry out this research a query collection has been designed; Several users have used the environment with clusters to resolve the queries. Several queries of popular knowledge were proposed, and for each query the users had to look for information about it and not an exactly and concrete answer. For each query 1000 results were retrieved[1]. The clustering solutions were computed using the complete documents because some authors [6], [7] consider that with complete documents the obtained results are better than with only the snippet. Furthermore the solution was computed with different number of clusters.

The system has been developed using the proposed architecture. The *Web Pages Retriever* has been developed with Google API[2]. In the component *Pre-Processor* several conversion tools have been used and an application was implemented to remove accents, punctuation marks and stopwords and to do stemming. The *Weight Calculator* computed the weight of index terms using *ntc* schema [13]. The *Cluster Computer* computes the solution of clustering using different algorithms; One of them computes a flat solution, other of them computes a hierarchical solution and the third computes a solution that is displayed in a flat format and in an hierarchical format. Finally the *Display Manager* displays to the user the Web pages grouped using the information in XML obtained from the component *Cluster Computer*.

3.1 Document Clustering

Document Clustering [2] consists in partitioning a group of documents in a preset number of k subsets, so that the documents assigned to each cluster are more

[1] This is the amount of Web pages that a search engine retrieves as answer to a user's query.

[2] http://www.google.com/apis

similar to the documents of their subset than to the documents assigned to other subset.

Flat clusters (Fig. 3) [2], [9] can be computed by means of partitional algorithms; in this case the number of groups or clusters are fixed previously and the documents are assigned to those partitions iteratively. There are several algorithms and they are differentiated by: the way to obtain the initial partition of the documents, the function of optimization, the reassignment process and the criterion of finalization.

Hierarchical algorithms [2], [9] compute a tree structure named dendogram (Fig. 4). In this structure the leaves are groups that they have only one document, so there are as many leaves as documents; the root is only one group that it contains all documents of the collection, and the intermediate levels are the possible configurations of clusters. These algorithms compute the solution of clustering merging (agglomeratives) or partioning (divisives) groups of documents.

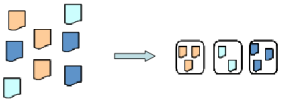


Fig. 3. Flat Clustering

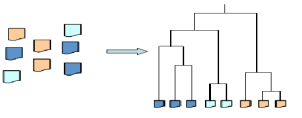


Fig. 4. Hierarchical clustering

To calculate the clustering solutions in the experiments both types of algorithms -flat and hierarchical- have been used [10], [16], [17]. Exactly 3 algorithms have been used. One of them (*Repeated Bisections*) has been used to compute a *flat* and an *hierarchical* solution:

– Classical partitional *Direct K-Way*, using as optimization function the next function,

$$max \sum_{r=1}^{k} \sum_{d_i \in S_r} sim(d_i, C_r) \tag{1}$$

Each cluster r is represented by means of its centroid vector, C_r, and the goal of this function is to find the clustering solution that maximizes the similarity between each document d_i and the centroid of the cluster that is assigned to, using the cosine function[3] [14] to compute de similarity.

With this algorithm the k clusters are computed simultaneously. In order to compute the solution k seed documents are chosen and then the similarity between each document and each one of the k seeds are computed; the document is assigned to the most similar cluster. After this, the initial solution

[3] $sim(d_i, C_r) = \frac{\boldsymbol{d_i} \bullet \boldsymbol{C_r}}{|\boldsymbol{d_i}| \times |\boldsymbol{C_r}|} = \frac{\sum_{t=1}^{m} (w_{t,i} * w_{t,r})}{\sqrt{\sum_{t=1}^{m} w_{t,i}^2} * \sqrt{\sum_{t=1}^{m} w_{t,r}^2}}$, being m the whole number of index terms and w the weigh of each index term in the document or in the centroide vector.

is refined by means of iterations. In each iteration the different documents are randomly visited. For each document d_i the value of the optimization function is computed supposing that the document d_i is located in each one of the others clusters. If a better value of the optimization function is obtained then the document d_i is assigned to one of the others clusters. If a better value is not obtained then the document remains in the same cluster.

The algorithm finishes when one iteration has been carried out with no documents have been assigned to other cluster.

This algorithm has been used to display the search results in a flat cluster format.

- Divisive (*Repeated Bisections*). In this case the solution is computed carrying out a repetitive sequence of divisions. The objects are divided into two groups, then one of the groups is selected and it is divided again, getting in this way 3 groups. This process is repeated until the number of clusters desired is obtained. The way in which this algorithm works is similar to the partitional one, but in this case only two documents are randomly selected and they are the seeds; In order to determine how to carry out the partition it uses an optimization function. In our experiments we have used the function 1. In each iteration the cluster selected to make the next partition is the one with the best results from the function have been obtained.

  This algorithm has been used to show the answer of a search engine using a flat format and an hierarchical format.
- Agglomerative, we have chosen the classical agglomerative. With this algorithm a pair of clusters are merged in each step, but in this case to determine what is the next cluster to group in each step the optimization function 1 has been used too.

  This algorithm has been used to display the documents grouped hierarchically.

In the Table 1 we resume the algorithms used for the experiments.

Table 1. Algorithms used in the experiments

	Algorithm
Flat	Direct k-way
	Repeated Bisections
Hierarchical	Repeated Bisections
	Agglomerative

3.2 Results

We have computed the average precision of the set of queries to carry out the evaluation. We have used the precision [1], [3], [4] of each query to compute the average precision. As it can be observed in the graph of the Fig. 5, in an environment with flat solutions we have obtained an average precision of 0.027 that it is

below to which we have obtained with hierarchical solutions (0.0472). So we can state that with a system based on hierarchical clusters the users obtain better results and they make better searches. Furthermore, an analysis was carried out to show the performance of the different algorithms and looking the graph of Fig. 6, we can see that for flat solutions the best algorithm is *Repeated Bisections* and for hierarchical solutions the best algorithm is *Agglomerative*.

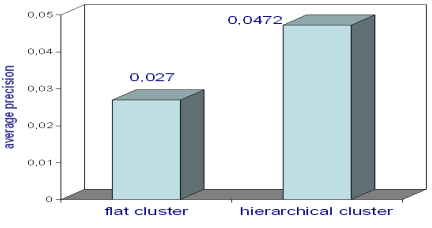


Fig. 5. Flat vs Hierarchical

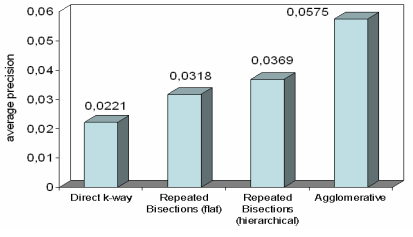


Fig. 6. 4 Algorithms used

4 Conclusions

In this work we have proposed and implemented a hybrid system for experimentation in order to apply clustering techniques in Web information retrieval.

With this system an experimentation has been carried out in order to evaluate the performance of some algorithms of clustering, but any other experiments could be carried out with a low cost. For example we could apply any other clustering algorithm or we could use any other weight schema or we could calculate other logical view.

On the other hand, we can state that better results have been obtained with the hierarchical paradigm, being the best of them the Agglomerative algorithm.

References

1. Baeza-Yates, R., Ribeiro-Neto, B.: Modern Information Retrieval. ACM PRESS, New York (1999)
2. Chakrabarti, S.: Mining the Web: discovering knowledge from hypertext data. Morgan Kaufmann, San Francisco (2003)
3. Cleverdon, C., Keen, E.: Factors determining the performance of indexing systems. Test results. ASLIB Cranfield Research Project, Cranfield, England, vol. ii (1996)
4. Cleverdon, C., Mills, J., Keen, E.: Factors determining the performance of indexing systems. Design. Cranfield Research Project, Cranfield, England, vol. i (1996)
5. Cui, H., Zaiane, O.R.: Hierarchical structural approach to improving the browsability of Web search engine results. In: Workshop on Digital Libraries (DLIB'2001), pp. 956–960 (2001)
6. Ferragina, P., Gulli, A.: A personalized search engie based on Web-snippet hierarchical clustering. In: 14th International Conference on WWW, pp. 801–810. ACM Press, New York (2005)

7. Giannoti, F., Nanni, M., Pedreschi, D., Samaritani, F.: WebCat: automatic categorization of Web search results. In: SEBD03, pp. 507–518. ACM Press, New York (2003)
8. He, A., Ding, C., Zha, H., Simon, H.: Automatic topic identification using webpage clustering. In: 2001 IEEE International Conference on Data Mining, IEEE Computer Society, Los Alamitos (2001)
9. He, Q.: A review of clustering algorithms as applied in IR. Technical report, University of Illinois at Urbana-Champaign (1999)
10. Karypis, G.: Cluto. A Clustering Toolkit. Technical report, Departament of Computer Science. University of Minnesota (2003)
11. Kummamuru, K., Lotlikar, R., Roy, S., Singal, K., Krishnapuram, R.: A hierarchical monothetic document clustering algorithm for summarization and browsing search results. In: 13th International Conference on WWW, pp. 658–665. ACM Press, New York (2004)
12. Lin, K., Kondadadi, R.: A word-based soft clustering algorithm for documents. In: 16th International Conference on Computers and their Applications (2001)
13. Salton, G., Buckley, C.: Term-weighting approaches in automatic text retrieval. Information processing and management. 24, 513–523 (1988)
14. Salton, G., McGill, M.: Introduction to Modern Information Retrieval. McGraw-Hill, New York (1983)
15. Zamir, O., Etzioni, O.: Web document clustering: a feasibility demonstration. In: 21 Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 46–54. ACM Press, New York (1998)
16. Zhao, Y., Karypis, G.: Criterion functions for document clustering. Experiment and analysis. Technical report, departament of computer science. University of Minnesota (2002)
17. Zhao, Y., Karypis, G.: Hierarchical clustering algorithms for document datasets. Data Mining and Knowledge Discovery, 141–168 (2005)

Automated Ham Quality Classification Using Ensemble Unsupervised Mapping Models

Bruno Baruque[1], Emilio Corchado[1], Hujun Yin[2], Jordi Rovira[3], and Javier González[3]

[1] Department of Civil Engineering. University of Burgos, Spain.
bbaruque@ubu.es, escorchado@ubu.es
[2] School of Electrical and Electronic Engineering. University of Manchester, UK
h.yin@manchester.ac.uk
[3] Department of Biotechnology and Food Science, University of Burgos, Spain
jrovira@ubu.es, javigonzalez77@hotmail.com

Abstract. This multidisciplinary study focuses on the application and comparison of several topology preserving mapping models upgraded with some classifier ensemble and boosting techniques in order to improve those visualization capabilities. The aim is to test their suitability for classification purposes in the field of food industry and more in particular in the case of dry cured ham. The data is obtained from an electronic device able to emulate a sensory olfative taste of ham samples. Then the data is classified using the previously mentioned techniques in order to detect which batches have an anomalous smelt (acidity, rancidity and different type of taints) in an automated way.

Keywords: topology preserving mappings, ensembles, electronic nose, food industry.

1 Introduction

Nowadays, the consumer's trust is very important when you want to consolidate and situate a product into the market. Dry cured ham is a product typically consumed in the Iberian Peninsula which has a high popular acceptation. We understand for "Serrano Ham" a ham salted and dried during more than 210 days and that can be displayed boned or deboned. In this type of product, during the storage period come about the development of rancidity and acid odours; most of then may increase significantly because of proteolysis and lipid oxidation [1]. It is important to find easy, quick and cheap techniques that allow, through the measure of simple parameters, to evaluate the quality of these products before the consumers purchase.

In order to help food industry analysis to support the subjective decisions of human testers, several recently appeared devices designed especially for that purpose can be used. The disadvantage with this alternative type of testing is that what humans interpret as tastes and smells, machines interpret as (usually complicated) numeric measures. Our objective is devising a system capable of interpreting the analyses

B. Apolloni et al. (Eds.): KES 2007/ WIRN 2007, Part II, LNAI 4693, pp. 435–443, 2007.
© Springer-Verlag Berlin Heidelberg 2007

made by an electronic device and making the results of those analyses more easily understandable for human experts.

Topology preserving models are mainly used for data visualization and inspection which is an interesting feature when assisting human in classification tasks such as the one presented in this work. Other features are pattern recognition and automated classification inherent to many of all the unsupervised learning techniques, which are especially useful in the present application. These models are given enhanced stability in this study with the use of ensemble techniques.

A combination of an electronic device for the analysis of volatile compounds (referred as "electronic nose" in the rest of the paper) and topology preserving mapping algorithms upgraded by the use of ensemble meta-algorithms is used to analyse a wide variety of samples of "Serrano" Ham to verify if hams with different olfative characteristics could be easily and reliably discriminated by this procedure.

2 Preliminary Ham Samples Analysis

2.1 Case Study Description

The objective of this study is to test the suitability of the combination of an "electronic nose" (or e-nose), used to measure the odour characteristics of the ham samples; with unsupervised learning data analysis and visualization models for a simple and reliable ham testing and classification. For that reason, this study has been conducted using several hams of different qualities and origins within the limits of the Spanish territory. The dataset used consists on measures obtained from several brands of seven types of Spanish ham, available in the Spanish market. The samples analysed included also some tainted and/or too rancid/acid samples. The tainted samples were randomly taken from among all the different quality types and origins of hams. The commercial brands from where the samples were extracted are not taken into account in this study.

2.2 The Use of the Electronic Nose for Odour Recognition

The main steps of odour recognition can be summarized as follows:

1. Heating the sample for a certain time generates volatile compounds in the headspace of the vial containing the sample.
2. The gas phase is transferred to a detection device which reacts to the presence of molecules.
3. The difference in sensor reactions is revealed using different statistical calculation techniques to classify the odours. The measures taken by each sensor are separated and stored in a simple database for further study.

In this study the analyses are performed by an Electronic Nose αFOX 4000 (AlfaMOS, Toulouse, France) with a sensor array of 18 metal oxide sensors. The e-nose takes dates each 0.5 seconds during 120 seconds of acquisition time and 600 seconds of acquisition delay. Only the highest value obtained for each sensor is stored in the database used for further analysis.

3 Competitive Learning Ensembles

3.1 The SOM and ViSOM Algorithms

The Self-Organizing Map (SOM) algorithm [2] and the Visualization Induced Self-Organizing Map (ViSOM) [3] are different types of Topology Preserving Mappings. This is a family of techniques with a common target: to provide a low dimensional representation of multi-dimensional datasets while preserving the topological properties of the input space. The SOM algorithm is based on a type of unsupervised learning called competitive learning; an adaptive process in which the neurons in a neural network gradually become sensitive to different input categories, sets of samples in a specific domain of the input space [4].

One interesting extension of this algorithm is the ViSOM proposed to directly preserve the local distance information on the map, along with the topology. The ViSOM constrains the lateral contraction forces between neurons and hence regularises the interneuron distances so that distances between neurons in the data space are in proportion to those in the input space [5].

The difference between the SOM and the ViSOM hence lies in the update of the weights of the neighbours of the winner neuron as can be seen from Eqs (1) and (2).

Update of neighbourhood neurons in SOM:

$$w_k(t+1) = w_k(t) + \alpha(t)\eta(v,k,t)\big(x(t) - w_v(t)\big) \tag{1}$$

Update of neighbourhood neurons in ViSOM:

$$w_k(t+1) = w_k(t) + \alpha(t)\eta(v,k,t)\left(\big[x(t) - w_v(t)\big] + \big[w_v(t) - w_k(t)\big]\left(\frac{d_{vk}}{\Delta_{vk}\lambda} - 1\right)\right) \tag{2}$$

where w_v is the winning neuron, α the learning rate of the algorithm, $\eta(v,k,t)$ is the neighbourhood function where v represents the position of the winning neuron in the lattice and k the positions of the neurons in the neighbourhood of this one, x is the input to the network and λ is a "resolution" parameter, d_{vk} and Δ_{vk} are the distances between the neurons in the data space and in the map space respectively.

Both models can be adapted for classification of new samples using a semi-supervised procedure [6].

3.2 Ensembles of Topology Preserving Maps

The ultimate goal of constructing an ensemble is to improve the performance obtained of a single working unit. When talking about classification it is generally accepted that the sets of patterns misclassified by the different classifiers would not necessarily overlap. This suggests that different classifier designs potentially offer complementary information about the patterns to be classified and could be harnessed to improve the performance of the selected classifier [7]. Many ensemble models and theories have been developed in the previous years that range from a

quite simple technique like Bagging [8] or AdaBoost [9] to the more sophisticated ones such as LPBoost [10] or and many other variants. These techniques have been mainly applied to models designed specifically for classification, specially supervised classifiers [11].

Bagging and AdaBoosting. Boosting meta-algorithms consist on training a simple classifier in several stages by incrementally adding new capacities to the current learned function.

Bagging (or bootstrap aggregating) [8] consists on training each of the classifiers composing the ensemble separately using a different subset of the main training dataset. This is accomplished by using re-sampling with replacement over the training set. The classification results are obtained by (weighted or majority) voting among its composing classifiers. The technique provides the ensemble with a balance between variability and similarity.

AdaBoost [9] is another algorithm that gives as a result an ensemble but its training method is a bit more developed. The difference with the previous method is that it is taken into accounts which of the training samples are not correctly classified by the current classifier. That way, the ensemble concentrates in the samples that are harder to classify, improving its learning capabilities. The final output of the ensemble is a weighted vote of all the classifiers composing it. Schapire and Freund [11] proposed two slightly different versions of the algorithm. In this work AdaBoost.M2 is used, as it is recommended for datasets with more than two different classes. Although this meta-algorithm has been previously used for supervised classification [12], a short number of studies [13], [14], [15] involves unsupervised one such as in the present work.

Combining Ensemble Results. It must be taken into account that the models used in this research were originally designed as visualization tools. Constructing classical ensembles can be considered as a good option when trying to boost their classification capabilities, stabilizing its learning algorithm and avoiding overfitting; but when dealing with its visualization feature an ensemble is not directly displayable [16], [15]. To overcome this problem, some “ensemble combination” algorithms have been devised [17] in order to obtain a unique network that somehow represents the information contained in the different networks composing the ensemble. Our objective is that this “combination” unites good classification accuracy with truthful representation of data for visual inspection.

The procedure is the same for all the tests involving this combination of networks. A simple n-fold cross-validation is used in order to employ all data available for training and testing the model and having several executions to calculate an average of its performance. In each step of the cross-validation, first an ensemble of networks must be obtained. The way in which the ensemble is trained does not affect the system in which the combination is computed. In the case of this study, this has been done using the bagging or the adaboost meta-algorithm. Then the computation of the combination is performed. Finally both, the ensemble and the combination generated from it are tested employing the test fold. In order to obtain the similarity needed to make networks comparable to combine them, the networks of the same ensemble are

initialized to the inter-neural weights obtained in the training of the previous network.

The different options studied for combining the network of the ensemble into a single network, summarizing its main characteristics are described in the following paragraphs. Further details can be found on [17], [18].

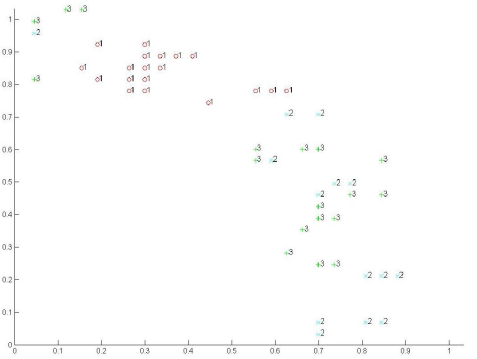
Ensemble Combination 1: Fusion [19] involves comparing the networks neuron by neuron in the input space. This implies that all the networks in the ensemble must have the same size. First, it searches for the neurons that are closer in the input space (selecting only one neuron in each network of the ensemble) then it "fuses" them to obtain the final neuron in the "fused" map. This process is repeated until all the neurons have been fused. In this case, a re-labelling of the neurons of the fused network, employing again the training dataset, is done in order to get a clear visualization of the map.

Ensemble Combination 2: Superposition [20] consists on "superposing" the maps formed by the networks composing the ensemble into a final map, on a neuron by neuron comparison (as is done in fusion). Note that the weights of each network in the ensemble are initialized in a way that makes the neurons in the same position of two (or more) networks comparable. On the contrary of fusion, this technique relies in the way the networks are trained to compute an acceptable summary, so the calculations on the combination phase are less time-consuming. The method ensures that the resultant "summarized" or "superposed" map represents visually what the majority of the maps composing the ensemble represent in a neuron-by-neuron basis. When using the resultant "superposed" map for classification purposes it returns the class represented by the neuron that is activated when the new sample is presented to the network.

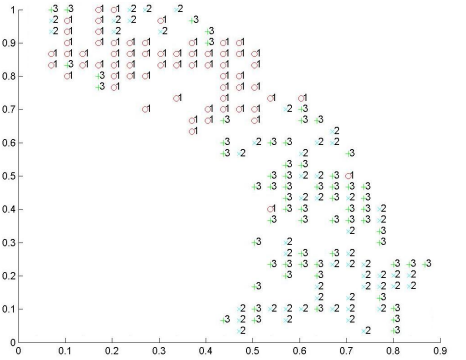
Ensemble Combination 3: Superposition + Re-labelling [20] has two main phases. The first one is the superposition explained before. The second one consists of testing which class actually recognizes better each neuron after the superposition, instead of relying on the recognition of the neurons in the same position done previously in the individual testing of each of the ensemble networks. So, after the superposition phase, the same dataset used for training is presented to the resultant network of the superposition to check which class is more consistently recognized by each neuron. Usually less number of neurons responds to this re-labelling, giving as a result a more compact map.

4 Experiments and Results

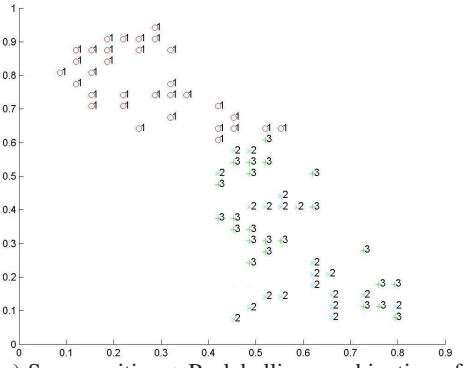
After having been recorded by the electronic nose, the measures of the 18 sensors are stored for each sample of ham cut in a database along with its corresponding human tester cataloguing. This cataloguing is normally more detailed, but for this initial study, it has been simplified to only three possible values: "non-spoiled", "rancid/acid" and "tainted". Thus, our final dataset consists in a total of 176 samples of ham, each of them composed of 18 different variables measured over it and 3 possible classes.



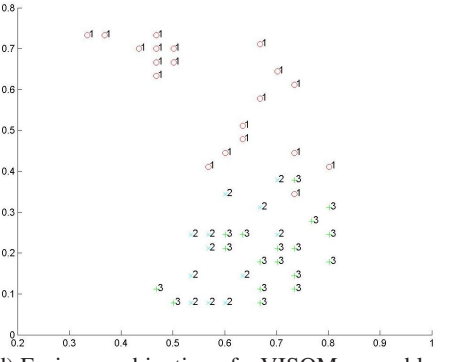
a) Single ViSOM map



b) Superposition combination of a ViSOM ensemble



c) Superposition + Re-labelling combination of a ViSOM ensemble



d) Fusion combination of a VISOM ensemble

Fig. 1. The four images (a), (b), (c) and (d) are obtained from VISOM models of a size of 30x30. The 'non-spoiled' samples are marked with a 1 (top of the images), 'acid/rancid' are marked as 2 and 'tainted' samples with 3.

Regarding to the visualization capabilities of the two topology preserving mapping models employed, the ViSOM is the one that obtains better results in this aspect, as it is specially devised for enhancing visualization. The use of ensemble and combination techniques can certainly improve the quality of the final map obtained, gaining in stability during the training phase and obtaining much less sparse maps, as the final representation can include all the responding neurons of the maps composing the ensemble. Among the ensemble combination methods, the Superposition + Re-labelling seems to capture better the internal structure of the dataset, obtaining a similar image as a single network. This can be observed in Fig. 1 a) to d).

What is interesting in this kind of analysis is that you get visual (and therefore immediate) information of the level of quality of a sample within its main group. In Fig. 2 neurons identified with 7, 10 and 12 are activated by the highest quality samples. Those identified 11 are activated by standard quality samples. Neurons marked with 2 and 3 are activated by samples that were previously classified as standard, but were not conserved properly and started to get spoiled. Neurons marked

with 1, 6, 8 and 9 are activated by 'acid/rancid' samples. Neurons marked with 4, 5 and 14 are the ones that are activated by definitively 'tainted' samples. As it can be seen in Fig 2 neurons recognising 'non-spoiled' samples are all situated at the top of the image (in Fig 1(c) marked with 1s). Within this group, the neurons activated by highest quality samples are above the ones activated by the standard quality ones, which appear closer to 'acid/rancid' group. The samples that were previously considered to be of standard quality (marked with 2 and 3) and got rancid are now recognized by neurons situated among the spoiled samples recognizing neurons.

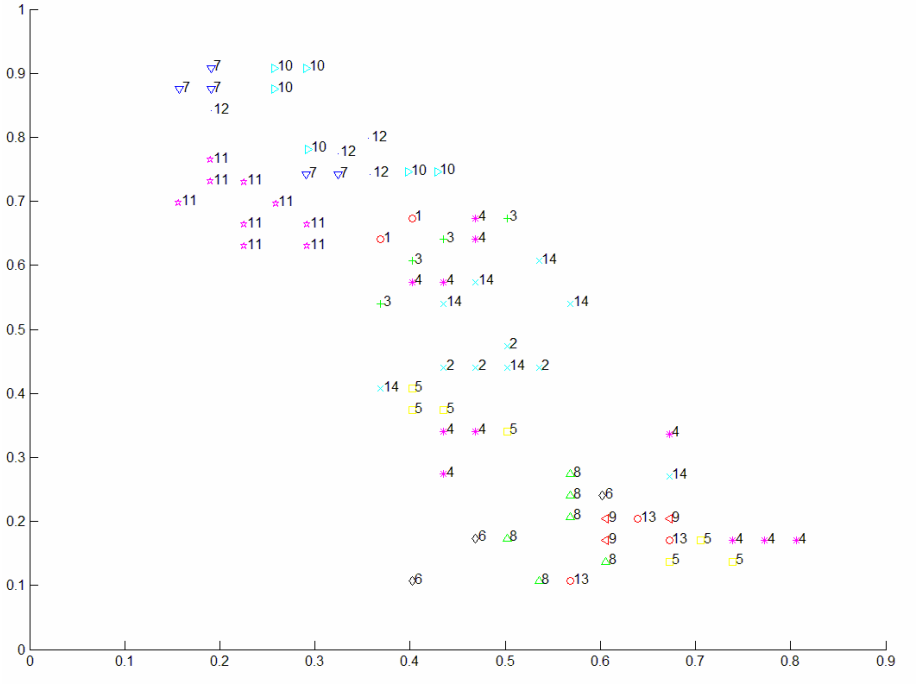


Fig. 2. The image is obtained form the same ensemble combination as Fig 1 (c), but using a more expressive classification on data samples, indicating different levels of quality among the previously described 3 groups

Table 1. Model classification accuracy using a 10 fold cross-validation over the dataset. The ensembles are all composed of 10 networks. The training was done employing the **bagging** meta-algorithm.

Model	*Best Single Netwk.*	*Ensemble (weighted voting)*	*Superp.*	*Superp. + Re-Labelling*	*Fusion*
SOM (10x10)	68.4%	84.1%	73%	69%	78.4%
SOM (20x20)	49.5%	78.7%	62.2%	57.7%	59.3%
SOM (30x30)	40.8%	67.2%	49.1%	52.6%	52.1%
ViSOM (20x20)	78%	90%	87.1%	83.9%	79.7%
ViSOM (30x30)	69.1%	82.8%	81.5%	61.7%	60.4%
ViSOM (40x40)	62.5%	86.9%	52.7%	70.4%	71.3%

Regarding the automated classification of samples, the SOM and ViSOM behave in a similar way. As it was expected, and can be seen both in Tables 1 and 2, the ensemble of maps obtains a much higher accuracy than its homologue single and combined networks. The three combination models are much more dependent of individual performances, so they obtain much lower results than the ensemble, but still significantly better classification rate than the single model. Adaboost obtains better results than Bagging in most of the cases, although is not a very significant difference, as can be seen comparing results of Table 1 and Table 2.

Table 2. Model classification accuracy using a 10 fold cross-validation over the dataset. The ensembles are all composed of 10 networks. The training was done employing the **adaboost.m2** meta-algorithm.

Model	*Best Single Netwk.*	*Ensemble (weighted voting)*	*Superp.*	*Superp. + Re-Labelling*	*Fusion*
SOM (10x10)	70.2%	82.9%	71.7%	73.4%	78%
SOM (20x20)	47.8%	78.3%	62.6%	61.7%	62.6%
SOM (30x30)	38%	66.7%	58.7%	49.4%	51.7%
ViSOM (20x20)	80.9%	91.1%	81.1%	82%	84.6%
ViSOM (30x30)	68%	90%	72.1%	70.9%	75.5%
ViSOM (40x40)	64.9%	85.6%	75.8%	66.3%	71%

5 Conclusions

In the present work a viable way of automated ham testing for quality based on ensembles of topology preserving maps is presented.

With the appropriate training and use of boosting techniques the classification can be very simple to the human expert, as the reliability of the classification can be inherently deduced from the position of the new sample in the map. For example, if the sample under test activates a neuron in a region of the map clearly identified as “non-spoiled” it can be concluded that the sample is definitively tainted. If the sample is represented in a region identified as “non-spoiled” but very close to the “acid/rancid” region it can be concluded that the sample is not as definitively “non-spoiled” as the previous one and will probably be a “non-spoiled” sample with a “rancid” scent.

This visual analysis can be complemented with the classification capabilities of the model that can achieve more than 90% of accuracy in explicit classification. Note that this measure is calculated using a hard discrimination, while the model is capable of giving a richer interpretation of the classification, providing a percentage of confidence in the inclusion of the sample into one class or another.

Concerning to future work, there are several methods that are regarded as possibly very useful in the present case study and will be tested. Some of them are more evolved boosting algorithms and some method for feature selection to reduce the dimensionality of the dataset in first instance. Also more refined accuracy measurements could be used to determine more exactly the classification capabilities of the models.

Acknowledgments. This research has been supported by the MCyT project TIN2004-07033.

References

1. Monahan, R.L., Brunton, N.P., Cronin, D.A., Durcan, R.: Determination of Hexanal in Cooked Turkey Using Solid Phase Microextraction (Spme)/Gc. In: 44th International Congress of Meat science and Technology (ICoMST), vol. 1, pp. 586–587 (1998)
2. Kohonen, T.: The Self-Organizing Map. Neurocomputing 21, 1–6 (1998)
3. Yin, H.: Data Visualisation and Manifold Mapping Using the Visom. Neural Networks 15, 1005–1016 (2002)
4. Kohonen, T., Lehtio, P., Rovamo, J., Hyvarinen, J., Bry, K., Vainio, L.: A Principle of Neural Associative Memory. Neuroscience 2, 1065–1076 (1977)
5. Yin, H.: Visom - a Novel Method for Multivariate Data Projection and Structure Visualization. Neural Networks, IEEE Transactions on 13, 237–243 (2002)
6. Kraaijveld, M.A., Mao, J., Jain, A.K.: A Nonlinear Projection Method Based on Kohonen's Topology Preserving Maps. Neural Networks, IEEE Transactions on 6, 548–559 (1995)
7. Heskes, T.: Balancing between Bagging and Bumping. In: Advances in Neural Information Processing Systems 9 - Proceedings of the 1996 Conference, vol. 9, pp. 466–472 (1997)
8. Breiman, L.: Bagging Predictors. Machine Learning 24, 123–140 (1996)
9. Freund, Y., Schapire, R.E.: A Decision-Theoretic Generalization of on-Line Learning and an Application to Boosting. Journal of Computer and System Sciences 55, 119–139 (1997)
10. Demiriz, A., Bennett, K.P., Shawe-Taylor, J.: Linear Programming Boosting Via Column Generation. Machine Learning 46, 225–254 (2002)
11. Freund, Y., Schapire, R.E.: Experiments with a New Boosting Algorithm. In: International Conference on Machine Learning, pp. 148–156 (1996)
12. Schwenk, H., Bengio, Y.: Boosting Neural Networks. Neural Computation. 12, 1869–1887 (2000)
13. Gabrys, B., Baruque, B., Corchado, E.: Outlier Resistant PCA Ensembles. In: Gabrys, B., Howlett, R.J., Jain, L.C. (eds.) KES 2006. LNCS (LNAI), vol. 4253, pp. 432–440. Springer, Heidelberg (2006)
14. Corchado, E., Baruque, B., Gabrys, B.: Maximum Likelihood Topology Preserving Ensembles. In: Corchado, E., Yin, H., Botti, V., Fyfe, C. (eds.) IDEAL 2006. LNCS, vol. 4224, pp. 1434–1442. Springer, Heidelberg (2006)
15. Petrakieva, L., Fyfe, C.: Bagging and Bumping Self-Organising Maps. Computing and Information Systems (2003)
16. Kaski, S.: Data Exploration Using Self-Organizing Maps. Department of Computer Science and Engineering. Helsinki University of Technology. Espoo, Finland (1997)
17. Baruque, B., Corchado, E., Yin, H.: ViSOM Ensembles for Visualization and Classification. In: International Work Conference on Artificial Neural Networks, San Sebastián, Spain, Springer, Heidelberg (2007) (in press)
18. Baruque, B., Corchado, E., Yin, H.: Boosting Unsupervised Competitive Learning Ensembles. In: International Conference of Neural Network (ICANN 2007) (2007)
19. Georgakis, A., Li, H., Gordan, M.: An Ensemble of SOM Networks for Document Organization and Retrieval. In: Int. Conf. on Adaptive Knowledge Representation and Reasoning (AKRR'05), p. 6 (2005)
20. Baruque, B., Corchado, E., Yin, H.: VISOM Ensembles for Visualization and Classification. In: International Work Conference on Artificial Neural Networks (IWANN'07), San Sebastián, Spain, Springer, Heidelberg (2007) (in press)

FreeEnCal: A Forward Reasoning Engine with General-Purpose

Jingde Cheng, Shinsuke Nara, and Yuichi Goto

Department of Information and Computer Sciences, Saitama University,
338-8570 Saitama, Japan
{cheng, nara, gotoh}@aise.ics.saitama-u.ac.jp

Abstract. A forward reasoning engine is an indispensable component in many advanced knowledge-based systems with purposes of creation, discovery, or prediction. This paper presents a forward reasoning engine with general-purpose, named "FreeEnCal", which can interpret and perform inference rules defined and given by its users, draw fragments of various classical and/or non-classical logic systems formalized as different formal systems, draw empirical theorems of various formal theories constructed based on various logic systems, and perform deductive, inductive, and abductive reasoning automatically. FreeEnCal can be used as a ready-made forward reasoning engine serving as a core and fundamental component in various advanced knowledge-based systems as well as an alone forward reasoning engine with general-purpose. The paper presents our basic ideas to design and implement FreeEnCal, facilities provided by FreeEnCal, and some applications of FreeEnCal.

Keywords: Creation, Discovery, Prediction, Deduction, Induction, Abduction, Advanced Knowledge-Based Systems.

1 Introduction

Forward reasoning engine is a computer program to automatically draw new conclusions by repeatedly applying inference rules, which are programmed in the reasoning engine or given by users to the reasoning engine as input, to given premises and obtained conclusions until some previously specified conditions are satisfied. Many advanced knowledge-based systems with purposes of creation, discovery, or prediction, such as theorem finding systems, epistemic programming systems, anticipatory reasoning-reacting systems, economic prediction systems, and various knowledge creation/discovery systems, require a forward reasoning engine as an indispensable component to draw new facts, theorems, hypothesis, or predictions from known facts, theorems, or hypothesis given as premises. Thus, we have a question: Can we design and implement a forward reasoning engine with general-purpose such that it can be used as a core and fundamental component to develop various advanced knowledge-based systems? If so, by using such a ready-made tool, one can design and implement various advanced knowledge-based systems easily.

B. Apolloni et al. (Eds.): KES 2007/ WIRN 2007, Part II, LNAI 4693, pp. 444–452, 2007.
© Springer-Verlag Berlin Heidelberg 2007

Based on our experiences of development and applications of EnCal [2], we have designed and implemented a forward reasoning engine with general-purpose, named "FreeEnCal", which can interpret and perform inference rules defined and given by its users, draw fragments of various classical and non-classical logic systems formalized as Hilbert style axiomatic systems, Gentzen natural deduction systems, or Gentzen sequent calculus systems, draw empirical theorems of various formal theories constructed based on various logic systems, and perform deductive, inductive, and abductive reasoning automatically [1]. FreeEnCal can be used as a ready-made forward reasoning engine serving as a core and fundamental component in various advanced knowledge-based systems as well as an alone forward reasoning engine with general-purpose. In particular, FreeEnCal can be used as a core and fundamental component in various hybrid intelligent systems with multiparadigms because it can provide its users with diverse, flexible, and powerful ways to define and perform various reasoning tasks. This paper presents our basic ideas to design and implement FreeEnCal, facilities provided by FreeEnCal, and some applications of FreeEnCal.

2 Requirement Analysis

Our primary purpose is to design and implement a forward reasoning engine with general-purpose such that it can be used as a core and fundamental component to develop various advanced knowledge-based systems. To this end, we analyzed essential requirements the forward reasoning engine should satisfy, and now present them as follows:

ER1: Because different knowledge-based systems may work based on different logical criteria of validity provided by different logic systems, the reasoning engine should be able to draw fragments of various logic systems in order to provide its users with various logical criteria.

ER2: Because different knowledge-based systems may work based on different representation forms of various logic systems, the reasoning engine should be able to accept various formalizations of logic systems in order to match the needs from various applications.

ER3: Because different knowledge-based systems may work based on different reasoning forms, the reasoning engine should be able to perform deductive, inductive, and abductive reasoning automatically in order to match the needs for creation, discovery, or prediction.

ER4: Because different knowledge-based systems may work based on different formal theories of various scientific and/or engineering disciplines, the reasoning engine should be able to draw empirical theorems in various formal theories based on various logic systems, from given empirical premises in order to match the needs from various disciplines.

From the above essential requirements, we can derive some concrete requirements the forward reasoning engine must satisfy as follows:

CR1: The reasoning engine must accept a formal language and interpret specifications written in the formal language such that its users can use the formal language to describe and represent formulas and inference rules of various logic

systems with all sorts of formalizations as well as various formal theories based on the logic systems.

CR2: The reasoning engine must provide an easy way for its users to customize reasoning tasks by providing different axioms, inference rules, and facts with all sorts of formalizations as premises to the reasoning engine to perform various reasoning tasks.

CR3: The reasoning engine must provide an easy way for its users to control the progress of any reasoning task such that a user can restrict the amount of conclusions and/or the time of reasoning task to suspend, resume, and stop the reasoning task.

3 Facilities Provided by FreeEnCal

We define some notions and notations at first in order to present the facilities provided by FreeEnCal formally and precisely.

In general, a ***formal logic system*** **L** consists of a formal language, called the ***object language*** and denoted by F(**L**), which is the set of all ***well-formed formulas*** of **L**, and a ***logical consequence relation***, denoted by meta-linguistic symbol $\vdash_{\mathbf{L}}$, such that for $P \subseteq F(\mathbf{L})$ and $c \in F(\mathbf{L})$, $P \vdash_{\mathbf{L}} c$ means that within the framework of **L**, c is a valid conclusion of premises P, or given P as premises, c as a valid conclusion follows from P. For a formal logic system $(F(\mathbf{L}), \vdash_{\mathbf{L}})$, a ***logical theorem*** t is a formula of **L** such that $\phi \vdash_{\mathbf{L}} t$ where ϕ is the empty set. We use Th(**L**) to denote the set of all logical theorems of L. Th(**L**) is completely determined by the logical consequence relation $\vdash_{\mathbf{L}}$. According to the representation of the logical consequence relation of a logic, the logic can be represented as a Hilbert style axiomatic system, a Gentzen natural deduction system, a Gentzen sequent calculus system, or other type of formal system. Let $(F(\mathbf{L}), \vdash_{\mathbf{L}})$ be a formal logic system and $P \subseteq F(\mathbf{L})$ be a non-empty set of ***sentences*** (i.e. ***closed well-formed formulas***). A ***formal theory*** with premises P based on **L**, called a ***L-theory with premises P*** and denoted by $T_{\mathbf{L}}(P)$, is defined as $T_{\mathbf{L}}(P) =_{df} Th(\mathbf{L}) \cup Th_{\mathbf{L}}^{e}(P)$, and $Th_{\mathbf{L}}^{e}(P) =_{df} \{et \mid P \vdash_{\mathbf{L}} et \text{ and } et \notin Th(\mathbf{L})\}$ where Th(**L**) and $Th_{\mathbf{L}}^{e}(P)$ are called the ***logical part*** and the ***empirical part*** of the formal theory, respectively, and any element of $Th_{\mathbf{L}}^{e}(P)$ is called an ***empirical theorem*** of the formal theory.

In logic, a sentence in the form of 'if ... then ...' is usually called a ***conditional proposition*** or simply ***conditional*** which states that there exists a relation of sufficient condition between the 'if' part and the 'then' part of the sentence. In general, a conditional must concern two parts which are connected by the connective 'if ... then ...' and called the ***antecedent*** and the ***consequent*** of that conditional, respectively. The truth of a conditional depends not only on the truth of its antecedent and consequent but also, and more essentially, on a necessarily relevant and conditional relation between them. The notion of conditional plays the most essential role in reasoning because any reasoning form must invoke it, and therefore, it is historically always the most important subject studied in logic and is regarded as the heart of logic. In a formal logic system with a primitive connective to represent the notion of conditional, a formula may have a structure of deeply nested conditional. With the depth of nested conditional, a formal logic system can be divided into different fragments about different degree of nested conditional.

For a formal logic system where the notion of conditional is represented by a primitive connective, say '⇒', a formula is called a ***zero degree formula*** if and only if there is no occurrence of ⇒ in it; a formula of the form $A⇒B$ or $Q(A⇒B)$ where Q is the quantifier prefix of $A⇒B$ is called a ***first degree conditional*** if and only if both A and B are zero degree formulas; a formula A is called a ***first degree formula*** if and only if it satisfies the one of the following conditions: (1) A is a first degree conditional, (2) A is in the form $+B$ or $Q(+B)$ where + is a one-place connective such as negation and so on and Q is the quantifier prefix of $+B$ such that B is a first degree formula, and (3) A is in the form $B*C$ or $Q(B*C)$ where * is a non-conditional two-place connective such as conjunction or disjunction and so on and Q is the quantifier prefix of $B*C$ such that both of B and C is first degree formulas, or one of B and C is a first degree formula and another is a zero degree formula.

The notion of degree of nested conditional can also be generally defined as follows. Let *A, B, C* are formulas. The degree of conditional in A, denoted by $D_{⇒}(A)$, is inductively defined as follows: (1) $D_{⇒}(A) = 0$ if and only if there is no occurrence of ⇒ in A, (2) if A is a conditional of the form $B⇒C$, then $D_{⇒}(A) = \max\{ D_{⇒}(B), D_{⇒}(C) \} + 1$, (3) if A is in the form $+B$ where + is a one-place connective such as negation and so on, then $D_{⇒}(A) = D_{⇒}(B)$, (4) if A is in the form $B*C$ where * is a non-conditional two-place connective such as conjunction or disjunction and so on, then $D_{⇒}(A) = \max\{ D_{⇒}(B), D_{⇒}(C) \}$, and (5) if A is in the form QB where Q is the quantifier prefix of B, then $D_{⇒}(A) = D_{⇒}(B)$. Let k be a natural number. A formula A is called a ***k^{th} degree formula*** if and only if $D_{⇒}(A) = k$, in particular, A is called a ***k^{th} degree conditional*** if it is a conditional.

Let $(F(\mathbf{L}), ⊢_{\mathbf{L}})$ be a formal logic system and k be a natural number. The ***k^{th} degree fragment*** of $\mathbf{L}$, denoted by $Th^k(\mathbf{L})$, is a set of logical theorems of $\mathbf{L}$ which is inductively defined as follows (in the terms of Hilbert style axiomatic system): (1) if A is an axiom of $\mathbf{L}$ and $D_{⇒}(A) ≤ k$, then $A ∈ Th^k(\mathbf{L})$, (2) if A is the result of applying an inference rule of $\mathbf{L}$ to some members of $Th^k(\mathbf{L})$ and $D_{⇒}(A) ≤ k$, then $A ∈ Th^k(\mathbf{L})$, and (3) Nothing else are members of $Th^k(\mathbf{L})$, i.e., only those obtained from repeated applications of (1) and (2) are members of $Th^k(\mathbf{L})$. Note that the kth degree fragment of logic $\mathbf{L}$ does not necessarily include all k^{th} degree logical theorems of $\mathbf{L}$ because it is possible for $\mathbf{L}$ that deductions of some k^{th} degree logical theorems of $\mathbf{L}$ must invoke those logical theorems whose degrees are higher than k. On the other hand, the following holds obviously: $Th^0(\mathbf{L}) ⊂ Th^1(\mathbf{L}) ⊂ ... Th^{k-1}(\mathbf{L}) ⊂ Th^k(\mathbf{L}) ⊂ Th^{k+1}(\mathbf{L}) ⊂ ...$.

Based on the notion of degree of nested conditional and the notion of degree of logic fragment about conditional, we can introduce the notion of degree of formal theory about conditional.

Let $(F(\mathbf{L}), ⊢_{\mathbf{L}})$ be a formal logic system, $P ⊂ F(\mathbf{L})$ ($P ≠ \phi$), and k and j be two natural numbers. A formula A is said to be ***j^{th}-degree-deducible from P based on $Th^k(\mathbf{L})$*** if and only if there is a finite sequence of formulas $f_1, ..., f_n$ such that $f_n = A$ and for all i ($i ≤ n$), (1) $f_i ∈ Th^k(\mathbf{L})$, or (2) $f_i ∈ P$, or (3) f_i is the result of applying an inference rule to some members $f_{j1}, ..., f_{jm}$ ($j1, ..., jm < i$) of the sequence and $D_{⇒}(f_i) ≤ j$. The set of all formulas which are j^{th}-degree-deducible from P based on $Th^k(\mathbf{L})$ is called the ***j^{th} degree fragment of the formal theory with premises P based on $Th^k(\mathbf{L})$***, denoted by $T^j_{Th^k(\mathbf{L})}(P)$. A formula is said to be ***j^{th}-degree-deductive from P based***

on Th^k(L) if and only if it is j^{th}-degree-deducible from P based on $\text{Th}^k(\mathbf{L})$ but not $(j-1)^{\text{th}}$-degree-deducible from P based on $\text{Th}^k(\mathbf{L})$. Note that in the above definitions, we do not require $j \leq k$. The notion of jth-degree-deductive can be regarded as a metric to measure the difficulty of deducing an empirical theorem from given premises P based on logic **L**. The difficulty is relative to the complexity of problem being investigated as well as the strength of underlying logic **L**.

The notion of degree of nested conditional can be generally extended to other logic connectives and modal operators as follows: Let θ be an arbitrary n-ary ($1 \leq n$) connective or modal operator of logic **L** and A be a formula of **L**, the degree of θ in A, denoted by $\text{D}_\theta(A)$, is defined as follows: (1) $\text{D}_\theta(A) = 0$ if and only if there is no occurrence of θ in A, (2) if A is in the form $\theta(a_1, a_2, \ldots, a_n)$ where $a_1, a_2, \ldots, a_n$ are formulas, then $\text{D}_\theta(\text{A}) = \max\{ \text{D}_\theta(a_1), \text{D}_\theta(a_2), \ldots, \text{D}_\theta(a_n) \} + 1$, (3) if A is in the form $\sigma(a_1, a_2, \ldots, a_n)$ where σ is a connective or modal operator different from θ and $a_1, a_2, \ldots, a_n$ are formulas, then $\text{D}_\theta(A) = \max\{ \text{D}_\theta(a_1), \text{D}_\theta(a_2), \ldots, \text{D}_\theta(a_n) \}$, and (4) if A is in the form QB where B is a formula and Q is the quantifier prefix of B, then $\text{D}_\theta(A) = \text{D}_\theta(B)$.

The notion of degree of logic fragment about conditional can also be generally extended to other logic connectives and modal operators as follows: Let $\theta_1, \theta_2, \ldots, \theta_n$ be connectives or modal operators of logic **L** and $k_1, k_2, \ldots, k_n$ be natural numbers, the fragment of **L** about $\theta_1, \theta_2, \ldots, \theta_n$ and their degrees $k_1, k_2, \ldots, k_n$, denoted by $\text{Th}^{(\theta_1, k_1, \theta_2, k_2, \ldots, \theta_n, k_n)}(\mathbf{L})$, is a set of logical theorems of **L** which is inductively defined as follows (in the terms of Hilbert style axiomatic system): (1) if A is an axiom of **L** and $\text{D}_{\theta_1}(\text{A}) \leq k_1, \text{D}_{\theta_2}(\text{A}) \leq k_2, \ldots, \text{D}_{\theta_n}(\text{A}) \leq k_n$, then $A \in \text{Th}^{(\theta_1, k_1, \theta_2, k_2, \ldots, \theta_n, k_n)}(\mathbf{L})$, (2) if A is the result of applying an inference rule of **L** to some members of $\text{Th}^{(\theta_1, k_1, \theta_2, k_2, \ldots, \theta_n, k_n)}(\mathbf{L})$ and $\text{D}_{\theta_1}(\text{A}) \leq k_1, \text{D}_{\theta_2}(\text{A}) \leq k_2, \ldots, \text{D}_{\theta_n}(\text{A}) \leq k_n$, then $A \in \text{Th}^{(\theta_1, k_1, \theta_2, k_2, \ldots, \theta_n, k_n)}(\mathbf{L})$, and (3) Nothing else are members of $\text{Th}^{(\theta_1, k_1, \theta_2, k_2, \ldots, \theta_n, k_n)}(\mathbf{L})$. Similarly, the notion of degree of formal theory about conditional can also be generally extended to other logic connectives and modal operators.

Based on the above notions and notations, we now present the major facilities provided by FreeEnCal as follows:

F1: FreeEnCal accepts a formal language we designed and can interpret specifications written in the formal language such that any user can use the formal language to describe and represent formulas and inference rules for deductive, inductive, and abductive reasoning. In the case of deductive reasoning, the formal language can be used to describe and represent formulas and inference rules of various logic systems (formalized as a Hilbert style axiomatic system, a Gentzen natural deduction system, or a Gentzen sequent calculus system) as well as various formal theories based on the logic systems.

F2: For a logic system **L**, its connectives and/or modal operators (e.g., $\theta_1, \theta_2, \ldots, \theta_n$), and degrees of the connectives and/or modal operators (e.g., $k_1, k_2, \ldots, k_n$), all of which are specified by users, FreeEnCal can reason out in principle all logical theorem schemata of the fragment $\text{Th}^{(\theta_1, k_1, \ldots, \theta_n, k_n)}(\mathbf{L})$, and reason out in practice all or a part of the logical theorem schemata under the control conditions attached to the reasoning task specified by users.

F3: For a formula A and the fragment $\mathrm{Th}^{(\theta_1, k_1, \ldots, \theta_n, k_n)}(\mathbf{L})$ of a logic system **L** specified by users, FreeEnCal can verify whether or not A is a logical theorem of $\mathrm{Th}^{(\theta_1, k_1, \ldots, \theta_n, k_n)}(\mathbf{L})$, if yes, then give the proof.

F4: For a logic system **L**, its connectives and/or modal operators (e.g., θ_1, θ_2, …, θ_n), degrees of the connectives and/or modal operators (e.g., k_1, k_2, …, k_n), empirical inference rules (which may be deductive, inductive, or abdactive), a set P of formulas as empirical premises, and degrees of connectives and/or modal operators in empirical theorems and facts, all of which are specified by users, FreeEnCal can reason out in principle all empirical theorems and facts satisfying the conditions about degrees from P based on $\mathrm{Th}^{(\theta_1, k_1, \ldots, \theta_n, k_n)}(\mathbf{L})$, and reason out in practice all or a part of the empirical theorems and facts under the control conditions attached to the reasoning task specified by users.

F5: For a formula A and the fragment of a formal theory with premises P based on the fragment $\mathrm{Th}^{(\theta_1, k_1, \ldots, \theta_n, k_n)}(\mathbf{L})$ of a logic system **L** specified by users, FreeEnCal can verify whether or not A is an empirical theorem of the fragment of the formal theory, if yes, then give the proof.

F6: For any reasoning task, FreeEnCal can restrict the amount of conclusions and/or the time of reasoning to suspend, resume, and stop the reasoning task according to instructions issued by users.

Basically, FreeEnCal can be regarded as a general-purpose interpreter and executor of inference rules, which may be deductive, inductive, or abdactive, for various logic systems as well as various formal theories based on the logic systems.

In principle, FreeEnCal can deal with various logic systems including classical mathematical logic, its various classical conservative extensions, and various philosophical logics [11-13]. However, from the viewpoint of application, the most effective way to use FreeEnCal is to adopt relevant (relevance) logics, in particular, strong relevant (relevance) logics [3, 4, 6] and their various conservative extensions as the fundamental logic systems to perform reasoning based on FreeEnCal. This is because FreeEnCal is a forward reasoning engine with general-purpose that may reason out a lot of conclusions irrelevant to the premises of reasoning if the fundamental logic system does not provide the guarantee about relevance relationship between the premises and conclusions of reasoning.

4 Applications of FreeEnCal

FreeEnCal can be used as a ready-made forward reasoning engine serving as a core and fundamental component in various advanced knowledge-based systems as well as an alone forward reasoning engine with general-purpose. We have used FreeEnCal in some different application systems.

The first application where FreeEnCal plays an important role is to implement anticipatory reasoning-reacting systems [9, 14, 15, 17].

An anticipatory reasoning-reacting system (ARRS for short) is a computing system containing a controller C with capabilities to measure and monitor the behavior of the whole system, a traditional reactive system RS, a predictive model PM of RS and its external environment, and an anticipatory reasoning engine ARE such that according

to predictions by ARE based on PM, C can order and control RS to carry out some operations with a high priority [5]. Anticipatory reasoning is reasoning to draw new, previously unknown and/or unrecognized conclusions about some future event or events whose occurrence and truth are uncertain at the point of time when the reasoning is being performed [5]. The most important components in ARRSs are the PM and the ARE because an ARRS cannot behave anticipatorily if it cannot predict the future circumstances. The PM must be dependent on the RS and its application area, while the ARE may be an application-independent general one. We used FreeEnCal to implement an anticipatory reasoning engine in our prototype implementation of an anticipatory reasoning-reacting system [15, 17].

The notion of anticipatory system, in particular, computing anticipatory system and anticipatory reasoning-reacting system, implies a fundamental assumption or requirement, i.e., to be anticipatory, a computing system must have the ability to make decision about its next action (i.e., to make present change of state) based on prediction about its potential future states. To establish a general methodology for designing and developing computing anticipatory systems, we have proposed a deontic relevant logic approach to reasoning about actions in computing anticipatory systems. We used FreeEnCal to implement an action-reasoning engine based on deontic relevant logic [14]. Our experiences lead us to conclude that FreeEnCal can also play an important role in a planning system in robot and/or agent applications where a decision maker is indispensable.

An ARRS with mobile objects, e.g., an ARRS for air traffic control, an ARRS for train control, an ARRS for elevator control, and so on, must concern not only the notion of time but also the notion of space. To let an ARRS have the ability to reason about mobile objects, we have proposed a spatio-temporal relevant logic approach to reasoning about mobile objects in ARRSs. We used FreeEnCal to implement a spatio-temporal reasoning engine based on spatio-temporal relevant logic [9].

The second application where FreeEnCal plays an important role is a prototype implementation of an epistemic programming system [18]. Conventional programming regards numeric values and/or character strings as the subject of computing, takes assignments as basic operations of computing, and regards algorithm as the subject of programming, while epistemic programming regards beliefs as the subject of computing, takes primary epistemic operations as basic operations of computing, and regards epistemic processes as the subject of programming [3]. We have designed an epistemic programming language, named "EPLAS", and implemented a prototype system for EPLAS [18]. All epistemic operations, in particular, deductive, inductive, and abductive reasoning, are implemented as system calls to FreeEnCal.

Our primary purpose to develop EnCal is to solve the well-known difficult problem of automated theorem finding [3]. However, because EnCal cannot deal with modal notions, we could not use it to solve the problems concerning modal notions. Now, having FreeEnCal as a powerful reasoning engine dealing with modal logics, we can use it to solve the problems concerning modal notions. We also proposed a cooperative grid computing approach to automated theorem finding and automated problem proposing: to build the Theory Grid as cooperatively shared formal theories within a virtual organization, and then to implement Grid Theorists as the wisdom of crowds working based on the Theory Grid such that the grid theorists have the ability

to find new theorems and propose new questions automatically or semi-automatically. In the approach, each agent and/or client use FreeEnCal to perform deductive, inductive or abductive reasoning tasks to find candidates of interesting theorems and questions [8].

We are developing an autonomous evolutionary information system for teaching and learning logic, named "HILBERT", where FreeEnCal also plays several important role as a reasoning engine, a prover, an inference rule generator, and a question generator [7].

Based on our experiences of design, development, and applications of FreeEnCal, we can say that FreeEnCal is a forward reasoning engine with general-purpose really useful in various advanced knowledge-based systems with purposes of creation, discovery, or prediction.

5 Concluding Remarks

The first forward deduction engine is "Logic Theory Machine", developed by Newell, Shaw and Simon in 1957, which tried to deduce all logical theorems of propositional classical mathematical logic [10]. As a well-known fact, the Logic Theory Machine was not successful due to the problem of computational complexity. This (and the resolution method discovered by Robinson) led almost all researchers to give up the approach of forward deduction but adopt the more efficient approach of backward deduction [16]. However, from the viewpoint of logic validity of reasoning, the failure of Logic Theory Machine is caused by classical mathematical logic rather than forward deduction. If we want to create, discover, or predict some new things rather than prove some things previously specified, then the only way is to ask forward reasoning [2-6].

We have presented our basic ideas to design and implement FreeEnCal, facilities provided by FreeEnCal, and some applications of FreeEnCal. This paper is the first formal presentation of FreeEnCal. There are many challenging interesting problems that have to be solved in order to widely apply FreeEnCal to more advanced knowledge-based systems in the real world.

References

1. Advanced Information Systems Engineering Laboratory, Saitama University: FreeEnCal Reference Manual. Saitama (2007)
2. Cheng, J.: EnCal: An Automated Forward Deduction System for General-Purpose Entailment Calculus. In: Terashima, N., Altman, E. (eds.) Advanced IT Tools, IFIP World Conference on IT Tools, IFIP '96 14th World Computer Congress, Canberra, Australia, September 2-6, 1996, pp. 507–514. Chapman & Hall, London (1996)
3. Cheng, J.: A Strong Relevant Logic Model of Epistemic Processes in Scientific Discovery. In: Kawaguchi, E., Kangassalo, H., Jaakkola, H., Hamid, I.A. (eds.) Information Modelling and Knowledge Bases XI. Frontiers in Artificial Intelligence and Applications. Frontiers in Artificial Intelligence and Applications, vol. 61, pp. 136–159. IOS Press, Amsterdam (2000)

4. Cheng, J.: Automated Knowledge Acquisition by Relevant Reasoning Based on Strong Relevant Logic. In: Palade, V., Howlett, R.J., Jain, L.C. (eds.) KES 2003. LNCS, vol. 2773, pp. 68–80. Springer, Heidelberg (2003)
5. Cheng, J.: Temporal Relevant Logic as the Logical Basis of Anticipatory Reasoning-Reacting Systems. In: Dubois, D.M. (ed.) Computing Anticipatory Systems: CASYS 2003 - Sixth International Conference, AIP Conference Proceedings Liege, Belgium, August 11-16, 2005. American Institute of Physics, vol. 718, pp. 362–375 (2004)
6. Cheng, J.: Strong Relevant Logic as the Universal Basis of Various Applied Logics for Knowledge Representation and Reasoning. In: Kiyoki, Y., Henno, J., Jaakkola, H., Kangassalo, H. (eds.) Information Modelling and Knowledge Bases XVII. Frontiers in Artificial Intelligence and Applications, vol. 136, pp. 310–320. IOS Press, Amsterdam (2006)
7. Cheng, J., Akimoto, N., Goto, Y., Koide, M., Nanashima, K., Nara, S.: HILBERT: An Autonomous Evolutionary Information System for Teaching and Learning Logic. In: Constantinou, C.P., Zacharia, Z.C. (eds.) Computer Based Learning in Science, Conference Proceedings 2003, 6th International Conference, CBLIS '03, July 5-10, 2003, vol. 1, pp. 245–254. University of Cyprus, Nicosia (2003)
8. Cheng, J., Nara, S., Koh, T., Goto, Y.: The Theory Grid and Grid Theorists. In: Proc. 2nd International Conference on Semantics, Knowledge and Grid, IEEE Computer Society, Los Alamitos (2006)
9. Cheng, J., Goto, Y., Kitajima, N.: Anticipatory Reasoning about Mobile Objects in Anticipatory Reasoning-Reacting Systems. In: 8th International Conference on Computing Anticipatory Systems, Liege (2007)
10. Davis, M.: The Early History of Automated Deduction. In: Robinson, J.A., Voronkov, A. (eds.) Handbook of Automated Reasoning, vol. 1, pp. 3–15. Elsevier Science B.V, Amsterdam (2001)
11. Gabbay, D., Guenthner, F. (eds.): Handbook of Philosophical Logic, 2nd edn., pp. 1–13. Kluwer Academic, Dordrecht (2001-2005)
12. Goble, L. (ed.): The Blackwell Guide to Philosophical Logic. Blackwell, Oxford (2001)
13. Jacquette, D. (ed.): A Companion to Philosophical Logic. Blackwell, Oxford (2002)
14. Kitajima, N., Nara, S., Goto, Y., Cheng, J.: A Deontic Relevant Logic Approach to Reasoning about Actions in Computing Anticipatory Systems. In: 8th International Conference on Computing Anticipatory Systems, Liege (2007)
15. Nara, S., Shang, F., Omi, T., Goto, Y., Cheng, J.: An Anticipatory Reasoning Engine for Anticipatory Reasoning-Reacting Systems. International Journal of Computing Anticipatory Systems, CHAOS, Leige 18, 225–234 (2006)
16. Robinson, J.A., Voronkov, A. (eds.): Handbook of Automated Reasoning, pp. 1–2. Elsevier Science B.V., Amsterdam (2001)
17. Shang, F., Nara, S., Omi, T., Goto, Y., Cheng, J.: A Prototype Implementation of an Anticipatory Reasoning-Reacting System. In: Dubois, D.M. (ed.): Computing Anticipatory Systems: CASYS 2005 - Seventh International Conference, Liege, Belgium, AIP Conference Proceedings August 8-13, 2005, American Institute of Physics, vol. 839, pp. 401–414 (2006)
18. Takahashi, I., Nara, S., Goto, Y., Cheng, J.: EPLAS: An Epistemic Programming Language for All Scientists. In: Shi, Y., et al. (eds.) Computational Science - ICCS 2007: 7th International Conference, Proceedings, Part I, Beijing, China, May 27-30, 2007. LNCS, vol. 4487, pp. 406–413. Springer, Heidelberg (2007)

Hybrid Agents Based Architecture on Automated Dynamic Environments

Dante I. Tapia, Javier Bajo, Juan M. Corchado, Sara Rodríguez, and Juan M. Manzano

Departamento Informática y Automática, Universidad de Salamanca
Plaza de la Merced s/n, 37008, Salamanca, Spain
{dantetapia, jbajope, corchado, srg}@usal.es,
jms.manzano@gmail.com

Abstract. This paper presents a hybrid agents based architecture that uses case-based reasoning and case-based planning systems as reasoning mechanism in deliberative BDI agents to manage daily tasks and hardware devices on dynamic environments. This architecture is also founded on Ambient Intelligence, joining a set of technologies, such as radio frequency identification, wireless networks and wireless control devices, which can be implemented on a wide diversity of environments.

Keywords: Hybrid Systems, Case-Based Reasoning, Case-Based Planning, Ambient Intelligence, Automation.

1 Introduction

Agents and multi-agent systems (MAS) have become increasingly relevant for developing distributed and dynamic open systems. This paper is focused on describing the main characteristics of an hybrid agents based architecture that use deliberative BDI (Believe, Desire, Intention) agents with Case-Based Reasoning (CBR) and Case-Based Planning (CBP) as reasoning mechanisms as a way to implement adaptive systems on automated dynamic environments. It is also examined Radio Frequency Identification (RFID), wireless networks (Wi-Fi), and ZigBee (802.15.4) devices that can provide the agents real time information about the environment to create plans to manage and automate daily tasks and physical services.

This architecture is also founded on Ambient Intelligence (AmI) environments, characterized by their ubiquity, transparency and intelligence. Ambient Intelligence proposes a new way to interact between people and technology, where this last one is adapted to individuals and their context, showing a vision where people are surrounded by intelligent interfaces merged in daily life objects [9], creating a computing-capable environment with intelligent communication and processing to the service of people by means of a simple, natural, and effortless human-system interaction for users [16], reason why to develop intelligent and intuitive systems and interfaces, capable to recognize and respond to the users necessities in a ubiquous way [8], considering people in the centre of the development [17], and creating

B. Apolloni et al. (Eds.): KES 2007/ WIRN 2007, Part II, LNAI 4693, pp. 453–460, 2007.
© Springer-Verlag Berlin Heidelberg 2007

technologically complex environments in medical, domestic, academic, etc. fields. [19]. Agents in this context must be able to respond to events, take the initiative according to their goals, communicate with other agents, interact with users, and make use of past experiences to find the best plans to achieve goals.

This work is aimed to develop an architecture using hybrid agents, implementing CBR and CBP reasoning mechanisms into deliberative BDI agents. CBP reasoning mechanism derivates from CBR systems [1] and are specially designed for planning construction. CBP-BDI facilitates learning and adaptation, and therefore a greater degree of autonomy than that found in pure BDI (Believe, Desire, Intention) architecture [4]. A CBR-BDI agent [6] uses case-based reasoning as a reasoning mechanism, which allows it to learn from initial knowledge, to interact autonomously with the environment as well as with users and other agents within the system, and to have a large capacity for adaptation to the needs of its surroundings. We shall refer to the CBR-BDI agents specialized in generating plans as CBP-BDI agents. BDI agents can be implemented by using different tools, such as Jadex [14]. Jadex agents deal with the concepts of beliefs, goals and plans; they are java objects that can be created and handled within the agent at execution time.

2 Developing a Hybrid Agents Based Architecture

The hybrid agents developed are based on the BDI (Belief, Desire, Intention) deliberative architecture model [4], where the internal structure and capabilities of the agents are based on mental aptitudes, using beliefs, desires and intentions. The hybrid architecture model presented in this paper facilitates the incorporation of CBR systems [2] as a deliberative mechanism within BDI agents, facilitating learning and adaptation and providing a greater degree of autonomy than pure BDI architecture. CBR is a type of reasoning based on the use of past experiences [13] to solve new problems by adapting solutions that have been used to solve similar problems in the past, and learn from each new experience. To introduce a CBR motor into a deliberative BDI agent it is necessary to represent the cases used in a CBR system by means of beliefs, desires and intentions, and implement a CBR cycle. The structure of a CBR system is composed of six elements, as seen on figure 1.

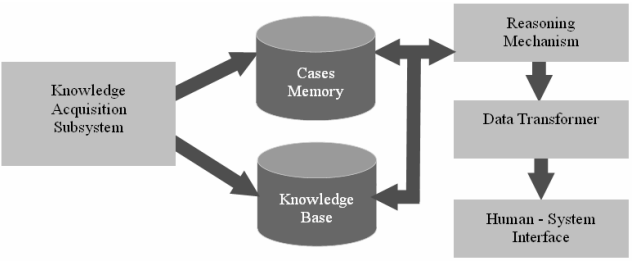


Fig. 1. CBR system structure

The primary concept when working with CBR systems is the concept of case, which is described as a past experience composed of three elements: an initial state or

problem description that is represented as a belief; a solution, that provides the sequence of actions carried out in order to solve the problem; and a final state that is represented as a set of goals. CBR manages cases (past experiences) to solve new problems. The way cases are managed is known as the CBR cycle, and consists of four sequential phases: retrieve, reuse, revise and retain. The retrieve phase starts when a new problem description is received. Similarity algorithms are applied in order to retrieve from the cases memory the cases with a problem description more similar to the current one. Once the most similar cases have been retrieved, the reuse phase begins, adapting the solutions for the retrieved cases to obtain the best solution for the current case. The revise phase consists of an expert revision of the solution proposed. Finally, the retain phase allows the system to learn from the experiences obtained in the three previous phases and consequently updates the cases memory.

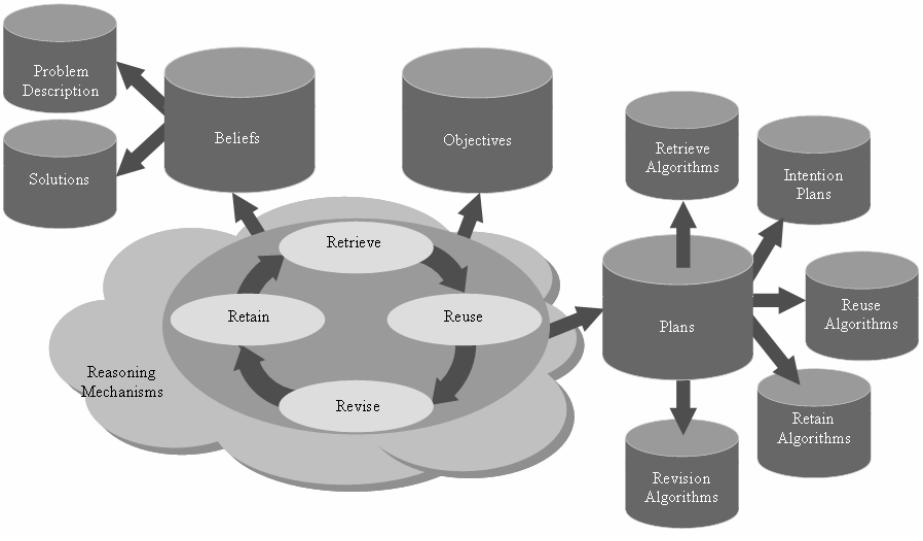


Fig. 2. CBP-BDI deliberative agent internal structure

In a planning agent, the reasoning motor generates plans using past experiences and planning strategies, so the concept of Case Base Planning is obtained [10]. CBP consists of four sequential stages: retrieve stage to recover the most similar past experiences to the current one; reuse stage to combine the retrieved solutions in order to obtain a new optimal solution; revise stage to evaluate the obtained solution; and retain stage to learn from the new experience. Case-based planning (CBP) is the idea of planning as remembering [11]. CBP is a specialization of case-based reasoning (CBR) which is a problem solving methodology based on using a library of solutions for similar problems [11]. In CBP, the solution proposed to solve a given problem is a plan, so this solution is generated taking into account the plans applied to solve similar problems in the past. The problems and their corresponding plans are stored in a plans memory. Figure 2 shows the internal structure of a CPB-BDI agent. Problem description (initial state) and solution (situation when final state is achieved) are

represented as beliefs, the final state as a goal (or set of goals), and the sequences of actions as plans. The CBP cycle is implemented through goals and plans. When the goal corresponding to one of the stages is triggered, different plans (algorithms) can be executed concurrently to achieve the goal or objective. Each plan can trigger new sub-goals and, consequently, cause the execution of new plans.

3 Dynamic Environment Automation: Technologies

The essential aspect in this work is the development of AmI based multi-agent systems that implement hybrid CBR-BDI and CBP-BDI deliberative agents over automated and dynamic environments. Thus the use of technologies that provide the agents automatic and real time information of the context, and allow them to react, is also important. Ambient Intelligence (AmI) provides an effective way to create systems with the ability to adapt themselves to the context and users necessities. The vision of AmI assumes seamless, unobtrusive, and often invisible but also controllable interactions between humans and technology. AmI provides new possibilities for resolving a wide range of problems and proposes a new way to interact between people and technology, where this last one is adapted to individuals and their context, showing a vision where people are surrounded by intelligent interfaces merged in daily life objects [9], creating a computing-capable environment with intelligent communication and processing to the service of people by means of a simple, natural, and effortless human-system interaction for users [16]. One of the most benefited segments of population with the appearance of AmI systems will be the elderly and people with disabilities, improving important aspects of their daily life, especially on health care [9]. Next, three main technologies merged in this architecture are described.

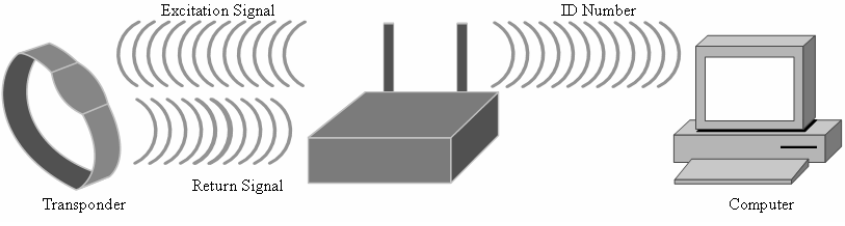


Fig. 3. Functioning of RFID technology

RFID technology is a wireless communications technology used to identify and receive information about humans, animals and objects on the move. An RFID system contains basically four components: tags, readers, antennas and software. Tags with no power system (batteries) integrated are called passive tags or “transponders”, these are much smaller and cheaper than tags with some kind of power system included, but have shorter read range. Figure 3 shows how these four elements combined enable the translation of information to a user-friendly format. The transponder is placed on the object itself (i.e. bracelet). As this object moves into the reader’s capture area, the reader is activated and begins signalling via electromagnetic waves (radio frequency).

The transponder subsequently transmits its unique ID information number to the reader, which transmit it to a device or a central computer where the information is processed and showed. This information is not restricted to the location of the object, and can include specific detailed information concerning the object itself. The most use is in industrial/manufacturing, transportation, distribution, etc., but there are other growth sectors including health care [18]. The configuration presented in this paper consists of a transponder mounted on a bracelet worn on the users' wrist or ankle, and several sensors installed over protected zones, with an adjustable capture range up to 2 meters, and a central computer where all the information is processed and stored.

Wireless LANs, also known as Wi-Fi (Wireless Fidelity) networks can be used to replace or as an extension of wired LANs. They provide reduced infrastructure and low installation cost, and also give more mobility and flexibility by allowing people to stay connected to the network as they roam among covered areas, increasing efficiency by allowing data to be entered and accessed on site [12]. The growing use of handheld devices in recent years has led to new necessities as well as to a great opportunity to extend traditional commerce techniques and apply new techniques. These new devices facilitate the use of new interaction techniques, for instance, some systems focus on facilitating users with guidance or location systems [7], [15] by means of their wireless devices. The architecture of the multi-agent system presented in this paper incorporates "lightweight" agents that can reside in mobile devices, such as phones, PDA's, etc. [3], and therefore support wireless communication (Wi-Fi, Bluetooth) which facilitates the portability to a wide range of devices.

ZigBee is a low cost, low power consumption, two-way, wireless communication standard developed by the ZigBee Alliance, based on IEEE 802.15.4 protocol, and operates at 868/915 & 2.4GHz spectrum. ZigBee is designed to be embedded in consumer electronics, home and building automation, industrial controls, PC peripherals, medical sensor applications, toys and games [20]. This technology is intended for home, building and industrial automation purposes, addressing the needs of monitoring, control and sensory network applications. ZigBee devices can be configured to act as network coordinator (control all devices), as router/repeater (send/receive/resend data to/from coordinator or end devices) or as end device (send/receive data to/from coordinator), connected by star, tree or mesh topologies.

4 Automated Environment System Prototype

All the elements described before are integrated into a generic multi-agent system prototype that can be re-modelled to be implemented on different scenarios. First, the essential hardware used and its basic characteristics are presented, and then all the system agents and their main functions are described.

The essential hardware used is: Sokymat's Q5 chip 125KHz RFID wrist bands and computer interface readers for people identification and location monitoring; Silicon Laboratories' C8051 chip-based 2.4GHz development boards for physical services automation (lights, door locks, alarms, etc.); mobile devices (PDA's) for interfaces and users' interaction; a Workstation where all the high demanding CPU tasks (planning and reasoning) are processed; and a basic Wi-Fi network for wireless communication between agents (in PDA's and Workstation). All the hardware is

some way integrated to agents, providing them real time information about the environment to create plans to manage and automate daily tasks and physical services.

The system is composed of five different agents, which have specific roles that can be changed according the location and the environment necessities where the system is implemented, for example in geriatric residences [5].

User Agent is a BDI agent that runs on mobile devices (PDA). It manages the users' personal data and behaviour (monitoring, location, daily tasks, and anomalies). The beliefs and goals used for every user depend on the plan or plans defined by the super-users. User Agent maintains continuous communication with the rest of the system agents, especially with the ScheduleUser Agent (through which the scheduled-users can communicate the result of their assigned tasks) and with the SuperUser Agent. The User Agent must ensure that all the actions indicated by the SuperUser are taken out, sending a copy of its memory base (goals and plans) to the Manager Agent in order to maintain backups.

SuperUser Agent is a BDI agent that runs on mobile devices (PDA). It inserts new tasks into the Manager Agent to be processed by the CBR mechanism. It also needs to interact with the User Agents to impose new tasks and receive periodic reports, and with the ScheduleUser Agents to ascertain plans' evolution.

ScheduleUser Agent is a hybrid CBP-BDI planner agent that runs on mobile devices (PDA). It schedules the users' daily activities obtaining dynamic plans depending on the tasks needed for each user. It manages scheduled-users profiles (preferences, habits, holidays, etc.), tasks, available time and resources. Every agent generates personalized plans depending on the scheduled-user profile.

Manager Agent is a hybrid CBR-BDI Agent that runs on a Workstation. It plays two roles: the security role that monitors the users' location trough RFID and manages the ZigBee control devices, and the manager role that handle the medical record database and the tasks assignation. It must provide security for the users and ensure the tasks assignments are efficient. This assignation is carried out through a CBR reasoning engine, which is incorporated within the Manager Agent. When a new assignation of tasks needs to be carried out, both past experiences and the needs of the current situation are recalled, allocating the respective and adequate task.

Devices Agent is a BDI agent that that runs on a Workstation. It monitors the users' location (continuously obtaining/updating data from the RFID readers), and interact with the ZigBee devices to receive information and control physical services (lights, door locks, etc.). The information obtained is sent to the Manager Agent to be processed.

5 Conclusions and Future Work

The hybrid architecture presented in this paper solves one of the problems of BDI (deliberative) architectures, which is the lack of learning capacity. The CBR systems reasoning cycle helps the agents to solve problems, adapt to changes in the environment, and identify new possible solutions.

The use of hybrid agents working in conjunction with RFID, Wireless, and ZigBee technologies, provide a ubiquous, intelligent and adaptable architecture that can be implemented in wide variety scenarios, such as hospitals, geriatric residences,

schools, homes or any dynamic environment where is a need to manage tasks and automate services, focusing on security and comfort. In fact, a prototype system, based on this architecture (figure 4), has been successfully implemented into a geriatric residence [5], improving the security and the health care efficiency through monitoring and automating medical staff's work and patients' activities, facilitating the assignation of working shifts and reducing time spent on routine tasks.

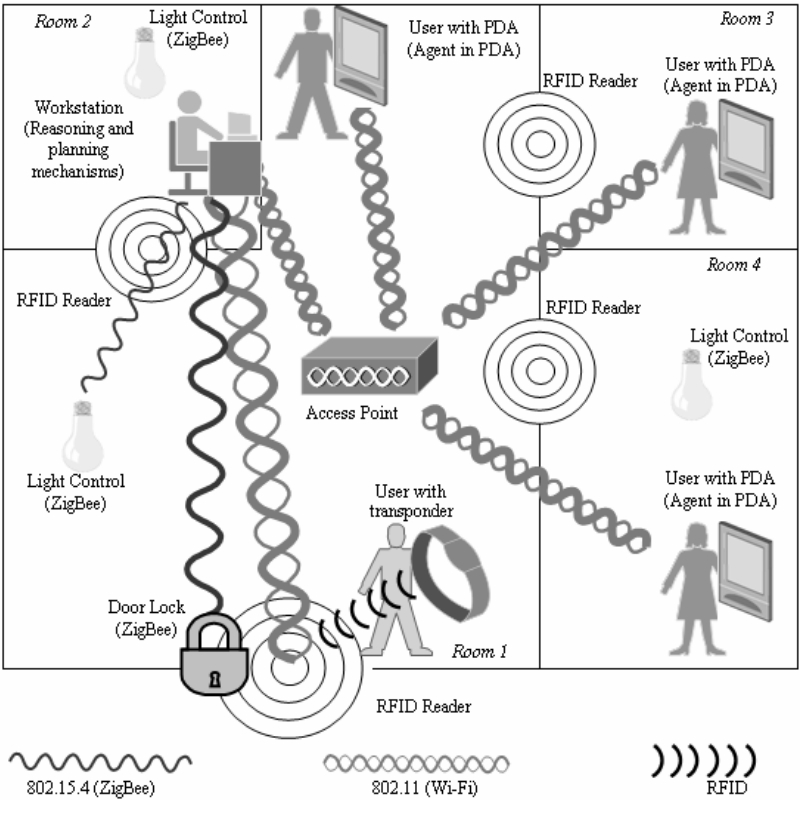


Fig. 4. Architecture applied on an automated environment

CBR-BDI and CBP-BDI agents, working in conjunction with RFID, wireless and automation devices, provide a ubiquous, non-invasive, high level interaction among users and the system, but it is necessary to continue developing and improving the architecture presented, adding new capabilities and integrating new technologies to create flexible and adaptable AmI-based systems to automate services and daily tasks.

Acknowledgements. This work has been partially supported by the MCYT TIC2003-07369-C02-02 and the JCYL-2002-05 project SA104A05. Special thanks to Sokymat by the RFID technology provided and to Telefónica Móviles (Movistar) for the wireless devices donated.

References

1. Aamodt, A., Plaza, E.: Case-Based Reasoning: foundational Issues, Methodological Variations, and System Approaches. AICOM 7, 39–59 (1994)
2. Allen, J.F.: Towards a general theory of action and time. Artificial Intelligence 23, 123–154 (1984)
3. Bohnenberger, T., Jacobs, O., Jameson, A.: DTP meets user requirements: Enhancements and studies of an intelligent shopping guide. In: Gellersen, H.-W., Want, R., Schmidt, A. (eds.) PERVASIVE 2005. LNCS, vol. 3468, Springer, Heidelberg (2005)
4. Bratman, M.E.: Intentions, Plans and Practical Reason. Harvard University Press, Cambridge, M.A. (1987)
5. Corchado, J.M., Bajo, J., Tapia, D.I.: ALZ-MAS: Alzheimer's special care multi-agent system. In: Proceedings of the workshop on Health Care. ECAI'06 (2006)
6. Corchado, J.M., Laza, R.: Constructing Deliberative Agents with Case-based Reasoning Technology. International Journal of Intelligent Systems. 18(12), 1227–1241 (2003)
7. Corchado, J.M., Pavón, J., Corchado, E., Castillo, L.F.: Development of CBR-BDI Agents: A Tourist Guide Application. In: Funk, P., González Calero, P.A. (eds.) ECCBR 2004. LNCS (LNAI), vol. 3155, pp. 547–559. Springer, Heidelberg (2004)
8. Ducatel, K., Bogdanowicz, M., Scapolo, F., Leijten, J., Burgelman, J.C.: That's what friends are for. Ambient Intelligence (AmI) and the IS in 2010. Innovations for an e-Society. Congress Pre-prints, Innovations for an e-Society. Challenges for Technology Assessment. Berlin, Germany (2001)
9. Emiliani, P.L., Stephanidis, C.: Universal access to ambient intelligence environments: opportunities and challenges for people with disabilities. IBM Systems Journal (2005)
10. Glez-Bedia, M., Corchado, J.M.: A planning strategy based on variational calculus for deliberative agents. Computing and Information Systems Journal. 10(1), 2–14 (2002)
11. Hammond, K.: Case-Base Planning: Viewing Planning as a Memory Task. Academic Press, New York (1989)
12. Hewlett-Packard.: Understanding Wi-Fi (2002), http://www.hp.com/rnd/library/pdf/
13. Kolodner, J.: Case-based reasoning. Morgan Kaufmann, San Francisco (1993)
14. Pokahr, A., Braubach, L., Lamersdorf, W.: Jadex: Implementing a BDI-Infrastructure for JADE Agents. In: EXP - In: Search of Innovation (Special Issue on JADE), Telecom Italia Lab, Turin, Italy, vol. 3, pp. 76–85 (September 2003)
15. Poslad, S., Laamanen, H., Malaka, R., Nick, A., Buckle, P., Zipf, A.: Crumpet: Creation of user- friendly mobile services personalised for tourism. In: Proceedings of 3G (2001)
16. Richter, K., Hellenschmidt, M.: Interacting with the Ambience: Multimodal Interaction and Ambient Intelligence. In: Position Paper to the W3C Workshop on Multimodal Interaction (July 19-20, 2004)
17. Schmidt, A.: Interactive Context-Aware Systems Interacting with Ambient Intelligence. In: Riva, G., Vatalaro, F., Davide, F., Alcañiz, M. (eds.) Ambient Intelligence, pp. 159–178. IOS Press, Amsterdam (2005)
18. Sokymat.: Sokymat (2006), http://sokymat.aaitg.com/
19. Susperregi, L., Maurtua, I., Tubío, C., Pérez, M.A., Segovia, I., Sierra, B.: Una arquitectura multiagente para un Laboratorio de Inteligencia Ambiental en Fabricación. 1er. Taller de Desarrollo de Sistemas Multiagente (DESMA). Málaga, España (2004)
20. ZigBee Standards Organization.: ZigBee Specification Document 053474r13. ZigBee Alliance (2006)

Hybrid Architecture for a Reasoning Planner Agent

Javier Bajo[1], Dante I. Tapia[2], Ana de Luis[2], Sara Rodríguez[2], Juan F. de Paz[2], and Juan M. Corchado[2]

[1] Universidad Pontificia de Salamanca,
Compañía 5, 37002, Salamanca, Spain
jbajope@upsa.es
[2] Departamento Informática y Automática, Universidad de Salamanca
Plaza de la Merced s/n 37008, Salamanca, Spain
{dantetapia,adeluis,srg,corchado}@usal.es, fcofds@gmail.com

Abstract. This paper presents a hybrid architecture that facilitates the incurporation of a case-based planning system as the reasoning motor for a deliberative agent. This architecture makes possible to solve a wide range of problems in terms of agents and multi-agent systems. The problems are resolved in terms of plans, using plans that have already been experienced.

Keywords: Case-based planning, Learning, Hybrid systems.

1 Introduction

Agents and multi-agent systems are adequate for developing applications in dynamic and flexible environments. Agents can be characterized through their capacities in areas such as autonomy, communication, learning, goal orientation, mobility, persistence, etc. Autonomy, learning and reasoning are especially important aspects for an agent. These capabilities can be modelled in different ways and with diverse tools [17], with the use of Case Based Reasoning (CBR) systems as a possibility. This paper focuses on presenting a hybrid CBR-based deliberative agent architecture that incorporates a specialized planning mechanism to implement the retrieve, reuse, revise and retain stages of the CBR system. Several architectures have been proposed for building deliberative agents, most of them based on the BDI (Belief, Desire, Intention) model [4], in which the internal structure of an agent and therefore its ability to choose a course of action is based on mental attitudes. The advantage of using mental attitudes in the design and construction of agents and multi-agent systems is the natural (human-like) modeling and the high abstraction level. The BDI model uses Beliefs as information attitudes, Desires as motivational attitudes and Intentions as deliberative attitudes for each agent. The method proposed in [7], [10] allows the incorporation of CBR systems as a reasoning engine in BDI agents, facilitating learning and adaptation and providing a greater degree of autonomy than pure BDI architecture [4].

This work is focused on the development of deliberative planner agents using case-based reasoning (CBR) systems, as a way to create a hybrid architecture that can be implemented in adaptive systems to improve assistance in a wide range of fields such

B. Apolloni et al. (Eds.): KES 2007/ WIRN 2007, Part II, LNAI 4693, pp. 461–468, 2007.
© Springer-Verlag Berlin Heidelberg 2007

as oceanography, health care, tourism, etc. [3], [6], [8]. This type of hybrid architecture meets the conditions needed to introduce a representation and a reasoning based on the action [2] because a CBR-BDI agent uses case-based reasoning as a reasoning mechanism, which allows it to learn from initial knowledge, to interact autonomously with the environment as well as with users and other agents within the system, and to have a large capacity for adaptation to the needs of its surroundings. The CBR-BDI agents specialized in generating plans are referred as CBP-BDI agents, in which a plan is defined as a sequence of document collection and delivery points.

In the next section, the concept of case based planning is explained. In the third section, the hybrid architecture which incorporates case based planning into a deliberative agent is presented. Finally, the conclusions and future work are shown.

2 Case Base Planning: Learning from Past Plans

Case-based Reasoning (CBR) is a type of reasoning based on the use of past experiences [13]. CBR systems solve new problems by adapting solutions that have been used to solve similar problems in the past, and learn from each new experience. The primary concept when working with CBR's is the concept of case. A case can be defined as a past experience, and is composed of three elements: A problem description, which describes the initial problem; a solution, which provides the sequence of actions carried out in order to solve the problem; and the final state, which describes the state achieved once the solution was applied. A CBR manages cases (past experiences) to solve new problems. The way cases are managed is known as the CBR cycle, and consists of four sequential phases: retrieve, reuse, revise and retain. The retrieve phase starts when a new problem description is received. In this phase a similarity algorithm is used to find the most cases in the cases memory. Once the most similar cases have been retrieved, the reuse phase begins, adapting the solutions for the retrieved cases to obtain the best solution for the current case. The revise phase consists of an expert revision for the solution proposed. Finally, the retain phase allows the system to learn from the experiences obtained in the three previous phases and consequently updates the cases memory.

Case-based planning (CBP) is the idea of planning as remembering [11]. CBP is a specialization of case-based reasoning (CBR) which is a problem solving methodology based on using a library of solutions for similar problems [11]. In CBP, the solution proposed to solve a given problem is a plan, so this solution is generated taking into account the plans applied to solve similar problems in the past. The problems and their corresponding plans are stored in a plans memory. In practice, what is stored is not only a specific problem with a specific solution, but also additional information about how the plans have been derived. As well as in case-based reasoning, the case representation, the plans memory organization and the algorithms used in every stage of the case-based planning cycle are essential to define an efficient planner. The formal description of a case-based planner [16] can be formalized as a 3-tuple $<I,G,Op>$:

- I is a set of formulae describing the initial state.
- G is a set of formulae describing the goal specification.

- *Op* is the set of operators (also called actions) that can be applied in a plan. Every action $a \in Op$ is described in terms of pre-conditions *Ca* (what has to be fulfilled in order to the action can be executed) and port-conditions *Ea* (what has to be fulfilled after the execution of the action).

A plan *P* is a tuple <*S,B,O,L*>:

- *S* is the set of plan actions. There are two special actions: t_I, those whose effects are *I*, that is, the initial state; and t_G, those actions whose pre-conditions are *G*, that is, the goal specification.
- *O* is an ordering relation on *S* allowing to establish an order between the plan actions. t_I is always the first action and t_G is the last action. If the ordering relation is total, P is a linear plan, whereas if it is a partial-order relation, *P* is a non-linear plan.
- *B* is a set that allows describing the bindings and forbidden bindings on the variables appearing in *P*.
- *L* is a set of casual links of the form $s \xrightarrow{p} s'$, where $s, s' \in S$, $p \in Es$ and $p \in Cs'$. That is, relations allowing to establish a link between plan actions.

A plan *P* constitutes the solution generated to solve a planning problem when for each action $s \neq t_I$, for each $p \in Cs$ there exists a causal link $s \xrightarrow{p} s'$ and for each action $s \neq t_G$ there exists at least a causal link $s \xrightarrow{q} s''$. In the case that the planner is interested in retaining the failures or unexpected situations during the plan, these failures or situations are represented as a set of formulae *F*.

3 Hybrid Constraint Satisfaction-Based Planner Agent

An autonomous deliberative BDI case-based planner (CBP-BDI) agent has been developed, which presents a hybrid architecture, based on the BDI model [4], where the internal structure and capabilities of the agents are based on mental aptitudes, using beliefs, desires, and intentions. The method presented in this paper facilitates the incorporation of CBR systems [2] as a deliberative mechanism within BDI agents, facilitating learning and adaptation and providing a greater degree of autonomy than pure BDI architecture. A deliberative CBP-BDI agent is specialized in generating plans and incorporates a case-based planning (CBP) reasoning mechanism. The purpose of CBR agents is to solve new problems by adapting solutions that have been used to solve similar problems in the past [2], and CBP agents are a variation of CBR agents, based on the plans generated from each case.

A *representation based on an action* requires an agent architecture in which the way to acquire and process the knowledge of the world (environment), at the reasoning stage, is closely related to the way in which plans are constructed and used in the execution phase. Figure 1 shows the internal structure of a CPB-BDI agent. Problem description (initial state) and solution (situation when final state is achieved) are represented as beliefs, the final state as a goal (or set of goals), and the sequences of actions as plans. The CBP cycle is implemented through goals and plans. When the goal corresponding to one of the stages is triggered, different plans (algorithms) can be executed concurrently to achieve the goal. Each plan can trigger new sub-goals and, consequently, cause the execution of new plans.

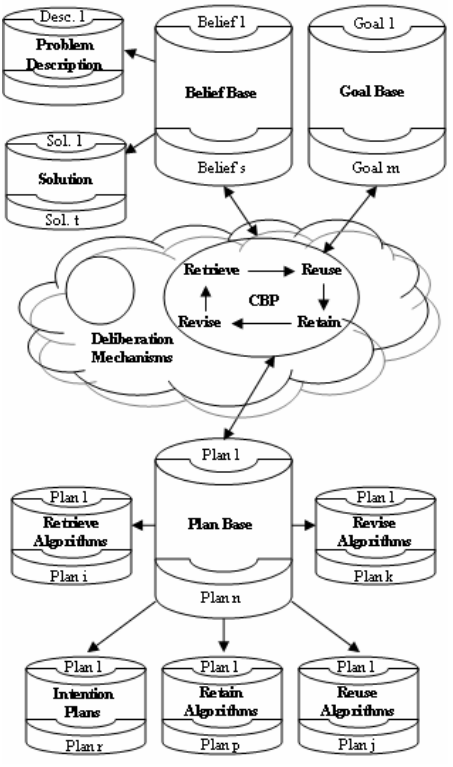


Fig. 1. CBR-BDI based on plans agent architecture

In this section, the CBP-BDI hybrid architecture is described. This architecture is achieved through a BDI agent model [4] in which a CBP reasoning motor has been incorporated [5], [11] using plans that have already been experienced. Firstly, the terminology used within the paper for a BDI agent is introduced:

- The *environment M* and the changes that are produced within it are represented from the point of view of the agent. Therefore, the world can be defined as a set of variables that influence a problem faced by the agent.

$$M = \{\tau_1, \tau_2, \cdots, \tau_s\} \text{ with } s < \infty \tag{1}$$

- The *beliefs* are objects of some (or all) of the world attributes taking a set of concrete values.

$$B = \{b_i / b_i = \{\tau_1^i, \tau_2^i, \cdots, \tau_n^i\}, n \leq s \quad \forall i \in N\}_{i \in N} \subseteq M \tag{2}$$

- A *state of the world* $e_j \in E$ is represented for the agent by a set of beliefs that are true at a specific moment in time *t*. Let $E = \{e_j\}_{j \in N}$ be the set of the world status for a fixed value of *t*, then (3).

$$e_j^t = \{b_1^{jt}, b_2^{jt}, \cdots b_r^{jt}\}_{r \in N} \subseteq B \quad \forall j,t \tag{3}$$

- The *desires* are imposed at the beginning. These are applications between a state of the current world and another that it is trying to reach.

$$d : \underset{e_0}{E} \underset{\rightarrow}{\rightarrow} \underset{e^*}{E} \tag{4}$$

- *Intentions* are the way the agent's knowledge is used in order to reach its objectives. They are reduced to the need to determine whether there is compatibility between the knowledge that the agent possesses and the requirements needed to be able to attain its desires. A desire is attainable if the application *i* exists:

$$i : \underset{(b_1, b_2, \cdots\cdots, b_n, e_0)}{BxBx\overset{n)}{\cdots}xBxE} \underset{\rightarrow}{\rightarrow} \underset{e^*}{E} \tag{5}$$

In our model, intentions guarantee that there is enough knowledge in the beliefs base for a desire to be reached via an action plan.

- We define an *agent action* as the mechanism that provokes changes in the world, making it change the state.

$$a_j : \underset{e_i}{E} \underset{\rightarrow}{\rightarrow} \underset{a_j(e_i)=e_j}{E} \tag{6}$$

- *Agent plan* is the name we give to a sequence of actions that, from a current state e_0, defines the path of states through which the agent passes in order to reach the other world state.

$$p_n : \underset{e_0}{E} \underset{\rightarrow}{\rightarrow} \underset{p_n(e_0)=e_n}{E}$$
$$p_n(e_0) = e_n = a_n(e_{n-1}) = \cdots = (a_n \circ \cdots \circ a_1)(e_0) \qquad p_n \equiv a_n \circ \cdots \circ a_1 \tag{7}$$

Below, the attributes that characterise the plans in the case base are presented, which allow us to relate BDI language with the interest parameters within a CBP. A constraint satisfaction problem (CSP) planning problem is considered in order to lend the model generality [15]. These kinds of problems do not only search for solutions but also have to conform to a series of imposed restrictions. Based on the theory of action [3], [9], the set of objectives for a plan and the resources available are selected as a variable upon which the CSP problems impose the restrictions. A plan *p* is expressed as *p=<E, O, O', R, R'>,* where:

- E is the environment, but it also represents the type of problem faced by the agent, characterised by $E = \{e_0, e^*\}$, where e_0 represents the starting point for the agent when it begins a plan, and e^* is the state or states that it is trying to attain.
- *O* indicates the objectives of the agent and *O'* are the results achieved by the plan.
- *R* are the total resources and *R'* are the resources consumed by the agent.

Table 1 shows the indicators derived from the attributes described above, used to identify and contrast the quality of the different plans (# means cardinal of a set).

Table 1. Indicators of plan quality

Indicator	Formulae
Efficacy of the plan: relationship between objectives attained and objectives proposed	$E_f = \frac{\#(O' \cap O)}{\#O}$
Cost of the plan: relationship between the resources used and the resources available	$C = \frac{\#R'}{\#R}$
Efficiency of the plan: relationship between the objectives attained and the resources consumed	$E_{ff} = \frac{\#(O' \cap O)}{\#R'}$

As a representation of a plan p in the case base, a more compact expression is used, $p = < E, F(O;R) = 0 >$, where $F(O;R) = 0$ is the **restriction function of the plan** or ***quality function***. Not all plans that obtain certain objectives by consuming resources are equally acceptable. This function demands some minimum objectives to be attained with a maximum consumption limit of resources in order for the plan to be considered acceptable. For simplicity, a single action plan $p_1 \equiv p_2$, $a_1 \equiv a_2$ is considered (8).

$$p \equiv a : \underset{e_0}{E} \underset{\rightarrow}{\rightarrow} \underset{a(e_0)=e_1}{E} \tag{8}$$

From the BDI agent point of view, in order to execute a plan p, beginning from e_0 and aiming to achieve the state e_1, the agent needs:

- to have characterised state e_0: $\{b_1^0, \cdots, b_n^0\} = B_1 \subset B$
- to have characterised state e_1: $\{b_1^1, \cdots, b_m^1\} = B_2 \subset B$
- to know how to execute the action a
- to know that the action a allows the agent to pass from e_1 to e_2, in other words, to be able to guarantee that the intention exists

If a problem $E = \{e_0, e_1\}$ has been defined, a plan p to solve the problem can be characterised by the relationships between the objectives reached and the resources consumed between both states. The general functioning process is derived by following the typical phases of a case based system [2] (eliminating the revision phase, since it can be external to the system needing the intervention of an expert). The name CBP system (or methodology) is used to describe a procedural information process in which past experiences are used to decide how to deal with a new problem which share similarities with previous problems [5], [11]. These experiences, stored as cases, act dynamically on each CBP cycle. The reasoning process of this kind of system carries out the following four sequential stages (noticing that the revision stage has been eliminated: it usually is carried out by an expert and is external to the system) as can be seen in Figure 2:

- **Retrieval:** Given a state of the perceived world e_0 and the desire that the agent encounters in a state $e_0 \neq e^*$, the system searches in the case base (previously checked by the intention function through which the system gains sufficient knowledge) for plans that have resolved similar problems in the past.

- **Adaptation/Reuse:** From the previous phase, a set of possible solutions for the agent $\{p_1,\ldots, p_n\}$ is obtained. In this phase, in accordance with the planning model G, the system uses the possible solutions to propose a solution p^* (9).

$$G(e_0, p_1, \cdots, p_n) = p^* \tag{9}$$

- **Learning/Retain:** The plan proposed may achieve its objective or fail in its development. The information on the quality of the final plan in the $w_f(p^*)$ cycle is stored for the future and is directly proportional (i) to the initial value of $w_i(p^*)$, and (ii) to the "rate of use" $\alpha(N)$, where N is the number of times that the plan has been used in the past [12].

$$w_f(p^*) = w_i(p^*)\alpha(N) \tag{10}$$

The model proposed conforms to the conditions required in order to obtain a representation and reasoning based on the action [14]. The capabilities of the hybrid system restrict what kind of plans can be generated. Plans structure and world representation can be easily adapted to a wide range of problems.

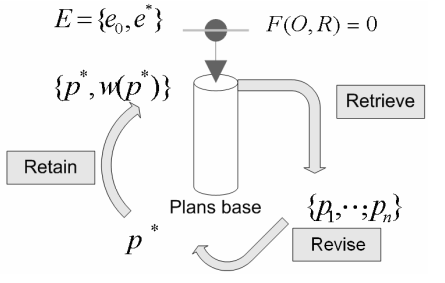


Fig. 2. CBP life cycle

4 Conclusions and Future Work

The CBP-BDI hybrid architecture presented in this paper solves one of the problems of BDI (deliberative) architectures, which is the lack of learning capacity. It also reduces the gap that exists between the formalization and the implementation of BDI agents. The CBR systems reasoning cycle helps the agents to solve problems, adapt to changes in the environment, and identify new possible solutions.

In this paper some fundamental ideas for the theoretical analysis of the agency and theory of planning have been introduced. Our future line work will attempt to generate different planning strategies according to the environment on concrete problems using this architecture, which will have great significance in the development of artificial intelligence systems with flexibility and adaptability to changes in the environment.

References

1. Aamodt, A., Plaza, E.: Case-Based Reasoning: Foundational Issues, Methodological Variations, and System Approaches. AICOM. 7, 39–59 (1994)
2. Allen, J.F.: Towards a General Theory of Action and Time. Artificial Intelligence 23, 123–154 (1984)
3. Bajo, J., Corchado, J.M.: Evaluation and Monitoring of the Air-Sea Interaction Using a CBR-Agents Approach. In: Muñoz-Ávila, H., Ricci, F. (eds.) ICCBR 2005. LNCS (LNAI), vol. 3620, pp. 50–62. Springer, Heidelberg (2005)
4. Bratman, M.E.: Intentions, Plans and Practical Reason. Harvard University Press, Cambridge, MA (1987)
5. Carbonell, J.: Learning by Analogy: Formulating and Generalizing Plans from Past Experience. In: Michalski, R., Carbonell, J., Mitchell, T. (eds.) Machine Learning: An Artificial Intelligence Approach. Cambridge, MA, pp. 137–162 (1983)
6. Corchado, J.M., Lees, B.: A Hybrid Case-based Model for Forecasting. Applied Artificial Intelligence 15(2), 105–127 (2001)
7. Corchado, J.M., Laza, R.: Constructing Deliberative Agents with Case-based Reasoning Technology. International Journal of Intelligent Systems 18(12), 1227–1241 (2003)
8. Corchado, J.M., Pavón, J., Corchado, E.S., Castillo, L.F.: Development of CBR-BDI Agents: A Tourist Guide Application. In: Funk, P., González Calero, P.A. (eds.) ECCBR 2004. LNCS (LNAI), vol. 3155, pp. 547–559. Springer, Heidelberg (2004)
9. Davis, L.: Theory of Action. Prentice-Hall, Englewood Cliffs, N.J. (1979)
10. Glez-Bedia, M., Corchado, J.M., Corchado, E.S., Fyfe, C.: Analytical Model for Constructing Deliberative Agents. Engineering Intelligent Systems 3, 173–185 (2002)
11. Hammond, K.: Case-Base Planning: Viewing Planning as a Memory Task. Academic Press, New York (1989)
12. Heckerman, D.: A Tutorial on Learning with Bayesian Networks. In: Jordan, M.I. (ed.) Learning in graphical models, Kluwer, Dordrecht, Netherlands (1998)
13. Kolodner, J.: Case-Based Reasoning. Morgan Kaufmann, San Francisco (1993)
14. Pollack, M.E.: The Uses of Plans. Artificial Intelligence. 57, 43–68 (1992)
15. Sam-Haround, D.: Constraints Consistency Techniques for Continuous Domains. PhD thesis, No.1423, Swiss Federal Institute of Technology in Laussanne, Switzerland (1995)
16. Spalazzi, L.: A Survey on Case-Based Planning. Artificial Intelligence Review 16(1), 3–36 (2001)
17. Wooldridge, M., Jennings, N.R.: Agent Theories, Architectures, and Languages: a Survey. In: Wooldridge, Jennings (eds.) Intelligent Agents, pp. 1–22. Springer, Heidelberg (1995)

Learning Topologic Maps with Growing Neural Gas

José García-Rodríguez, Francisco Flórez-Revuelta, and Juan Manuel García-Chamizo

Department of Computer Technology. University of Alicante. Apdo. 99. 03080 Alicante, Spain
{jgarcia,florez,juanma}@dtic.ua.es

Abstract. Self-organising neural networks try to preserve the topology of an input space by means of their competitive learning. This capacity has been used, among others, for the representation of objects and their motion. In this work we use a kind of self-organising network, the Growing Neural Gas, to represent different objects shape. As a result of an adaptive process the objects are represented by a topology representing graph that constitutes an induced Delaunay triangulation of their shapes. This feature can be used to learn and represent topologic maps that mobile devices use to navigate in different environments.

Keywords: Self-organising maps, topology preservation, topologic maps, robot navigation, computational geometry.

1 Introduction

Self-organising neural networks, by means of a competitive learning, make an adaptation of the reference vectors of the neurons, as well as, of the interconnection network among them; obtaining a mapping that tries to preserve the topology of an input space. Besides, they are able of a continuous re-adaptation process even if new patterns are entered, with no need to reset the learning.

These capacities have been used for the representation of objects [1] (Figure 1) and their motion [2] by means of the Growing Neural Gas (GNG) [3] that has a learning process more flexible than other self-organising models, like Kohonen maps [4].

These two applications, representation of objects and their motion, have in many cases temporal constraints, reason why it is interesting the acceleration of the learning process. In computer vision applications the condition of finalization for the GNG algorithm is commonly defined by the insertion of a predefined number of neurons. The election of this number can affect the quality of the adaptation, measured as the topology preservation of the input space [5].

In this work GNG neural networks are combined with graph representation and interpretation techniques to form a hybrid intelligent system used to learn and represent topologic maps that mobile devices use for navigation tasks.

The remaining of the paper is organized as follows: section 2 provides a detailed description of the GNG topology learning algorithm. In section 3 an explanation on how GNG can be applied to represent objects is given. A set of experimental results using GNG to represent different topologic maps is presented in section 4, followed by our major conclusions and further work.

B. Apolloni et al. (Eds.): KES 2007/ WIRN 2007, Part II, LNAI 4693, pp. 469–476, 2007.
© Springer-Verlag Berlin Heidelberg 2007

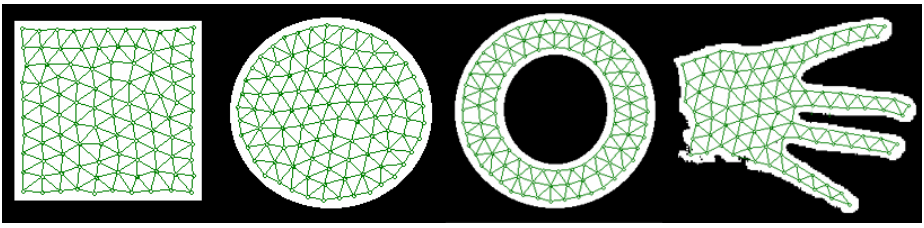


Fig. 1. Representation of two-dimensional objects with a self-organising network

2 Topology Learning

One way of selecting points of interest in 2D shapes is to use a topographic mapping where a low dimensional map is fitted to the high dimensional manifold of the shape, whilst preserving the topographic structure of the data. A common way to achieve this is by using self-organising neural networks where input patterns are projected onto a network of neural units such that similar patterns are projected onto units adjacent in the network and vice versa. As a result of this mapping a representation of the input patterns is achieved that in post-processing stages allows one to exploit the similarity relations of the input patterns. Such models have been successfully used in applications such as speech processing [4], robotics [7,8] and image processing [9]. However, most common approaches are not able to provide good neighborhood and topology preservation if the logical structure of the input pattern is not known a priori. In fact, the most common approaches specify in advance the number of neurons in the network and a graph that represents topological relationships between them, for example, a two-dimensional grid, and seek the best match to the given input pattern manifold. When this is not the case the networks fail to provide good topology preserving as for example in the case of Kohonen's algorithm.

The approach presented in this paper is based on self-organising networks trained using the Growing Neural Gas learning method [3], an incremental training algorithm. The links between the units in the network are established through competitive hebbian learning [10]. As a result the algorithm can be used in cases where the topological structure of the input pattern is not known a priori and yields topology preserving maps of feature manifold [5].

2.1 Growing Neural Gas

With Growing Neural Gas (GNG) [3] a growth process takes place from minimal network size and new units are inserted successively using a particular type of vector quantisation [4]. To determine where to insert new units, local error measures are gathered during the adaptation process and each new unit is inserted near the unit which has the highest accumulated error. At each adaptation step a connection between the winner and the second-nearest unit is created as dictated by the competitive hebbian learning algorithm. This is continued until an ending condition is fulfilled, as for example evaluation of the optimal network topology based on some measure. Also the ending condition could it be the insertion of a predefined number of neurons or a

temporal constrain. In addition, in GNG networks learning parameters are constant in time, in contrast to other methods whose learning is based on decaying parameters.

In the remaining of this Section we describe the growing neural gas algorithm. The network is specified as:

- A set N of nodes (neurons). Each neuron $c \in N$ has its associated reference vector $w_c \in R^d$. The reference vectors can be regarded as positions in the input space of their corresponding neurons.
- A set of edges (connections) between pairs of neurons. These connections are not weighted and its purpose is to define the topological structure. An *edge aging scheme* is used to remove connections that are invalid due to the motion of the neuron during the adaptation process.

The GNG learning algorithm to approach the network to the input manifold is as follows:

1. Start with two neurons a and b at random positions w_a and w_b in $\mathcal{R}^d$.
2. Generate a random input pattern ξ according to the data distribution $P(\xi)$ of each input pattern. In our case since the input space is 2D, the input pattern is the (x, y) coordinate of the points belonging to the object. Typically, for the training of the network we generate 1000 to 10000 input patterns depending on the complexity of the input space.
3. Find the nearest neuron (winner neuron) s_1 and the second nearest s_2.
4. Increase the age of all the edges emanating from s_1.
5. Add the squared distance between the input signal and the winner neuron to a counter error of s_1 such as:

$$\Delta error(s_1) = \left\| w_{s_1} - \xi \right\|^2 \tag{1}$$

6. Move the winner neuron s_1 and its topological neighbours (neurons connected to s_1) towards ξ by a learning step ε_w and ε_n, respectively, of the total distance:

$$\Delta w_{s_1} = \varepsilon_w (\xi - w_{s_1}) \tag{2}$$

$$\Delta w_{s_n} = \varepsilon_n (\xi - w_{s_n}) \tag{3}$$

7. If s_1 and s_2 are connected by an edge, set the age of this edge to 0. If it does not exist, create it.
8. Remove the edges larger than a_{max}. If this results in isolated neurons (without emanating edges), remove them as well.
9. Every certain number λ of input signals generated, insert a new neuron as follows:
   - Determine the neuron q with the maximum accumulated error.
   - Insert a new neuron r between q and its further neighbour f:

$$w_r = 0.5(w_q + w_f) \tag{4}$$

- Insert new edges connecting the neuron r with neurons q and f, removing the old edge between q and f.
- Decrease the error variables of neurons q and f multiplying them with a constant α. Initialize the error variable of r with the new value of the error variable of q and f.

10. Decrease all error variables by multiplying them with a constant β.
11. If the stopping criterion is not yet achieved, go to step 2. (In our case the criterion is the number of neurons inserted)

3 Representation of 2D Objects with GNG

Given an image $I(x,y) \in \mathcal{R}$ we perform the transformation $\psi_{\mathcal{T}}(x,y) = \mathcal{T}(I(x,y))$ that associates to each one of the pixels its probability of belonging to the object, according to a property $\mathcal{T}$. For instance, in figure 2, this transformation is a threshold function.

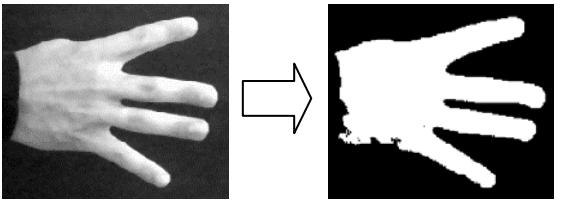


Fig. 2. Silhouette extraction

If we consider $\xi = (x,y)$ and $\mathcal{P}(\xi) = \psi_{\mathcal{T}}(\xi)$, we can apply the learning algorithm of the GNG to the image I, so that the network adapts its topology to the object. This adaptive process is iterative, so the GNG represents the object during all the learning.

As a result of the GNG learning we obtain a graph, the Topology Preserving Graph $\mathcal{TPG} = \langle \mathcal{N}, \mathcal{C} \rangle$, with a vertex (neurons) set $\mathcal{N}$ and an edge set $\mathcal{C}$ that connect them (figure 1). This $\mathcal{TPG}$ establishes a Delaunay triangulation induced by the object [6].

Also, the model is able to characterize the diverse parts of an object, or several present objects in the scene with the same values for the visual property $\mathcal{T}$, without initialize different data structures for each one of the objects. This is due to the capacity of GNG to divide itself when removing neurons.

4 Experiments

To illustrate the capacity of representation of the GNG we have tested its adaptation to some different mazes obtaining graphs that represent the different environments that appear in the images. Since we obtain information of the neurons coordinates and

the neighbours of any of them, we can apply the results to different fields like the representation and tracking of single or multiple objects using techniques based on graph theory or algorithms for robot navigation among others.

4.1 Representing and Learning Topologic Maps

For this experiment we have obtained the representing graph of two different mazes that represent the same environment with some modifications.

In both cases we have considered the possibility that the obstacles in the path of a navigating robot were part to the background, for images on the left and belong to the foreground, for images on the right.

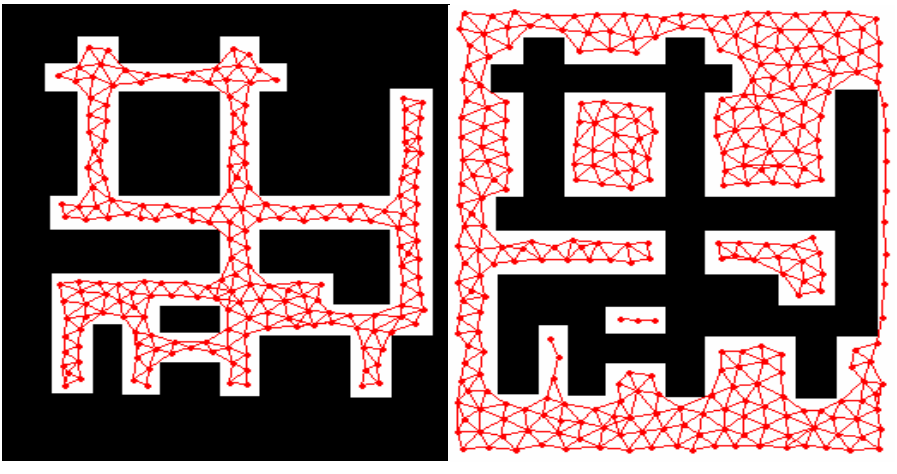


Fig. 3. Topologic maps based on GNG adaptation

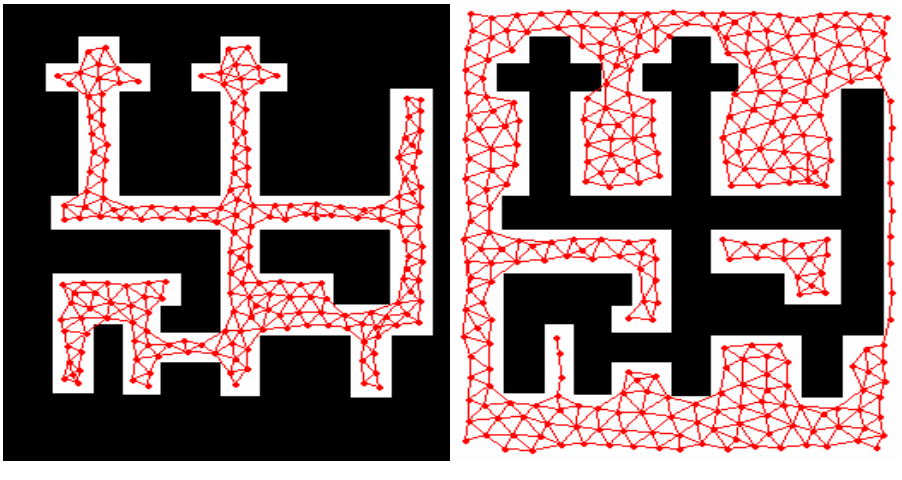


Fig. 4. GNG adaptation to changes in the environment

Figure 3 shows the first maze. On the left, we have used as input space to be learned by the self-organising maps pixels in white as a visual property. On the right we have used the same visual property but in this case reversing the image colours before. In this last case the space that represents paths in image on the left appears as obstacles on image on the right and vice-versa

In figure 4 we have modified the original maze. As in the first case we have studied the cases of obstacles as a foreground and background obtaining the expected representation of the image that could be useful for navigation tasks.

The induced Delaunay triangulation could be used in this case, for example to calculate the shortest path from one point to a defined target following the nodes and edges of the obtained graph with a simple algorithm like Dijkstra or even defining a minimal spanning tree graph of the induced Delaunay triangulation.

In table 1 can be observed topographic product, a measure for topology preservation [5], mean square error and time spent for maps learning with different number of neurons. Increasing the number of neurons we obtain representations with better topology preservation but spending more time.

Table 1. Comparative study of topology preservation depending of number of neurons

Neurons	Topographic product	Quantization error	Time
50	0,0064	10,45	5
100	0,001791	7,59	16
200	0,000146	5,30	39
300	0,000104	4,33	71

When changes in the environment are small like occlusion of some paths is not necessary to reset the algorithm to represent the new environment and the last obtained map can be used as a based to learn changes in the new one. This feature accelerates the learning process improving the performance of the system and providing the representation system with real time capacities. In the case of figures 3 and 4, learning process it takes about 39 seconds inserting 200 neurons while using the map created for figure 3 as starting point to learn figure 4 the time spent to adapt it is less than 1 second.

4.2 Obtaining the Data

We have considered two possibilities to obtain the information about the environment to create de topologic maps depending on the available information.

a) If external information of the environment is available, for example images from an external camera, we use them to adapt the map to this image and obtain the possible paths for the mobile devices.
b) If only local information from the robot sensors and cameras is available, we obtain it exploring with the mobile device the environment and using the information obtained during the training process to learn the map, updating this information along the time to detect changes in the environment, like in figures 3 and 4.

4.3 Using Topologic Maps for Path Planning

The obtained Delaunay triangulation can be use for path planning tasks. Since we keep the map updated along the time we can modify the mobile device trajectory to avoid obstacles.

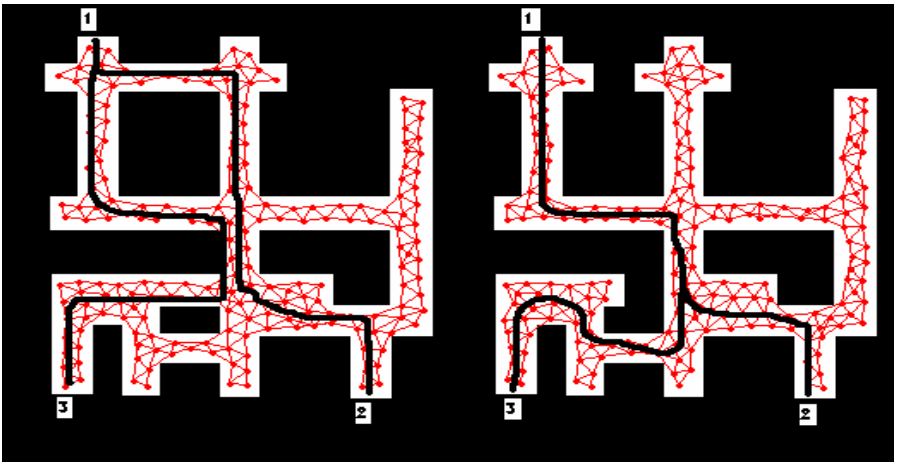


Fig. 5. Path planning adaptation to modifications in environment

In figure 5 we show an example based on the updated trajectories depending on changes in the environment. When some paths are deleted, the topologic map is updated and a new trajectory is designed to arrive from point 1 to points 2 and 3. Applying to the obtained Delaunay triangulation an algorithm to calculate minimal length paths like Dijkstra we would obtain these results.

The GNG parameters used for the simulations are: $\lambda = 10000$, $\varepsilon_w = 0.1$, $\varepsilon_n = 0.001$, $\alpha = 0.5$, $\beta = 0.95$, $\alpha_{\max} = 250$.

5 Conclusions and Further Work

In this paper, we have demonstrated the capacity of representation of bi-dimensional objects by a self-organising neural network. Establishing a suitable transformation function, the model is able to adapt its topology to images of the environment. Then, a simple, but very rich representation is obtained.

The model, by its own adaptation process, is able to divide itself so that it can characterize different fragments from an object or different objects with the same visual property.

We have obtained a representation model that can be useful to learn topologic maps and that has a fast adaptation to small changes in these maps providing the system with real time capabilities. Also, the graph representation can be used to applied algorithms that optimize trajectories for mobile devices navigation.

Finally, the iterative and parallel performance of the presented representation model is the departure point for the development of high performance architectures that supply a characterization of an object depending on the time available.

References

1. Flórez, F., García, J.M., García, J., Hernández, A.: Representation of 2D Objects with a Topology Preserving Network. In: Proceedings of the 2nd International Workshop on Pattern Recognition in Information Systems (PRIS'02), Alicante. ICEIS Press, pp. 267–276 (2001)
2. Flórez, F., García, J.M., García, J., Hernández, A.: Hand Gesture Recognition Following the Dynamics of a Topology-Preserving Network. In: Proc. of the 5th IEEE Intern. Conference on Automatic Face and Gesture Recognition, pp. 318–323. IEEE, Inc., Washington, D.C. (2001)
3. Fritzke, B.: A Growing Neural Gas Network Learns Topologies. In: Tesauro, G., Touretzkyy, D.S., Leen, T.K. (eds.) Advances in Neural Information Processing Systems, vol. 7, pp. 625–632. MIT Press, Cambridge (1995)
4. Kohonen, T.: Self-Organising Maps. Springer, Heidelberg (1995)
5. Martinetz, T., Schulten, K.: Topology Representing Networks. Neural Networks 7(3), 507–522 (1994)
6. O'Rourke, J.: Computational Geometry in C. Cambridge University Press, Cambridge (2001)
7. Ritter, H., Schulten, K.: Topology conserving mappings for learning motor tasks. In: Neural Networks for Computing, AIP Conf. Proc. (1986)
8. Martinez, T., Ritter, H., Schulten, K.: Three dimensional neural net for learning visuomotor-condination of a robot arm. IEEE Transactions on Neural Networks 1, 131–136 (1990)
9. Nasrabati, M., Feng, Y.: Vector quantisation of images based upon kohonen self-organising feature maps. In: Proc. IEEE Int. Conf. Neural Networks, pp. 1101–1108. IEEE Computer Society Press, Los Alamitos (1998)
10. Martinez, T.: Competitive hebbian learning rule forms perfectly topology preserving maps. In: ICANN (1993)

Using a Neurofuzzy Approach in a Medical Application

Constantinos Koutsojannis and Ioannis Hatzilygeroudis

Department of Computer Engineering & Informatics, School of Engineering,
University of Patras, 26500 Rion, Patras, Hellas, Greece
{ckoutsog,ihatz}@ceid.upatras.gr

Abstract. Today hybrid computing is a popular framework for solving complex problems such as in medical domain. Hybrid intelligent systems are systems that combine two or more intelligent techniques. Medicine and health care are closely related domains, where the types of problems faced are suitable for application of hybrid intelligent techniques. In this paper we present the initial evaluation of FUzzy NEUrule System which is a Neuro Fuzzy approach based on fuzzy Adaline neurons and uses Differential Evolution for optimization of membership functions. According to our previous Neuro-fuzzy approaches and a well-defined hybrid system HYMES, FUNEUS is an attempt to the direction for integration of neural and fuzzy components with Differential evolution. Despite the fact that it remains difficult to compare neurofuzzy systems conceptually and evaluate their performance, early experimental results in a medical database proved a promising performance and the need for further evaluation in other medical applications.

1 Introduction

Hybrid systems mix different methods of knowledge engineering and make them "work together" to achieve a better solution to a problem, compared to using a single method for the same problem [1],[2]. Hybrid connectionist production systems incorporate artificial neural networks (ANNs) into production rules to enhance them with approximate reasoning and learning capabilities [2], [3]. Fuzzy inference modules are incorporated into production rules in a similar way [4], [5]. Today, Neuro-Fuzzy (NF) computing is a popular framework for solving complex problems. While the learning capability is an advantage from the viewpoint of FIS, the formation of linguistic rule base will be advantage from the viewpoint of ANN [4], [5]. Interestingly this synergy still becomes a target yet to be satisfied. In most NF systems [FuzzyCOPE, ANFIS, NEFCLASS, FUNN etc] one of the two technologies becomes predominant resulting a commonly visible accuracy-interpretability trade-off [3], [4], [5]. Generally, because the approximation abilities are easier to be quantified and eventually to be scientifically or technologically realized, the usual tendency in most NF systems is to pay attention to the neural side. Thus, the approximation capabilities are highly "glorified" and even supported with evolutionary programming techniques, but the interpretation abilities are implicitly reduced (e.g. FUZZNET, FNES, FALCON, ANFIS, NEFCON, FINEST, FuNN, NEFCLASS, etc) [4], [3, [5]. For NF systems, to be highly intelligent, some of the major requirements are: fast

B. Apolloni et al. (Eds.): KES 2007/ WIRN 2007, Part II, LNAI 4693, pp. 477–484, 2007.
© Springer-Verlag Berlin Heidelberg 2007

learning (memory based-efficient storage and retrieval capacities), on-line adaptability (accommodating new features like inputs, outputs, nodes, connections etc), achievement a low global error rate and computationally inexpensive [3], [4], [7]. The data acquisition and pre-processing training data are also quite important for the success of almost all NF systems [8], [9]. The underlying conjecture in the latter is that the future NF systems should be constructed based on a simple possessing unit [4]: *Fuzzy logic Neuron* (FN). FN includes fuzzy data and fuzzy logic operations in its' unit; its inputs and/or weights are also expressed in terms of membership functions whose transparency and learning abilities are accentuated to highest possible level, as already proposed in the literature [3], [10]. Fuzzy neurules are a kind of integrated rules that combine symbolic rules and a neurofuzzy unit, the fuzzy adaline unit. The system implementing fuzzy neurules is called FUNEUS [11]. Fuzzy neurules give pre-eminence to the symbolic component. Thus, the constructed knowledge base retains the modularity of classical fuzzy rules, since it consists of autonomous units (*fuzzy neurules*), and their naturalness, since they look much like traditional rules. Fuzzy neurules is an effort to introduce fuzziness into neurules, which combine symbolic rules and the adaline neural unit [2]. Medicine and health care are closely related domains, where the types of problems faced are suitable for application of hybrid intelligent techniques. In this paper, we use FUNEUS in a medical domain to produce fuzzy neurules from empirical data and test their classification capabilities for some kind of heart diseases. Preliminary experimental results are promising as far as both the naturalness and classification capabilities of the rules are concerned.

The structure of the paper is as follows. Section 2 presents the hybrid connectionist system and the basic ideas and architecture of fuzzy neurules. In Section 3, the system architecture and parameter adjustment of FUNEUS is described. In Section 4 the hybrid inference mechanism is presented. Section 5 contains evaluation results in a medical database. Finally Section 6 discusses related work.

2 Low Level Hybrid Systems and Fuzzy Neurules

A hybrid system is a mixture of methods and tools used in one system which are loosely or tightly coupled [1], [4] from the functional point of view and hierarchical [3], flat, sequential or parallel [3], [4] from the architectural point of view. Additionally there are systems that are blended at a low structural and functional level that they are not separable from functional and structural point of view [3], [4, [6]. Building a connectionist rule base is possible not only by training a neural network with a set of data examples but by inserting existing rules into a neural network structure [5], [7]. This approach retains the advantages of connectionism, such as: learning, generalization, modularity, robustness, massive parallelism etc., into the elegant methods of symbolic processing, logical inferences and global-driven reasoning. Both paradigms can be blended at a low, neuronal level and structural knowledge can built up in a neuron and in a NN realized as a connectionist rule-based system [3], [4]. The underlying conjecture of the previous is that the new NF systems should be constructed in a simple possessing unit: *Fuzzy logic Neuron* (FN). FN includes fuzzy data and fuzzy logic operations in its' unit; the inputs and/or the

weights are also expressed in terms of membership functions whose transparency and learning abilities are accentuated to the highest possible level [4]. Types of fuzzy neurons have been successfully applied to prediction and classification problems [5], [9]. The Wirdow-Hoff's ADALINE can be thought of as the smallest, linear building block of the artificial neural networks. This element has been extensively used in science, statistics (in the linear regression analysis), engineering (the adaptive signal processing, control systems), and so on. Recently, Fuzzy ADALINE neurons were developed introducing the following modifications to ADALINEs [6]. In the proposed model, with the input and output constrained in the range [-1, +1], the unity slope in the nonlinearity of the neuron will fail to yield the desired output whenever the target signal is higher than the average of the inputs. In order to circumvent this situation, an adaptive algorithm for adjustment of the slope of the linear portion in the non-linear clipper function has been developed. In order the *fuzzy rules* to work in a fuzzy environment, we incorporated a "fuzzy" neuron to produce sets of rules called *fuzzy neurules*. Each *fuzzy neurule* is now considered as a fuzzy adaline unit that has replaced a number of fuzzy rules in the rule base of a fuzzy system. However, pre-eminence is still given to the fuzzy component. Thus, the constructed knowledge base retains the modularity of production rules, since it consists of autonomous units (*fuzzy neurules*), and their naturalness, since fuzzy *neurules* look much like fuzzy rules.

3 FUNEUS: A Hybrid Fuzzy Connectionist Production System

NF systems could be broadly classified in two types: a) weakly coupled systems and b) tightly coupled systems [6]. A weakly coupled NF system employs both a neural net unit and a fuzzy system unit in cascade. In a tightly coupled system neurons, the basic elements of a NN, are constructed amalgamating the composite characteristics of both a neuronal element and fuzzy logic [4], [11]. Additionally a Fuzzy-Genetic algorithm synergism is used for membership functions optimizations, which are intuitively chosen in a fuzzy system [6]. Given that the optimization of fuzzy membership functions may involve many changes to many different functions, and that a change to one function may affect others, the large possible solution space for this problem is a natural candidate for a GA based approach despite many NF approaches that use a gradient descent-learning algorithm to fine-tune the parameters of the fuzzy systems (Fig. 1). The GA module for adapting the membership function parameters acts as a stand-alone system that already has the if-then rules. GA optimizes the antecedent and consequent membership functions. Differential Evolution (DE) algorithm, which is used here, is a very simple population based, stochastic function minimizer, which is very powerful and at the same time turned out to be the best genetic type of algorithm for solving the real-valued test functions [14].

3.1 Fuzzy Neurule Base Construction

We use the method in [2] in combination with the proposed one by Wang and Mendel [16] for generating fuzzy rules from numerical input-output training data. The task here is to generate a set of fuzzy rules from the desired input-output pairs and then use

these fuzzy rules to determine the complete structure of the rule base. The algorithm for constructing a hybrid rule base from training data is outlined below:

1. Determine the input and output variables (fuzzy) and use dependency information to construct an initial fuzzy rule for each intermediate and output variable.
2. Determine the training set for each initial fuzzy rule from the training data pairs, train each initial fuzzy rule using its training input-output set and produce the corresponding fuzzy neurule(s).
3. Put the produced fuzzy neurules into the fuzzy neurule base.

In the sequel, we elaborate on the first two steps of the algorithm.

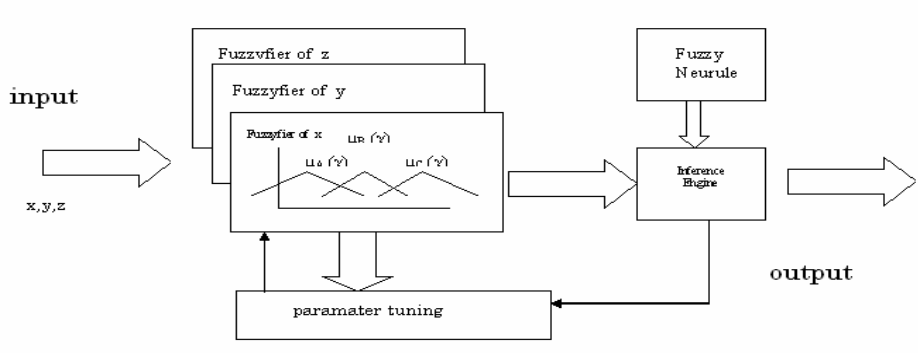


Fig. 1. Membership functions adjustment by GA

3.2 Constructing and Training the Initial Fuzzy Neurules

To construct *initial neurules*, first we need to know or determine the input, intermediate and output variables. Then, we need *dependency information*. Dependency information indicates which intermediate variables (concepts) and output variables (concepts) depend on which input/intermediate variables. If dependency information is missing, then output variables depend only on input variables, as indicated by the *training data*. In constructing an initial fuzzy neurule, all the conditions including the input, intermediate and the output variables that contribute in drawing a conclusion (which includes an intermediate or an output variable) constitute the inputs of the initial *fuzzy neurule* and the conclusion its output. Each fuzzy neurule is a *fuzzy Adaline* neuron with inputs the fuzzified values, weights the membership functions and additional weights with initial values of '1'. From the initial training data, we extract as many (sub)sets as the initial fuzzy neurules. Each such set, called a *training set*, contains *training examples* in the form $[x_1\ x_2\ \ldots\ x_n\ d]$, where x_i, $i= 1, \ldots, n$ are their *component values*, which correspond to the n inputs of the fuzzy neurule, and d is the *desired output*. So, step 2 of the algorithm for each initial *fuzzy neurule* is analyzed as follows:

2.1 From the initial training data, produce as many initial training sets as the number of the initial fuzzy neurules .

2.2 Assign each example in the set to the region that has maximum degree resulting one value with one membership function for each input parameter
2.3 For each set of desired input-output obtained, do the following:
2.3.1 Produce one fuzzy neurule of the form if x_1 is val1 and x_2 val2 …. then d is val
2.3.2 Assign an additional weight to each fuzzy neurule, which is defined as $CFi = \mu_A(x_1)\mu_B(x_2)\ldots\mu_C(d)$. This step is further performed to delete redundant rules and obtain a concise fuzzy neurule base.
2.3.3 Produce the corresponding fuzzy neurule, which is of the form if x_1 is A and x_2 is B …. then d is C (CF)

If two or more generated fuzzy neurules have the same conditions and consequents, then the rule that has maximum degree in the desired output is used. In this way, assigning the additional weight to each rule, the fuzzy rule base can be adapted or updated by the relative weighting strategy: *the more task related the rule becomes, the more weight degree the rule gains*. As a result, not only is the conflict problem resolved, but also the number of rules is reduced significantly [2]. Note, that if two or more generated fuzzy rules have the same preconditions and consequents, then the rule that has maximum degree is used. In this way, assigning the degree to each rule, the fuzzy rule base can be adapted or updated by the relative weighting strategy: the more task related the rule becomes, the more weight degree the rule gains. As a result, not only is the conflict problem resolved, but also the number of rules is reduced significantly. After the structure-learning phase *(if-then* rules), the whole network structure is established, and the network enters the second learning phase to optimally adjust the parameters of the membership functions using the GA algorithm to minimise the error function.

4 Inference Through FUNEUS

A functional block diagram of the FUNEUS model consists of two phases of learning processes: a) The first phase is the structure-learning *(if-then* rules) phase using the knowledge acquisition module producing the *fuzzy neurule base. b)* The second phase is the parameter-learning phase for tuning membership functions to achieve a desired level of performance with the use of a Genetic Algorithm to tune membership functions to fine-tune the parameters of the fuzzy membership functions. In the connectionist structure, the input and output nodes represent the input states and output decision signals, respectively, and in the hidden layers, there are nodes functioning as quantification of membership functions (MFs) and *if-then* rules. As soon as the initial input data set is given and put in the Working Memory (WM), the resulting output *fuzzy neurules* are considered for evaluation. One of them is selected for evaluation. Selection is based on textual order. A rule succeeds if the output of the corresponding fuzzy adaline unit is computed to be '1', after evaluation of its conditions. A condition evaluates to 'true', if it matches a fact in the WM, where there is a fact with the same variable, predicate and value. A condition cannot be evaluated if there is no fact in the WM with the same variable. In this case, either a question is

made to the user to provide data for the variable, in case of an input variable, or an intermediate fuzzy neurule in Fuzzy Neurule Base (FNRB) with a conclusion containing that variable is examined, in case of an intermediate variable. A condition with an input variable evaluates to 'false', if there is a fact in the WM with the same variable, predicate and different value. A condition with an intermediate variable evaluates to 'false' if additionally to the latter there is no unevaluated intermediate fuzzy neurule in the FNRB that has a conclusion with the same variable. Inference stops either when one or more output fuzzy neurules are fired (success) or there is no further action (failure). In the training phase input membership functions and desired output values of the training data-set are used in a text-form to fine tune the parameters of the fuzzy sets with the offline use of the GA component.

5 Model Validation in Medical Domain

The experimental data set was taken from Takumi Ichimura and Katsumi Yoshida [13] that prepared a medical database named Coronary Heart Disease DataBase (CHD_DB), which makes it possible to assess the effectiveness of classification methods in medical data. It includes more than 10,000 records related to the development of coronary heart disease (CHD). We used FUNEUS to develop a diagnostic system that outputs whether each record is a non-CHD case or a CHD case. Data in coronary heart diseases database are divided into two classes: non-oronary heart disease cases (non-CHD) and coronary heart disease cases (CHD). Each patient's disorder is diagnosed according to the results of eight test items. The eight items tested are Cholesterol (TC), Systolic Blood Pressure (SBP), Diastolic Blood Pressure (DBP), Drinking (ALCOHOL) that are used as inputs with fuzzy values, Left Ventricular Hypertrophy (LVH), Origin (ORIGIN), Education (EDUCATE), Smoking (TABACCO), and that used as inputs with non-fuzzy value. The fuzzy inference system was created using the FUNEUS. We used triangular membership functions and each input variable were assigned three MFs. Ten fuzzy rules were created using the methodology mentioned in Section 3 (Rule format: IF (…) THEN CHD) :

Rule 4 : *(SBP medium) AND (DBP > high) AND (LVH = 0)AND (EDUCATE < 2) AND (TABACCO > 0) AND (TABACCO < 3) …*
Rule 10 : *(TC high) AND (SBP high) AND (DBP > high) AND (LVH = 0) AND (TABACCO > 0)AND (TABACCO < 3) AND (ALCOHOL medium)*

We also explored the fine-tuning of membership functions using DE algorithm. We started with a population size 10, tournament selection strategy, mutation rate 0.01 and implemented a one point crossover operator. After a trial and error approach by increasing the population size and the number of iterations (generations), we finalized the population size and number of iterations as 50. We used Train_Z, set consisted of 400 CHD cases, and 3600 non-CHD cases. The judgment accuracy for 4000 training data was as follows: results were correct for 310 cases of 400 CHD cases, and were false for 2683 of 3600 non-CHD cases. Therefore, the recognition rate to the set of training data was 77.3%, comparable with machine learning method used in [15].

6 Discussion and Related Work

A number of NF systems have been used for various real life problems. Among others, three neuro-fuzzy systems that use hybrid learning for rules generation and parameter tuning like FUNEUS presented in this paper have already been described [3]. The first one is called NEFCON (NEuro-Fuzzy CONtroller) and used for control applications. The next one is NEFCLASS (NEuro-Fuzzy CLASSifier) and used for classification problems and pattern recognition [5]. The third one is NEFPROX (Neuro-Fuzzy function for approximation) and used for function approximation. The three systems are based on generic fuzzy perceptron which is a kind of FNN. Each of these systems can learn a rule-base and then tune the parameters for the membership functions. Additionally, fuzzy weights are associated with linguistic terms. Connections between them are weighted with fuzzy sets instead real numbers. In NEFCON, inputs are state variables and the only one output neuron outputs control action applied to a technical system. In NEFCLASS, the rule-base approximates a function for a classification problem and maps an input pattern to proper class using no membership functions in the rule's consequents [3]. There are also examples of NF systems that do not employ any algorithms for rule generating, so the rule base must be known in advance. They can only adjust parameters of the antecedent and consequent fuzzy sets. The most popular system of this kind is ANFIS (Adaptive-Network-based Fuzzy Inference System), which is one of the first hybrid NF systems for function approximation. The architecture is five-layer feed-forward, implementing Tagaki-Sugeno type of rules [12]. According to Abraham [3], Sugeno type systems are high performers but often require complicated learning procedures and are computationally expensive. An optional design of a NF system can only be achieved by the adaptive evolution of membership functions, rule base and learning rules. For each architecture, evolution of membership functions proceeds at a faster time scale in an environment decided by the problem, the architecture and the inference system. Thus, global search of fuzzy rules and membership functions provide the fastest possible time scale [3]. The main problem here is that the resulting rules are usually not acceptable by the experts, because they are meaningless. So, it is desirable to tune a set of predefined rules rather to produce a set from data [15]. In this paper, we present FUNEUS, which is a NF system based on fuzzy Adaline neuron and uses Genetic Algorithms for optimization of membership functions. Experimental results proved an acceptable performance of the system, but further evaluation in domains such as, Medical Diagnosis [17] and Intelligent Educational Systems [18], is required. The strongest point of our approach lies in the fact that, contrary to other approaches, it produces natural and meaningful rules. Further work could focus in direct experimental comparison, if possible, with other hybrid approaches.

References

1. Bonissone, P.P., Chen, Y.T., Goebel, T., Khedkar, P.: Hybrid Soft Computing Systems: Industrial and Commercial Applications. Proc. of the IEEE 87(9), 1641–1667 (1999)
2. Hatzilygeroudis, I., Prentzas, J.: HYMES: A HYbrid Modular Expert System with Efficient Inference and Explanation. In: Proceedings of the 8th Panhellenic Conference on Informatics, Nicosia, Cyprus, vol. 1, pp. 422–431 (2001)

3. Abraham, A., Nath, B.: Hybrid Intelligent Systems Design-A Review of a Decade of Research, Technical Report, School of Computing and Information Technology, Monash University, Australia (2000)
4. Pedrycz, W.: Heterogeneous Fuzzy Logic Networks: Fundamentals and Development Studies. IEEE Trans. on Neural Networks 15(6), 1466–1481 (2004)
5. Kasabov, N.: Foundations of Neural Networks, Fuzzy Systems and Knowledge Engineering. MIT Press, Cambridge (1996)
6. Konar, A.: Computational Intelligence. Springer, Heidelberg (2005)
7. Gallant, S.I.: Neural Network Learning and Expert Systems. MIT Press, Cambridge (1993)
8. Shavlik, J.: Combining Symbolic & Neural Learning. Machine Learning 14, 321–331 (1994)
9. Rutkowska, D.: NF Architecture and Hybrid Learning. Physica-Verlag Press, Heidelberg (2002)
10. Yamakawa, T.: Pattern recognition hardware system employing fuzzy neuron. In: Proceedings of the International Conf. of Fuzzy Logic and Neural Networks, Lizuka Japan, July 1990, pp. 943–948 (1990)
11. Koutsojannis, C., Hatzilygeroudis, I.: FYNEUS: A Neurofuzzy Approach Based on Fuzzy Adaline Neurons. In: Penserini, L., et al. (eds.) STAIRS 2006, vol. 1, pp. 96–107. IOS Press, Amsterdam (2006)
12. Jang, J.: ANFIS: Adaptive-Network-based-Fuzzy-Inference System. IEEE Trans on Systems, Man and Cybernetics 23, 665–685 (1993)
13. Suka, M., Ichimura, T., Yoshida, K.: Development of Coronary Heart Disease Database. In: Negoita, M.G., Howlett, R.J., Jain, L.C. (eds.) KES 2004. LNCS (LNAI), vol. 3214, pp. 1081–1088. Springer, Heidelberg (2004)
14. Storn, R.: System Design by Constraint Adaptation and Differential Evolution. IEEE Trans. on Evolutionary Computation 3(1), 22–34 (1999)
15. Hara, A., Ichimura, T.: Extraction of Rules from Coronary Heart Disease Database Using Automatically Defined Groups. In: Negoita, M.Gh., Howlett, R.J., Jain, L.C. (eds.) KES 2004. LNCS (LNAI), vol. 3214, pp. 1089–1096. Springer, Heidelberg (2004)
16. Wang, L.X., Mendel, J.M.: Generating Fuzzy Rules by Learning from Examples. IEEETranscation on System, Man and Cybernetics 22(6), 1414–1427 (1992)
17. Dounias, G.D., Linkens, D.A. (eds.): Adaptive Systems and Hybrid Computational Intelligence in Medicine. In: Special Session Proceedings of the EUNITE 2001 Symposium, Tenerife, Spain, December 13-14, 2001, A Publication of the University of the Aegean (2001) ISBN 960-7475-19-4
18. Hatzilygeroudis, I., Giannoulis, C., Koutsojannis, C.: Combining Expert Systems and Adaptive Hypermedia Technologies in a Web Based Educational System. In: ICALT 2005, Kaohsiung, Taiwan (July 5-8, 2005)

A Likelihood Ratio Test for Differential Metabolic Profiles in Multiple Intensity Measurements

Frank Klawonn[1], Claudia Choi[2], Beatrice Benkert[2], Bernhard Thielen[3], Richard Münch[2], Max Schobert[2], Dietmar Schomburg[3], and Dieter Jahn[2]

[1] Department of Computer Science
University of Applied Sciences Braunschweig /Wolfenbüttel
Salzdahlumer Str. 46/48
38302 Wolfenbüttel, Germany
[2] Institute of Microbiology
Technical University of Braunschweig
Spielmannstraße 7
38106 Braunschweig, Germany
[3] Institute of Biochemistry
University of Köln
Zülpicher Straße 47
50674 Köln, Germany

Abstract. High throughput technologies like transcriptomics using DNA arrays or metabolomics employing a combination of gas chromatography with mass spectrometry provide valuable information about cellular processes. However, the measurements are often highly corrupted with noise of the experimental data which makes it sometimes difficult to draw reliable conclusions. Therefore, suitable statistical methods are needed for the evaluation of the experimental data to distinguish changes caused by biological phenomena from random variations due to noise. This paper introduces a likelihood ratio test to multiple metabolome measurements. The method was tested to differentiate differential metabolite compositions obtained from the pathogenic bacterium *Pseudomonas aeruginosa* grown under various environmental conditions.

Keywords: Likelihood ratio test, Pseudomonas aeruginosa, metabolomics.

1 Introduction

Various high throughput technologies enable biologists to access important classes of cellular bio-molecules in order to gain insight into the corresponding biological processes [1]. DNA microarray chips for measuring gene expressions are one popular example of such high throughput technologies. Mass spectrometry is another often employed method that provides information on the presence of molecules of interest such as proteins or metabolites [2]. The measured values provided by these high throughput technologies are usually displayed as peak

B. Apolloni et al. (Eds.): KES 2007/ WIRN 2007, Part II, LNAI 4693, pp. 485–492, 2007.
© Springer-Verlag Berlin Heidelberg 2007

areas or relative intensities. These values are usually compared between different conditions encountered by the analyzed organisms or tissue like variable environmental settings (e.g. aerobic versus anaerobic growth conditions). From the experimental point of view it is important to identify variations in the measured intensities under different conditions. Obtained data are usually subjected to initial normalization steps. Nevertheless, the measurements are usually corrupted by significant noise. Therefore, simply looking at the raw data and comparing intensities might lead to wrong conclusions, if the effect of the noise is ignored. Consequently, statistical methods are needed that enable the researcher to distinguish between significant differences in measured intensities and variations that are caused solely by noise. In order to estimate the influence of noise, it is necessary to have multiple measurements of the same intensities under identical experimental conditions. Unfortunately, a reliable statistical estimation requires a much larger sample size than biological experiments can deliver for reasonable costs. Repetitions of the same noisy measurement usually range between two and ten. However, by taking into account that the strength of the observed noise in the measured intensities does vary randomly, but follows a certain pattern, more reliable statistical conclusions can be drawn. Consequently, the noise estimation is not carried out separately for each analysis and measured bio-molecule, rather for all at the same time. This requires a global noise model.

Although the approach presented in this paper might be applicable to other high throughput technologies, this communication will focus on metabolome data measured by gas chromatography/mass spectrometry (GC/MS, [3]). The analysis of complete metabolomes, usually called metabolomics, has gained much attention during the past few years as an integral part of modern systems biology [4] [5]. Systems biology attempts to deduce computational models of complex cellular processes from high throughput data such as metabolomes, transcriptomes, proteomes and other related data [1].

2 Biological Context

Pseudomonas aeruginosa is a versatile soil bacterium and an important opportunistic pathogen causing persistent infection of immunocompromised patients. Particularly the lung of cystic fibrosis patients is commonly infected by *P. aeruginosa* growing as biofilm-like microcolonies. Due to this microcolony formation along with other molecular strategies *P. aeruginosa* is resistant to antibiotics, impeding patient treatment. In order to understand antibiotics resistance and the adaption to involved environmental settings, two different clinical isolates PA14 and PAO1 were grown under aerobic conditions as biofilms. Strain PA14 causes disease in a broader host range than strain PAO1 due to additional gene products [6]. Additionally planktonic cultures of strain PAO1 were collected in the early stationary phase, late stationary phase, and exponential growth phase as summarized in Table 1 (a).

5 replicates of 150 mg wet weight of the various *P. aeruginosa* strains grown under indicated conditions were collected (Table 1). Cells were harvested,

Table 1. (a) Description of all conditions, in which *P. aeruginosa* was grown and tested in this paper. (b) Number of metabolites in *P. aeruginosa*, which were different (p-value $< 10^{-10}$) between the given conditions.

Condition	Strain	Growth conditions
1	PA14	aerobic, biofilm
2a	PAO1	aerobic, biofilm
2b	PAO1	anaerobic, planktonic, early stationary phase
2c	PAO1	anaerobic, planktonic, late stationary phase
2d	PAO1	anaerobic, planktonic, exponential growth phase

(a)

	1	2a	2b	2c	2d
1		17	11	26	17
2a			17	7	27
2b				24	20
2c					35

(b)

metabolites extracted, derivatized, and analyzed by GC/MS as outlined before [3]. Identification and peak areas were deconvoluted with the software AMDIS [3]. Ribitol served as internal reference standard. Identified key metabolites with significantly altered intensity pattern for the various *P. aeruginosa* strains grown under indicated conditions will be subjected to further investigation of the metabolic network and flux analysis.

3 The Noise Model

The biological experiments considered here provided data in the following form. A number of c conditions were considered where c varies between two and five in the experiments. Even if there were more than two conditions analyzed, the pairwise comparison of conditions was the essential point for the statistical analysis. Clearly, it was sufficient to consider the situation where measurements from two conditions were available. For each condition, the intensity of each metabolite was measured k times. Typical values for k range between three and ten. Out of 176 detected metabolites from *P. aeruginosa* 107 were chemically identified. This number represents only a fraction of all metabolites present in *P. aeruginosa* due to the current limitation of the detection technology. In comparison, for the baker yeast *Saccharomyces cerivisiae* about 560 low molecular weight metabolites have been detected [7].

As mentioned above, it was not possible to deduce reliable statistical statements about the noise or the variance for intensity measurements for a single metabolite under a single condition tested, since a sample size of ten or less is too small. However, under the assumption that the variance followed a certain pattern depending on the true intensity, statistical inference has been carried out in a more reliable way. Before this assumption is described in more detail, necessary preprocessing steps have to be explained.

The first preprocessing step deleted all zero intensities. These intensities do not correspond to measurements, rather to missing values. Therefore, the number of available measurements for the various metabolites from one of the analyzed conditions may vary. If there were too many missing values for a certain metabolite from one analyzed condition – i.e. less than two measurements were left after removing the zeros – this metabolite was excluded from statistical analysis.

Instead of considering the measured intensities directly, a logarithmic transformation was carried out in advance. This preprocessing step is also commonly used in the context of gene expression data evaluation derived from microarray experiments [8,9]. The measured intensities in the experiments considered here ranged from values close to zero up to more than 600,000. A statistical analysis of such data tends to overemphasize the extremely large values and to neglect the smaller values. The logarithmic transformation reduces this effect. Another reason for applying the logarithmic transformation was that the original measured intensities (of one metabolite under one analyzed condition) do not follow a normal distribution, since they always yield positive values. Although the number of data for single metabolites under a single condition was too small to carry out statistical tests for the assumption of normal distribution, our interpretation of the data looked reasonable. After these preprocessing steps, the structure of the data set for one condition was as follows: $\left(x_1^{(1)}, \ldots, x_1^{(k_1)}, \ldots, x_n^{(1)}, \ldots, x_n^{(k_n)}\right)$. There were n metabolites and each tuple

$$x_i^{(1)}, \ldots, x_i^{(k_i)} \tag{1}$$

represented k_i noisy measurements of the same unknown logarithmic intensity μ_i. As outlined before, the basic assumption was that the subsample (1) originated from independent samples with normally distributed data, with unknown mean μ_i and unknown variance σ_i. Figure 1 compares the mean logarithmic intensities on the x-axis with the empirical variances on the y-axis. Because of the small sample sizes, the variances still differ strongly. Nevertheless, there was a tendency that small (logarithmic) intensities were less reliable or more noisy than larger ones. This effect is in accordance with previous experiences and with data from other mass spectrometry experiments [10].

In the context of microarray expression data, Bayesian approaches are very popular to estimate posterior probabilities of differential expressions in order to determine whether observed differences in expressions are significant or not [11,12,13]. The approach outlined in this paper is based on the classical frequentistic approach in statistics applying a likelihood ratio test. For this test a suitable model for the noise was required taking the above mentioned decreasing noise in the logarithmic measurements into account. This was modelled by assuming that a sample of the form (1), representing k_i measurements of the same true logarithmic intensity μ_i, followed a normal distribution with unknown mean μ_i and unknown standard deviation $\sigma_i = \sigma(\mu_i) = \sigma(\mu_i; v)$. The deviation σ_i is a function of the true logarithmic intensity μ_i and a parameter vector v. In [10], it was proposed to choose $\sigma(\mu_i; v) = \sigma(\mu_i; a, r, \lambda) = a + re^{-\lambda\mu_i}$. This means the parameter vector $v = (a, r, \lambda)$ has three components. Component a represents the absolute noise independent of the true underlying logarithmic intensity μ_i, r reflects a relative portion of the noise depending on μ_i and λ determines how fast the relative portion of noise decreases. Of course, other functions are also possible to model the property of decreasing noise with increasing logarithmic intensities. Note that neither the true logarithmic intensities μ_i nor the values for the parameters a, r and λ were known in the considered experiments.

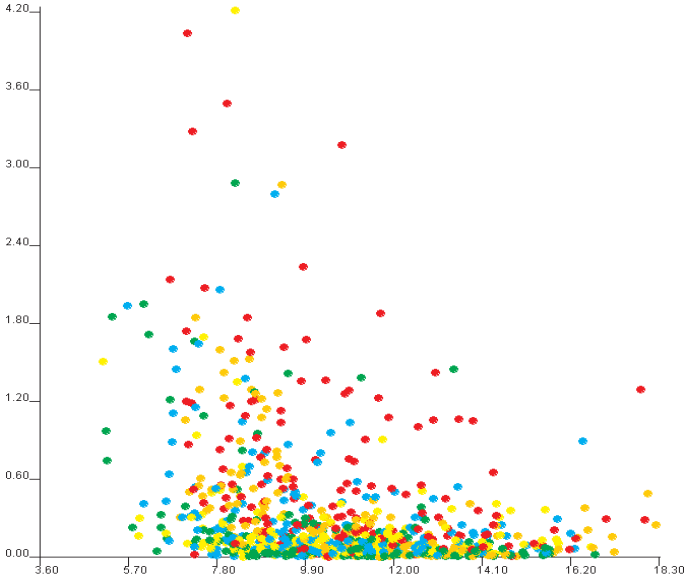


Fig. 1. Means and variances of the logarithmic intensities

4 A Statistical Test for Differential Occurrence

Considering the statistical model for the noise described in the previous section, the task was the following. There were two samples of the form (1) coming from the identified metabolite, however, measured under different conditions. The question was whether the true underlying (logarithmic) intensities μ_1 and μ_2 were identical or not. If there was enough statistical evidence that the two intensities were not identical, then this was a clear indication for a biological cause of the observed difference.

In this section a statistical test is described that provides a p-value for the null hypothesis $\mu_1 = \mu_2$ meaning that the expression levels for the considered metabolite were identical under the two conditions versus the alternative hypothesis that true intensities differed. A low p-value meant that the null hypothesis (no differential expression) had to be rejected with high certainty. The test was derived from the application of the likelihood ratio method based on Wilks' theorem [14], stating that under certain general regularity conditions, the statistic $-2\ln T$ where

$$T = \frac{f_Y\left(Y \mid \tilde{\lambda}^{(\Lambda_0)}\right)}{f_Y\left(Y \mid \tilde{\lambda}^{(\Lambda_1)}\right)} \tag{2}$$

has an approximate χ^2-distribution. $f_Y\left(Y \mid \tilde{\lambda}^{(\Lambda_0)}\right)$ and $f_Y\left(Y \mid \tilde{\lambda}^{(\Lambda_1)}\right)$ are the likelihoods for the sample Y given the parameter vectors $\tilde{\lambda}^{(\Lambda_0)}$ and $\tilde{\lambda}^{(\Lambda_1)}$,

respectively. These parameter vectors are the maximum likelihood estimators under the constraints Λ_0 and Λ_1, respectively. In the case considered here, the parameter vector was $(a, r, \lambda, \mu_1, \mu_2)$, where μ_1 and μ_2 were the maximum likelihood estimates for the true logarithmic intensities of the metabolite from the two considered conditions. For Λ_0 the constraint $\mu_1 = \mu_2$ applies, i.e. the two true intensities are assumed to be identical. In this case the χ^2-distribution of the test statistic has one degree of freedom. The maximum likelihood estimation of the parameters involves numerical methods that are described in [10]. Here, the statistic T in (2) can be written in the form

$$T = \frac{\left(a + re^{-\lambda\mu_1}\right)^{k_1} \cdot \left(a + re^{-\lambda\mu_2}\right)^{k_2}}{\left(a + re^{-\lambda\mu}\right)^{k_1+k_2}} \cdot \frac{\left(\prod_{i=1}^{k_1} \exp\left(\frac{(x_i^{(1)} - \mu)^2}{2\left(a + re^{-\lambda\mu}\right)^2}\right)\right) \cdot \left(\prod_{i=1}^{k_2} \exp\left(\frac{(x_i^{(2)} - \mu)^2}{2\left(a + re^{-\lambda\mu}\right)^2}\right)\right)}{\left(\prod_{i=1}^{k_1} \exp\left(\frac{(x_i^{(1)} - \mu_1)^2}{2\left(a + re^{-\lambda\mu_1}\right)^2}\right)\right) \cdot \left(\prod_{i=1}^{k_2} \exp\left(\frac{(x_i^{(2)} - \mu_2)^2}{2\left(a + re^{-\lambda\mu_2}\right)^2}\right)\right)}.$$

$(x_1^{(1)}, \ldots, x_{k_1}^{(1)})$ and $(x_1^{(2)}, \ldots, x_{k_2}^{(2)})$ are the measured (logarithmic) intensities of the considered metabolite under conditions 1 and 2, respectively. Note that the estimation of the parameters a, r and λ is based on all metabolites and not only on the metabolite under consideration, leading to a much larger sample size.

The statistical test is not only applied to one metabolite, but to a number of n metabolites in a way that several statistical tests are carried out simultaneously. A p-value considered as significant for a single metabolite could be too large in the case of multiple testing. If α is considered as a sufficiently high confidence level for a single test – i.e. p-values smaller than α lead to the rejection of the null hypothesis – a correction has to be carried out for multiple testing. Various methods are available for such corrections [15], for instance the conservative Bonferroni correction using $\frac{\alpha}{n}$ as the corrected confidence level, when n tests are carried out simultaneously. Another possible choice for the corrected confidence level is $1-(1-\alpha)^{1/n}$, which is applicable for two-sided hypotheses, multivariate normal statistics, and positive orthant dependent statistics, but is usually not correct in the general case [16]. In order to be on the safe side, the conservative Bonferroni correction is recommended here, especially since the likelihood ratio test is only an asymptotic, not an exact test.

5 Results and Discussion

Biofilm-like microcolonies of *P. aeruginosa* are resistant to antibiotics due to low oxygen concentration in the microcolony resulting in nearly non-growing cells amongst other factors [17]. Likewise metabolites of an aerobic biofilm PAO1 showed the highest similarity to those from anaerobic planktonic late stationary phase cells (see Table 1 (b)). Metabolomes measured for biofilms of PA14

appeared to be similar to early stationary phase cells, which indicated a more intense growth behaviour of PA14 biofilms compared to PAO1 biofilm cultures. However, the metabolome of strain PA14 in general might be different to that of strain PAO1, hence overinterpretation of these results should be avoided. With the aid of the statistical tests outlined here, we were able to find analogies between the different analyzed conditions.

Table 2. Differences in metabolite composition between late stationary phase cells versus exponential growth phase cells of *P. aeruginosa* strain PAO1 (condition 2c vs 2d) identified by the outlined likelihood ratio test. Metabolites with p-value $< 10^{-10}$ are shown. Ratio and inverse ratio show the n-fold differential occurrence of each metabolite.

Condition 2c vs 2d	Ratio r	$1/r$	p-value	Condition 2c vs 2d	Ratio r	$1/r$	p-value
Trehalose	54.38	0.02	1.8 E-14	1-6-Anhydro-beta	0.12	8.66	5.2 E-15
Quebrachitol	12.21	0.08	3.1 E-15	-D-glucose			
Mannose	8.10	0.12	7.8 E-16	Glucose-6-phosphate	0.11	8.81	3.1 E-15
Isomaltose	5.77	0.17	2.0 E-14	Hexadecanoic acid	0.11	8.92	2.6 E-15
Palmitic acid amide	0.21	4.78	4.4 E-14	Mannose-6-phosphate	0.11	9.39	3.8 E-15
1-Monostearoylglycerol	0.18	5.46	6.4 E-14	Xylose	0.09	11.46	1.2 E-15
Phenylalanine	0.18	5.49	7.1 E-15	Lyxose	0.09	11.57	<1.0 E-16
Oleic acid amide	0.16	6.34	1.3 E-15	Uridine-5'-monophos.	0.07	13.53	5.4 E-15
Xylulose-5-phosphate	0.15	6.56	7.1 E-14	Fructose-6-phosphate	0.07	14.66	4.1 E-15
Phospho-ethanolamine	0.15	6.64	8.9 E-16	Xylulose	0.07	15.30	2.1 E-11
Homoserine	0.15	6.66	5.2 E-14	Isoleucine	0.05	19.71	<1.0 E-16
Glutamic acid	0.14	7.20	3.1 E-15	Diethyleneglycol	0.05	20.49	1.3 E-15
Serine	0.14	7.22	4.4 E-16	Proline	0.05	21.12	<1.0 E-16
Threonine	0.13	7.65	<1.0 E-16	2-Monooleoylglycerol	0.04	23.76	<1.0 E-16
5-Deoxy-5-Methylthioad.	0.13	7.69	5.6 E-16	1-Monopalmitoylglyc.	0.04	26.17	<1.0 E-16
Valine	0.12	8.15	<1.0 E-16	1-Monooleoylglycerol	0.02	47.41	7.5 E-15
Ribulose-5-phosphate	0.12	8.28	1.1 E-15	Shikimic acid	0.01	121.35	1.8 E-14
Adenine	0.12	8.49	3.3 E-16	Leucine	0.01	134.68	3.8 E-15

The most pronounced difference (35 metabolites) was observed between resting cells (2c) and exponentially growing cells (2d). Details are shown in Table 2. Naturally, the activity of metabolic processes is much higher in growing cells compared to those in resting cells as reflected by in the differential metabolite profile. The di- and monosaccharides trehalose, isomaltose, and mannose were found less abundant in growing cells, since these carbon sources are usually rapidly metabolized as energy sources. Furthermore, certain sugars such as sucrose and trehalose are supposed to play a role in the adaption of *P. aeruginosa* to environmental stresses [18]. Metabolites, which are essential for amino acid biosynthesis and subsequent protein biosynthesis (leucine, proline, shikimic acid, isoleucine, valine, serine, glutamic acid, phenylalanine), for DNA and RNA formation (adenine, uridine-5'-monophosphate) and for energy production (glucose-6-phosphate, fructose-6-phosphate, mannose-6-phosphate) were found elevated in growing cells. Furthermore, lipids (palmitic acid amide, phosphoethanolamine, 2-monooleoyl-, 1-monopalmitoyl-, and 1-monooleoylglycerol) were more pronounced in growing cells, since they are essential for membrane biosynthesis.

The statistical tests for metabolome data from *P. aeruginosa* made clear differences accessible between various distinct growth conditions. Obtained results

are in good accordance to common biological interpretation. Further application of this novel statistical approach is also possible for MS/MS-peptide identification and microarray data.

References

1. Joyce, A.R., Palsson, B.Ø.: The model organism as a system: integrating 'omics' data sets. Nat. Rev. Mol. Cell. Biol. 7, 198–210 (2006)
2. Hollywood, K., Brison, D.R., Goodacre, R.: Metabolomics: current technologies and future trends. Proteomics 6, 4716–4723 (2006)
3. Strelkov, S., von Elstermann, M., Schomburg, D.: Comprehensive analysis of metabolites in Corynebacterium glutamicum by gas chromatography/mass spectrometry. Biol. Chem. 385(9), 853–861 (2004)
4. Kell, D.B.: Systems biology, metabolic modelling and metabolomics in drug discovery and development. Drug Discov. Today 11, 1085–1092 (2006)
5. Weckwerth, W., Morgenthal, K.: Metabolomics: from pattern recognition to biological interpretation. Drug Discov. Today 10(22), 1551–1558 (2005)
6. Choi, J.Y., Sifri, C.D., Goumnerov, B.C., Rahme, L.G., Ausubel, F.M., Calderwood, S.B.: Identification of virulence genes in a pathogenic strain of Pseudomonas aeruginosa by representational difference analysis. J. Bacteriol. 184, 952–961 (2002)
7. Förster, J., Gombert, A.K., Nielsen, J.: A functional genomics approach using metabolomics and in silico pathway analysis. Biotechn. Bioeng. 79, 703–712 (2002)
8. Spellman, P.T.: Cluster analysis and display. In: Bowtell, D., Sambrook, J. (eds.) DNA Microarrays, Cold Spring Harbor Lab, pp. 569–581 (2002)
9. Georgieva, O., Klawonn, F., Härtig, E.: Fuzzy clustering of macroarray data. In: Reusch, B. (ed.) Computational Intelligence, Theory and Application, pp. 83–94. Springer, Heidelberg (2005)
10. Klawonn, F., Hundertmark, C., Jänsch, L.: A maximum likelihood approach to noise estimation for intensity measurements in biology. In: Tsumoto, S., Clifton, C., Zhong, N., Wu, X., Liu, J., Wah, B., Cheung, Y.M. (eds.) Proc. 6th IEEE Intern. Conf. on Data Mining: Workshops, pp. 180–184. IEEE, Los Alamitos (2006)
11. Baldi, P., Long, A.D.: A Bayesian framework for the analysis of microarray expression data: Regularized *t*-test and statistical inferences of gene changes. Bioinformatics 17, 509–519 (2001)
12. Cui, X., Hwang, J.T.G., Qiu, J., Blades, N.J., Churchill, G.A.: Improved statistical tests for differential gene expression by shrinking variance components. Biostatistics 6, 59–75 (2005)
13. McLachlan, G.J., Bean, R.W., Ben-Tovim Jones, L.: A simple implementation of a normal mixture approach to differential gene expression in multiclass microarrays. Bioinformatics 22, 1608–1615 (2006)
14. Wilks, S.S.: The large-sample distribution of the likelihood ratio for testing composite hypotheses. Ann. Math. Statistics 9, 60–62 (1938)
15. Miller, R.G.J.: Simultaneous Statistical Inference. Springer, New York (1991)
16. Shaffer, J.P.: Multiple hypothesis testing. Ann. Rev. Psych. 46, 561–584 (1995)
17. Borriello, G., Werner, E., Roe, F., Kim, A.M., Ehrlich, G.D., Stewart, P.S.: Oxygen limitation contributes to antibiotic tolerance of Pseudomonas aeruginosa in biofilms. Antimicrob. Agents Chemother. 48, 2659–2664 (2004)
18. Cioni, P., Bramanti, E., Strambini, G.B.: Effects of sucrose on the internal dynamics of azurin. Biophys. J. 88(6), 4213–4222 (2005)

Combining Bagging, Boosting and Dagging for Classification Problems

S.B. Kotsianti and D. Kanellopoulos

Educational Software Development Laboratory
Department of Mathematics
University of Patras, P.A. Box: 1399, Rio 26 500
sotos@math.upatras.gr, dkanellop@teipat.gr

Abstract. Bagging, boosting and dagging are well known re-sampling ensemble methods that generate and combine a diversity of classifiers using the same learning algorithm for the base-classifiers. Boosting algorithms are considered stronger than bagging and dagging on noise-free data. However, there are strong empirical indications that bagging and dagging are much more robust than boosting in noisy settings. For this reason, in this work we built an ensemble using a voting methodology of bagging, boosting and dagging ensembles with 8 sub-classifiers in each one. We performed a comparison with simple bagging, boosting and dagging ensembles with 25 sub-classifiers, as well as other well known combining methods, on standard benchmark datasets and the proposed technique had better accuracy in most cases.

Keywords: Machine learning, data mining, ensembles of classifiers.

1 Introduction

Both empirical observations and specific machine learning applications confirm that a given learning algorithm outperforms all others for a specific problem or for a specific subset of the input data, but it is unusual to find a single expert achieving the best results on the overall problem domain [5]. As a consequence multiple learner systems (an ensemble of classifiers) try to exploit the local different behavior of the base learners to enhance the accuracy and the reliability of the overall inductive learning system. Numerous methods have been suggested for the creation of ensemble of classifiers [16]. Mechanisms that are used to build ensemble of classifiers include: i) Using different subset of training data with a single learning method, ii) Using different training parameters with a single training method, iii) Using different learning methods. An accessible and informal reasoning, from statistical, computational and representational viewpoints, of why ensembles can improve results is provided in [5]. The key for success of ensembles is whether the classifiers in a system are diverse enough from each other, or in other words, that the individual classifiers have a minimum of failures in common. If one classifier makes a mistake then the others should not be likely to make the same mistake.

B. Apolloni et al. (Eds.): KES 2007/ WIRN 2007, Part II, LNAI 4693, pp. 493–500, 2007.
© Springer-Verlag Berlin Heidelberg 2007

Three of the most popular ensemble algorithms are bagging [4], boosting [6] and dagging [14]. There are two major differences between bagging and boosting. Firstly, boosting changes adaptively the distribution of the training set based on the performance of previously created classifiers while bagging changes the distribution of the training set stochastically. Secondly, boosting uses a function of the performance of a classifier as a weight for voting, while bagging uses equal weight voting. On the other hand, dagging creates a number of disjoint, stratified folds out of the data and feeds each chunk of data to a copy of the supplied base learner, while predictions are made via majority vote. Boosting algorithms are considered stronger than bagging and dagging on noise-free data; however, bagging and dagging are much more robust than boosting in noisy settings. For this reason, in this work, we built an ensemble combing bagging, boosting and dagging version of the same learning algorithm using the sum voting methodology. We performed a comparison with simple bagging, boosting and dagging ensembles as well as other known ensembles on standard benchmark datasets and the proposed technique had better accuracy in most cases. For the experiments, decision stump algorithm was used as base learner.

Section 2 presents the most well known methods for building ensembles that are based on a single learning algorithm, while section 3 discusses the proposed ensemble method. Experiment results using a number data sets and comparisons of the proposed combining method with other ensembles are presented in section 4. We conclude with summary and further research topics in Section 5.

2 Ensembles of Classifiers

There is a growing realization that combinations of classifiers can be more effective than single classifiers. Why rely on the best single classifier, when a more reliable and accurate result can be obtained from a combination of several? This essentially is the reasoning behind the idea of multiple classifier systems. This section provides a brief survey of methods for constructing ensembles using a single learning algorithm. This set of ensemble creating techniques relies on varying the data in some way. Methods of varying the data include; sampling, use of different data sources, use of different pre-processing methods, distortion, and adaptive re-sampling.

Probably the most well-known sampling approach is that exemplified by bagging [4]. Given a training set, bagging generates multiple bootstrapped training sets and calls the base model learning algorithm with each of them to yield a set of base models. Given a training set of size t, bootstrapping generates a new training set by repeatedly (t times) selecting one of the t examples at random, where all of them have equal probability of being selected. Some training examples may not be selected at all and others may be selected multiple times. A bagged ensemble classifies a new example by having each of its base models classify the example and returning the class that receives the maximum number of votes. The hope is that the base models generated from the different bootstrapped training sets disagree often enough that the ensemble performs better than the base models.

Breiman [4] made the important observation that instability (responsiveness to changes in the training data) is a prerequisite for bagging to be effective. A committee of classifiers that all agree in all circumstances will give identical performance to any

of its members in isolation. If there is too little data, the gains achieved via a bagged ensemble cannot compensate for the decrease in accuracy of individual models, each of which now sees an even smaller training set. On the other end, if the data set is extremely large and computation time is not an issue, even a single flexible classifier can be quite adequate.

Another method that uses different subsets of training data with a single learning method is the boosting approach [6]. It assigns weights to the training instances, and these weight values are changed depending upon how well the associated training instance is learned by the classifier; the weights for misclassified instances are increased. Thus, re-sampling occurs based on how well the training samples are classified by the previous model. Since the training set for one model depends on the previous model, boosting requires sequential runs and thus is not readily adapted to a parallel environment. After several cycles, the prediction is performed by taking a weighted vote of the predictions of each classifier, with the weights being proportional to each classifier's accuracy on its training set. AdaBoost is a practical version of the boosting approach [6].

Dagging [14] creates a number of disjoint, stratified folds out of the data and feeds each chunk of data to a copy of the supplied base learner. Predictions are made via majority vote. MultiBoosting [15] is another method of the same category that can be considered as wagging committees formed by AdaBoost. Wagging is a variant of bagging; bagging uses re-sampling to get the datasets for training and producing a weak hypothesis, whereas wagging uses re-weighting for each training example, pursuing the effect of bagging in a different way.

Melville and Mooney [9] present a new meta-learner (DECORATE, Diverse Ensemble Creation by Oppositional Re-labeling of Artificial Training Examples) that uses an existing "strong" learner (one that provides high accuracy on the training data) to build a diverse committee. This is accomplished by adding different randomly constructed examples to the training set when building new committee members. These artificially constructed examples are given category labels that disagree with the current decision of the committee, thereby directly increasing diversity when a new classifier is trained on the augmented data and added to the committee.

3 Proposed Methodology

Recently, several authors [1], [4], [14] have proposed theories for the effectiveness of bagging, boosting and dagging based on bias plus variance decomposition of classification error. In this decomposition we can view the expected error of a learning algorithm on a particular target function and training set size as having three components:

1. A bias term measuring how close the average classifier produced by the learning algorithm will be to the target function;
2. A variance term measuring how much each of the learning algorithm's guesses will vary with respect to each other (how often they disagree); and
3. A term measuring the minimum classification error associated with the Bayes optimal classifier for the target function.

Unlike bagging and dagging, which is largely a variance reduction method, boosting appears to reduce both bias and variance. After a base model is trained, misclassified training examples have their weights increased and correctly classified examples have their weights decreased for the purpose of training the next base model. Clearly, boosting attempts to correct the bias of the most recently constructed base model by focusing more attention on the examples that it misclassified. This ability to reduce bias enables boosting to work especially well with high-bias, low-variance base models. As mentioned in [1] the main problem with boosting seems to be robustness to noise. This is expected because noisy examples tend to be misclassified, and the weight will increase for these examples. Bagging uses a voting technique which is unable to take into account the heterogeneity of the instance space. When the majority of the base classifiers give a wrong prediction for a new instance then the majority vote will result in a wrong prediction. The problem may consist in discarding base classifiers (by assigning small weights) that are highly accurate in a restricted region of the instance space because this accuracy is swamped by their inaccuracy outside the restricted area. It may also consist in the use of classifiers that are accurate in most of the space but still unnecessarily confuse the whole classification committee in some restricted areas of the space. The advantage of boosting over bagging is that boosting acts directly to reduce error cases, whereas bagging works indirectly.

For additional improvement of the prediction of a classifier, we suggest combing bagging, boosting and dagging methodology with sum rule voting (Vote B&B&D). When the sum rule is used each sub-ensemble has to give a confidence value for each candidate. In our algorithm, voters express the degree of their preference using as confidence score the probabilities of sub-ensemble prediction. Next all confidence values are added for each candidate and the candidate with the highest sum wins the election. It has been observed that for bagging, boosting and dagging, an increase in committee size (sub-classifiers) usually leads to a decrease in prediction error, but the relative impact of each successive addition to a committee is ever diminishing. Most of the effect of each technique is obtained by the first few committee members [1], [4], [6]. We used 8 sub-classifiers for each sub-ensemble for the proposed algorithm. The proposed ensemble is effective owing to representational reason. The hypothesis space h may not contain the true function f (mapping each instance to its real class), but several good approximations. Then, by taking weighted combinations of these approximations, classifiers that lie outside of h may be represented. It must be also mentioned that the proposed ensemble can be easily parallelized (one machine for each sub-ensemble). This parallel execution of the presented ensemble can reduce the training time.

4 Comparisons and Results

For the comparisons of our study, we used 36 well-known datasets mainly from many domains from the UCI repository [2]. These data sets were hand selected so as to come from real-world problems and to vary in characteristics. Thus, we have used data sets from the domains of: pattern recognition (anneal, iris, mushroom, zoo),

image recognition (ionosphere, sonar), medical diagnosis (breast-cancer, breast-w, colic, diabetes, heart-c, heart-h, heart-statlog, hepatitis, lymphotherapy, primary-tumor) commodity trading (autos, credit-g) music composition (waveform), computer games (kr-vs-kp, monk1, monk2), various control applications (balance), language morphological analysis (dimin) [3] and prediction of student dropout (student) [8].

In order to calculate the classifiers' accuracy, the whole training set was divided into ten mutually exclusive and equal-sized subsets and for each subset the classifier was trained on the union of all of the other subsets. Then, cross validation was run 10 times for each algorithm and the average value of the 10-cross validations was calculated [16]. In the following Table, we represent with "*" that the specific ensemble looses from the base classifier. That is, the specific algorithm performed statistically better than the specific ensemble according to t-test with $p<0.01$. In addition, in Table, we represent with "v" that the base classifier looses from the specific ensemble according to t-test with $p<0.01$. In all the other cases, there is no significant statistical difference between the results (Draws). It must be mentioned that the conclusions are based on the resulting differences for $p<0.01$ because a p-value of 0.05 is not strict enough, if many classifiers are compared in numerous data sets [12]. In the last rows of the Table one can also see the aggregated results in the form (a/b/c). In this notation "a" means that the specific ensemble algorithm is significantly more accurate than the base algorithm in a out of 36 data sets, "c" means that the base algorithm is significantly more accurate than the specific ensemble in c out of 36 data sets, while in the remaining cases (b), there is no significant statistical difference between the results.

For bagging, boosting and dagging, much of the reduction in error appears to have occurred after ten to fifteen classifiers. But Adaboosting continues to measurably improve their test-set error until around 25 classifiers [11]. For this reason, we used 25 sub-classifiers for our experiments. The time complexity of the proposed ensemble is about the same with simple bagging, boosting and dagging with 25 sub-classifiers. This happens because we use 8 sub-classifiers for each sub-ensemble (totally 24). The proposed ensemble also use less time for training than both Multiboost and Decorare combining methods. In the following subsection, we present the experiment results using Decision stump [7] as base classifier.

4.1 Using Decision Stump as Base Classifier

Decision stumps (DS) are one level decision trees [10] that classify instances by sorting them based on feature values [7]. Each node in a decision stump represents a feature in an instance to be classified, and each branch represents a value that the node can take. Instances are classified starting at the root node and sorting them based on their feature values. At worst a decision stump will reproduce the most common sense baseline, and may do better if the selected feature is particularly informative.

We compare the proposed methodology with bagging, boosting, dagging and MultiBoost version of DS (using 25 sub-classifiers). In addition, we compare the presented ensemble with DECORATE combining method using DS as base classifier. In the last raw of the Table 1 one can see the aggregated results.

Table 1. Comparing the proposed ensemble with other well known ensembles that uses as base classifier the DS

Datasets	DS	Bagging DS	Dagging DS	Adaboost DS	Multiboost DS	Decorate DS	VOTE B&B&D DS
anneal	77.17	82.62 v	83.63 v	83.63 v	83.63 v	76.89	83.43 v
audiology	46.46	46.46	36.34 *	46.46	46.46	46.46	46.46
autos	44.9	44.9	44.52	44.9	44.9	51.81	44.9
badges	100	100	100	100	100	100	100
balance-scale	56.72	68.88 v	81.1 v	71.77 v	71.77 v	81.25 v	75.64 v
breast-cancer	69.27	73.44	72.53	71.55	71.76	75.18 v	73.9 v
breast-w	92.33	92.63	95.05 v	95.28 v	95.05 v	95.78 v	95.14 v
colic	81.52	81.52	79.13	82.72	82.9	82.03	82.14
credit-g	70	70	70.16	72.6	71.8	69.71	70.33
diabetes	71.8	72.55	75.56	75.37	75.19	76.09 v	75.03
dimin	59.31	59.31	59.53	59.31	59.31	64.75 v	59.31
glass	44.89	44.99	47.21 v	44.89	44.89	53.12 v	50.69 v
haberman	71.57	72.74	73.01	74.06	73.8	71.61	72.85
heart-c	72.93	75.64	80.67 v	83.11 v	83.54 v	72.43	82.8 v
heart-h	81.78	81.37	81.71	82.42	81.91	81.78	81.78
heart-statlog	72.3	75.04	81.19 v	81.81 v	82.89 v	81.48 v	81.81 v
hepatitis	77.62	80.72	79.45	81.5	82.21	80.02	80.64
hypothyroid	95.39	95.39	94.97	92.97 *	92.97 *	95.39	95.33
ionosphere	82.57	82.54	81	92.34 v	90 v	90.4 v	86.73 v
iris	66.67	70.33	78.13 v	95.07 v	94.73 v	93.93 v	95.07 v
kr-vs-kp	66.05	66.05	66.94	95.08 v	93.9 v	90.43 v	94.09 v
lymphography	75.31	74.63	73.61	75.44	74.96	72.25 *	75.5
monk1	73.41	73.41	65.68 *	69.79 *	70.37	70.94 *	73.09
monk2	59.58	61.31	55.88	53.99	54.19	61.95	58.28
primary-tumor	28.91	28.91	27.37	28.91	28.91	29.09	29.32
segment	28.52	56.54 v	61.05 v	28.52	28.52	53.91 v	56.66 v
sick	96.55	96.55	96.5	97.07	97.14	96.57	96.5
sonar	72.25	73.16	70.29	81.06 v	77.58	72.91	75.26
soybean	27.96	27.88	43.51 v	27.96	27.96	41.42 v	30.76
students	87.22	87.22	87.16	87.16	86.95	87.1	87.22
titanic	77.6	77.6	77.6	77.83	77.62	77.6	77.6
vote	95.63	95.63	95.61	96.41	95.63	95.59	95.52
vowel	17.47	23.52 v	34.84 v	17.47	17.47	32.08 v	27.75 v
waveform	56.82	57.41	67.5 v	67.68 v	66.44 v	68.7 v	60.45 v
wine	57.91	85.16 v	71.21 v	91.57 v	91.17 v	96.45 v	91.74 v
zoo	60.43	60.63	39.51 *	60.43	60.43	61.96	60.43
W/D/L		*5/31/0*	*12/21/3*	*11/23/2*	*10/25/1*	*15/19/2*	*14/22/0*

The proposed ensemble is significantly more accurate than single DS in 14 out of the 36 data sets, while it has significantly higher error rate in none data set. In addition, the Bagging DS is significantly more accurate than single DS in 5 out of the 36 data sets, whilst it has significantly higher error rate in none data set. Furthermore, Adaboost DS and Decorate DS have significantly lower error rates in 11 and 15 out of the 36 data sets than single DS, respectively whereas they are significantly less accurate in two data sets. Multiboost DS has significantly lower error rates in 10 out of the 36 data sets than single DS, whereas it is significantly less accurate in one data set. Dagging DS has significantly lower error rates in 12 out of the 36 data sets than single DS, whereas it is significantly less accurate in 3 data sets.

To sum up, the performance of the proposed ensemble is more accurate than the other well-known ensembles that use only the DS algorithm (more significant wins than looses in relation to the base algorithm in the used data sets). The proposed ensemble can achieve a reduction in error rate about 21% compared to simple DS.

5 Conclusion

An ensemble of classifiers is a set of classifiers whose individual decisions are combined in some way (typically by weighted or unweighted voting) to classify new examples. One of the most active areas of research in supervised learning has been to study methods for constructing good ensembles of classifiers. The main discovery is that ensembles are often much more accurate than the individual classifiers that make them up. The main reason is that many learning algorithms apply local optimization techniques, which may get stuck in local optima. For instance, decision trees employ a greedy local optimization approach, and neural networks apply gradient descent techniques to minimize an error function over the training data. As a consequence even if the learning algorithm can in principle find the best hypothesis, we actually may not be able to find it. Building an ensemble may achieve a better approximation, even if no assurance of this is given [13].

Boosting algorithms are considered stronger than bagging and dagging on noise-free data; however, bagging and dagging are much more robust than boosting in noisy settings. In this work we built an ensemble using a voting methodology of bagging, boosting and dagging ensembles. It was proved after a number of comparisons with other ensembles, that the proposed methodology gives better accuracy in most cases. The proposed ensemble has been demonstrated to (in general) achieve lower error than either boosting or bagging or dagging when applied to a base learning algorithm and learning tasks for which there is sufficient scope for both bias and variance reduction. The proposed ensemble can achieve an increase in classification accuracy of the order of 21% compared to the tested base classifier.

Our approach answers to some extent such questions as generating uncorrelated classifiers and control the number of classifiers needed to improve accuracy in the ensemble of classifiers. While ensembles provide very accurate classifiers, too many classifiers in an ensemble may limit their practical application. To be feasible and competitive, it is important that the learning algorithms run in reasonable time. In our method, we limit the number of sub-classifiers to 8 in each sub-ensemble.

Finally, there are some open problems in ensemble of classifiers, such as how to understand and interpret the decision made by an ensemble of classifiers because an ensemble provides little insight into how it makes its decision. For learning tasks such as data mining applications where comprehensibility is crucial, voting methods normally result in incomprehensible classifier that cannot be easily understood by end-users. These are the research topics we are currently working on and hope to report our findings in the near future.

References

1. Bauer, E., Kohavi, R.: An empirical comparison of voting classification algorithms: Bagging, boosting, and variants. Machine Learning 36, 105–139 (1999)
2. Blake, C.L., Merz, C.J.: UCI Repository of machine learning databases. Irvine, CA: University of California, Department of Information and Computer Science (1998), http://www.ics.uci.edu/m learn/MLRepository.html
3. Bosch, A., Daelemans, W.: Memory-based morphological analysis. In: proceedings of 37th Annual Meeting of the ACL, University of Maryland, pp. 285–292 (1999), http://ilk.kub.nl/ãntalb/ltuia/week10.html
4. Breiman, L.: Bagging Predictors. Machine Learning 24(3), 123–140 (1996)
5. Dietterich, T.G.: Ensemble methods in machine learning. In: Kittler, J., Roli, F. (eds.) MCS 2000. LNCS, vol. 1857, pp. 1–15. Springer, Heidelberg (2000)
6. Freund, Y., Schapire, R.E.: Experiments with a New Boosting Algorithm. In: proceedings of ICML'96, pp. 148–156 (1996)
7. Iba, W., Langley, P.: Induction of one-level decision trees. In: proceedings of Ninth International Machine Learning Conference (1992). Aberdeen, Scotland (1992)
8. Kotsiantis, S., Pierrakeas, C., Pintelas, P.: Preventing student dropout in distance learning systems using machine learning techniques. In: Palade, V., Howlett, R.J., Jain, L. (eds.) KES 2003. LNCS, vol. 2774, pp. 267–274. Springer, Heidelberg (2003)
9. Melville, P., Mooney, R.: Constructing Diverse Classifier Ensembles using Artificial Training Examples. In: proceedings of IJCAI-2003, Acapulco, Mexico, pp. 505–510 (August 2003)
10. Murthy: Automatic Construction of Decision Trees from Data: A Multi-Disciplinary Survey. Data Mining and Knowledge Discovery 2, 345–389 (1998)
11. Opitz, D., Maclin, R.: Popular Ensemble Methods: An Empirical Study. Artificial Intelligence Research 11, 169–198 (1999)
12. Salzberg, S.: On Comparing Classifiers: Pitfalls to Avoid and a Recommended Approach. Data Mining and Knowledge Discovery 1, 317–328 (1997)
13. Schapire, R.E., Freund, Y., Bartlett, P., Lee, W.S.: Boosting the margin: A new explanation for the effectiveness of voting methods. The Annals of Statistics 26, 1651–1686 (1998)
14. Ting, K.M., Witten, I.H.: Stacking Bagged and Dagged Models. In: Fourteenth international Conference on Machine Learning, San Francisco, CA, pp. 367–375 (1997)
15. Webb, G.I.: MultiBoosting: A Technique for Combining Boosting and Wagging. Machine Learning 40, 159–196 (2000)
16. Witten Ian, H., Eibe, F.: Data Mining: Practical machine learning tools and techniques, 2nd edn. Morgan Kaufmann, San Francisco (2005)

Functional Clustering and Functional Principal Points

Nobuo Shimizu[1] and Masahiro Mizuta[2]

[1] The Institute of Statistical Mathematics, 4-6-7 Minami-Azabu, Minato-ku, Tokyo 106-8569 Japan
[2] Information Initiative Center, Hokkaido University, Kita 11 Nishi 5, Kita-ku, Sapporo 060-0811, Japan

Abstract. In this paper, we deal with the relation between functional clustering and functional principal points. The k principal points are defined as the set of k points which minimizes the sum of expected squared distances from every points in the distribution to the nearest point of the set, and are mathematically equivalent to centers of gravity by k-means clustering[3]. The concept of principal points can be extended for functional clustering. We call the extended principal points *functional principal points*. *Random function*[5] is defined in a probability space, and functional principal points of a random function have a close relation to functional data analysis. We derive functional principal points of polynomial random functions using orthonormal basis transformation. For functional data following Gaussian random functions, we discuss the relation between the optimum clustering of the functional data and the functional principal points.

Keywords: Functional data analysis, k-means algorithm, Optimization.

1 Introduction

Functional data analysis are proposed as methods to analyze complex data [9][11]. Many classical multidimensional data analysis methods are extended for functional data [1][6][7][8][12].

On the other hand, principal points [3] are equivalent to optimal 'cluster means' in a probability distribution by k-means clustering. Tarpey and Kinateder extended the concept of principal points for functional data analysis [13]. We call the extended principal points *functional principal points*. Whereas principal points are defined for a probability distribution, functional principal points are defined for a random function.

In this paper, we deal with the relation between functional clustering and functional principal points. In Section 2, we introduce the definitions of principal points and functional principal points, functional clustering algorithm and orthonormal basis transformation of random functions. In Section 3, we classify n functional data the random function ξ into k clusters, and compare the k centers of the optimal clusters with the k functional principal points of ξ.

B. Apolloni et al. (Eds.): KES 2007/ WIRN 2007, Part II, LNAI 4693, pp. 501–508, 2007.
© Springer-Verlag Berlin Heidelberg 2007

2 Functional Principal Points

2.1 Definition of Principal Points

In this section, we describe the definitions of principal points and the domain of attraction according to Flury [3][4].

Let $W = \{\boldsymbol{y}_1, \cdots, \boldsymbol{y}_k\}, \boldsymbol{y}_j \in R^p$ $(1 \leq j \leq k)$ denote a set of k points in p-dimensional space. The minimal distance between $\boldsymbol{x} \in R^p$ and W is defined as

$$d(\boldsymbol{x}|\boldsymbol{y}_1, \cdots, \boldsymbol{y}_k) = \min_{1 \leq j \leq k} \{(\boldsymbol{x} - \boldsymbol{y}_j)'(\boldsymbol{x} - \boldsymbol{y}_j)\}^{1/2}.$$

Definition 1 (Flury[3])
$\boldsymbol{\xi}_j \in R^p$ $(1 \leq j \leq k)$ are called k principal points of p-variate random variable X, which is distributed according to p-variate probability distribution F if

$$E_F\{d^2(X|\boldsymbol{\xi}_1, \cdots, \boldsymbol{\xi}_k)\} = \min_{\boldsymbol{y}_1, \cdots, \boldsymbol{y}_k} E\{d^2(X|\boldsymbol{y}_1, \cdots, \boldsymbol{y}_k)\}.$$

Let

$$A_j = \{\boldsymbol{x} \in R; (\boldsymbol{x} - \boldsymbol{y}_j)'(\boldsymbol{x} - \boldsymbol{y}_j) < (\boldsymbol{x} - \boldsymbol{y}_i)'(\boldsymbol{x} - \boldsymbol{y}_i), {}^{\forall} i \neq j\} \tag{1}$$

denote *the domain of attraction of* $\boldsymbol{y}_j$ [4]. Flury defined *self-consistent* as follows:

Definition 2 (Flury[4])
The set of k points $\{\boldsymbol{y}_1, \cdots, \boldsymbol{y}_k\}$ is called self-consistent (with respect to the p-variate random variable X) if

$$E[X|X \in A_j] = \boldsymbol{y}_j \quad (1 \leq j \leq k) \tag{2}$$

where the A_j are defined as in expression (1).

2.2 Definition of Functional Principal Points of Random Functions

In this section, we describe the definition of functional principal points according to Tarpey and Kinateder [13].

Let (Ω, F, P) be a probability space. Let E denote an integration with respect to the probability measure P. Let T be a real interval and we consider $L^2(T)$, the complete separable Hilbert space consisting of all square-integrable measurable functions on T. The associated norm and the inner product on L^2 will be denoted by $||\cdot||$ and $(\cdot, \cdot)$ respectively. Let ν denote the space of measurable maps from Ω to L^2, an element of ν will be called *a random function* [5].

Let $y_1, \cdots, y_k$ denote k distinct nonrandom functions in $L^2(T)$ and define $A_j = \{y \in L^2(T); ||y_j - y||^2 < ||y_i - y||^2, {}^{\forall} i \neq j\}$. A_j is called *the domain of attraction of* y_j.

Functional principal points can be defined as the extension of Definition 1 as follows:

Definition 3 (Tarpey and Kinateder [13])
Let $\xi \in L^2(T)$ denote a continuous random function. For a given set of k distinct nonrandom functions $x_1, \cdots, x_k$, with respective domains of attraction A_j, $j = 1, \cdots, k$, define another random function x as $x = \sum_{j=1}^{k} x_j I_{\{\xi \in A_j\}}$. The functions $\{x_1, \cdots, x_k\}$ are called the set of k functional principal points of ξ if $E||\xi - x||^2 \leq E||\xi - y||^2$ for all y of the form $y = \sum_{j=1}^{k} y_j I_{\{\xi \in D_j\}}$, with k distinct nonrandom functions $y_1, \cdots, y_k$ and respective domains of attraction $D_1, \cdots, D_k$.

2.3 *K*-Means Functional Clustering

We suppose functional data consist of n functions: $f_1(t)$, $\cdots$, $f_n(t)$, with $t \in [T_L, T_U]$. To cluster the functions using k-means algorithm[2], one begins with k arbitrary and distinct initial cluster means: $m_1(t)$, $\cdots$, $m_k(t)$. Let C_j denote the cluster formed by centering on $m_j(t)$. Then each function $f_i(t)$ would be assigned to the cluster based on the closest cluster mean according to the squared $L^2[T_L, T_U]$ distance:

$$\min_j [\int_{T_L}^{T_U} (m_j(t) - f_i(t))^2 dt]. \tag{3}$$

After all of the functions $f_i(t)$ have been assigned to clusters, the cluster means are updated as follows:

$$m_j(t) = \sum_{f_i(t) \in C_j} f_i(t)/n_j, \tag{4}$$

where n_j is the number of functions in C_j. This procedure continues until no point (function) is assigned to the other cluster.

When $m_1(t)$, $\cdots$, $m_k(t)$ minimize

$$q(k) = \sum_{j=1}^{k} \sum_{y_i(t) \in C_j} \int_{T_L}^{T_U} (m_j(t) - f_i(t))^2 dt, \tag{5}$$

$m_1(t)$, $\cdots$, $m_k(t)$ are called *optimal cluster centers.*

2.4 Orthonormal Basis Transformation of Random Functions

When $g_1(t)$, $\cdots$, $g_p(t)$ are defined as p orthonormal bases of $\xi(t)$ in $L^2[T_L, T_U]$, we can write $\xi(t)$ as

$$\xi(t) = \sum_{i=1}^{p} b_i g_i(t) \tag{6}$$

with p coefficients b_1, $\cdots$, b_p. Here, $g_1(t)$, $\cdots$, $g_p(t)$ satisfy

$$\int_{T_L}^{T_U} \{g_i(t)\}^2 dt = 1 \quad (i = 1, \cdots, p) \tag{7}$$

and

$$\int_{T_L}^{T_U} g_i(t)g_j(t)dt = 0 \quad (i \neq j). \tag{8}$$

The p-dimensional coefficient vector $(b_1, \cdots, b_p)'$ follows p-variate distribution. By using the orthonormal basis transformation, we can correspond the k principal points in p-dimensional space spanned by $\{g_1(t), \cdots, g_p(t)\}$ to the k functional principal points of $\xi(t)$.

3 Optimal Functional Clustering and Functional Principal Points

On functional data analysis, we often have the case that observed functional data are distributed according to an unknown random function. To confirm how the functional principal points of unknown random functions are reliable, we classify nonrandom n functional data following the random function ξ into k clusters. Based on this result, we compare the k center functions of the optimal clusters and the k functional principal points of ξ.

Example. Suppose the polynomial random functions in $L^2[0,1]$:

$$\xi(t) = a_1 + a_2 t + a_3 t^2, \tag{9}$$

where $(a_1, a_2, a_3)'$ following trivariate normal distribution $N_3((0,0,0)', \Phi_3)$. If $g_1(t)$, $g_2(t)$, $g_3(t)$ satisfy

$$g_1(t) = 1, \quad g_2(t) = \sqrt{12}(t - 1/2), \quad g_3(t) = \sqrt{180}(t^2 - t + 1/6). \tag{10}$$

The polynomial random function (9) can be rewritten by using (6), with (10) and

$$b_1 = a_1 + a_2/2 + a_3/3, \quad b_2 = (a_2 + a_3)/\sqrt{12}, \quad b_3 = a_3/\sqrt{180}. \tag{11}$$

The 3-dimensional coefficient vector $(b_1, b_2, b_3)'$ follows $B_3 \Phi_3 A_3 B_3^{-1}$ where

$$A_3 = \begin{pmatrix} 1 & 1/2 & 1/3 \\ 1/2 & 1/3 & 1/4 \\ 1/3 & 1/4 & 1/5 \end{pmatrix}, \qquad B_3 = \begin{pmatrix} 1 & 1/2 & 1/3 \\ 0 & 1/\sqrt{12} & 1/\sqrt{12} \\ 0 & 0 & 1/\sqrt{180} \end{pmatrix}.$$

Suppose Φ_3 satisfies

$$B_3 \Phi_3 A_3 B_3^{-1} = \mathrm{diag}(\varsigma_1^2, \varsigma_2^2, \varsigma_3^2). \tag{12}$$

We generate n functional data following $\xi(t)$ and classify them into $k = 3$ clusters. In functional clustering, the initial values of $\{m_1(t), \cdots, m_k(t)\}$ have a big influence on the result of the optimal clusters. Then, we prepared $N = 1000$ sets of initial values.

Fig.1 shows a simulated set of $n = 250$ functional data following $\xi(t)$ with $\varsigma_1^2 = \varsigma_2^2 = \varsigma_3^2 = 1$. We classify the functional data set into $k = 3$ clusters by using functional clustering algorithm (see Section 2.3). We could get many local

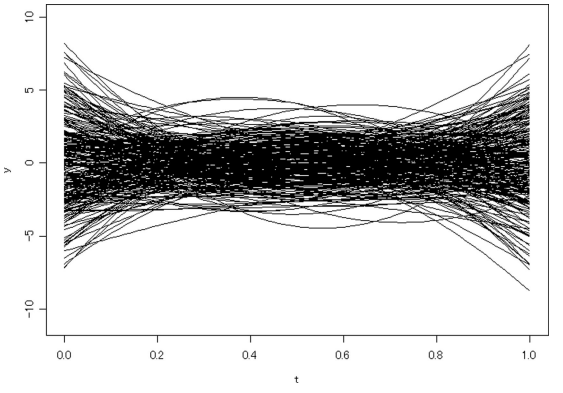


Fig. 1. A simulated set of $n = 250$ functional data following $y = \xi(t) = a_1 + a_2 t + a_3 t^2$

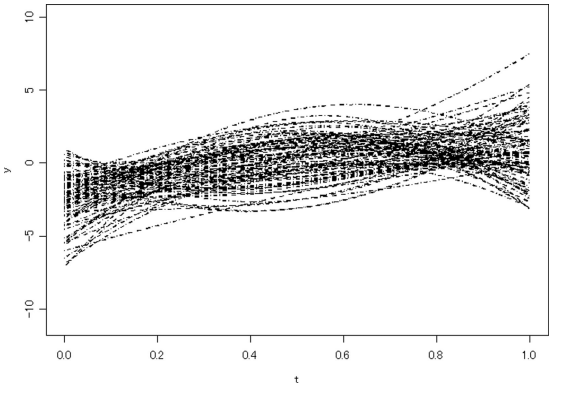


Fig. 2. Functional data in Cluster 1 (at the optimal clusters)

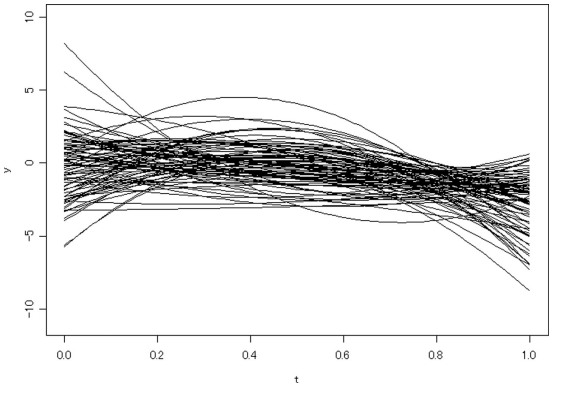


Fig. 3. Functional data in Cluster 2 (at the optimal clusters)

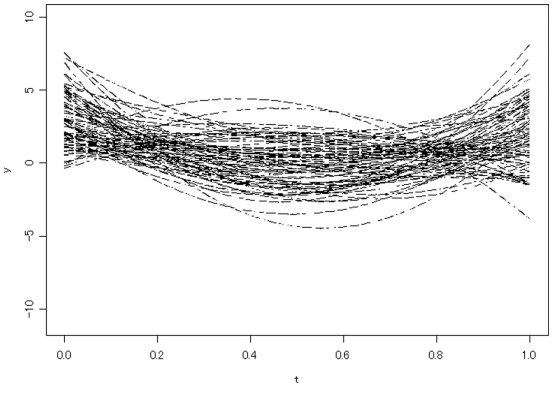


Fig. 4. Functional data in Cluster 3 (at the optimal clusters)

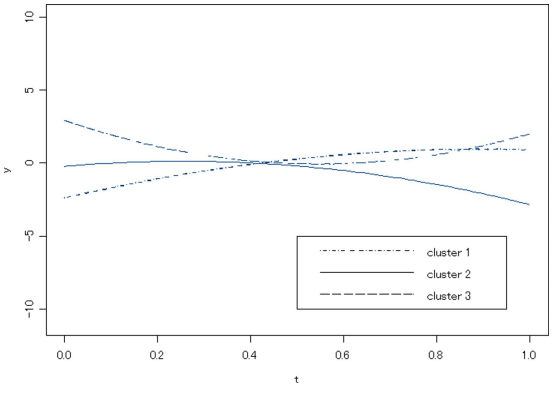


Fig. 5. $k = 3$ center functions of the optimal clusters

minima by using different initial values, and chose the clusters which minimized loss function $q(k)$ (see (5)). Fig.2, 3, 4 are the optimal clusters of the functional data by k-means functional clustering. We show $k = 3$ center functions of the optimal clusters in Fig.5.

Let (9) transform into (6) with (10) and (11). Then, k functional principal points are the k functions corresponding to k principal points in the space spanned by $\{g_1(t), g_2(t), g_3(t)\}$. If $\varsigma_1^2 = \varsigma_2^2 = \varsigma_3^2 = 1$ are satisfied, $k = 3$ principal points in the space spanned by $\{g_1(t), g_2(t), g_3(t)\}$ are not unique, because $(b_1, b_2, b_3)'$ follows the trivariate standard normal distribution and the set of $k = 3$ principal points has freedom of rotation that centers on the expectation of the distribution. Fig.6 shows an example set of $k = 3$ functional principal points of ξ as the nearest $k = 3$ center functions of the optimal clusters in Fig.5. The 3 functions in Fig.5 and the 3 functional principal points in Fig.6 are very close. Therefore, when nonrandom functional data following ξ are classified into

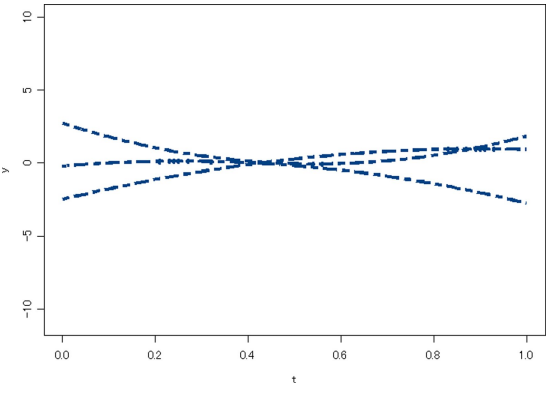


Fig. 6. $k = 3$ functional principal points of $\xi(t)$

k clusters, the k center functions of the optimal clusters can be considered a good approximation of the k functional principal points of ξ.

4 Conclusion

k-means functional clustering is 'equivalent' to clustering the coefficients of the orthonormal basis transformation of the functions [13]. However, as for theoretical researches of functional clustering and functional principal points, more detailed researches are requested.

We showed a close relationship between k functional principal points of a Gaussian random function and k center functions of the optimal clusters of nonrandom functional data following random functions.

This means that for observed functional data according to an unknown random function, k-means functional clustering to the functional data can estimate functional principal points of the unknown random function.

References

1. Abraham, C., Cornillion, P.A., Matzner-Lober, E., Molinari, N.: Unsupervised curve-clustering using B-splines. Scandinavian Journal of Statistics 30, 581–595 (2003)
2. Forgy, E.: Cluster analysis of multivariate data: Efficiency versus interpretability of classification. Biometrics 21, 768–780 (1965)
3. Flury, B.: Principal points. Biometrika 77(1), 33–41 (1990)
4. Flury, B.: Estimation of principal points. Journal of the Royal Statistical Society, Series C, Applied Statistics 42(1), 139–151 (1993)
5. Ibragimov, I.A., Rozanov, Y.A.: Gaussian Random Processes. Springer, Heidelberg (1978)
6. James, G.M., Hastie, T.J., Sugar, C.A.: Principal component models for sparse functional data. Biometrika 87, 587–602 (2000)

7. Mizuta, M.: Functional multidimensional scaling. In: Proceedings of the 10th Japan and Korea Joint Conference of Statistics, pp. 77–82 (2000)
8. Mizuta, M.: K-means method for functional data. Bulletin of International Statistical Institute, 54th Session, Book 2, 69–71 (2003)
9. Ramsay, G.O.: When the data are functions. Psychometrika 47, 379–396 (1982)
10. Ramsay, G.O., Silverman, B.W.: Applied Functional Data Analysis: Methods and Case Studies. Springer, Heidelberg (2002)
11. Ramsay, G.O., Silverman, B.W.: Functional Data Analysis, 2nd edn. Springer, Heidelberg (2005)
12. Rossi, F., Conan-Guez, B., El Golli, A.: Clustering functional data with the SOM algorithm. In: Proceedings of European Symposium on Artificial Neural Networks, pp. 305–312 (2004)
13. Tarpey, T., Kinateder, K.: Clustering functional data. Journal of Classification 20, 93–114 (2003)

Prescreening of Candidate Rules Using Association Rule Mining and Pareto-optimality in Genetic Rule Selection

Hisao Ishibuchi, Isao Kuwajima, and Yusuke Nojima

Department of Computer Science and Intelligent Systems, Graduate School of Engineering, Osaka Prefecture University
1-1 Gakuen-cho, Naka-ku, Sakai, Osaka 599-8531, Japan
hisaoi@cs.osakafu-u.ac.jp, kuwajima@ci.cs.osakafu-u.ac.jp, nojima@cs.osakafu-u.ac.jp

Abstract. Genetic rule selection is an approach to the design of classifiers with high accuracy and high interpretability. It searches for a small number of simple classification rules from a large number of candidate rules. The effectiveness of genetic rule selection strongly depends on the choice of candidate rules. If we have hundreds of thousands of candidate rules, it is very difficult to efficiently search for their good subsets. On the other hand, if we have only a few candidate rules, rule selection does not make sense. In this paper, we examine the use of Pareto-optimal and near Pareto-optimal rules with respect to support and confidence as candidate rules in genetic rule selection.

Keywords: Data mining, classifier design, genetic rule selection, evolutionary multiobjective optimization, multiobjective machine learning.

1 Introduction

Association rule mining [2] is a well-known data mining technique. In its basic form, all association rules satisfying the minimum support and confidence are extracted. Its application to classification problems is often referred to as classification rule mining [17]. Classification rule mining consists of two phases: rule discovery and rule selection. In the rule discovery phase, a large number of classification rules are extracted from numerical data using an association rule mining technique. A part of extracted rules are selected to design a classifier in the rule selection phase using a heuristic rule sorting criterion. In this paper, we use genetic algorithms for rule selection.

Genetic algorithm-based rule selection was first proposed to design an accurate and compact fuzzy classifier in [15] where a weighted sum fitness function was used to maximize the classification accuracy and minimize the number of fuzzy rules. An evolutionary multiobjective optimization (EMO) algorithm was used to search for Pareto-optimal fuzzy classifiers with respect to these two objectives in [11]. The total number of antecedent conditions was introduced as the third objective in [12] to minimize not only the number of fuzzy rules but also their

B. Apolloni et al. (Eds.): KES 2007/ WIRN 2007, Part II, LNAI 4693, pp. 509–516, 2007.
© Springer-Verlag Berlin Heidelberg 2007

length. The use of a memetic EMO algorithm was examined in [16]. Genetic rule selection was also used for non-fuzzy classification rule mining in [13,14].

Whereas the basic idea of association rule mining is to extract all association rules that satisfy the minimum support and confidence [2], other measures have also been proposed to evaluate an association rule. Among them are gain, variance, chi-squared value, entropy gain, gini, laplace, lift, and conviction [3]. It is shown in [3] that the best rule according to any of these measures is a Pareto-optimal rule with respect to support and confidence. The use of an EMO algorithm was proposed to search for Pareto-optimal rules with respect to support and confidence for partial classification [5]. In [19], an EMO algorithm was modified to search for not only Pareto-optimal rules but also near Pareto-optimal rules using the concept of ε-dominance.

In this paper, we examine the use of Pareto-optimal and near Pareto-optimal rules as candidate rules in genetic rule selection. We slightly modify the concept of ε-dominance to remove a large number of (near) weak Pareto-optimal rules.

2 Multiobjective Genetic Rule Selection

Let us assume that we have m training patterns $\mathbf{x}_p = (x_{p1}, x_{p2}, ..., x_{pn})$, $p = 1, 2, ..., m$ from M classes in the n-dimensional pattern space. For this classification problem, we use classification rules of the following type:

$$\text{Rule } R_q: \text{ If } x_1 \text{ is } A_{q1} \text{ and } ... \text{ and } x_n \text{ is } A_{qn} \text{ then Class } C_q \text{ with } CF_q, \quad (1)$$

where R_q is the label of the q-th rule, $\mathbf{x} = (x_1, x_2, ..., x_n)$ is an n-dimensional pattern vector, A_{qi} is an antecedent interval for the i-th attribute, C_q is a class label, and CF_q is a rule weight (i.e., certainty grade). Each antecedent condition "x_i is A_{qi}" in (1) means the inclusion condition "$x_i \in A_{qi}$". We denote the classification rule R_q in (1) as "$\mathbf{A}_q \Rightarrow C_q$" where $\mathbf{A}_q = (A_{q1}, A_{q2}, ..., A_{qn})$. It should be noted that R_q in (1) does not always have n antecedent conditions. Any conditions can be "*don't care*".

Let us denote the support count of $\mathbf{A}_q \Rightarrow C_q$ by $SUP(\mathbf{A}_q \Rightarrow C_q)$, which is the number of patterns compatible with both the antecedent part $\mathbf{A}_q$ and the consequent class C_q. $SUP(\mathbf{A}_q)$ and $SUP(C_q)$ are the number of patterns compatible with $\mathbf{A}_q$ and C_q, respectively. The support and the confidence of the classification rule $\mathbf{A}_q \Rightarrow C_q$ are calculated as follows [2,3]:

$$Support(\mathbf{A}_q \Rightarrow C_q) = \frac{SUP(\mathbf{A}_q \Rightarrow C_q)}{m}, \quad (2)$$

$$Confidence(\mathbf{A}_q \Rightarrow C_q) = \frac{SUP(\mathbf{A}_q \Rightarrow C_q)}{SUP(\mathbf{A}_q)}. \quad (3)$$

In order to generate candidate rules, we first divide the domain interval of each attribute into K disjoint subintervals using an optimal splitting method [8] based on the class entropy measure [18]. Since the choice of an appropriate partition for each attribute is not easy, we simultaneously use four different

partitions with $K = 2, 3, 4, 5$. Thus we have 14 antecedent intervals in total for each attribute.

A large number of candidate rules are generated from the given training patterns using an association rule mining technique based on three threshold values: the minimum support, the minimum confidence, and the maximum rule length. The maximum rule length specifies the upper limit on the number of antecedent conditions in each classification rule. The confidence of each rule is used as the rule weight CF_q.

Let us assume that N candidate rules are generated. We denote its subset by S. In genetic rule selection, S is represented by a binary string of length N. As in [13,14], we use the following three objectives to evaluate each string:

$f_1(S)$: The number of correctly classified training patterns by S,
$f_2(S)$: The number of selected rules in S,
$f_3(S)$: The total number of antecedent conditions over selected rules in S.

The first objective is maximized while the second and third are minimized. When the first objective is to be calculated, each pattern is classified by a single winner rule with the maximum rule weight among activated (i.e., fired) rules in S. We use NSGA-II [6,7] to search for Pareto-optimal rule sets with respect to these three objectives.

3 Prescreening of Candidate Rules

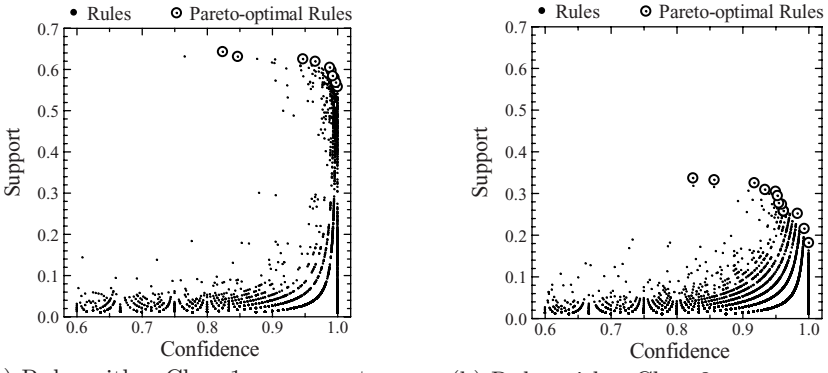
In Fig. 1, we show extracted candidate rules for the Wisconsin breast cancer data set in the UCI machine learning repository. This data set has 683 patterns (excluding incomplete patterns with missing values), which were used to extract candidate rules with the minimum support 0.01, the minimum confidence 0.6, and the maximum rule length 3. The number of extracted candidate rules is 4,729 for Class 1 and 11,480 for Class 2. Appropriate specifications of the three threshold values are not easy (e.g., the minimum support 0.4 seems to be good in Fig. 1 (a), but it is not good in Fig. 1 (b)).

In Fig. 1, we also show Pareto-optimal rules with respect to support and confidence for each class by open circles. The number of Pareto-optimal rules is 10 for Class 1 in Fig. 1 (a) and 14 for Class 2 in Fig. 1 (b). The number of candidate rules becomes too small when we use only Pareto-optimal rules as candidate rules in rule selection.

Using a dominance margin ε, we define ε-dominance in a similar manner to [19], which was also used in fuzzy data mining in our former study [10]. A rule R_i is said to be ε-dominated by another rule R_j when both the two inequalities in (4) hold and at least one of the two inequalities in (5) holds:

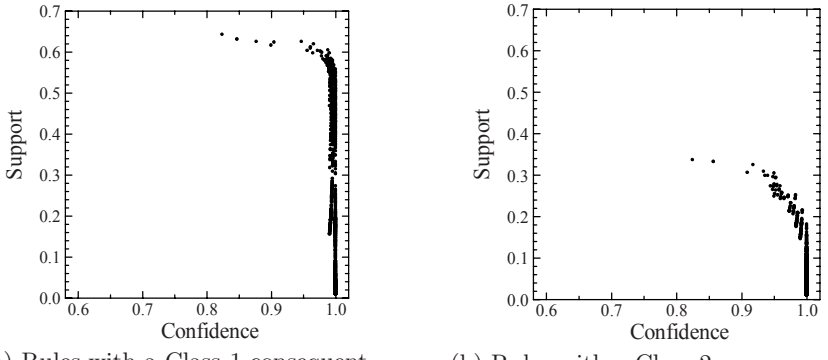
$$Confidence(R_i) + \varepsilon \leq Confidence(R_j),\ Support(R_i) + \varepsilon \leq Support(R_j) \quad (4)$$

$$Confidence(R_i) + \varepsilon < Confidence(R_j),\ Support(R_i) + \varepsilon < Support(R_j) \quad (5)$$



(a) Rules with a Class 1 consequent. (b) Rules with a Class 2 consequent.

Fig. 1. Extracted candidate rules for the Wisconsin breast cancer data. Pareto-optimal rules for each class are shown by open circles.



(a) Rules with a Class 1 consequent. (b) Rules with a Class 2 consequent.

Fig. 2. Candidate rules for the Wisconsin breast cancer data when $\varepsilon = 0.01$

When a rule R_i is not dominated by any other rules in the sense of ε-dominance in (4) and (5), we call R_i an ε-non-dominated rule. Pareto dominance is exactly the same as ε-dominance with $\varepsilon = 0$. When $\varepsilon = \infty$, all rules are ε-non-dominated rules.

We choose only the ε-non-dominated rules with $\varepsilon = 0.01$ as candidate rules from Fig. 1. The obtained candidate rules are shown in Fig. 2. The number of the candidate rules is 2,483 for Class 1 in Fig. 2 (a) and 5,457 for Class 2 in Fig. 2 (b). Whereas we use a very small value of ε, the number of the candidate rules is still very large.

From Fig. 2, we can see that there exist a large number of weak Pareto-optimal rules with the maximum confidence 1.0. From the definition of ε- dominance in (4) and (5), those weak Pareto-optimal rules are never removed from candidate rules unless ε is specified as $\varepsilon = 0$. As we have already shown, the number of candidate rules seems to be too small when we use only Pareto-optimal rules

(i.e., when $\varepsilon = 0$). In order to remove weak Pareto-optimal rules with small support values, we propose an idea to specify the minimum support for each class depending on Pareto-optimal rules and the value of ε as follows (This idea was not used in our former study [10]):

$$s_{\text{Class } h} = \max\{\mathit{Support}(R_q) \mid C_q = \text{Class } h, \mathit{Confidence}(R_q) = 1\} - \varepsilon. \quad (6)$$

All the candidate rules with smaller support values than the minimum support in (6) are removed. When we apply the minimum support with $\varepsilon = 0.01$ in (6) to Fig. 2, the number of the candidate rules is drastically decreased from 2,483 for Class 1 in Fig. 2 (a) to 86 and from 5,457 for Class 2 in Fig. 2 (b) to 121. While the number of candidate rules is drastically decreased, it is still much larger than the number of Pareto-optimal rules in Fig. 1. This means that not only Pareto-optimal rules but also near Pareto-optimal rules are used as candidate rules. The effect of the removal of weak and near weak Pareto-optimal rules by (6) is further examined in Section 4.

4 Computational Experiments

We examined the effect of the proposed prescreening method of candidate rules through computational experiments on benchmark data sets from the UCI machine learning repository. We used NSGA-II [6,7] as an EMO algorithm for multiobjective genetic rule selection (see [1,4,6] for other EMO algorithms and their application areas). Classification rates of obtained non-dominated rule sets on training and test data were evaluated by five independent iterations of the two-fold cross-validation (2-CV) procedure.

The average number of generated candidate rules over five iterations of 2-CV is shown in Table 1 where ε-dominance in (4) and (5) was used with no

Table 1. Average number of candidate rules before the modification of ε-dominance

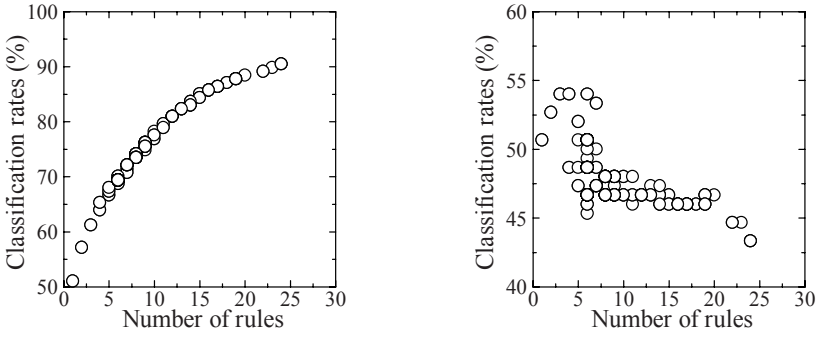
Data set	Value of ε				
	0	0.001	0.01	0.1	∞
Breast W	18	10083	10312	12573	**16828***
Glass	240	6755	7098	10597	**13930***
Heart C	65	4318	4543	7789	**14577***
Wine	27	20465	20489	23984	**35530***

Table 2. Average number of candidate rules after the modification of ε-dominance

Data set	Value of ε				
	0	0.001	0.01	0.1	∞
Breast W	18	34	105	897	**16828***
Glass	240	425	953	7884	**13930***
Heart C	65	118	506	4565	**14577***
Wine	27	199	223	2722	**35530***

Table 3. Average CPU times for multiobjective genetic rule selection from Table 2 (minutes)

Data set	Value of ε				
	0	0.001	0.01	0.1	∞
Breast W	0.8	0.9	1.2	4.8	**81.2***
Glass	2.0	2.3	3.2	16.2	**27.4***
Heart C	1.2	1.3	2.2	11.9	**35.8***
Wine	1.1	1.4	1.5	5.4	**60.2***



(a) Training data accuracy. (b) Test data accuracy.

Fig. 3. Non-dominated rule sets obtained by a single run of NSGA-II for the Cleveland heart disease data set. Candidate rules were prescreened by the proposed method in (6) with $\varepsilon = 0.05$.

modification. On the other hand, the minimum support in (6) was used in Table 2. From Table 1 and Table 2, we can see that the use of the minimum support in (6) drastically decreased the number of candidate rules. The average CPU time for a single run of NSGA-II for multiobjective genetic rule selection from candidate rules in Table 2 is shown in Table 3. All computational experiments were performed on a PC with Pentium D 3.6 GHz (Dual Core) where multiple runs were not executed simultaneously. Table 2 and Table 3 show that the number of candidate rules had a large effect on the computation load for multiobjective genetic rule selection by NSGA-II. The average CPU time was decreased from the case of $\varepsilon = \infty$ by an order or two in Table 3.

A number of non-dominated rule sets were obtained by a single run of NSGA-II. In Fig. 3, we show obtained non-dominated rule sets by its single run for the Cleveland heart disease data set. We chose a single rule set with the highest accuracy on training data in each run of NSGA-II (e.g., the rightmost point in Fig. 3 (a)). The classification rate of the chosen rule set was also calculated for test data (It should be noted that the chosen rule set is not the best rule set with respect to the generalization ability as shown in Fig. 3). Average accuracy of the chosen rule set over five iterations of 2-CV is summarized in Table 4 for training data and Table 5 for test data. In Table 4, training data accuracy was

Table 4. Average classification rates of rule sets with the highest training data accuracy (%)

Data set	Value of ε				
	0	0.001	0.01	0.1	∞
Breast W	97.6	98.0	98.6	99.2	**99.7***
Glass	81.8	83.9	86.0	94.7	**97.0***
Heart C	77.6	79.6	87.7	**96.1***	95.6
Wine	99.5	99.4	99.4	**100.0***	**100.0***

Table 5. Average classification rates on test data of rule sets chosen in Table 4 (%)

Data set	Value of ε				
	0	0.001	0.01	0.1	∞
Breast W	95.6	95.3	95.5	**95.9***	95.0
Glass	61.9	61.4	61.2	62.5	**62.7***
Heart C	**52.1***	51.6	49.7	43.9	40.5
Wine	**92.8***	90.0	89.9	89.9	90.5

gradually degraded by decreasing the value of ε (i.e., by decreasing the number of candidate rules). On the other hand, test data accuracy was not always degraded by decreasing the value of ε as shown in Table 5. The effect of the value of ε on test data accuracy is totally problem-dependent. This is because each problem has a different accuracy-complexity tradeoff curve for test data [9,13].

5 Conclusions

In this paper, we proposed an idea of using Pareto-optimal and near Pareto-optimal rules as candidate rules in genetic rule selection. The proposed idea is based on two tricks: One is the use of ε-non-dominated rules with respect to support and confidence [10], and the other is the specification of the minimum support based on Pareto-optimal rules and the value of ε. A large number of (near) weak Pareto-optimal rules with small support values were removed by the second trick, which is a new idea introduced in this paper. The effectiveness of the proposed prescreening method was demonstrated through computational experiments.

This work was partially supported by Grant-in-Aid for Scientific Research on Priority Areas: KAKENHI (18049065) and Grant-in-Aid for Scientific Research (B): KAKENHI (17300075).

References

1. Abraham, A., Jain, L., Goldberg, R. (eds.): Evolutionary Multiobjective Optimization. Springer, London (2005)
2. Agrawal, R., Mannila, H., Srikant, R., Toivonen, H., Verkamo, A.I.: Fast Discovery of Association Rules. In: Fayyad, U.M., Piatetsky-Shapiro, G., Smyth, P., Uthurusamy, R. (eds.) Advances in Knowledge Discovery and Data Mining, pp. 307–328. AAAI Press, Menlo Park (1996)

3. Bayardo, Jr., R.J., Agrawal, R.: Mining the Most Interesting Rules. In: Proc. of 5th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 145–153 (1999)
4. Coello, C.A.C., Lamont, G.B. (eds.): Applications of Multi-Objective Evolutionary Algorithms. World Scientific, Singapore (2004)
5. de la Iglesia, B., Philpott, M.S., Bagnall, A.J., Rayward-Smith, V.J.: Data Mining Rules using Multi-Objective Evolutionary Algorithms. In: Proc. of 2003 Congress on Evolutionary Computation, pp. 1552–1559 (2003)
6. Deb, K.: Multi-Objective Optimization Using Evolutionary Algorithms. John Wiley & Sons, Chichester (2001)
7. Deb, K., Pratap, A., Agarwal, S., Meyarivan, T.: A Fast and Elitist Multiobjective Genetic Algorithm: NSGA-II. IEEE Trans. on Evolutionary Computation 6, 182–197 (2002)
8. Elomaa, T., Rousu, J.: General and Efficient Multisplitting of Numerical Attributes. Machine Learning 36, 201–244 (1999)
9. Ishibuchi, H., Kuwajima, I., Nojima, Y.: Relation between Pareto-Optimal Fuzzy Rules and Pareto-Optimal Fuzzy Rule Sets. In: Proc. of 2007 IEEE Symposium on Computational Intelligence in Multicriteria Decision Making, pp. 42–49. IEEE Computer Society Press, Los Alamitos (2007)
10. Ishibuchi, H., Kuwajima, I., Nojima, Y.: Use of Pareto-optimal and Near Pareto-optimal Rules as Candidate Rules in Genetic Fuzzy Rule Selection. In: Proc. of IFSA 2007 World Congress (2007) (in press)
11. Ishibuchi, H., Murata, T., Turksen, I.B.: Single-Objective and Two-Objective Genetic Algorithms for Selecting Linguistic Rules for Pattern Classification Problems. Fuzzy Sets and Systems 89, 135–150 (1997)
12. Ishibuchi, H., Nakashima, T., Murata, T.: Three-Objective Genetics-Based Machine Learning for Linguistic Rule Extraction. Information Sciences 136, 109–133 (2001)
13. Ishibuchi, H., Namba, S.: Evolutionary Multiobjective Knowledge Extraction for High-Dimensional Pattern Classification Problems. In: Yao, X., Burke, E.K., Lozano, J.A., Smith, J., Merelo-Guervós, J.J., Bullinaria, J.A., Rowe, J.E., Tiňo, P., Kabán, A., Schwefel, H.-P. (eds.) Parallel Problem Solving from Nature - PPSN VIII. LNCS, vol. 3242, pp. 1123–1132. Springer, Heidelberg (2004)
14. Ishibuchi, H., Nojima, Y.: Accuracy-Complexity Tradeoff Analysis by Multiobjective Rule Selection. In: Proc. of ICDM 2005 Workshop on Computational Intelligence in Data Mining, pp. 39–48 (2005)
15. Ishibuchi, H., Nozaki, K., Yamamoto, N., Tanaka, H.: Selecting Fuzzy If-Then Rules for Classification Problems using Genetic Algorithms. IEEE Trans. on Fuzzy Systems 3, 260–270 (1995)
16. Ishibuchi, H., Yamamoto, T.: Fuzzy Rule Selection by Multi-Objective Genetic Local Search Algorithms and Rule Evaluation Measures in Data Mining. Fuzzy Sets and Systems 141, 59–88 (2004)
17. Liu, B., Hsu, W., Ma, Y.: Integrating Classification and Association Rule Mining. In: Proc. of 4th International Conference on Knowledge Discovery and Data Mining, pp. 80–86 (1998)
18. Quinlan, J.R.: C4.5: Programs for Machine Learning. Morgan Kaufmann, San Mateo (1993)
19. Reynolds, A., de la Iglesia, B.: Rule Induction using Multi-Objective Metaheuristics: Encouraging Rule Diversity. In: Proc. of 2006 International Joint Conference on Neural Networks, pp. 6375–6382 (2006)

Term Clustering in Texts Based on Fuzzy Neighborhoods and Kernel Functions

Sadaaki Miyamoto[1], Satoshi Hayakawa[2], and Yuichi Kawasaki[2]

[1] Department of Risk Engineering
School of Systems and Information Engineering
University of Tsukuba, Ibaraki 305-8573, Japan
miyamoto@risk.tsukuba.ac.jp
[2] Graduate School of Systems and Information Engineering
University of Tsukuba, Ibaraki 305-8573, Japan
{satossu,ttyrkwsk}@soft.risk.tsukuba.ac.jp

Abstract. A fuzzy neighborhood model for analyzing information systems having topological structures on occurrences of keywords is proposed and associated kernel functions are derived. Sufficient conditions when a neighborhood defines a kernel are described. Clustering algorithms with and without a kernel function are developed. Illustrative examples are given.

1 Introduction

Text mining is a hot topic of today and many methods have been proposed. Yet there are rooms for further investigation and existing methods are still considered to be insufficient.

We have proposed a model of fuzzy neighborhoods for text mining and term clustering [9]. This method generalizes and refines the vector space model [11] for text analysis. A feature of this model is to handle terms of which occurrences are dispersed on a universal set with a natural topology.

In this paper we contrast two methods of c-means clustering for this model. One uses a direct definition of a similarity measure and cluster center in accordance with the measure, whereby we develop crisp and fuzzy c-means algorithms. The other is a kernel-based method that induces an inner-product space on the set of terms, whereby we derive another algorithms of the c-means clustering. We show a sufficient condition that the fuzzy neighborhoods provide kernel functions using a theorem by Pólya [10], together with additional propositions.

Illustrative examples are given to show these algorithms correctly work. Finally, implications of the kernel function for the present framework are mentioned.

2 Term Relations and Neighborhoods in Text Sets

We assume two sets: T is called a term set in which elements are denoted by $t, t', t_i, \ldots \in T$; O is called an occurrence space in which elements are denoted by

B. Apolloni et al. (Eds.): KES 2007/ WIRN 2007, Part II, LNAI 4693, pp. 517–524, 2007.
© Springer-Verlag Berlin Heidelberg 2007

$a, b, c, \ldots \in O$. Note also that $|T|$ is the number of elements in T and moreover we put $L = |T|$.

Let the set of documents be $\mathcal{D} = \{d_1, d_2, \ldots, d_n\}$. A document d consists of a sequence of occurrences of terms. We may suppose an occurrence corresponds to a unique term, and the correspondence of a term t to an occurrence a is denoted by $R(t, a)$. Thus a corresponds to a unique t such that $R(t, a) = 1$, while $R(t, \cdot)$ corresponds to multiple occurrences in general.

We handle a sequence of occurrences, and accordingly we define $Sqnc(d)$ is the sequence of occurrences. For example assume $Sqnc(d) = abcde$ and term of a, b, d corresponds to t; term of c and e is t'. Then $Sqnc(d) = ttt'tt'$ using the term symbols.

From technical reason we define concatenation of two document sequences by $Sqnc(d)|Sqnc(d')$. Thus if $Sqnc(d) = abcd$ and $Sqnc(d') = vwxyz$, then

$$Sqnc(d)|Sqnc(d') = abcdvwxyz.$$

The whole sequence X of the document set $\mathcal{D}$ is

$$X = Sqnc(d_1)|Sqnc(d_2)|\cdots|Sqnc(d_n).$$

A natural distance D is defined on X:

$$D(a, b) = \{ \text{ number of occurrences between } a \text{ and } b \} + 1, \quad (a \neq b) \qquad (1)$$

and we obviously set $D(a, a) = 0$. For the above d and d', $D(v, w) = 1$ and $D(v, z) = 4$.

A fuzzy neighborhood $N(a, b)$ is a fuzzy relation of X satisfying the following properties.

(i)[Reflexivity]

$$N(a, a) = 1.$$

(ii)[Symmetry]

$$N(a, b) = N(b, a).$$

(iii)[Implication of a neighborhood–I]

$$\text{If} \quad D(a, b) > D(a, c) \quad \text{then} \quad N(a, b) \leq N(a, c).$$

(iv)[Implication of a neighborhood–II]

$$\lim_{D(a,b)\to+\infty} N(a, b) = 0.$$

We sometimes write $N_a(b) = N(a, b)$ to emphasize b is in the neighborhood of a with the membership $N_a(b)$.

A natural way to define $N_a(b)$ is to use a monotonically decreasing function $f\colon [0, \infty) \to [0, 1]$ such that $\lim_{x\to\infty} f(x) = 0$. We define

$$N_a(b) = f(D(a, b)). \qquad (2)$$

Thus for this distance the fuzzy neighborhood is naturally defined. The next two are typical examples.

Crisp and fuzzy neighborhoods N_{C_K} and $N^{\beta}_{F_K}$

$$N_{C_K}(a,x) = \begin{cases} 1 & (D(a,x) \le K), \\ 0 & (D(a,x) > K). \end{cases} \tag{3}$$

$$N_{F^{\beta}_K}(a,x) = \begin{cases} 1 - \{D(a,x)/K\}^{\beta} & (D(a,x) \le K), \\ 0 & (D(a,x) > K). \end{cases} \tag{4}$$

A neighborhood should not intersect different documents. For this purpose we put dummy occurrences which do not correspond to any term between different documents.

Note also that the above model generalizes the ordinary term-document frequency model [11]. To see this, suppose $N(a,x) = 1$ when a and x are in the same document and $N(a,x) = 0$ when a and x are in different documents. Then it is easily seen that the neighborhood model coincides with the term-document frequency model.

3 Kernel Functions Based on the Neighborhood

Let us define a proximity measure $p(t,t')$ between two terms $t, t' \in T$ and a normalized measure $s(t,t')$ associated with the former.

$$p(t,t') = \sum_{a \in O} \sum_{b \in O} R(t,a) N(a,b) R(t',b) \tag{5}$$

$$s(t,t') = \frac{p(t,t')}{\sqrt{p(t,t)p(t',t')}}. \tag{6}$$

We first consider when $p(t,t')$ becomes an inner product. Note that $s(t,t')$ is a cosine correlation [2], if $p(t,t')$ is an inner product.

Recall that $p(t,t')$ is called positive semi-definite, if and only if all the eigenvalues of P are nonnegative, or in other words,

$$\sum_{t,t' \in T} z_t z_{t'} p(t,t') \ge 0, \quad \forall z_t, z_{t'} \in \boldsymbol{R}^L$$

Furthermore, $p(t,t')$ is called positive definite, if all eigenvalues are positive.

Note 1. In the discussion below, the distinction between the positive definiteness and the positive semi-definiteness is not very important in applications. That is, when $p(t,t')$ is positive semi-definite but not positive definite, we can take an approximation $p(t,t') + \varepsilon I$ instead, where I is the identity matrix and ε is a small positive parameter. Notice $p(t,t') + \varepsilon I$ is obviously positive definite.

3.1 Kernel Functions on Terms and Fuzzy Neighborhoods

Let us consider sufficient conditions when a neighborhood defines a kernel function. It should be noted that a positive definite function defines a kernel. We therefore consider conditions for the positive definiteness of $p(t,t')$ and $s(t,t')$.

To begin with, consider a simple case when $T = O$ and $R(t,a) = 1$ if and only if $t = a$. That is, an occurrence itself is a term. In this case (5) reduces to $p(t,t') = N(t,t')$. Thus, $p(t,t')$ is positive definite if and only if $N(t,t')$ is positive definite.

We proceed to the general case of (5).

Proposition 1. *A sufficient condition such that $p(t,t')$ is positive definite is that $N(a,b)$ is positive definite for all $a,b \in O$.*

The proof is not difficult and is omitted here. From this proposition we should study when $N(a,b)$ is positive definite.

Recall that the fuzzy neighborhood is defined by $N(a,b) = N_a(b) = f(D(a,b))$. We have the following sufficient condition.

Proposition 2. *A sufficient condition such that $N(a,b)$ is positive definite is that the function f is convex on $[0,+\infty)$ and $f(x) \to 0$ as $x \to +\infty$.*

Proof. Take an arbitrary $c \in O$. Define $x(a)$ by $x(a) = D(a,c)$ when a is the left hand of c; $x(a) = -D(a,c)$ when a is right hand of c. Then the conclusion is immediately obtained from the Pólya's theorem [10] which states that

$$\sum_{a,b} z_a z_b f(|x(a) - x(b)|) \geq 0, \quad \forall z_a, z_b \in \boldsymbol{R},$$

when f is convex on $[0,+\infty)$ and $f(x) \to 0$ as $x \to +\infty$. □

For example, the fuzzy neighborhood $N_{F_K^\beta}(a,x)$ of (4) is convex if $0 < \beta \leq 1$, while the crisp neighborhood (3) is not guaranteed to provide a positive definite kernel. Finally, we note the next proposition whose proof is easy and is omitted.

Proposition 3. *A necessary and sufficient condition such that $p(t,t')$ is positive definite is that $s(t,t')$ is positive definite.*

4 Clustering Algorithms

We consider two types of clustering algorithms. The first does not use a kernel, while the second is based on the inner product space induced by a positive definite kernel.

4.1 Hard and Fuzzy *c*-Means Clustering Without a Kernel Function

Let us consider hard and fuzzy c-means in which the measure $s(t,t')$ is used (cf. [9]). For this purpose assume G_i is a fuzzy cluster:

$$G_i = \sum_{t_k \in T} u_{ki}|t_k,$$

in which $\sum u_{ik}|t_k$ is an abbreviated notation of the fuzzy set G_i where the membership is $\mu_{G_i}(t_k) = u_{ki}$, $t_k \in T$. Put also

$$\|G_i\|^2 = \sum_k u_{ki}^2 \|t_k\|^2.$$

A cluster center, denoted by z_i, is assumed to be a fuzzy set of terms

$$z_i = \sum_{t_k \in T} w_{ki}|t_k \tag{7}$$

We should determine w_{ki} that satisfies

$$\max_z s(G_i, z_i).$$

where we put

$$s(G_i, z) = \frac{\sum_{t_k \in T} \sum_{a \in O} \sum_{b \in O} u_{ki} R(t_k, a) N(a, b) R(z, b)}{\|G_i\| \sqrt{\sum_{t_k \in T} w_{ki}^2 \|t_k\|^2}}.$$

The detailed calculation of z_i is omitted but the solution of w_{ki} is given by

$$w_{ki} = \left(\sum_{t_r \in T} \frac{\|t_k\|^4 (\sum_{t_s \in T} u_{si} X_r)^2}{\|t_r\|^2 (\sum_{t_s \in T} u_{si} X_k)^2} \right)^{-\frac{1}{2}},$$

where

$$X_k = \sum_{a \in O} \sum_{t' \in T} \sum_{b \in O} R(t_k, a) N(a, b) R(t', b).$$

For crisp c-means, the membership is determined by the nearest center allocation rule:

$$u_{ki} = \begin{cases} 1 & (i = \arg \max\limits_{1 \le j \le c} s(t_k, z_j)), \\ 0 & (\text{otherwise}). \end{cases}$$

Hence iterative calculation of the last equation for u_{ki} and the cluster center z_i until convergence leads to the solution of the c-means clustering.

Note 2. If the entropy-based method should be used, u_{ki} is given by the next formula:

$$u_{ki} = \frac{\exp(\lambda s(t_k, z_i))}{\sum_j \exp(\lambda s(t_k, z_j))}$$

Note also if we use the standard method of fuzzy c-means [1], we should not use the similarity measure $s(t, t')$ but to use a dissimilarity such as $D(t, t') = 1 - s(t, t')$ is necessary. We omit the detail but in the next section we will use this dissimilarity measure for all kernel-based algorithms.

4.2 Kernel-Based Method

Kernel-based hard and fuzzy c-means algorithms have been derived in [6,7,8]. Hence $p(t,t')$ can be used as a kernel for the c-means clustering [1]. Frequently the cosine correlation $s(t,t')$ gives more satisfactory results than $p(t,t')$, however.

Let us state the algorithm of the standard fuzzy c-means clustering [1] using $s(t,t')$ as the kernel function. The use of a kernel function implies that we transform data into a high-dimensional feature space H by a mapping $\Phi(x)\colon T \to H$. This means that we should use the dissimilarity measure

$$\begin{aligned} D(t,t') &= \|\Phi(t) - \Phi(t')\|^2 = \langle \Phi(t), \Phi(t)\rangle + \langle \Phi(t'), \Phi(t')\rangle - 2\langle \Phi(t), \Phi(t')\rangle \\ &= s(t,t) + s(t',t') - 2s(t,t'). \end{aligned}$$

Noting $s(t,t) = s(t',t') = 1$, we have the dissimilarity $D(t_k, t_l) = 2 - 2s(t_k, t_l)$. The initial cluster centers $z_1, \ldots, z_c$ are given by random selection from T and accordingly we can calculate $D(t_k, z_i)$ by the last formula.

The membership is defined by

$$u_{ki} = \left[\sum_{j=1}^{c} \left(\frac{D(t_k, z_i)}{D(t_k, z_j)} \right)^{\frac{1}{m-1}} \right]^{-1} \tag{8}$$

Then the distance $D(t_k, z_i)$ should be updated by the next formula (cf. [6,8]):

$$D(t_k, z_i) = 2 - \frac{2}{\sum_{l=1}^{L} (u_{li})^m} \sum_{l=1}^{L} (u_{li})^m s(t_k, t_l), \quad (L = |T|). \tag{9}$$

These equations (8) and (9) should be repeated until convergence.

Note 3. We can derive the algorithms of the crisp c-means and the entropy-based fuzzy c-means clustering likewise. For example (8) should be replaced by

$$u_{ki} = \frac{\exp(-\lambda D(t_k, z_i))}{\sum_{j=1}^{c} \exp(-\lambda D(t_k, z_j))}$$

and $m = 1$ in (9) when the entropy-based method is used [5,8]. If we use the crisp c-means clustering,

$$u_{ki} = \begin{cases} 1 & (i = \arg \min\limits_{1 \le j \le c} D(t_k, z_j)), \\ 0 & (\text{otherwise}). \end{cases}$$

and $m = 1$ in (9).

5 Illustrative Examples

A toy example that shows the proposed algorithms correctly work and another example of a Japanese text have been analyzed.

Example 1. A toy example that consists of the next sequence is considered, where a character stands for a term. The next example consists of three classes: $\{A, B, C, D, C\}$, $\{i, j, k, l\}$, and $\{X, Y, Z\}$. Four data sets, one of which is shown next, have been prepared. In this example, a colon shows separation of sections such that a neighborhood does not intersect between two sequences with a colon. Moreover, a document is identified with such a sequence as below.

```
E,C,A:i,j,l:X,Y:j,i,k:D,E,B:X,Z,Y:i,l,j:E,C,B:
X,Z,Z:j,l:B,D:k,i:X,Z:B,A:i,l:B,A,A,C,D:X,Y,Y
```

The neighborhood has been given by $N_{F_K^\beta}(a, x)$ with $\beta = 5$.

Note that a vector space model cannot separate the three classes, since a sequence as above include all characters of the three classes.

The first method of crisp c-means without a kernel has produced sometimes correct results and sometimes not, according to the variation of initial values, The percentage of the correct result ranges from $3\% \sim 36\%$, while the kernel-based standard fuzzy c-means algorithm has given 100% correct result except one example. Moreover the kernel-based entropy fuzzy c-means algorithm has given 100% correct results for all four examples.

Note 4. For confirmation we have applied the agglomerative hierarchical clustering [2,4] to the same example using the similarity $s(t, t')$ and confirmed that the correct results are obtained.

Example 2. A real text in Japanese of an address by Mr. Asoh, the Foreign Minister (see http://www.mofa.go.jp/mofaj/press/enzetsu/, 2006), has been analyzed by the same algorithms using $N_{F_K^\beta}(a, x)$ with $\beta = 5$. Those words that appear three times or more have been taken to be the terms for clustering. Five hundred trials have been made and the solution of the minimum (or maximum) value of the objective function has been taken. When we set the number of clusters $c = 3$, we had the same clusters by the two methods of crisp c-means without a kernel and the standard fuzzy c-means with the kernel:

```
cluster 1: minister, Japan, our country, nation;
cluster 2: SDF, Iraq, value, world, diplomacy;
cluster 3: economics, system, embassy.
```

Note that the original words are in Japanese.

Also, we have not obtained the same results from other methods such as the fuzzy c-means without a kernel, the entropy-based method with the kernel, and so on. We moreover have confirmed that the above three clusters have been obtained by an agglomerative hierarchical clustering algorithm [2].

Note 5. More results based on real data are shown in Japanese in the Master Thesis of Hayakawa [3], where it is shown by several examples that the neighborhood model is superior to the vector space model.

6 Conclusions

In this paper, a model of fuzzy neighborhoods with application to text clustering has been shown and c-means algorithms with and without a kernel have been developed. The examples show that the both methods work correctly and the kernel-based algorithms work better overall.

As noted above, the fuzzy neighborhood model is a generalization of the method of the term-document frequency matrix. Hence the present model has many applications in document/text mining and analysis.

Moreover the approach of kernel functions is important in the sense that various methods in data analysis can be applied to the inner product space defined by a kernel. We therefore have found the way to use many algorithms to the model of fuzzy neighborhood other than data clustering. A typical method is the support vector machine [12]. There are many research opportunities in and around this model.

References

1. Bezdek, J.C.: Pattern Recognition with Fuzzy Objective Function Algorithms. Plenum, New York (1981)
2. Everitt, B.S.: Cluster Analysis, 3rd edn. Arnold, London (1993)
3. Hayakawa, S.: Automatic Classification Algorithms Based on the Concept of Fuzzy Neighborhoods, Master Thesis, Graduate School of Systems and Information Engineering, University of Tsukuba, March (in Japanese) (2007), http://www.soft.risk.tsukuba.ac.jp/miyamoto/lib/exe/fetch.php?id=publications3Ah18&cache=cache&media=publications:hayakawa_masterthesis.pdf
4. Miyamoto, S.: Fuzzy Sets in Information Retrieval and Cluster Analysis. Kluwer, Dordrecht (1990)
5. Miyamoto, S., Mukaidono, M.: Fuzzy c-means as a regularization and maximum entropy approach. In: Proc. of the 7th International Fuzzy Systems Association World Congress (IFSA'97), June 25-30, 1997, Prague, Czech, vol. II, pp. 86–92 (1997)
6. Miyamoto, S., Suizu, D.: Fuzzy c-means clustering using transformations into high dimensional spaces. In: Proc. of FSKD'02: 1st International Conference on Fuzzy Systems and Knowledge Discovery, November 18-22, Singapore, vol. 2, pp. 656–660 (2002)
7. Miyamoto, S., Nakayama, Y.: Algorithms of hard c-means clustering using kernel functions in support vector machines. Journal of Advanced Computational Intelligence and Intelligent Informatics 7(1), 19–24 (2003)
8. Miyamoto, S., Suizu, D.: Fuzzy c-means clustering using kernel functions in support vector machines. Journal of Advanced Computational Intelligence and Intelligent Informatics 7(1), 25–30 (2003)
9. Miyamoto, S., Hayakawa, S.: A Fuzzy Neighborhood Model for Clustering, Classification, and Approximations. In: Greco, S., Hata, Y., Hirano, S., Inuiguchi, M., Miyamoto, S., Nguyen, H.S., Słowiński, R. (eds.) RSCTC 2006. LNCS (LNAI), vol. 4259, pp. 882–890. Springer, Heidelberg (2006)
10. Pólya, G.: Remarks on characteristic functions. In: Proc. First Berkeley Conf. on Math. Stat. and Prob., pp. 115–123. Univ. of Calf Press, Berkeley (1949)
11. Salton, G., McGill, M.J.: Introduction to Modern Information Retrieval. McGraw-Hill, New York (1983)
12. Vapnik, V.N.: Statistical Learning Theory. Wiley, Chichester (1998)

Variable Based Fuzzy Blocking Regression Model

Mika Sato-Ilic

Faculty of Systems and Information Engineering, University of Tsukuba
Tsukuba, Ibaraki 305-8573, Japan
mika@risk.tsukuba.ac.jp

Abstract. A fuzzy clustering based blocking regression model is proposed considering fuzzy intercepts and weights for each variable. The fuzzy intercepts and weights are obtained by using two fuzzy clustering results. One is a conventional fuzzy clustering over all variables and the other uses variable based fuzzy clustering. By involving the fuzzy clustering results, we can implement a regression model including nonlinear spatial data structures which are observed in a space consisting of all of the variables and a space consisting of each variable.

Keywords: Fuzzy clustering, Fuzzy intercept, Spatial analysis, Nonlinear regression, Variable-based clustering.

1 Introduction

Recently, we are faced with the challenge of analyzing vast amounts of complicated data. In order to extract efficient information from the complicated data, several regression models that exploit the spatial classification structure of data have been proposed. Fuzzy c-regression model [5], geographically weighted regression model [2], and fuzzy cluster loading models [6] are typical examples. These models use weights which show the degree of contribution of objects to clusters and represent the distribution of the weights by using fuzzy membership function or probabilistic density function.

A fuzzy regression model has been proposed which attempts to solve this problem by adding different increments according to different clusters (or blocks) [8]. This model is an extension of the conventional blocking regression model [3]. In this model, identification of clusters (or blocks) are represented by dummy variables. The values of the dummy variables are $\{0, 1\}$ and have to be given in advance by using external information to identify the blocks. However, usually, data is observed as numerical continuous values, if we use the conventional blocking regression model with the dummy values, then we need to analyze the mixed data including continuous values and discrete values. Therefore, the fitness is decreased. In order to solve these problems, we have proposed the fuzzy blocking regression model [8] by using fuzzy clustering for identifying the clusters (or blocks). In this model, we exploit the feature of fuzzy clustering so that we can obtain the result as continuous values. Therefore, we can avoid the decrease of the fitness. Moreover,

B. Apolloni et al. (Eds.): KES 2007/ WIRN 2007, Part II, LNAI 4693, pp. 525–532, 2007.
© Springer-Verlag Berlin Heidelberg 2007

we do not need any external information to identify the clusters (or blocks) and so we can consider the substantial reliance on data itself.

In this paper we propose an extended fuzzy blocking regression model involving weights which can show the variable based fuzzy clustering structure for the data. That is, weighting can be done in which the classification structure is clearer, which makes the variable more significant. That is, this weight works to represent how each variable contributes to the classification by using the dissimilarity structure at each variable.

This paper consists of eight sections. The next section describes a fuzzy clustering method. In section 3, we show the variable based fuzzy c-means method and the use to define weights for each variable. In section 4, we briefly mention an ordinal blocking regression analysis. In section 5, we state a regression model using fuzzy block increments. Section 6 proposes a weighted fuzzy blocking regression model. Section 7 shows several numerical examples and section 8 states the conclusions.

2 Fuzzy C-Means Method

The state of fuzzy clustering is represented by a partition matrix $U = (u_{ik})$ whose elements show the grade of belongingness of the objects to the clusters, $u_{ik},\ i = 1, \cdots, n,\ k = 1, \cdots, K$, where n is the number of objects and K is the number of clusters. In general, u_{ik} satisfies the following conditions:

$$u_{ik} \in [0, 1],\ \sum_{k=1}^{K} u_{ik} = 1. \tag{1}$$

Fuzzy c-means (FCM) [1] is one of the fuzzy clustering methods. FCM is the method which minimizes the weighted within-class sum of squares:

$$J(U, v) = \sum_{i=1}^{n} \sum_{k=1}^{K} u_{ik}^{m} \sum_{a=1}^{p} (x_{ia} - v_{ka})^2, \tag{2}$$

where $v = (v_{ka}),\ k = 1, \cdots, K,\ a = 1, \cdots, p$ shows a matrix of centers whose element v_{ka} denotes the value of the center of a cluster k with respect to a variable a. x_{ia} is i-th object with respect to a variable a. p is the number of variables. The exponent m which determines the degree of fuzziness of the clustering is chosen from $[1, \infty)$ in advance. The purpose is to obtain the solutions U and v which minimize equation (2). From conditions shown in equation (1), the local extrema of equation (2) can be obtained as follows:

$$u_{ik} = 1 \Big/ \sum_{t=1}^{K} \left[\frac{\sum_{a=1}^{p} (x_{ia} - v_{ka})^2}{\sum_{a=1}^{p} (x_{ia} - v_{ta})^2} \right]^{\frac{1}{m-1}}, \tag{3}$$

$$v_{ka} = \sum_{i=1}^{n} u_{ik}^{m} x_{ia} \Big/ \sum_{i=1}^{n} u_{ik}^{m},\ \ i = 1, \cdots, n,\ \ k = 1, \cdots, K,\ \ a = 1, \cdots, p.$$

3 Variable Based Fuzzy C-Means Method and Weights for Variables

We define the following objective function [7].

$$F(U_a, v) = \sum_{i=1}^{n}\sum_{k=1}^{K}\sum_{a=1}^{p} u_{ika}^{m}(x_{ia} - v_{ka})^2. \tag{4}$$

Where u_{ika} shows degree of belongingness of an object i to a cluster k with respect to a variable a and is assumed to satisfy the following conditions:

$$u_{ika} \in [0,1], \ \sum_{k=1}^{K} u_{ika} = 1. \tag{5}$$

If $\exists i, k$, u_{ika} is a constant for all a, then this is the same objective function as the fuzzy c-means shown in equation (2). The purpose is to estimate $U_a = (u_{ika})$ and v which minimize equation (4) under conditions shown in equation (5). The local extrema of equation (4) can be obtained as follows:

$$u_{ika} = 1/\sum_{t=1}^{K}\left[\frac{x_{ia} - v_{ka}}{x_{ia} - v_{ta}}\right]^{\frac{2}{m-1}}. \tag{6}$$

$$v_{ka} = \sum_{i=1}^{n} u_{ika}^{m} x_{ia} / \sum_{i=1}^{n} u_{ika}^{m}.$$

Using u_{ika} shown in equation (6), we define the weight for each variable as follows:

$$w_a = \frac{1}{n}\sum_{i=1}^{n}\sum_{k=1}^{K} u_{ika}^{2}, \ a = 1, \cdots, p. \tag{7}$$

Equation (7) shows a partition coefficient [1] of a classification using a dissimilarity matrix with respect to a fixed variable. That is, a value of w_a becomes larger when the classification structure is clearer and becomes smaller when the classification structure is more unclear. w_a satisfies $\frac{1}{K} \leq w_a \leq 1, \ \ \forall a$. In order to normalize the weight w_a into $[0,1]$, we put w_a as follows:

$$w_a \equiv \frac{w_a - \frac{1}{K}}{1 - \frac{1}{K}}. \tag{8}$$

4 Blocking Regression Model

Suppose

$$X = \begin{pmatrix} \boldsymbol{x}_1 \\ \vdots \\ \boldsymbol{x}_n \end{pmatrix}, \quad \boldsymbol{x}_i = (x_{i1}, \cdots, x_{ip}), \ \ i = 1, \cdots, n$$

is a $n \times p$ data matrix with respect to p independent variables and

$$\boldsymbol{y} = (y_1, \cdots, y_n)^t$$

shows values of a dependent variable. If an observed data has K distinct blocks, a blocking regression model [3] is defined as follows:

$$\boldsymbol{y} = \hat{X}\hat{\boldsymbol{\beta}} + \hat{\boldsymbol{e}}, \tag{9}$$

where the $n \times (p + K + 1)$ matrix $\hat{X}$ is defined as:

$$\hat{X} = (\mathbf{1} \mid \bar{X} \mid X).$$

$\mathbf{1}$ is a $n \times 1$ vector whose elements are all of 1. The $n \times K$ matrix $\bar{X}$ shows K dummy variables which represent K blocks of data. That is,

$$\bar{X} = (\bar{\boldsymbol{x}}_1, \cdots, \bar{\boldsymbol{x}}_K), \quad \bar{\boldsymbol{x}}_k = (\bar{x}_{1k}, \cdots, \bar{x}_{nk})^t,$$

$$\bar{x}_{ik} \in \{0, 1\}, \quad \sum_{k=1}^{K} \bar{x}_{ik} = 1, \quad \forall i, k. \tag{10}$$

In equation (10), for example, if there are three distinct blocks in the data, each row of $(\bar{\boldsymbol{x}}_1, \bar{\boldsymbol{x}}_2, \bar{\boldsymbol{x}}_3)$ has values of the following three types: $(1, 0, 0)$ indicates data which belongs to the first block, $(0, 1, 0)$ is the data for the second block, and $(0, 0, 1)$ shows the data of the third block. The blocks of data are determined by the external information of the data.

$$\hat{\boldsymbol{\beta}} = (\hat{\beta}_0, \hat{\beta}_1, \cdots, \hat{\beta}_K, \hat{\beta}_{K+1}, \cdots, \hat{\beta}_{p+K})^t, \tag{11}$$

shows the regression coefficients. In equation (11), $\hat{\beta}_1, \cdots, \hat{\beta}_K$ are regression coefficients for K blocks and $\hat{\beta}_{K+1}, \cdots, \hat{\beta}_{p+K}$ are regression coefficients for p independent variables. $\hat{\boldsymbol{e}} = (\hat{e}_1, \cdots, \hat{e}_n)^t$ shows residuals. The estimate of the regression coefficient shown in equation (11) can be obtained as follows:

$$\hat{\boldsymbol{\beta}} = (\hat{X}^t\hat{X})^{-1}\hat{X}^t\boldsymbol{y}.$$

For each object i, a blocking regression model shown in equation (9) can be rewritten as follows under the conditions shown in equation (10):

$$y_i = \hat{\beta}_0 + \hat{\beta}_l + \hat{\beta}_{K+1}x_{i1} + \cdots + \hat{\beta}_{p+K}x_{ip} + \hat{e}_i, \tag{12}$$

$$\text{if } \boldsymbol{x}_i \in C_l = \{\boldsymbol{x}_i \mid \bar{x}_{il} = 1\}, \quad l = 1, \cdots, K, \quad i = 1, \cdots, n.$$

So, in this model, the difference of blocks is represented by the difference of intercepts $\hat{\beta}_l$ in equation (12).

5 Fuzzy Blocking Regression Model

If we assume independency for each cluster (or block), we can redefine equation (12) as follows:

$$\boldsymbol{y} = \tilde{\hat{X}}^{(l)}\hat{\boldsymbol{\beta}}^{(l)} + \hat{\boldsymbol{e}}^{(l)}, \quad l = 1, \cdots, K, \tag{13}$$

where the $n \times (p+2)$ matrix $\tilde{\hat{X}}^{(l)}$ is defined as:

$$\tilde{\hat{X}}^{(l)} = (\mathbf{1} \mid \bar{\boldsymbol{x}}_l \mid X), \tag{14}$$

$\hat{\boldsymbol{\beta}}^{(l)}$ and $\hat{\boldsymbol{e}}^{(l)}$ are as follows:

$$\hat{\boldsymbol{\beta}}^{(l)} = (\hat{\beta}_0^{(l)}, \hat{\beta}_l, \hat{\beta}_{K+1}^{(l)}, \cdots, \hat{\beta}_{p+K}^{(l)})^t, \tag{15}$$

$$\hat{\boldsymbol{e}}^{(l)} = (\hat{e}_1^{(l)}, \cdots, \hat{e}_n^{(l)})^t, \ \ l = 1, \cdots, K. \tag{16}$$

In equation (13), the estimate of the regression coefficient shown in equation (15) can be obtained as follows:

$$\hat{\boldsymbol{\beta}}^{(l)} = ({\tilde{\hat{X}}^{(l)}}^t \tilde{\hat{X}}^{(l)})^{-1} {\tilde{\hat{X}}^{(l)}}^t \boldsymbol{y}, \ \ l = 1, \cdots, K. \tag{17}$$

As a natural extension of the model (12), we propose the following model using fuzzy block increments:

$$y_i = \beta_0 + \beta_l u_{il} + \beta_{K+1} x_{i1} + \cdots + \beta_{p+K} x_{ip} + e_i, \ \ l = 1, \cdots, K, \tag{18}$$

where

$$u_{il} \equiv \max_{1 \le k \le K} u_{ik}, \ \ i = 1, \cdots, n, \ \ \exists l. \tag{19}$$

Then if we assume independency for each cluster (or block), we can redefine as follows:

$$\boldsymbol{y} = \tilde{\hat{X}}^{(l)} \boldsymbol{\beta}^{(l)} + \boldsymbol{e}^{(l)}, \ \ l = 1, \cdots, K, \tag{20}$$

where the $n \times (p+2)$ matrix $\tilde{\hat{X}}^{(l)}$ is defined as:

$$\tilde{\hat{X}}^{(l)} = (\mathbf{1} \mid \tilde{\boldsymbol{u}}_l \mid X), \ \ \tilde{\boldsymbol{u}}_l = (\tilde{u}_{1l}, \cdots, \tilde{u}_{nl})^t. \tag{21}$$

$\tilde{\boldsymbol{u}}_l$ shows l-th column vector of a $n \times K$ matrix $\tilde{U}$ which is defined as

$$\tilde{U} = (\tilde{u}_{ik}), \ \ i = 1, \cdots n, \ \ k = 1, \cdots, K. \tag{22}$$

In equation (22), $\tilde{u}_{ik}$ is assumed to be satisfied as:

$$\begin{cases} \tilde{u}_{il} = u_{il} \text{ if } u_{il} = \max\limits_{1 \le k \le K} u_{ik} \\ \tilde{u}_{ik} = 0 \quad \text{otherwise} \end{cases}. \tag{23}$$

$\boldsymbol{\beta}^{(l)}$ and $\boldsymbol{e}^{(l)}$ are as follows:

$$\boldsymbol{\beta}^{(l)} = (\beta_0^{(l)}, \beta_l, \beta_{K+1}^{(l)}, \cdots, \beta_{p+K}^{(l)})^t, \tag{24}$$

$$\boldsymbol{e}^{(l)} = (e_1^{(l)}, \cdots, e_n^{(l)})^t, \ \ l = 1, \cdots, K. \tag{25}$$

In equation (20), the estimate of the regression coefficient shown in equation (24) can be obtained as follows:

$$\boldsymbol{\beta}^{(l)} = (\tilde{\tilde{X}}^{(l)^t}\tilde{\tilde{X}}^{(l)})^{-1}\tilde{\tilde{X}}^{(l)^t}\boldsymbol{y}, \;\; l = 1, \cdots, K. \tag{26}$$

The difference between equations (12) and (18) are intercepts, $\hat{\beta}_l$ or $\beta_l u_{il}$. In equation (18), we use the intercepts multiplied by the fuzzy grade which are obtained as a result of fuzzy clustering. Moreover, in equation (18), β_l shows the regression coefficient between the obtained fuzzy cluster and dependent variable. So, in this model, we can capture the relationship between the fuzzy clusters and a dependent variable and the relationship between independent variables and a dependent variable simultaneously. Also, the clusters (or blocks) in equation (18) are obtained objectively using a numerical fuzzy clustering. However, in equation (12), we need external information in order to identify the clusters (or blocks). Moreover, we apply a regression model to the mixed data $\tilde{\tilde{X}}^{(l)}$ shown in equation (14) and this causes a decrease of the fitness. However, in equation (21), the fuzzy clustering result, $\tilde{u}_{il}$, can be obtained as continuous values so we can obtain a better performance of fitness for the proposed model.

For the identification of solutions shown in equation (26), the following are assumed to be satisfied:

$$p \leq n - 2, \;\; n \geq 3.$$

6 Variable Based Weighted Fuzzy Blocking Regression Model

An extended model of the fuzzy blocking regression model shown in equation (18) is defined as follows:

$$y_i = \beta_0 + \beta_l u_{il} + \beta_{K+1} w_1 x_{i1} + \cdots + \beta_{p+K} w_p x_{ip} + e_i, \;\; l = 1, \cdots, K, \tag{27}$$

where

$$u_{il} \equiv \max_{1 \leq k \leq K} u_{ik}, \;\; i = 1, \cdots, n, \;\; \exists l.$$

In equation (27), $w_a, \; a = 1, \cdots, p$ are the weights shown in equation (8). Then if we assume independence for each cluster (or block), we can redefine as follows:

$$\boldsymbol{y} = \tilde{X}^{(l)}\tilde{\boldsymbol{\beta}}^{(l)} + \tilde{\boldsymbol{e}}^{(l)}, \;\; l = 1, \cdots, K, \tag{28}$$

where the $n \times (p+2)$ matrix $\tilde{X}^{(l)}$ is defined as:

$$\tilde{X}^{(l)} = (\mathbf{1} \mid \tilde{\boldsymbol{u}}_l \mid XW), \;\; \tilde{\boldsymbol{u}}_l = (\tilde{u}_{1l}, \cdots, \tilde{u}_{nl})^t, \;\; W = \mathrm{diag}(w_1, \cdots, w_p), \tag{29}$$

where $\tilde{u}_{ik}$ satisfy the condition shown in equation (23). $\tilde{\boldsymbol{\beta}}^{(l)}$ and $\tilde{\boldsymbol{e}}^{(l)}$ are as follows:

$$\tilde{\boldsymbol{\beta}}^{(l)} = (\tilde{\beta}_0^{(l)}, \tilde{\beta}_l, \tilde{\beta}_{K+1}^{(l)}, \cdots, \tilde{\beta}_{p+K}^{(l)})^t, \tag{30}$$

$$\tilde{\boldsymbol{e}}^{(l)} = (\tilde{e}_1^{(l)}, \cdots, \tilde{e}_n^{(l)})^t, \;\; l = 1, \cdots, K. \tag{31}$$

In equation (28), the estimate of the regression coefficient shown in equation (30) can be obtained as follows:

$$\tilde{\boldsymbol{\beta}}^{(l)} = (\tilde{X}^{(l)^t}\tilde{X}^{(l)})^{-1}\tilde{X}^{(l)^t}\boldsymbol{y}, \quad l = 1, \cdots, K. \tag{32}$$

7 Numerical Examples

The data is the Fisher iris data [4] which consists of 150 samples of iris flowers with respect to four variables, sepal length, sepal width, petal length, and petal width. The samples are observed from three kinds of iris flowers, iris setosa, iris versicolor, and iris virginica. We use the data with respect to three variables, sepal width, petal length, and petal width for the data of independent variables and the data of sepal length is used for a dependent variable. The number of clusters is assumed to be 3.

Figure 1 shows the result of the fuzzy c-means method when $m = 2$ in equation (2) for the data of independent variables. In this figure, three axes show the three variables which are sepal width, petal length, and petal width. The locations of symbols show the observational values of each iris flower and the cubes show the observation space. The gray tone shows the degree of belongingness for the clusters and the darker color means larger degree of belongingness for the cluster. Each symbol shows each cluster. Black dots show the result for cluster 1, triangles show the result for cluster 2, and squares are the result for cluster 3. We only show the maximum values of degree of belongingness for a fixed object over the three fuzzy clusters. That is, we show the degree of belongingness of $\max_{1 \le k \le 3} u_{ik}, \quad i = 1, \cdots, 150$.

From this figure, it can be seen that there are three clusters which can show that the three kinds of iris flowers and the three clusters are vaguely situated in relation to each other. Using a result of variable based fuzzy c-means shown in equation (6) when $m = 2$, we obtain the weights shown in equation (8) as

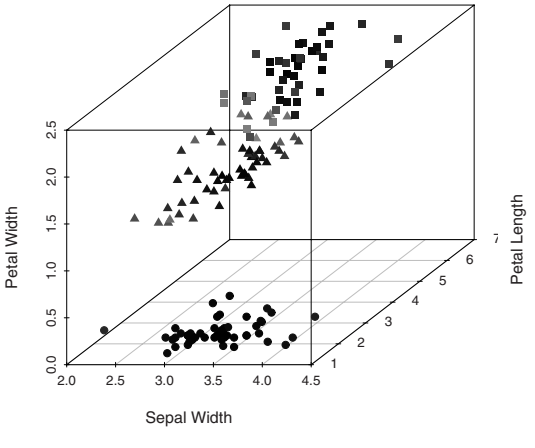


Fig. 1. Result of Fuzzy c-Means Method

follows: $w_1 = 0.85$, $w_2 = 0.90$, $w_3 = 0.95$. Using these weights and results shown in figure 1, we apply this data to the proposed model shown in equation (28).

Table 1 shows the comparison of the residuals of the blocking regression model and weighted fuzzy blocking regression model shown in equations (16) and (31). That is we compare the values of $s = \hat{s}^{(1)} + \hat{s}^{(2)} + \hat{s}^{(3)}$, $\hat{s}^{(l)} \equiv \hat{\boldsymbol{e}}^{(l)^t}\hat{\boldsymbol{e}}^{(l)} = \hat{e}_1^{(l)^2} + \cdots + \hat{e}_n^{(l)^2}$ and $\tilde{s} = \tilde{s}^{(1)} + \tilde{s}^{(2)} + \tilde{s}^{(3)}$, $\tilde{s}^{(l)} \equiv \tilde{\boldsymbol{e}}^{(l)^t}\tilde{\boldsymbol{e}}^{(l)} = \tilde{e}_1^{(l)^2} + \cdots + \tilde{e}_n^{(l)^2}$, $l = 1, 2, 3$. From this table, it can be seen that our proposed weighted fuzzy blocking regression model can obtain a better result.

Table 1. Comparison of Residuals

Weighted Fuzzy Model	Hard Model
42	43

8 Conclusion

We present a weighted fuzzy blocking regression model which is an extension of a fuzzy blocking regression model. When compared to an ordinal blocking regression model, these fuzzy blocking models have several merits. Regression coefficients for both the independent variables and all fuzzy clusters can be obtained simultaneously. Any external information is not needed to identify the clusters (or blocks). We can obtain a better performance of fitness for the fuzzy blocking regression models. Moreover, when compared to an ordinal fuzzy blocking regression model, the weighted fuzzy blocking regression model can consider the difference of classification structure for each variable.

References

1. Bezdek, J.C., Keller, J., Krisnapuram, R., Pal, N.R.: Fuzzy Models and Algorithms for Pattern Recognition and Image Processing. Kluwer Academic Publishers, Dordrecht (1999)
2. Brunsdon, C., Fotheringham, S., Charlton, M.: Geographically Weighted Regression Modeling Spatial Non-Stationarity. Journal of the Royal Statistical Society 47, 431–443 (1998)
3. Draper, N.R., Smith, H.: Applied Regression Analysis. John Wiley & Sons, Chichester (1966)
4. Fisher, R.A.: The Use of Multiple Measurements in Taxonomic Problems. Annals of Eugenics 7, 179–188 (1936)
5. Hathaway, R.J., Bezdek, J.C.: Switching Regression Models and Fuzzy Clustering. IEEE Trans. Fuzzy Syst. 1(3), 195–204 (1993)
6. Sato-Ilic, M., Jain, L.C.: Innovations in Fuzzy Clustering. Springer, Heidelberg (2006)
7. Sato-Ilic, M.: Weighted Fuzzy Clustering on Subsets of Variables. In: International Symposium on Signal Processing and its Applications with the International Conference on Information Sciences, Signal Processing and its Applications (2007)
8. Sato-Ilic, M.: Regression Analysis Considering Fuzzy Block Intercepts. In: International Symposium Management Engineering, T03 (2007)

A Design and Testing Technology for Ubiquitous Robot Companion Software Component

Sun-Myung Hwang[1], Yun-Koo Chung[2], and Hang-Kon Kim[3]

[1] Department of Computer Engineering Daejeon University, Daejeon, Korea
sunhwang@dju.ac.kr
[2] Electronics and Telecommunications Research Institute (ETRI), Daejeon, Korea
ykchung@etri.re.kr
[3] Department of Computer Engineering Catholic University, Daegu, Korea
hangkon@cu.ac.kr

Abstract. This paper focuses on the design and testing technology for URC component quality. The test case design using scenario is an effective methodology to cover all user requirements. And we are continuously studying about functional design standard for URC component Software by design specification.

1 Introduction

Since 1980s, robot market for industry has developed as an intelligent robot, a service robot, which is closely related to the human, searching for the new industrial formation and technical development in the 1990s. In this background, KOREA MIC suggests the least price of robot, and the robot based on the new ground, URC (Ubiquitous Robot Companion) that diversifies the function and service is open to the market.

URC(Ubiquitous Robot Companion) is defined as "a robot that offers necessary service all the time, anywhere at all," and by using the network to the robot, expanding the applicable service, keeping all the function in itself, it undertakes functions for the causing problems through the network.

This paper focuses on the testing technology for URC software component quality between URC infra-system technology, the core technology of URC and URC software component technology.

2 URC Robot Software Components

The components consisting of URC robot software is composed of HRI (Human-Robot Interaction) and robot smart-action technology.

2.1 HRI Technology

URC robot values the interaction of skillful users different from other information machinery. Robot technology requires the cognition and response of user's voice, the verbal transmission through voice composition on the information, the cognition to user's

B. Apolloni et al. (Eds.): KES 2007/ WIRN 2007, Part II, LNAI 4693, pp. 533–540, 2007.
© Springer-Verlag Berlin Heidelberg 2007

facial expression and gesture from the image signal through cameras. Chart 6 contains the contents of the human-robot interaction technology which is under development. This software module is offered by URC service, and it is developed by the aim focused on the reaction to 100 clients within one second with simultaneous contact.

Table 1. HRI Technology

HRI component technology	Infra-Image HRI component technology	User Recognition	User's cognition module by an quasi-organism information, facial expression at a long distance in the situation of robot's upward sight
		Gesture Recognition	User's pursuit and gesture cognition module development Caller Identification module
		Expression Recognition	Facial component detection and special feature extraction module
	Infra-Voice HRI component technology	Voice Recognition	Cognitive module of user's voice information transmitted from robot
		Voice Decomposition	Conversion module changing arbitrary sentence with composition voice
		Voice Pursuit	Exclusion module removing noises in the robot and outside noises for tracking the original sound
		Speaker Recognition	Speaker cognition module by the context speaker independent method

2.2 URC Robot's Smart Action Technology

Smart Action Technology can be divided into the intellectual covering technology and the intellectual manipulating technology for realizing the physical action of URC (Table 2).

Table 2. Smart Action Technology

Smart Action Technology	The Intellectual Covering Technology	Position-Deciding Technology	The correctly searching technology of robot's position
		Map-Forming Technology	Map-forming technology presented by digital data that grasps the robot's covering environment
		Route-Framing Technology	Route-controlling technology that make the robot plan its course, make it move as the original course
	The Intellectual Manipulating Technology	Environment-Object Cognitive Technology	The cognitive technology that makes the robot recognize the manipulating objects and circumference-environment with using camera, distance sensor
		Manipulation-Control Technology	The technology that manipulates the special object with controlling robot's arms articulation

3 Test Case Design Base on Scenario

The Key issue of testing becomes what subset of all possible test cases has the highest probability of detecting the most errors?

As one with answers to tin question, we propose a test case design methodology base on base scenario which is made up user requirements.

A well selected test case should also have two other properties:

1) It reduces, by more than a count of one, the number of other test cases that must be developed to achieve some predefined goal of reasonable testing.
2) It covers a large set of other possible requirements. That is, t tell us something abut the presence or absence of errors over and above this specific scenario.

We define the 5 steps that test-case design by scenario coverage as follows:

1) Identify the workable user requirements in the specification.
2) Assign a unique number to each user requirements.
3) Design Test scenario by combining requirement each requirement should be included by least one scenario.
4) Until all requirements have been covered by scenario, make new scenario.
5) Design Test cases which can satisfy each scenario.

Table 3. User Requirement of HRI Component

component	Requirement id.	Requirement description
Image recognition	SFR.HI.01	Provide a learning tool using “Template Update Algorithm”.
	SFR.HI.02	Provide a “back lighted detect function, size of face, as well as GPS verification”.
	SFR.HI.03	Provide a “street measuring information” using supersonic sensor.
	SFR.HI.04	Provide a “register and non-register distinction function” through boundary value.
	SFR.HI.05	Provide a “detect function” in 15 degree (up & down, left & right, front & back).
	SNR.HI.PE.01	The target users’ recognition rate must reach more than 95%, in three meters of light
	SNR.HI.PE.02	Processing speed should be reached P4, 3G CPU, 1G RAM at 5 frm/sec
Voice Recognition	SFR.HS.01	Provide an easy way on-line registration using GUI (speaker registration)
	SFR.HS.02	Provide “independent sentence speaker recognition”
	SFR.HS.03	Provide a “detect function” and “speaker recognition” between family members.
	SFR.HS.04	Provide “strength of signal regularity function” and “choice of microphone” for the long distance transaction purpose.
	SFR.HS.05	Provide “speaker recognition functions” using three or more microphone.
	SFR.HS.06	Provide a “noise filter” and “noise detect function”.
	SFR.HS.07	Provide a guidance system after detecting noise.
	SFR.HS.08	Provide a sex and age identification system.
	SFR.HS.09	Provide a adoptable speaker recognition
	SNR.HS.PE.01	“Speaker recognition” and “verification rate” must reach more than 95% when noise level reaches 5~15dB

Table 3. (*countinued*)

	SNR.HS.PE.02	While someone using on-line registration, this system must be able to register using two or less sentences.
	SNR.HS.PE03	This system must be able to recognize categories in all four directions within three meters.
Image interaction	SFR.HR.01	Provide a "face detect function" (with in 3m).
	SFR.HR.02	Provide a (head detect function), based on Omega.
	SFR.HR.03	Provide a "follower identification function" using colored clothes.
	SFR.HR.04	Provide "follower identification function" using GPS location system between previous screen and current screen.
	SFR.HR.05	If the Robert looses its follower, then provide a voice message and alarm using TTS.
	SFR.HR.06	If the Robert looses its follower, this system will save pursuit info automatically. After a while, if this system find exact same follower, then he/she will follow the Robert continuously.
	SNR.HR.PE.01	Frontal verification rate must reach more than 95% in one to three meters.
	SNR.HR.PE.02	Personal identification efficiency must reach more than 95% when it does not have occlusion.
	SNR.HR.PE.03	Flank and rear verification rate must reach more than 90%.
	SNR.HR.PE.04	
	SNR.HR.PE.05	

4 Approach Method for Component Test Evaluation Standard Specification

4.1 Component Scope

This paper contains the following components.

1) Navi-Guide component
2) Map-Design component
3) Automatic Running component
4) Image cognition component
5) Phonetics cognition component
6) Image base interaction
7) SA-reference component
8) Common component
9) The rest

4.2 Framing Method to the Component Test Evaluation Standard Specification

The specification is composed of 4 items as follows:

1) GR.CN.nn: This item is partly defined as a general content to the Component Test Capacity Appraisal, which means GR (Generic Rule). CN consists of two characters as an abbreviated form, which attaches serial numbers of two figures.
2) Basis Format: Appraisal item to the Component Capacity within tasks performing by this reach group. As Generic Rule consists of embodiment contents

within present task performance system, Appraisal Standard and test method are determined by these standards.

3) Appraisal Standard: This is described as quantitative analysis to capacity appraisal item by applicable appraisal standard as much as possible and contains function and capacity of each component and should define more than one.
4) Test Method: This is defined as concretely applicable test methods to the above appraisal standard and can make out separate test case.

4.3 Case Study for Test Evaluation Standard Specification

As I said, there are some of components for URC, in this paper; we describe the specification for Nevi-Guide Component and Automatic Running Component.

4.3.1 Navi-Guide Component

1) GR.NG.01 is Navi-Guide should measure the length of line and course data by real time.
   Basis Format Navi-Guide should measure the length of pulled line and course data by real time. Measure speed should regulate by programs.
   Appraisal Standard is Measure speed <= 5 Hz (5 times per second) (Efficiency-Average response time)
   Test Method is the measure on the length of pulled line as following 50cm, 1M, 3M, maximum permissible distance (4 cases) and the course measure on the length of pulled line as following 0', ±45', ±90', maximum permissible angle(6 cases) and the measure more than five times by each set More than 95% by 5Hz (5 times per second) in speed is eligible.
2) GR.NG.02 is Navi-Guide is made as a module type and should be installed in the given robot platform.
   Basis Format Navi-Guide is made as a module type and should be installed in multi-robot platforms used in ETRI.
   Appraisal Standard is Installation Potentiality = 100% (Portability Installation rate)
   Test Method is Navi-Guide can be installed in the given robot platform. If possible, It is eligible.
3) GR.NG.03 is Navi-guider should be within permitted power.
   Basis Format is the utility voltage of Navi-guider should be operated be among 5~24V.
   Appraisal Standard is 5V <= permissible electric power <= 24V (Functionality-boundary value process rate)
   Test Method is checking whether or not Navi-Guide operates to the following electric power 5V, 10V, 15V, 20V, 24V and Repeating Measure more than 5 times to each electric power. If 100% operation is possible to permissible electric power, it is eligible.

4.3.2 Autonomic Running Component (AR)

1) GR.AR.01 Autonomic running component should provide the function that can use map data.

Basis Format Autonomic running component should use international standard interface with environment map. Measure speed should regulate by programs.

Appraisal Standard is Backup potentiality of standard map data = 100%

Test Method Identify the items of standard map data And Check the potentialities on each item of standard map data. If all items are available, it is eligible.

2) GR.AR.02 Autonomic running component should provide the function that can make the optimum plan of the transfer course of robot to the appointed destination.

Basis Format Autonomic running component should make the optimum plan of the transfer course of robot when the destination is arranged.

Appraisal Standard is the transfer course project potentiality =100% whether or not algorithms can be realized when searching for the optimum course.

Test Method is Select 5 the optional starting points and destinations each and identifies the setting-up of transfer course on starting point-destination set each. If the transfer courses are optimum to all sets, it is eligible.

3) GR.AR.03 Autonomic running component should provide the function that can stop against emergencies taking precautions against elucidated situation.

Basis Format is Autonomic running component should have the emergency stop function when communication jamming happens with clients.

Appraisal Standard emergency stop potentiality =100 %.

Test Method is the communication can be stopped more than 5 times to the different other clients. If the emergency stop is possible to all cases, it is eligible.

4) GR.AR.04 Autonomic running component should provide the function that can navigate to appointed destination, with the transfer robot shunning static and kinetic obstacles.

Basis Format is the transfer robot can navigate with its shunning obstacles when the destination is appointed.

Appraisal Standard obstacle shunning potentiality = 100% (faculty-function realization completeness)

Test Method is the different other destinations can be established more than 5 times in the different other maps more than 5 maps. The establishment of different kinetic obstacles to appointed destination more than 5 times. The establishment of different static obstacles to appointed destination more 5 times. If the transfer robot can navigate with its shunning obstacles to each situation, it is eligible.

5) GR.AR.05 Autonomic running component should provide the function that the supersonic wave sensor of robot can detect obstacles in proceeding courses in real time.

Basis Format is the supersonic wave sensor of robot should detect, stop, and shun obstacles in proceeding courses by real time.

Appraisal Standard obstacle-detecting time <= 5Hz (5 times per second) (faculty-function realization accuracy).

Test Method is Measure detecting time to different obstacles more than 5 times. If more than 5Hz (5 times per second) can be detected, it is eligible.

6) GR.AR.06 Autonomic running component should provide the function that prepare against the collision by a sudden appointed outbreak.
   Basis Format is Autonomic running component should provide the function that prepare against the collision by sudden obstacles. Autonomic running component should prepare against the collision by unreliable information of supersonic waves.
   Appraisal Standard outbreak situation preparatory rate=100% (faculty-function realization completeness)
   Test Method is Identify collision preparation, with sudden obstacles happening more than 5 times. Identify collision preparation, with unreliable information happening more than 5 times. If collision does not happen to all cases, it is eligible.
7) GR.AR.07 Autonomic running component should provide the function that informs server of impracticable transfer.
   Basis Format is Autonomic running component should informs server of impracticable transfer when it can't shun obstacles and it should be stopped.
   Appraisal Standard potentiality on informing server of impracticable transfer = 100% (faculty-function realization accuracy)
   Test Method Robot generates impracticable transfer more than 5 times. If autonomic running component informs server of impracticable transfer, it is eligible.
8) GR.AR.08 Autonomic running component should provide the function that can shun obstacles when it recognizes them in the place where has permissible breadth.
   Basis Format is Autonomic running component should provide the function that can shun obstacles when it recognizes them in the place where robot has permissible breadth more than 1 meter.
   Appraisal Standard Sunning potentiality = 100 % (faculty-function realization accuracy)
   Test Method Establish the place of the breadth with 0.9m, 1m, 1,1m, 1.5m, 2m and Identify shunning when recognizes obstacles. If autonomic running component shuns obstacles more than 5 times according to each place, it is eligible
9) GR.AR.09 Autonomic running component should provide the power that braking distance can be within permissible scope when transferring robot with permissible speed recognizes obstacles.
   Basis Format is Autonomic running component should provide the power that braking distance can be within 0.8m when transferring robot with 0.3m/s speed recognizes obstacles.
   Appraisal Standard Braking distance scope < = 0.8m (faculty-function realization accuracy)
   Test Method Robot should move 0.3m/s, and establish different obstacles more than 5 times. Measure the braking distance in each situation. If the breaking distance can be within 0.8m to all cases, it is eligible.

Like these, URC Component can add clean service, home monitoring service, taking picture service, video message service, news services, weather service, schedule service, learning service to its standard function, and it can provide clients with the useful service of ubiquitous robot in practical life.

5 Conclusion

URC robot offers necessary service all the time, anywhere at all, and by using the network to the robot, expanding the applicable service, keeping al the function in itself, it undertakes functions for the causing problems through thee network.
This paper focuses on the testing technology for URC component quality. The test case design using scenario is an effective methodology to cover all user requirements. And we are continuously studying about standards for URC component Software testing based an ISO/IEC 9126.

References

[1] Kim, H.: Network based intelligent service robot – URC, ETRI CEO information No.15 (2004)
[2] Kim, H., Cho, Y.J., Oh, S.R.: CAMUS: A Middleware Supporting Context-aware Services for Network-based Robots. In: IEEE Workshop on Advanced Robotic and its Social Impacts (2005)
[3] ISO/IEC TR 9126, Software engineering – Product quality
-Part1: Quality model (2003)
-Part2: External Metrics (2003)
-Part3: Internal Metrics (2003)
-Part4: Quality in Use Metrics (2003)

A Multiagent Based Vehicle Engine Fault Diagnosis

Xiaobing Wu[1], Xueshan Gao[1], and Dharmendra Sharma[2]

[1] Beijing Institute of Technology, 100081, Beijing, China
Xiaobing.Wu, Xueshan.Gao}@bit.edu.cn
[2] University of Canberra, 2601, Canberra, Australia
Dharmendra.Sharma@canberra.edu.au

Abstract. A special two-level neural network Multiagent system is used in vehicle engine fault diagnosis. Each agent contains a neural network (NN) for its intelligent. The first level is used to acquire the reasons of the fault and the second level is used to classify the fault. There are a few advantages for using two-level NN multiagent system. When new knowledge of faults is acquired, only the first level needs to be retrained. An agent system is added into the second level of the whole system, and the main structure of the network need not be changed. Under the cooperation of the relative agent system, we can diagnose out complex faults and synthetic faults with less resource and better performance.

Keywords: Fault diagnosis, multiagent, neural network, vehicle engine.

1 Introduction

Intelligent fault diagnosis has been an important field since the 1990s [1]. Multiagent system is an efficient tool for fault diagnosis as it has many advantages such as its distributed processing mode on the large scale, association memory, self-organized learning, robustness and so on. It can also process uncertain knowledge or imprecise data and classify complex modes. The faults can be mapped to the symptoms where there exist non-linear relations between them based on the trained ANN.

The multiagent system has the local knowledge. For example, some agents have the knowledge about the parts of the vehicle engine. An NN agent keeps, collects, and uses this information as local knowledge and provides a decision when needed [2].

An agent can specialize in one specific area. Some information is only useful to some faults. For example, only parts of the feature eigenvectors are useful for someone fault. The agent using these eigenvectors is important to this special fault while is not important to other faults.

A multiagent system is also survivable. The agents are distributed and autonomous. So the multiagent system can survive the loss of some agents. In a large system consisting of many components, the components have some degree of reliability.

A multiagent system is scalable too. In a multiagent system, an agent contacts only the other relative agents when necessary. An agent is not aware of how other agent works. Adding one more agent into the system needs not notify every single agent in the system. Only the involved agents are aware of the dropping in of the new agent.

B. Apolloni et al. (Eds.): KES 2007/ WIRN 2007, Part II, LNAI 4693, pp. 541–546, 2007.
© Springer-Verlag Berlin Heidelberg 2007

A multiagent system has good communication ability. In the system, the agents share the knowledge together to keep the whole system work synchronic. It should be pay attention that "Expressions in given agent's native language should be understood by some other agent that uses different implementation and domain assumptions" [3].

2 Two-Level NN for Agent Intelligence

When we use an usual one-level (one hidden layer) NN, we can detect only the fault type without the degree of fault. When both the fault type and the degree are need, the traditional one-level NN is time-costly due to large amount of computation. For example, if there are four types of faults and each type of fault has four degrees, the NN will have three input data with 16 outputs. It involves a lot of computation and training. Besides, when there is a new type of fault, the structure of the NN will possibly change and the NN will have to be re-trained because the old NN can not recognize the new fault with its knowledge. We propose a two-level NN multiagent system to solve this problem, in where each agent is a NN. The first level of the two-level NN is used for recognizing the types of the faults and the other level is for the degree of the faults. Figure 1 is an example of our two levels NN multiagent system. The first level has three input data and four output data. The second level has four sub-NNs which have four input data and four output data individually. It is shown in Figure 1, with only the sub-NN agent of fault 1.

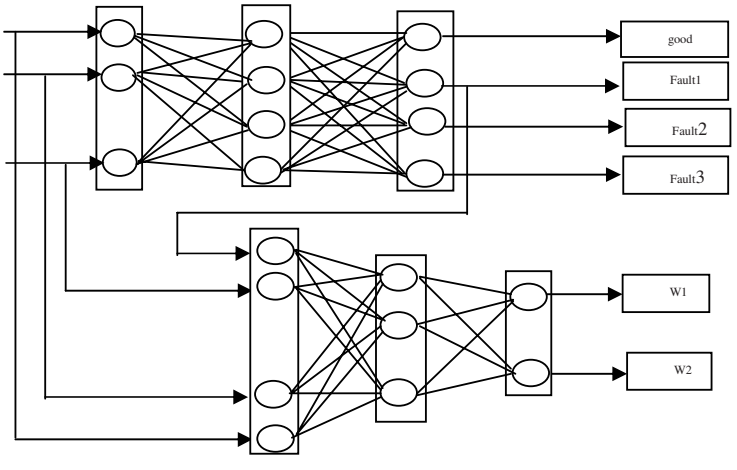


Fig. 1. Structure of two-level NN multiagent system

3 A Two-Level NN Multiagent Based Vehicle Engine Fault Diagnosis

For experiment purpose, we apply this two-level NN multiagent system in fault diagnosis of diesel engine S195. There are eight input data and three output faults.

Each type of fault has two degrees: degree I means worse and degree II means worst. Therefore the first level of the NN has eight input data and four output data (including one without fault). The eight input data are mean square values, wave indexes, impulse indexes, kurtosises, gradients, average square frequencies, spectral barycenters, and peak values. The second level has three sub-NN. Each agent has four or five input data which includes one corresponding output from the first level and the data from the original input that respondent to some one fault pattern. Table 1 shows the sensitivity of the data to the three fault patterns. After trained, it is applied to recognizing the faults of the S195 engine.

In the application of the vehicle engine fault diagnosis, the three types of faults are the wearingness of linker's bearing, wearingness of piston ring, and wearingness of main bearing.

Table 1. Sensitivity of fault to different feature data

False type / Eigenvector	Fault1	Fault2	Fault3
Mean square value (p3)	high	high	
Wave index (p4)	high		
Impulse index (p5)			high
Kurtosis (p6)	high		
Gradient (p7)	high		
Average square frequency		high	high
spectral barycenter (p10)		high	high
Peak value (p14)		high	

With the help of Matlab [4], we calculate the testing result as shown in Tables 2 and 3. Table 4 shows the usual one-level NN result. With the two-level NN multiagent system, we train the first level NN 684 times and the second one 212 times. The total is 1320 times. While for one-level BP NN, we have to train the NN 2173 times to get the same result. Not to mention that the two-level NN has a higher precision of diagnosis.

Table 2. Result of the first level NN agent

Engine	good	Fault 1	Fault 2	Fault 3
Good	0.999837	0.001996	0.016169	0.003228
Good	0.999633	0.002642	0.001102	0.006593
Good	0.998927	0.000034	0.000382	0.008650
Fault 1	0.001775	0.988289	0.009371	0.005227
Fault 1	0.000504	0.992281	0.001235	0.003145
Fault 1	0.000100	0.982054	0.002177	0.013234
Fault 2	0.002482	0.003666	0.977279	0.024275
Fault 2	0.002282	0.013136	0.976203	0.017955
Fault 2	0.002512	0.017359	0.980134	0.013091
Fault 3	0.004470	0.015895	0.018165	0.967384
Fault 3	0.001687	0.007290	0.005094	0.988020
Fault 3	0.002187	0.004301	0.002291	0.995403

Table 3. Result of the second level NN agent

	I (worse)	II (worst)
Fault 1	0.996737	0.004008
Fault 1	0.996747	0.003997
Fault 1	0.009333	0.992129
Fault 2	0.992410	0.009048
Fault 2	0.003150	0.997462
Fault 2	0.000593	0.998381
Fault 3	0.000597	0.998401
Fault 3	0.997142	0.001210
Fault 3	0.997793	0.000857

Table 4. The result of one-level NN system

	Good	Fault 1	Fault 2	Fault 3
Good	0.951329	0.008896	0.009281	0.004836
Good	0.954914	0.005330	0.008093	0.005779
Good	0.975340	0.002078	0.021755	0.030689
Fault 1	0.012732	0.963921*	0.040245	0.008392
Fault 1	0.038283	0.943021*	0.024863	0.016298
Fault 1	0.028263	0.953976**	0.047584	0.027521
Fault 2	0.038876	0.018273	0.969270*	0.027568
Fault 2	0.028947	0.037622	0.953862**	0.042725
Fault 2	0.014839	0.042836	0.953836**	0.026151
Fault 3	0.032725	0.026258	0.013635	0.963736**
Fault 3	0.028240	0.049723	0.033726	0.942726*
Fault 3	0.017269	0.038361	0.013872	0.973728*

Where * means worse fault, ** means worst fault

When we add one more type of fault into the NN, the one-level NN will be changed totally, but the two-level NN will only be changed partly by adding one input and one sub-NN agent. So, the two-level NN can be used in realtime fault diagnosis system, where we do not know in advancee how many types of the faults a vehicle enginee has. The following test of synthetic fault diagnosis shows that in the multiagent system, the first level NN of the system will be retrained, but the second level NN needs only to add one agent. The total retrained time is 754 while the 1-level NN need 2500 times to get the same result.

Besides, in the multiagent system, each agent is responsible for only one fault pattern, the time used in diagnosis is only about 79% of the one-level NN system. The results of diagnosis are also much better compared with the one-level NN system.

The agents can also cooperate together to diagnose a synthetic fault pattern. We have another test to demonstrate it. Table 5 and 6 shows the result of synthetic faults. In this test, we add one synthetic fault pattern which is the wearingness of the main bearing of vehicle engine and piston ring.

Table 5. Result of the first level NN agent

Engine	good	Fault 1	Fault 2	Fault 3
Good	0.999825	0.001767	0.019899	0.004228
Good	0.999763	0.002121	0.001990	0.009654
Good	0.998677	0.000030	0.001355	0.010901
Fault 1	0.001765	0.988200	0.018371	0.010001
Fault 1	0.000500	0.992787	0.003236	0.003213
Fault 1	0.000120	0.973229	0.004331	0.018734
Fault 2	0.002422	0.012145	0.978891	0.024270
Fault 2	0.002275	0.009897	0.975468	0.017951
Fault 2	0.002521	0.017359	0.980004	0.013111
Fault 3	0.002997	0.019980	0.016508	0.976784
Fault 3	0.001755	0.010989	0.100094	0.988221
Fault 3	0.002020	0.007677	0.012899	0.992780
Fault4	0.000436	0.032099	0.007790	0.000721
Fault4	0.000923	0.047385	0.076442	0.009002
Fault4	0.086116	0.000165	0.000637	0.000002

Table 6. Result of the second level NN agent

	I (worse)	II (worst)
Fault 1	0.996737	0.004008
Fault 1	0.996747	0.003997
Fault 1	0.009333	0.992129
Fault 2	0.992410	0.009048
Fault 2	0.003150	0.997462
Fault 2	0.000593	0.998381
Fault 3	0.000597	0.998401
Fault 3	0.997142	0.001210
Fault 3	0.997793	0.000857
Fault 4	0.989011	0.000973
Fault 4	0.999720	0.000955
Fault 4	0.000021	0.967454

4 Conclusion

The paper presents a multiagent absed system for vehicle engine fault diagnosis. Each agent is a NN. The multiagent system has a higher precision in fault diagnosis for the diesel engine S195. It has a higher efficiency as well. When adding a new fault, the multiagent system can learn it online.

References

1. Yongji, W., Jian, T.: Neural Network controlling. Mechanics Press, Beijing (1998)
2. Bradshaw, J.M.: An Introduction to Software Agents. In: Bradshaw, J.M. (ed.) Software Agents, AAAI Press/ The MIT Press (1997)

3. Finin, T., Labrou, Y., Mayfield, J.: KQML as an agent Communication Language. In: Bradshaw, J. (ed.) software agents, MIT Press, Cambridge (1995)
4. Shuang, C.: The Neural Network tool box and its application based on Matlab. China Science and Technology University Press, Hefei (1998)

A Multi-agent Security Framework for e-Health Services

Rossilawati Sulaiman, Dharmendra Sharma, Wanli Ma, and Dat Tran

School of Information Sciences and Engineering, University of Canberra, Australia
{Rossillawati.Sulaiman,Dharmendra.Sharma,Wanli.Ma,
Dat.Tran}@canberra.edu.au

Abstract. Nowadays, information and communication technologies are widely used in e-health to deliver services. The most popular communication choice is the Internet, as it offers cheap and worldwide coverage. However, as the technologies grow, the threats also grow in parallel. The threats put the patient's privacy at risk. Therefore, it is important to make sure that security technologies can cater for the privacy and security needs, whenever communications occur through the Internet. This paper presents a multi-agent based security framework for e-health services. The actors and the type of communications are described and a multi-level communications for the users is introduced. The security requirements and the multi-agent based security approach are modeled and presented in the framework.

Keywords: Multi-agent, security framework, e-health.

1 Introduction

Generally, the Internet is not a secure place to communicate and share sensitive information as many threats are coming from there. The organizations must make sure that their systems are secured, especially for systems that deal with highly sensitive information. Undisputedly, there have been numerous trends of threats reported regarding these systems. A survey conducted by Computer Security Institute involving 616 computer practitioners in U.S. shows that virus attacks, laptop/mobile theft, unauthorized use of computer systems from the inside of the organizations, unauthorized access to information, denial of services, system penetration and abuse of wireless network are among 13 top ranking trends of threats in the networks [1]. In spite of the threats reported, the Internet however, is still being used as a mean of communication in various fields such as in universities, military defense systems, business corporation, financial institutions, or medical institutions. Therefore, it is important to convince the users of the systems that their information is safe and their privacy is protected. In e-health, the information technology and communication is used to enhance healthcare services. E-health improves the relationship between patients and doctors and provides online education for both patient and physician through online resources, and sharing information especially for remote patients [2]. Moreover, e-health also improves efficiencies, reduces cost and improves the quality of health service delivery [3].

B. Apolloni et al. (Eds.): KES 2007/ WIRN 2007, Part II, LNAI 4693, pp. 547–554, 2007.
© Springer-Verlag Berlin Heidelberg 2007

Current practices in e-health encompass many aspects with information and telecommunication technology as the backbone. For example, elderly and disable people use assistive technology [4] to help them with their daily lives such as walking aid and electrical equipment remote control. There are also wearable devices worn by patients such as portable electrocardiogram (ECG), which is monitored remotely by GSM cellular network [5]. Apart from that, there are also e-health software and tools that are used to assist medical experts. For example, nurses in Royal District Nursing Service in Victoria, Australia are provided with Personal Digital Assistant (PDA) and smart phone to help the nurses to get information through email while moving from one house to another [6].

The online communication in e-health is currently being accepted and exercised to be part of the medical practices. Videoconferencing is taking the lead in some cases. For example, the Renal Unit Staffs at the Queen Elizabeth Hospital in Adelaide are having weekly meeting using video conferencing with other staffs from other hospitals [6] to discuss issues regarding renal transplantation using images captured from previous operation. Another option for communication in e-health is the use of email. Email is a convenience tool for communication between patient and physician [7]. In addition, email can be used for consultation, prescription renewal and to obtain test result [8]. There are also efforts given to developed web-based application for e-health. Web-based messaging is a better way for communication compared with standard email applications because the message can be automatically categorized and routed to the appropriate staffs [9]. In addition, with web-based application, functions like electronic delivery for laboratory results, appointments, online referral, prescription renewal, and billing [10][11] can be developed and used easily as long as the users are connected to the Internet.

The communications among users in e-health involve sensitive information either through wired or wireless media. However, although the use of information and communication technologies are undeniably helpful to enhance the delivery of health services, the threats to the technologies is unavoidable. Therefore, there is a need to secure the communication mediums between the medical experts and patients to protect the privacy of the information transmitted. In this paper, we discuss about online services specifically in e-health, with multi-agent based approach. The security requirements and the proposed security framework are presented, where we introduced multi-level communications and present the proposed security framework for secure communication environment. The rest of this paper describes the actors and the type of communications. Next, the multi-level communication approach for the users is explained and finally the security requirements and the multi-agent based security framework are presented.

2 E-Health Services

Generally, in e-health, the communications occur among actors over the network. The communications occur either from within the organization or from the organization to the outside network. For example, a doctor in a hospital communicates with another doctor from another hospital through the Internet. Another example is a paramedic at an accident spot communicates with a system coordinator at the hospital to update

current status of a patient using his personal data assistance (PDA). The roles or actors involved in the communications can be identified as Doctor, Patient, Nurse, Social Worker (SW), Paramedic, System Coordinator (SC), and System Administrator (SA) labeled as 1 to 7 as described in Fig. 1.

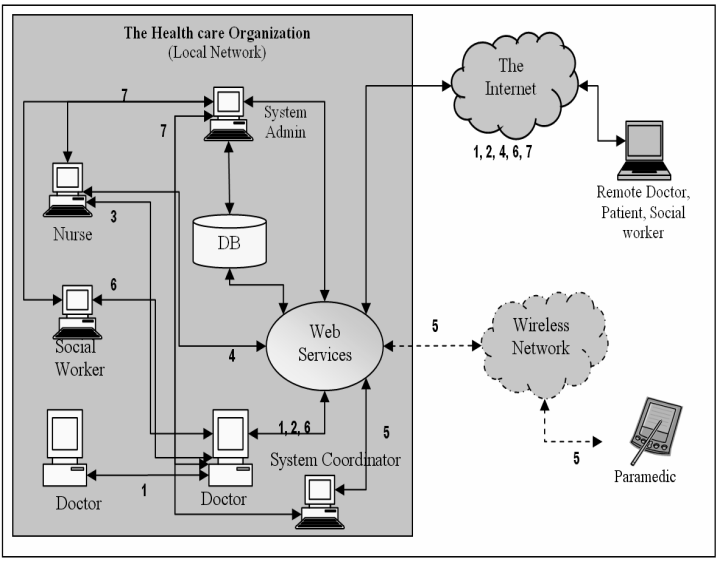


Fig. 1. The actors and the communications in e-health

The seven main types of communications are identified as (1) Doctor ⇔ Doctor, (2) Doctor ⇔ Patient, (3) Doctor ⇔ Nurse, (4) Nurse ⇔ Patient, (5) Paramedic ⇔ SC (6) Social workers ⇔ Doctor, and (7) SA ⇔ (Doctor, Nurse, Patient, SC, SW, and Paramedic). The shaded area shows communications that occur within the organization. The communication can also occur from within the organization to the outside network particularly for remote users such as patients and SW at home, doctors at another hospital, or paramedic at an accident spot. For the paramedic and SC, the information from the paramedic is sent through the wireless network using PDA and received by the SC in the hospital for further action. For example, the SC can prepare for a medical team while waiting for the patient to arrive at the hospital. There are also information for public such as information on diseases, health care annual reports, available services and other public announcement.

2.1 Classifying the Communication

There are types of communications, which involve highly sensitive information compared to the others. These types of communications must be secured with stronger security protection. For example, Doctor ⇔ Doctor and Doctor ⇔ Patient communications transmit more sensitive information compared to Doctor ⇔ SW communication. This is because a SW does not have to know every single detail of the patient illness. On the other hand, the doctor and patient communicate with each

other regarding the patient's diseases, current condition or medication history that must be treated as highly confidential. Another example is paramedic and SC communication. This type of communication is also considered as sensitive, since it involves patient's information such as blood and heart conditions, allergies information and any other injuries that should not be revealed to public. Moreover, the wireless technology is used in this communication, which is vulnerable and can be easily attacked. Therefore, higher security protection should be applied. The SA in contrast, communicates to the other users simply by telephone or the SA can communicate through secure email in an emergency case. Therefore, only the email communication must be made secure. For open channel communication, the public can communicate with the healthcare organization through the organization's website using user id and password to protect the user account. The user account is an alternative way for the public to get information that is not made public. For example, a researcher can ask for the right person to get medical data for his/her research.

It is now obvious that each type of communication requires different level of security protection, since the information has different levels of sensitivity. In this research, we classify the communication levels in order to provide different security technologies to secure different levels of communications. We also intend to demonstrate the capability of the system to flexibly cater for all kinds of users' needs. In addition, devices with low processing power (e.g. PDA and smart phone) can be facilitated with appropriate security protection. The classification can be labeled like the following:

1. *Level 1: Extremely Sensitive Communication.* This type of communication is extremely sensitive and highest protection must be provided to preserve privacy, confidentiality and integrity of the medical information transmitted such as personal details and health information. The communications include Doctor ⇔ Doctor, Doctor ⇔ Patient, Nurse ⇔ Doctor and Nurse ⇔ Patient.
2. *Level 2: Highly Sensitive Communication.* The SC ⇔ Paramedic communication is labeled as Level 2, as it involves wireless communication. GSM service is chosen to provide wider range of coverage. The information transferred is related to current patient's condition and considered as highly sensitive and therefore the privacy of the information must be protected.
3. *Level 3: Medium Sensitive Communication.* Level 3 type of communication includes Social Worker ⇔ Doctor. Only general information involved and requires medium security protection.
4. *Level 4: Low Sensitive Communication*. Level 4 type of communication involves the SA ⇔ all other users. These communications are considered as low sensitive and therefore low security protection is needed. The SA communicates with other users regarding user account and password, which are highly confidential. The SA can either call or email the users using unsecured email application to ask users to collect the Id and password at the office. Alternatively, the system administrator can send the information using encrypted email.
5. *Level 5: Public.* This type of communication is open for public. Usually, the organization uses this communication channel to broadcast relevant announcement as well as to educate the public. Similar concept can be found for instance at http://www.nhsdirect.nhs.uk/ and http://www.arc.org.uk/ for information provision.

The public channel also gives opportunity for the public to establish contact and obtain legal medical information according to the organization's policy.

This classification can be used to determine the level of security processes for each level. In addition, with this classification the organization can save CPU processing power and increase system performance, because only the most sensitive information requires the highest security processes.

2.2 Agent-Based Technologies

An agent is defined as a computer system that can receive input and output from the environment, autonomous and flexible [12]. There are many other advantages of the agents. For example, agents can be given knowledge to do specific task, collaborate with each other, and extendible which are very useful when developing secure systems.

An agent can represent a user and do tasks on behalf of the user. For instance, if a doctor wants to send a message to a nurse, the Doctor-agent knows how to handle the communication processes and initiate the communication with the Nurse-agent. Since the information is sensitive, the Doctor-agent can collaborate with the Encryption-agent to make sure that the message is encrypted and digitally signed before sending it to the Nurse-agent. If the message has been sent and if the Doctor-agent found out that the Nurse-agent is not available, the Doctor-agent will take necessary steps to keep the message and try to send it again later, or delete it within a certain period of time. This shows that the agent is autonomous and it does not need user's intervention and can do tasks on its own. In addition, agent systems are extendible. A new agent can be created instantly and added to the existing system to represent a new user such as a researcher, without reconfiguring the whole system.

2.3 Security Requirements

It is important to identify the security requirements to create a secure system environment. The requirements for a system used in a health care are authentication to enter the system, confidentiality to prevent disclosure of sensitive information, and anonymity [14]. These requirements are important in order to protect patient's privacy during and after the communication where the data must be kept safe. This research focuses on creating a secure communication environment and therefore, the requirements include (1) securing the communication processes between users, (2) analyzing and monitoring the network to protect against security breaches and (3) security policies that include access control and site policy, hardened host policy, incident and disaster response, and auditing policy [15] These policies varies from organization to organization and can be added or modified appropriately.

To secure the communication process, the client and the server side as well as the media of communications must be secured. The communication channel must be secured either through wired or wireless media. The wireless technologies are not robust and easily cracked by the attacker. Therefore, the messages must be encrypted or digitally signed before transmission. While it is important that the information is

safe in transit and is going to the right server, it is equally important that the information is not stolen when it reaches the server. Web servers are exposed to attack and it is possible for an attacker to perform malicious action on the system. At the server-side where the data source is located, the user is authenticated and all authorized users must follow the access control policy to access the data. Any unauthorized access must be logged for auditing purposes. At the client-side, two aspects are investigated, which are authentication to enter the system and communication manager. The communication manager handles two types of communications, which are communication with email application and communication over web-based technology, either using wired or wireless transmission. The communication manager also prepares and determines the security processes needed for a communication to take place. Before using the email software to transmit messages, digital certificate is installed, which is needed to sign and encrypt the message. Alternatively, if the user chooses to use web-based application, the messages must be encrypted and digitally signed.

The network monitoring and analysis are important to evaluate current status of the network. All activities from firewall, IDS, and software patches and updates must be logged in log files to be analyzed for any possible abnormality in the network environment. To make sure that the security mechanisms are functioning properly, the organization must enforce the security policy to get the best result which according to [15], "*In security, you get what you enforced*". Therefore, the organization must make sure that the security policy is enforced for satisfactory results.

3 The Multi-agent Based Security Framework

Fig. 2 presents the security framework for secure communication environment in e-health based on multi-agent approach. The framework is divided into three main parts that are the Organization Policies, Communication Manager Agent (CMA), and Network Analysis and Monitoring Agent (NAMA). As discuss previously, the organization policies must be enforced to get the most out of the security processes implemented. Therefore, all security processes must follow the policies. The id and password are used as authentication process by the Authentication Agent. The CMA manages other agents based on services needed. For example, Multi-level Communication Agent is used to identify the level of communication for a user, based on the user's role (Doctor, Nurse, Patient, etc). The CMA will make sure the type of security processes needed (e.g. encryption/digital signature) for a communication before transmitting the message. For example, the encryption algorithm given in Level 1 is stronger than Level 3. If a user chooses to use the existing email application like Microsoft Outlook, the Email Agent will make sure that there is a certificate installed at the user's machine to encrypt the message before sending the message out to the network. Multi-agent approach can be used to integrate other services to create a secure computing environment [13]. The NAMA integrates security services such as the intrusion detection system, firewall,

anti-virus, and patch server (as a proxy for software and security updates) to observe and analyze the network traffic for any malicious activity and thus protect the communication processes from security breach.

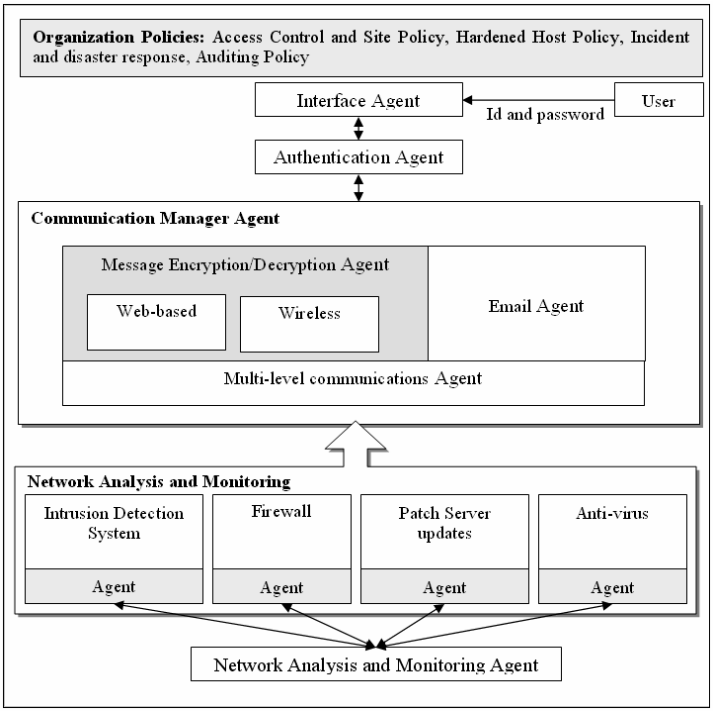


Fig. 2. The architecture of the e-health services security framework

4 Discussion and Future Work

A multi-agent security framework for communication in e-health is presented. The advantage of this framework is that it provides an alternative way to secure the communication between users in e-health by providing multi-level communications and secure environment for the communication to occur. The multi-level communications approach is introduced to cater for different levels of users as well as flexible enough to facilitate different kinds of user's needs in order to gain users' trust and confidence to communicate private data over the network. The framework will be implemented based on multi-agent approach for future works in order to study the capability of the agents to coordinate and cooperate together to improve the performance of the security processes.

Acknowledgments. Rossilawati Sulaiman is supported by the Public Service Department (PSD) of Malaysia.

References

1. Gordon, L.A., Loeb, M.P., Lucyshyn, W., Richardson, R.: Computer Crime and Security Survey, 2006 (retrieved on October 1st, 2006), Link www.gocsi.com
2. Eysenbach, G.: What is e-health? J. Med. Internet Res. 3(2), 20 (2001)
3. Joseph Wen, H., Tan, J.: The Evolving Face of TeleMedicine & E-Health: Opening Doors and Closing Gaps in E-Health Services Opportunities & Challenges. In: Proceedings of the 36th Annual Hawaii International Conference on System Sciences (HICSS'03) (2003)
4. Cowan, D., Turner-Smith, A.: The role of assistive technology in alternative models of care for older people. In: Royal Commission on Long Term Care. With respect to old age: research, London Stationery Office (Cm 4192-II/2), vol. 2 (1999)
5. Scanlon, W.G.: Using wireless technology to develop an effective personal telemedicine service. Centre for Communication Engineering, University of Ulster (2000), Link http://www.cyberfab.net/Documents/EMS/univulsterGSM.pdf
6. Mitchell, J.: From Telehealth to E-health: The Unstoppable Rise of E-health. John Mitchell & Associates for the Federal Australian Department of Communications, Information Technology and the Arts (DOCITA) (1999)
7. Shou, L.L., Gingrich, D., Lewis, P.R., Mauger, D.T., George, J.H.: Enhancing Doctor-Patient Communication Using Email: A Pilot Study. The Journal of the American Board of Family Practice 18, 180–188 (2005)
8. Couchman, G.R., Forjuoh, S.N., Rascoe, T.G.: E-mail Communications in Family Practice What Do Patients Expect? Journal of Famaliy Practice 50, 414–418 (2001)
9. Liederman, E.M., Zimmerman, E.M., Athanasoulis, M.A., Young, M.A.C.: Systemwide rollout of doctor-patient secure web messaging: The university of california, davis, virtual care experience. In: Whitten, P., Cook, D. (eds.) Understanding health communication technologies, 1st edn., pp. 244–250. A Wilery Imprint, San Francisco Jossey-Bass (2004)
10. Pagliari, C., Donnan, P., Morrison, J., Ricketts, I., Gregor, P., Sullivan, F.: Adoption and perception of electronic clinical communications in Scotland. Informatics in Primary Care 13(2), 97–104(8) (2005)
11. Tang, P., Black, W., Buchanan, J., Young, C., Hooper, D., Lane, S., Love, B., Mitchell, C., Smith, N., Turnbull, J.: PAMFOnline: Integrating EHealth with an Electronic Medical Record System. In: Proceedings of the American Medical Informatics Association Annual Symposium, pp. 649–653 (2003)
12. Jennings, N.R., Sycara, K., Wooldridge, M.: A roadmap of agent research and development. Autonomous Agents and Multi-Agent Systems 1(1), 7–38 (1998)
13. Sharma, D., Ma, W., Tran, D.T.: On an IT Security Framework. In: Khosla, R., Howlett, R.J., Jain, L.C. (eds.) KES 2005. LNCS (LNAI), vol. 3681, pp. 226–232. Springer, Heidelberg (2005)
14. Bleumer, G.: Security of decentralized health information systems. International Journal of Biomed. Comp., 139–145 (February 1994)
15. Panko, R.R.: Corporate Computer and Network Security. Prentice Hall, Upper Saddle River, New Jersey (2003)

A Proactive Routing Algorithm Based on Bitmap Tables for Wireless Sensor Networks

Sangjoon Jung and Younky Chung

School of Computer Engineering, Kyungil University
712-701, 33 Buho-ri, Hayang-up, Gyeongsan-si, Gyeongsang buk-do, Korea
{sjjung,ykchung}@kiu.ac.kr
http://computer.kiu.ac.kr

Abstract. Current routing protocols in the sensor network focus on finding energy-efficient route setup so that the lifetime of the network is maximized. They do not use a routing table to decide the route due to increasing memory and updating costs. However, this paper proposes a new routing algorithm to broadcast a routing table to reduce the number of messages transferred to setup the routing path. The routing table is formed with a two-dimensional array, where each row and column is represented by one bit. When a node receives a request message, it updates its routing table after broadcasting. Thus, each node has a routing table with partial link information, which allows them to give much alternative route to the sink node. The proposed routing algorithm reduces the number of required messages for the routing establishment and preserves larger remained energy in compared to the previous algorithms.

1 Introduction

A sensor network is consisted with a large number of tiny sensing devices, deployed in a region of interest. Each device has computing and wireless communication capabilities, which enable it to gather information from the environment and to generate and deliver report with sensed data to the remote base station[1-4]. The base station aggregates and analyzes the reported messages through the established path. To aggregate and collect data efficiently, routing in sensor networks is an essential consideration due to several characteristics that sensor nodes are constrained in terms of limited energy supply, transmission power, processing capacity, and communication ability[5-8].

Recently, many routing algorithms have been proposed to establish a routing path supporting limited constrains in sensor networks [9-15]. They are commonly classified into three categories, i.e. the data-centric algorithm, the hierarchical algorithm, and the location based algorithm. In data-centric algorithms, the sink sends queries to certain regions and waits for data from the sensors located in the selected regions. Since data is being requested through queries, attribute-based naming is necessary to specify the properties of data[11,12]. In hierarchical algorithms, network clustering has been proposed to improve the system without additional load and degrading the service [13,14]. Location based algorithms use geographically informed neighbor

B. Apolloni et al. (Eds.): KES 2007/ WIRN 2007, Part II, LNAI 4693, pp. 555–564, 2007.
© Springer-Verlag Berlin Heidelberg 2007

selection heuristics to route a packet towards the target region[15]. The path should be established by querying data from a sink node to transfer the sensed data from nodes. However, establishing and maintaining a path is the most time consuming tasks. So, designing a new routing algorithm is necessary to reduce the overload in the network.

Most of the current sensor network environment assumes that sensor nodes are stationary while sink supporting the mobility [5-9]. Building a network that each sensor node has a routing table is considerable while the size of table is minimized. Thus, we propose a new routing algorithm that uses a bitmap to reduce the number of routing messages. Each node has a routing table with a bitmap, which contains link information. A bitmap is formed with a two-dimensional array, where each row and column is represented by one bit. Each node only updates its own bitmap if it receives a bitmap from adjacent nodes after broadcasting. The routing algorithm using the routing table should be concerned if the requirement of memory is restricted.

2 Related Routing Algorithms for WSN

In sensor networks, many researches have been processed to establish a route from sensor nodes to a sink. We briefly describe the routing algorithms under the categories.

In many applications of sensor networks, it is not feasible to assign global identifiers to each node due to the sheer number of nodes deployed. Such lack of global identification along with random deployment of sensor nodes make it hard to select a specific set of sensor nodes to be queried[5-10]. This consideration has led to data-centric routing, which is different from traditional address-based routing where routes are created between addressable nodes. In data-centric routing algorithms, the sink sends queries to certain regions and waits for data from the sensors located in the selected regions. Since data is being requested through queries, attribute-based naming is necessary to specify the properties of data[5,10].

Directed diffusion[11,12] has been suggested by using a naming scheme for the data. A sink floods interests for named data, setting up gradients within the network designed to draw event. Nodes able to satisfy the interest disseminate information along the reverse path interest propagation. Nodes receiving the same interest from multiple neighboring nodes may propagate events along the corresponding multiple links. Interests specify a low rate of data flow initially, but the sink starts receiving events it will reinforce one neighbor by requesting higher data rate events. This process proceeds recursively until it reaches the nodes generating the events, causing them to generate events at a higher data rate[11,12].

Directed diffusion is a good method with no need for a node addressing mechanism. Each node can do aggregation, caching, and sensing without maintaining global network topology. However, Directed diffusion cannot be applied to several applications since it is based on query-driven data delivery model. So, The applications that request continuous monitoring data to the sink will not work efficiently.

A single-tier network can cause the gateway to overload with too much data aggregation. Such overload might cause latency in communication and inadequate tracking of events. To improve the system without additional load and degrading the service, network clustering has been proposed[5,13]. The use of clusters for transmitting data to the sink is to efficiently maintain the energy consumption of sensor nodes by

involving them in multi-hop communication within a particular cluster and by performing data aggregation and fusion in order to decrease the number of transmitted messages.

Low-energy adaptive clustering hierarchy (LEACH)[13,14] is one of the most popular hierarchical routing algorithms for sensor networks. The main idea in the LEACH is to form clusters of sensor nodes based on the received signal strength and use local cluster heads as routers to the sink. LEACH is able to perform to local computation in each cluster to reduce the amount of data that must be transmitted to the sink. This achieves a large reduction in the energy dissipation since the transmissions will only be done by such cluster heads rather than all sensor nodes. LEACH is able to distribute energy dissipation evenly throughout the sensors, since cluster heads are selected randomly so that dynamic clustering increases lifetime of the networks. However, in the LEACH algorithm, it assumes adjustable transmitting power and assumes that the cluster head talks directly to the sink node. We assume each node has fixed transmission power, so the LEACH algorithm is outside scope of this paper.

In sensor networks, although there is no addressing scheme, most location information is needed so that sensor nodes report to the sink[5,15]. If the region is known, the query can be diffused only a particular region or node with using the location of sensors, which will eliminate the number of transferred messages. Some of the algorithms are proposed using these concepts. Geographic and energy aware routing (GEAR)[15] technique uses energy aware and geographically informed neighbor selection heuristics to route a packet towards the target region. Within a region, it uses a recursive geographic forwarding technique to disseminate the packet. In GEAR, each node keeps an estimated cost by learning distance to the destination through its neighbors[15]. Geographic routing is designed to improve efficiency relative to flooding data. However, many of these location-based algorithms are not applicable to sensor networks where there is mobility.

3 Routing Algorithm Based on Bitmap

The next section shows a structure of bitmap which contains links about neighbors in order for minimizing the size of routing table. The following describes how a sensor node broadcasts a bitmap.

3.1 Routing Algorithm by Broadcasting Bitmap

The main reason of using a bitmap is to reduce the number of messages when establishing routing paths and maintaining a small size of routing table. General routing algorithm using a routing table exchanges and updates routing information with neighbors. After updating is complete, each node has same routing table. This mechanism is not appropriate in a sensor network due to constrained energy. So, we design in a different way, and all nodes in our algorithm execute the same process as follows.

- Each node has a routing table in the form of bitmap.
- Each node only broadcasts it routing table once while diffusing an interest.
- After broadcasting, if a node then receives a routing table from other adjacent nodes, it only updates its own bitmap without propagation.

Each node contains a routing table in the form of bitmap, which consists of each row and column represented by one bit where a node has a path between its neighbors. The following equation shows the architecture of bitmap.

$$map[i][j] = \begin{bmatrix} id & id & \cdots & id \\ id & 0 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ id & 0 & \cdots & 0 \end{bmatrix} \quad (1)$$

From the equation 1, *map*[0][0] has the node own *id*. If the cross point between node *i* and node *j* has value '1', then there exists a link connection between them. This structure reduces the size of routing table.

To compose a route with using a bitmap in a node, the sink node broadcasts an interest. Each node then determines the attributes of its sensing and broadcasts its bitmap to neighbors.

3.1.1 Bitmap Updating Algorithm

A sink node disseminates an interest which involves in sensing tasks. The interest is propagated to flow towards multiple paths. The following describes an interest.

```
type = four-legged animal    // target task
interval = 20ms              // send back events each
                                20ms
duration = 10 seconds        // for next 10 seconds
rect = [-100, 100, 200, 400]// target location
sink node id = 0x01          // the sink node id is 1
```

We add a sink node *id* in the interest because the sink node *id* is necessary to expand sensor fields. The detail description is mentioned in Section 3.2. Each node investigates the interest whether it matches the attributes to its sensing tasks. If the node does not find a proper task about the attributes, then it broadcasts the received interest and its own bitmap. After broadcasting, the node updates its bitmap without propagation. Updating mechanism occurs when a node receives an interest from its adjacent as follows.

Step 1: When a node first receives a broadcast message from its neighbor, the node determines whether its data matches the attributes in the message by identifying the interest.

Step 2: The node updates its own bitmap as follows:

```
if (the received node id does not exist)
       { 1. Insert node id at the ith-position.
         2. Set 1 into the i and jth-position.}
   else
       { bitmap = own bitmap || received bitmap
         where || denotes or-operation.}
end if
```

Step 3: Each node broadcasts its own bitmap to its neighbors.

Step 4: If the node receives other request messages from adjacent nodes, it then just updates its own bitmap information without propagation.

3.1.2 Propagation of Bitmap

Each node broadcasts its bitmap to neighbors and updates its own bitmap free from propagating when the node receives another messages from neighbors. The total number of routing messages is reduced in our algorithm, because the node only broadcasts routing information once. As a result, each node has a bitmap with partial link information rather than total link information. Although each node has partial information, it does not matter to determine paths to report sensing tasks. As a sink node triggers all paths to a sensing area, so each node finds a path to reach to the sink. Routing steps by broadcasting are shown in Fig. 1.

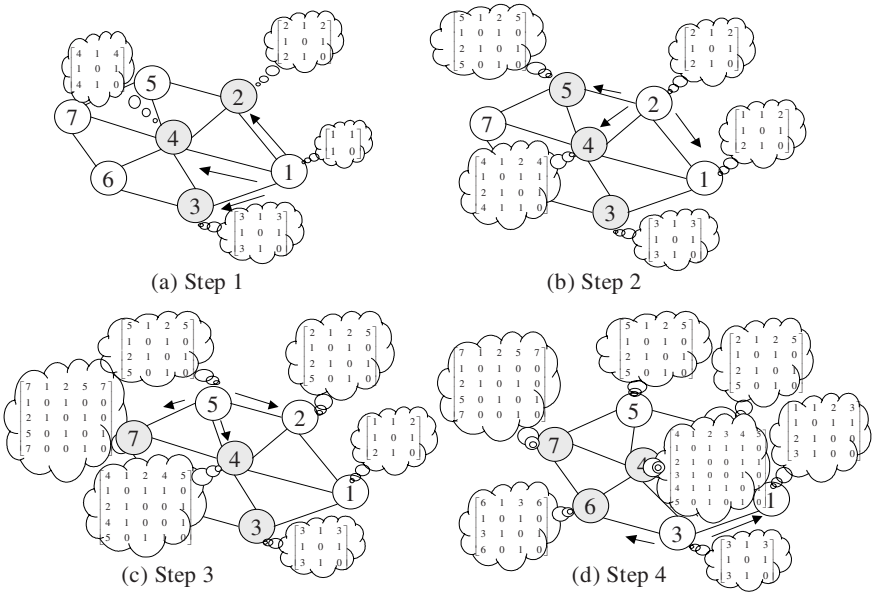


Fig. 1. Propagation of a bitmap (From step 1 to step 4)

Above Fig. 1, gray circles are ready to transmit link information. By using the above algorithm, a sink sends its own bitmap to neighbors to compose a routing table in the first step. The adjacent nodes then keep the bitmap from the node 1 and determine to broadcast its routing table(in Fig. 1(a)). As soon as each node receives the bitmap from the neighbor, it checks the interest and broadcasts the bitmap. After broadcasting, the node updates its own bitmap without propagation(in Fig. 1(b)). By doing this repeatedly, each node keeps a routing table(in Fig. 1(c) and (d)). All nodes participating in the sensor field have a bitmap after completing the routing algorithm.

3.1.3 Reporting an Interest to a Sink Node

After completing our algorithm, each node involving in sensing tasks identifies a sink node and path by investigating the bitmap. Each node searches the shortest path to the sink by using the BFS algorithm to report sensing tasks. An advantage of the proposed method is to find easily alternative paths when the shortest path is failed.

3.2 Structure for Reducing the Size of Bitmap

As a sensor node has limited resources, it is important to reduce the size of a bitmap. Fig. 2 shows a way to reduce the size of bitmap.

$$Node_ID[i] = \begin{bmatrix} 0 & 1 & \cdots & i & \cdots & n\text{-}1 \\ id_0 & id_1 & \cdots & id_i & \cdots & id_{n-1} \end{bmatrix}$$

$$link[i][j] = \begin{array}{c} \\ 0 \\ 1 \\ \vdots \\ j \\ \vdots \\ n\text{-}1 \end{array} \begin{array}{c} \begin{array}{cccccc} 0 & 1 & \cdots & i & \cdots & n\text{-}1 \end{array} \\ \begin{bmatrix} 0 & 0 & \cdots & 0 & \cdots & 0 \\ 0 & 0 & \cdots & 0 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 0 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 0 & \cdots & 0 \end{bmatrix} \end{array}$$

Fig. 2. Structure for reducing the size of bitmap

Each sensor node has two arrays which contain links information. One array has node *id*s and the other array has link information. Array of node *id* has the *i*th information that is related to the bit about the *i*th link of an adjacent node. If node *id* is allowed to be 1 byte, the size of node *id* array is 256bytes and the size of bitmap array is 8192bytes (65536 bits = 256 bits × 256 bits). So, each node contains routing table with 8448bytes.

3.3 Expansion of Sensing Area

As the number of nodes increases, the size of bitmap grows indispensably in spite of expressing by bits. To apply a routing algorithm using a routing table in a large sensor field, the routing table must be maintained a constant size. Thus, when creating a bitmap at the beginning of broadcasting, the number of adjacent nodes in the bitmap is limited by *l*. If the number of adjacent nodes is larger than *l*, the node broadcasts an interest where the sink node *id* in the interest is substituted with its own *id*. Thus, a node that receives the interest containing the substituted node *id* regards the propagating node as the sink node. The next shows the interests which changes the original sink node *id* to own node *id*. Each node has thirty node *id*s in its bitmap. Node *id* with 30 broadcasts an interest as depicted as following.

```
type = four-legged animal   // target task
interval = 20ms           // send back events each 20ms
duration = 10 seconds       // for next 10 seconds
rect = [-100, 100, 200, 400]// target location
sink node id = 0x1E         // the sink node id is 30
```

The node only propagates the received interest without bitmap. Thus, the next node that receives the interest regards a transferred node as a sink node, allowing it to create a new bitmap.

4 Performance Evaluations

To evaluate the performance of the proposed algorithm, we use NS(Network simulator)-2 to compare it with Directed diffusion algorithm.

4.1 Simulation Environment

We placed 50, 75, 100, 125, 150, 175, and 200 nodes in 1000m*1000m and executed 10 times respectively, because all node located randomly. To measure the remained energy, a sensor node include the following parameters in the sensor node's `node-config` routine:

```
-energyModel EnergyModel \
-rxPower 0.175 \
-txPower 0.175 \
-sensePower 0.00000175 \
-idelPower 0.0 \
-initialEnergy 5
```

where

- `rxPower` 0.175 indicates *175mW* consumed for receiving a packet,
- `txPower` 0.175 indicates *175mW* consumed for transmitting a packet,
- sensePower 0.00000175 indicates 1.75μW consumed for matching the attributes to the tasks, and
- `initialEnergy` 5 indicates a total energy reserve of *5J*.

4.2 Results

Two algorithms have been simulated to obtain the total number of routing messages and remained energy. First, we check the number of routing messages, because each node can preserve its energy by reducing the number of routing messages. Fig. 3 shows the total number of routing messages between the Directed diffusion and the proposed algorithm.

In Fig. 3, we notice that the number of routing messages of the proposed algorithm is smaller than that of the Directed diffusion. Therefore, our algorithm reduces consumption of energy by reducing the broadcasted messages. Fig. 4 shows results about remained energy over two algorithms.

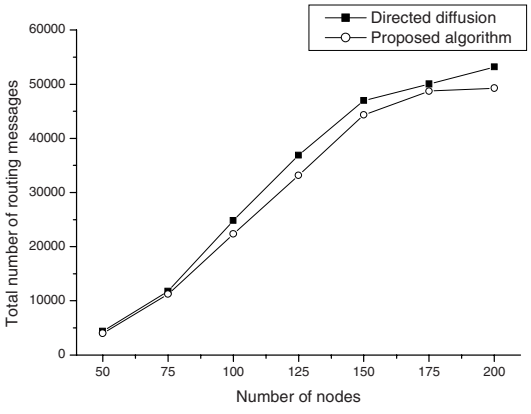


Fig. 3. Total number of routing messages

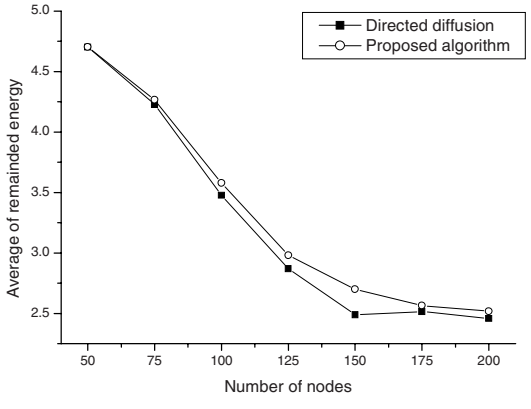


Fig. 4. Average of remained energy

In Fig. 4, all nodes using our algorithm have more energy in compared to the Directed diffusion algorithm. As all nodes in our simulation located randomly, so differences of two algorithms are not constant. It is found that our algorithm can achieve lower dissipation value of energy, better preserving dissipated network energy and effectively postpone network lifetime.

5 Conclusions

In this paper, we have proposed a routing algorithm to broadcast a bitmap in sensor networks. The main focus to design an algorithm is to reduce the battery power by

reducing the number of routing messages. Our consideration is to optimize the path selection using a routing table and to prolong a network lifetime. When establishing routing paths, a sensor node can preserve more resources by reducing the number of routing messages. Each node has a bitmap which contains link information about neighbor nodes. Upon receiving a bitmap, each node updates and only broadcasts the bitmap to the neighbors once. After broadcasting, the node just updates its own bitmap without propagation. By using the proposed algorithm, each node has routing information to find a sink node while restricting the number of routing messages and prolonging the network lifetime. Our future plan is to implement the algorithm in a sensor node.

References

[1] Tian, D., Georganas, N.D.: Energy efficient routing with guaranteed delivery in wireless sensor networks. In: Wireless Communications and Networking, WCNC 2003, March 16-20, 2003, vol. 3, pp. 1923–1929. IEEE, Los Alamitos (2003)
[2] Akyildiz, I.F., Su, W., Sankarasubramaniam, Y., Cayirci, E.: Wireless Sensor Networks: A Survey. Computer Networks 38(4), 393–422 (2002)
[3] Sohrabi, K., Gao, J., Ailawadhi, V., Pottie, G.J.: Protocols for Self-Organization of a Wireless Sensor Network. IEEE Personal Communications 7(5), 16–27 (2000)
[4] Min, R., Bhardwaj, M., Cho, S., Shih, E., Sinha, A., Wang, A., Chandrakasan, A.: Low-Power Wireless Sensor Networks. In: Proceedings of International Conference on VLSI Design, Bangalore, India (January 2001)
[5] Akkaya, K., Younis, M.: A survey on routing protocols for wireless sensor networks. Ad Hoc Networks 3(3), 325–349 (2005)
[6] Rodoplu, V., Meng, T.H.: Minimum Energy Mobile Wireless Networks. IEEE Journal of Selected Areas in Communication 17(8), 1333–1344 (1999)
[7] Li, L., Halpern, J.: Minimum-Energy Mobile Wireless Networks Revisited. In: Proceedings of IEEE Conference of Communications (ICC '01), Helsinki, Finland (June 2001)
[8] Krishnan, R., Starobinski, D.: Efficient clustering algorithms for self-organizing wireless sensor networks, Ad Hoc Networks, Corrected Proof, Available online (March 22, 2005) (in press)
[9] Tilak, S., Abu-Ghazaleh, N.B., Heinzelman, W.: A taxonomy of wireless microsensor network models. Mobile Computing and Communications Review 6(2), 28–36 (2002)
[10] Heinzelman, W., Kulik, J., Balakrishnan, H.: Adaptive protocols for information dissemination in wireless sensor networks. In: Proceedings of 5th Annual ACM/IEEE International Conference on Mobile Computing and Networking (MobiCom'99), Seattle, WA (August 1999)
[11] Intanagonwiwat, C., Govindan, R., Estrin, D.: Directed diffusion: a scalable and robust communication paradigm for sensor networks. In: IEEE/ACM Mobicom, pp. 56–67 (2000)
[12] Estrin, D., Govindan, R., Heidemann, J., Kumar, S.: Next century challenges: Scalable coordination in sensor networks. In: Mobile Computing and Networking, pp. 263–270 (1999)

[13] Heinzelman, W., Chandrakasan, A., Balakrishnan, H.: Energy-efficient communication protocol for wireless sensor networks. In: Proceeding of the Hawaii International Conference System Sciences, Hawaii, pp. 3005–3014 (January 2000)
[14] Subramanian, L., Katz, R.H.: An architecture for building self configurable systems. In: Proceedings of IEEE/ACM Workshop on Mobile Ad Hoc Networking and Computing, Boston, MA (August 2000)
[15] Yu, Y., Estrin, D., Govindan, R.: Geographical and energy-aware routing: a recursive data dissemination protocol for wireless sensor networks, UCLA Computer Science Department Technical Report, UCLA-CSD TR-01-0023 (May 2001)

A Study on Mobile Agent Based Resource Management in Grid

Chen Zhao[1,3], Jian Yu[2,3], and Bencheng Chai[1]

[1] School of Computer Science and Information Technology, Zhejiang Wanli University
Ningbo, Zhejiang 315100, P.R. China
[2] College of Information and Engineering, China Jiliang University
Hangzhou, Zhejiang 310018, P.R. China
[3] College of Computer and Science, Zhejiang University
Hangzhou, Zhejiang 310027, P.R. China
{ChenZhao,csczhao}@sohu.com

Abstract. The cost models for estimating the response time have been presented, and the architecture of grid resource management based on mobile agent is given in this paper. An integration of mobile agent technology with OGSA is working effectively towards reducing the network traffic during the grid resource processing.

Keywords: Mobile Agent, Resources Management, Grid, OGSA.

1 Introduction

A grid offers a simple, dependable and pervasive access across wide area networks and presents users with an integrated global resource. Resources in grid are distributive and dynamic and must be accessible to all users on the grid. In this distributive environment, the problems to be resolved include the assignment and adjustment of the resources and tasks, security, network bandwidth, and resources discovery and monitoring.

Mobile agent as a new network technology is developing very rapidly nowadays. An agent is flexible, mobile, and autonomous. Integrating mobile agent technology and grid technology to make most of the advantages of the both will effectively reduce the cost of communications in grid resources discovery and monitoring [1]. Although existing frameworks and distributed systems for mobile computing, such as Rover [2], tackle distributed consistency and replication management, their communication models and programming models are unsatisfactory. However, agent based communication paradigms have shown enormous potential for operating in unpredictable, metamorphic environments, such as mobile computing networks.

This paper is organized as follows. In section 2, we review related work in the field of grid structure. In section 3, the cost formulae are presented, and the design of Mobile Agent-based Resources Information Management is described. Finally in section 4, we present the current status of our project and the future directions of this work.

B. Apolloni et al. (Eds.): KES 2007/ WIRN 2007, Part II, LNAI 4693, pp. 565–570, 2007.
© Springer-Verlag Berlin Heidelberg 2007

2 The Structure of Grid

There are two main architectures of grid: one is the layered hourglass model that focuses on protocol [3]. The other is OGSA (Open Grid Service Architecture) that focuses on service [4]. OGSA is the most important and new kind of grid architecture after five layers grid architecture considered next generation grid architecture. OGSA focuses on shared services. Service concept that includes all sorts of computing resources, memory resources, network, procedure and database is widely used in OGSA. From resource to service, it unifies resources, information and data and is benefit for realizing flexible, consistent and dynamic share mechanism. This makes distributed system administration have standard interface and behavior. All are considered as grid service in OGSA. So grid is extensible set of Grid Services, namely grid={grid service}. In the Grid Services there are teams of interfaces that resolve problems of discovering services, creating services dynamically, managing life period and notifying.

Grid in the narrow sense refers to Globus (OGSA) and its application platform, but in the broad sense it refers to the highly integrated, highly synthesized, highly safe and flexible and efficient platform, which fully absorbs and converses various computation resources into reliable, standard and economical computation capacity acquirable everywhere.

3 Mobile Agent-Based Resources Information Management

A traditional resources information management system on a set of network computers requires mapping each drive on the local file system. This is a tedious task. During the operation, the results are frequently returned to the requestor on a continuous basis consuming valuable network resources. One possible solution to this problem is to develop an agent utility that will only do local work. This utility is then deployed on each of the desired machines.

We used mobile agent technology to solve this problem. With this technology, a requestor creates a mobile agent and dispatches it to a desired remote machine on the network. The search criteria are defined in the agent that carries the related code to the remote machine. After the work is done, the results are dispatched to the sender. In general, mobile agent reduces the load on network, reduces the code design risk, and provides a robust application over unreliable networks. Mobile agent technology has a few disadvantages. One is the requirement for an agent server on each participating machine. The other disadvantage is protection against the malicious code and general code and data security. These disadvantages will affect the system response time.

3.1 Cost Agent Model

In this section, we present our cost models for distributed resources management in grid. The cost formulae are apriori estimates of the distributed response time for a given task using a specified architectural model when environmental factors are taken into consideration.

The response time for distributed resources management is as follows:

$$T = t_{drp} + t_{kj} \tag{1}$$

Where T is the response time, t_{drp} is the time taken to perform distributed resources processing in a distributed environment and t_{kj} is the time taken to perform knowledge integration.

This part is characterized by a given distributed resources using the mobile agent paradigm. The core steps involved include: submission of a task by a user, dispatching of mobile agents to the respective resources servers, return of mobile agents from the resources with processing results. This agent model is characterized by a set of mobile agents traversing the relevant resources servers to perform algorithm. In general, this can be expressed as *m* mobile agents traversing *n* resources. The algorithm used across the different resources can be uniform or varied. The model dispatches a mobile agent encapsulating the algorithm to each of the resources servers participating in the distributed resources analysis.

Before considering the general case of *m* agents and *n* resource, we first develop the cost model for the situation where there is only one mobile agent for *n* distributed resources sites. The agent is dispatched from the agent center to the first of the *n* resources sites. From there on, the agent completes the task and carries the results obtained to the next resources site and so on, until each of the *n* sites have been completed. Therefore, the response time from distributed resources is as stated in equation (1). In this scenario, t_{drp} is as expressed in equation (2) as follows:

$$\begin{aligned} t_{drp} &= t_{drAgent_{(AC,1)}} + \\ &\sum_{i=1}^{n-1} t_{dr}(i) + \sum_{i=1}^{n-1} t_{drAgent}(i, i+1) + t_{dr}(n) + t_{drAgent(n,AC)} \end{aligned} \tag{2}$$

In the above cost estimate, the term $t_{drAgent(AC,1)}$ is the time taken for the resources agent to travel from the agent center to the first resources site (the data servers being numbered from *1* through *n*). The second term is the total time taken by the agent to perform resources at server *i* and then travel to server *i+1* as *i* ranges form *1* to *(n-1)*.

Now we extend the above cost estimate for the general case involving *m* mobile agents and *n* resources sites. The *n* resources sites are numbered from *1* through *n*. Let *ds* be labeling of a set of resources sites. Therefore $ds = \{1,2,\cdots,n\}$. Let *m* be the number of agents available. The set *ds* are divided into *m* subsets – $d_{s1}, d_{s2}, \cdots, d_{sm}$ with the following property: $d_{si} \subseteq ds(1 \le i \le m)$. The sets $d_{si}(1 \le i \le m)$ have the additional property of independence. That is, $d_{si} \cap d_{sj} = \phi, i \ne j$. A corollary of the above is that the sum of the cardinalities of the subsets $d_{s1}, d_{s2}, \cdots, d_{sm}$ is equal to the cardinality of *ds*. This implies that $|d_{s1}| + |d_{s2}| + \cdots |d_{sm}| = |d_s|$. Thus,

the n resources servers are divided into m subsets and the ith resources agent is assigned the task of the resources sites in the subsets i (where $1 \le i \le m$). This means that m different agents can operate at the same time. Thus the ith agent has to travel to and perform in the number of sites specified in $|d_{si}|$. The total time taken to analysis is therefore the time taken by the agent, which takes the maximum time interval to complete its task. The cost estimate for the ith agent's response time is as follows in equation (3):

$$t_{drp}(i) = \left(\sum_{j=1}^{|d_{si}|-1} t_{dr}(d_{si}(j) + t_{drAgent}(j, j+1)\right) + t_{dr}(d_{si}(|d_{si}|)) + t_{drAgent}(d_{si}(|d_{si}|), AC) + t_{drAgent}(AC, d_{si}(1)) \tag{3}$$

In the above equation, the first term is the time taken by the resources agent to travel from the agent center to the first resources server in its path (i.e. the ith subset d_{si}). The term involving the summation is the time taken for the agent to analysis and to travel to the respective data sites within the set assigned to it (excluding the final site). The second last term is the time taken to analysis at the last resources site in its path. The final term in the expression is the time taken for the agent to travel from the last site on its path to the agent center.

Since there are m agents operating concurrently, the time taken for completion of the distributed resources process is the time taken by the agent requiring the longest completion time. Thus,

$$t_{drp} = \max(t_{drp}(i)), \text{ where } i = 1 \cdots m \tag{4}$$

where $t_{drp}(i)$ is estimated from equation (3).

3.2 Resources Information Management Module

In order to improve the response time, here we propose a Hybrid Model for Resources Information Management. This Model is based on Globus Generic and Data Grid Services, and composed of three parts: Information Server, Agent Center, and Optimiser. The Information Server creates corresponding Mobile Agent on the basis of resources features, and designates a specific binding condition and route rule, process data brought back by Mobile Agent. Agent Center dispatches the Mobile Agent, transfers according to routes and strategy predetermined by the Information Server between the nodes of the resources and alternates with the system agent in each node, and process dynamic information of resources. Resources node provides executive environment and system agent for Mobile Agent, while executive environment accepts Agent Center, and provides access to local resources. The alternation between resources node and Agent Center finishes the task of information collection set by the resources Information Server. The model is shown in figure 1.

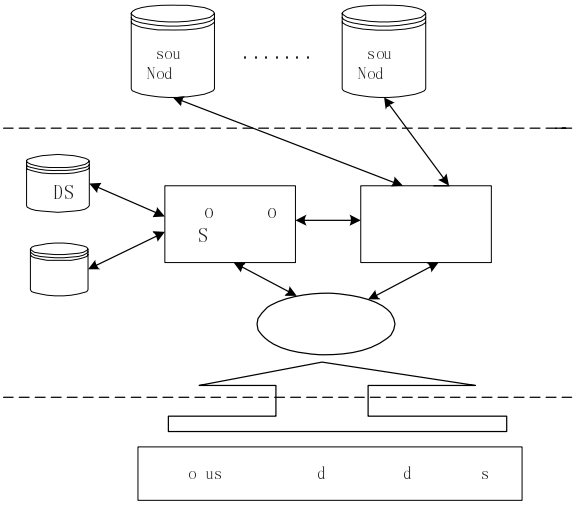


Fig. 1. Mobile Agent-based Resources Information Management Module

Let i be the number of resources sites to be processed. The Optimiser decides on the basis of the cost estimates that it builds that m_i sites are to be processed using the Agent Model. The response time in the Hybrid Model assuming that the Mobile Agent-Based processing components are done in parallel is the time taken by the technique requiring a longer duration.

$$t_{hybrid} = \max(t_i) + t_{IS}, \text{ where } i = 1 \cdots m \tag{5}$$

Where t_i is the time taken to process the m_i sites using the Agent Model and t_{IS} is the time to process the Information Server. The detailed design of the Hybrid Model can be found in [5].

The Hybrid Model are built on top of Globus [6], a middleware for Grid platforms, and simplify the task of managing computations that access distributed and large data sources. The Knowledge Directory Service (KDS) extends the basic Globus Metacomputer Directory Service (MDS), the metadata managed by the KDS are represented through XML documents stored in the Knowledge Metadata Repository (KMR). Metadata regard the following kind of objects: data sources characteristics, data management tools, data mining tools, mined data, and data visualization tools. Metadata representation for output data models may also adopt the (PMML) standard [7].

4 Conclusions and Future Work

In this paper we presented cost agent models for distributed resources information management. The part of our work provides theoretical basis for cost Agent model. The cost model will need experimental validation, so we are currently the hybrid models for resources information. Using the cost estimates, we aim to establish that an integrated model using the agent-based grid can be more effective in reducing the response time of distributed resources analysis.

Mobile Agent-based resources information collection method transmits a small quantity of running code and status to resources nodes, analyzes and predicts about the resources function locally at the resources nodes, and bring the results back to the information server, thus avoiding transmitting a large quantity of resources dynamic information back to the resources information server and resulting in the decrease of network load.

Moreover, when a resources node goes out of the grid abnormally, Mobile Agent cannot reach the node. This status can be recorded in the interior status. When Mobile Agent returns to the information server, the information servers cancel the node so that the resources information does not have to maintain the heartbeat signal in every resources node. The advantage of this kind model is obvious when the membership of the grid is expanded.

Grid resources discovery and management plays a vital role in application environment in grids. The Mobile Agent-based grid resources information management model proposed in this paper is a seamless integration of grid technology and mobile agent technology, which is built upon the heterogeneity and variety of grid resources and the flexibility, mobility, autonomy and agency function which will reduce the cost of the system and improve the efficiency of resources management.

Acknowledgments. This work was supported by Ningbo Science Foundation under grant 2006A610031.

References

1. Peine, H., Stolpmann, T.: The architecture of the Ara platform for mobile agents. In: Rothermel, K., Popescu-Zeletin, R. (eds.) MA 1997. LNCS, vol. 1219, Springer, Heidelberg (1997)
2. Joseph Anthony, D., Alan, F., DeLespinasse, J.A., Tauber David, K.: Gifford.: Rover: A Toolkit for Mobile Information Access. In: Proceedings of the Fifteenth Symposium on Operating System Principles (December 1995)
3. Zhihui, D., Lisan, D., Chen, Y., Peng, L.: Grid Computing and its prototype execution research. Computer Science 29(8), 1–5 (2002)
4. Foster, I., Gannon, D.: Open Grid Services Architecture Platform (OGSA) (2003), http://forge.gridforum.org/projects/ogsa-wg
5. Krishnaswamy, S., Zaslavsky, A., Loke, S.W.: An Architecture to Support Distributed Data Mining Services in E-Commerce Environments. In: the 2nd Intl. Workshop on Advanced Issue. E-Commerce and Web-Based Information Systems, San Jose, Californinia, July 8-9 (to appear, 2000)
6. Foster, I., Kasselman, C.: Globus: A metacomputing infrastructure toolkit. Intl. J. Supercomputer Applications 2(11), 115–128 (1997)
7. The Data Mining Group. PMML 2.0, http://www.dmg.org/pmmlspecs_vs/pmml_v2_0.html

An Agent-Based Evolutionary Robotic System for Its Reconfiguration

Xueshan Gao[1], Koki Kikuchi[2], and Xiaobing Wu[1]

[1] Beijing Institute of Technology, 100081, Beijing, China
{Xueshan.Gao,Xiaobing.Wu}@bit.edu.cn
[2] Chiba Institute of Technology, 275-0016, Narashino, Japan
kouki.kikuchi@it-chiba.ac.jp

Abstract. In this paper, the evolutionary algorithm is applied to generate the morphology and intelligence of an individual agent robot in the multi-agent robotic system. Each robot agent is given the functions emerges from intelligence and balance between morphology and intelligence from the other agent robot and task of following and contacting. Both of the morphology and intelligence are automatically generated and evolved by genetic programming for the task. Furthermore, we performed evolution simulation experiments, and discussed the symmetry properties.

Keywords: Robot agent, Multi-agent robot system, Symmetry properties, Emergent gene.

1 Introduction

In recent years, multi-agent system (MAS) approaches to multi-robot system design have drawn much attention in the robotics community [1]-[4]. In an agent-based architecture, multiple robot agents can communicate, coordinate and negotiate to meet their goals in a framework suitable for task execution[5].The architecture supports both centralized (agent facilitator model) and distributed (agent to agent model) arrangement [6]. It has been reported that the multi-agent robot systems which can do many kids of tasks efficiently by training the rules between environments and actions using evolutionary algorithm [7]-[10].

For many complex tasks in robotic system, such as spanning obstacle or load transport, it is desirable to take advantage of the flexibility provided by using multi-agent robots[11][12]. It would also be interesting to adopt a multi-robot approach in

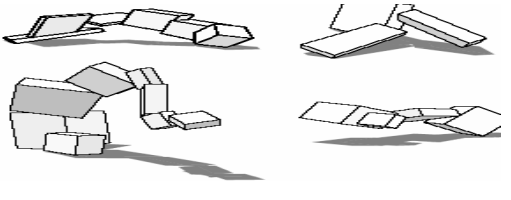


Fig. 1. Image of the reconfigurable robot system with robot agents

B. Apolloni et al. (Eds.): KES 2007/ WIRN 2007, Part II, LNAI 4693, pp. 571–580, 2007.
© Springer-Verlag Berlin Heidelberg 2007

assisting disaster rescue, for example in an earthquake or fire occurs. A multi-robot cooperative system appears to be more effective and adaptive to accomplish various complex tasks, relative to a single robot approach.

In this kind of situation, to obtain the multi-agent robot system ability, the robot agent is required to have the ability to recognize the morphology of itself with the other robot agent.

From living creatures, it can be realized that characteristic of symmetry or asymmetry is extremely important for physical structure and functionality, including control [13]-[16]. Hence, contributed to the development of a methodology for a robot agent, its own the morphology and intelligence depending on a changing environment can be designed. In this paper, we focus on symmetry properties and perform computational simulations by genetic programming (GP) [17], [18] using an evolutionary robot agent with reconfigurable morphology and intelligence. Here, the morphology is represented by the arrangement of two visual sensors, and the intelligence is described as a decision tree consisting of sensor inputs and motor outputs. And then two experiments are performed: Case A and Case B, differing in the relative velocities of a robot and an object. Because our studies are focused on individual robot agent, the object is considered as the other robot agent.

2 Robotic System Setup

2.1 Modeling

In this study, the robot is modeled in two dimensions and has two visual sensors, two wheels with motors, and a round shape. The motion of the robot is calculated based on a two-dimensional dynamic model including realistic physical conditions such as collision and friction. The equations of motion are given by

$$\begin{aligned} M\ddot{x} &= \sum_{i=0}^{1} \frac{T_i}{R_t} \sin\theta + F_x + P_x \\ M\ddot{y} &= \sum_{i=0}^{1} \frac{T_i}{R_t} \cos\theta + F_y + P_y \end{aligned} \tag{1}$$

where M is the mass of the robot, x and y are the coordinates of the center of the robot, Ti is the torque of the motor, Rt is the wheel radius, F is the friction force with the floor, P is the impact force with the object or a wall, I is the moment of inertia, θ

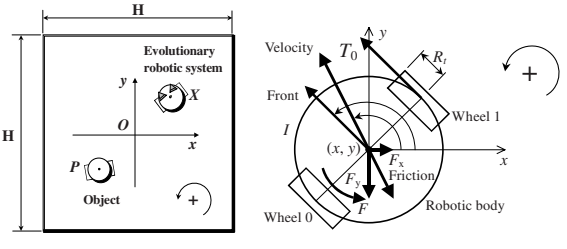


Fig. 2. Simulation arena and robotic system

is the direction of the robot, and i is the wheel ID with left =0 and right =1, and the origin of the centre of the arena, as shown in Fig.2. Using these equations, the motions of the robot and of the object are simulated by the Runge-Kutta method.

In the beginning, there is a certain distance for the robot and an object, which are in an arena surrounded by walls, as shown in Fig.2. The object moves randomly and the robot behaves using a morphology and intelligence automatically generated by GP. A fitness value F is then calculated for the robot, according to its performance and defined as:

$$F = \frac{1}{N_1}\sum_{i=1}^{N_1}\left\{\alpha_1 \times \frac{1}{T}\int_0^T |(|\overrightarrow{X_{t,i}P_{t,i}}| - D)|^2 \, dt + \alpha_2 \times \frac{(N_2)^2}{T}\right\} \tag{2}$$

Where X is the center of the robot, P is the center of the object, T is the evaluation time, N_1 is the number of trials, and N_2 is the number of collision between the robot and the object. The weighting factors are $\alpha_1 = 0.7$ and $\alpha_2 = 0.3$. The first term is the time integral of the distance between the robot and object, and so the robot gets a high evaluation if it maintains the target distance D. The second term depends on the number of collisions over time, so the robot gets a high evaluation if it does not collide with the object. In this simulation, the fitness value is the average of a number of trials N_1. We define the actual fitness value Fa for $N_1 = 10{,}000$, for high reliability, for the behavior analysis.

The evolutionary robotic system is optimized through the processes of GP: (1) developing, (2) evaluation, (3) selection and multiplication, and (4) crossover and mutation. Under GP, each robot is treated as an individual coded by a morphology gene and an intelligence gene.

2.2 Morphology and Intelligence Genes

The morphology is represented by the physical arrangement of two flexible visual sensors, two fixed motors, and its round shape, as illustrated in Fig.3.

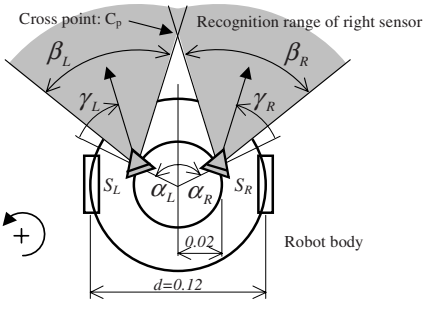


Fig. 3. Parameters of morphology

Here, two visual sensors S_L and S_R have three degrees of freedom: alpha, beta, and gamma. Alpha corresponds to the arrangement angle of the sensors on a circumference of 0.02m radius ($0° \le \alpha_L, \alpha_R \le 90°$). Beta corresponds to the range of

the field of view ($0° \le \beta_L, \beta_R \le 50°$). Gamma corresponds to the direction of the visual axis ($-90° \le \gamma_L, \gamma_R \le 90°$). Thus, the robotic system has six degrees of freedom. Note that the shaded areas show the recognition area for the object; the sensor becomes "ON" when the object is recognized. The intelligence of this robot is a computer program expressed as a decision tree that represents the relationship between the sensor inputs and the motor outputs. The decision tree is created by using two kinds of nodes: terminal nodes and non-terminal nodes, as shown in Table1. The sensor nodes *L* and *R* correspond to the state of the two sensors S_L and S_R shown in Fig.3, with "true" and "false" assigned to "ON" and "OFF".

Table 1. Nodes for Decision Tree

	Terminal nodes	Non-terminal nodes
Sensor nodes	L, R	
Motor nodes	*MOVE_F, MOVE_B, TURN_L, TURN_R, STOP*	
Function nodes		IF, AND, OR, NOT

The motor nodes have the action functions *MOVE_F* to move forward, *TURN_L* to turn left, *TURN_R* to turn right, *MOVE_B* to move back, and *STOP* to stop. Fig. 4 shows the behaviors of these functions. The non-terminal nodes are function nodes, etc. To avoid bloating [19], the maximum number of nodes is set to 200 and if the intelligence gene exceeds this limit during crossover or mutation.

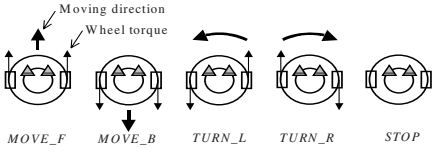


Fig. 4. Robotic behavior for each motor node

3 Simulation

3.1 Simulation Conditions

To investigate the relationship between the morphology and the intelligence, we carried out two evolutionary experiments. In Case A, the robot is twice as fast as the object and in Case B, the object is twice as fast as the robot.

The physical conditions: the length of one side of the arena *H* was 4.0 m, the diameter of the robot and the object d was 0.12 m, the target distance *D* was 0.5 m, the evaluation time *T* was from 30.0 s to 45.0 s (randomly determined), the maximum speed for the robot and the object was 0.07 m/s, the sensor sampling time was 0.02 s, the robot weight and the object *M* was 0.4 kg and the recognition error of the sensors was from –3.0º to 3.0º (randomly determined). The parameters of GP were set as

follows: the population size was 500, the generation was 150, the selection ratio was 0.2, the crossover ratio was 0.3, and the mutation ratio was 0.1. The initial positions and directions of the robot and object were randomly determined from a uniform distribution within a center area of 2.0 m × 2.0 m. Here, the probability that the two were arranged by an initial distance shorter than D was 15%.

3.2 Indices of Symmetry Properties

To analyze the symmetry properties of the robotic system, we defined three indices that represent the symmetry properties of the structure: $|\alpha_L - \alpha_R|$, $|\beta_L - \beta_R|$, and $|\gamma_L - \gamma_R|$. In the first GP process, these values were uniformly generated.

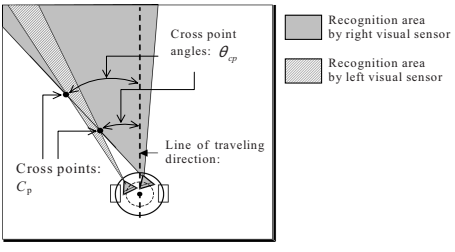


Fig. 5. Definition of cross points and cross point angles

We defined further indices for the state space created by the visual sensors. The values for C_p represent the cross points of the recognition areas of the two visual sensors, as illustrated in Fig. 5 and the values of $_{cp}$ represent the angles between the line of the traveling direction and the lines connecting the cross points and the center of robot. Note that the maximum number which a robot can create is four. We denote the cross points that are employed for action assignment as the actual cross points C_p. Similarly, $_{cp}$ represent actual cross point angles. And 200 simulations were performed for both Case A and Case B.

4 Discussion

4.1 Robotic Systems Description

Table2 shows the average of the actual fitness Fa, of the elite individuals at the final generation and the standard deviation for each case. Table 3, 4, and 5 detail typical genotypes (i.e., both morphology and intelligence genes): Type I and Type II obtained in Case A and Type III in Case B. The left and right diagrams in Fig.6 and Fig.7 show the corresponding phenotypes presented by each genotype and illustrate the state-action spaces.

The robotic system Type I (Table3 and Fig.6) made up 37% of the elite individuals in Case A. The state space for this robot is divided into four areas, with each of four actions assigned to each space. This robot accomplished the task using the following strategy. As shown in Table3, if both *L* and *R* are false, then select *TURN_L*, namely,

if either left or right visual sensors recognize the object, the robot turns left (State1 in the left diagram of Fig.6). If *L* is false and *R* is true, then select *MOVE_F*, namely, if the left visual sensor does not recognize the object while the right does, the robot moves forward (State 2 in the left diagram of Fig.6). If both *L* and *R* are true, then select *TURN_R*, namely, if both sensors recognize the object, the robot turns right (State 3 in the left diagram of Fig.6).

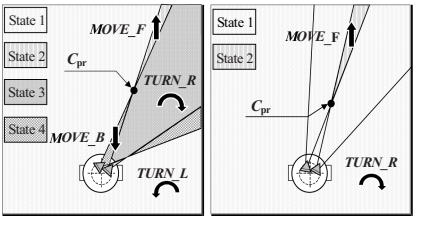


Fig. 6. Phenotypes: state-action spaces of Type I (left) and Type II (right)

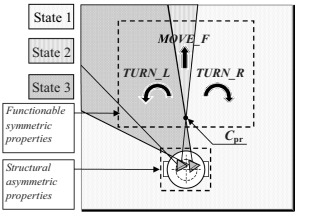


Fig. 7. Phenotypes: state-action space of Type III and two kinds of properties

If *L* is true and *R* is false, then select *MOVE_B*, namely, if the right sensor does not recognize the object while the left does, the robot moves backwards (State 4 in the left diagram of Fig.6). The Type I robot approached the object by using these behaviors and achieved a high evaluation, an average F_a of 1.60. It could track down the object to the cross point under any arrangement of the robot and the object.

Similarly the robotic system Type II (Table4 and Fig.6) made up 12% of the elite individuals in Case A. The state space for this robot is divided into two areas with two actions assigned to them. But, as the period during which the robot lost the object was long, this robot did not get a high evaluation, achieving an average F_a of 3.85.

And the robotic system Type III (Table5 and Fig.7) made up 52% of the elite individuals in Case B. The state space for this robot is divided into three areas, with three assigned actions. Its strategy allowed a high-speed chase unlike the Type I robot, and shortened the period during which the robot lost the object, unlike the Type II robot. The Type III was an average F_a of 3.17. The Type III robot also made up 37% of Case A. However, it did not get a high evaluation for Case A because it was not able to go backwards when the robot agent moved too close to the object considering its own high-speed. The Type III and Type I robots accomplish the task by employing parallax of its left and right visual sensors.

Table 2. Actual fitness *F*a, obtainned in Case A and Case B

Heading level	Ave.	Std.Dev.
Case A	2.95	1.66
Case B	8.60	0.56

Table 3. Type I: Genotype of evolutionary robotic system, obtained in Case A

Morphology genes [deg]	Intelligence genes
α_L=26, β_L=44, γ_L=74	(if (or L (not R)) (if R *TURN_R* *MOVE_B*) (if R *MOVE_F* *TURN_L*)
α_R=78, β_R=36, γ_R=39	

Table 4. Type II: Genotype of the robotic system obtained in Case A

Morphology genes [deg]	α_L=67, β_L=19, γ_L=80 α_R=89, β_R=30, γ_R=60
Intelligence genes	(*if* (*and* *L* *R*) *MOVE_F* *TURN_R*)

Table 5. Type III: Genotype of the robotic system. obtained in Case B

Morphology genes [deg]	α_L=82, β_L=43, γ_L=67 α_R=44, β_R=50, γ_R=77
Intelligence genes	(*if* (*or* *R* *L*) (*if* *R* *TURN_L* *MOVE_F*) *TURN_R*))

Table 6. Results of structural indices obtained in Case A and Case B

	Case A		Case B	
	Ave.	Std.Dev	Ave.	Std.Dev
$\lvert\alpha_L-\alpha_R\rvert$	34.4	22.2	38.1	23.1
$\lvert\beta_L-\beta_R\rvert$	16.9	12.0	17.0	12.7
$\lvert\gamma_L-\gamma_R\rvert$	32.0	24.0	38.3	33.0

Table 7. Corelation coefficients between structural indices and actual fitness

	Correlation Coefficients	
	Case A	Case B
$\lvert\alpha_L-\alpha_R\rvert$	0.06	0.11
$\lvert\beta_L-\beta_R\rvert$	0.14	0.32
$\lvert\gamma_L-\gamma_R\rvert$	0.48	0.48

4.2 Symmetric Properties

Table 6 shows the average of the indices of the symmetry properties of the structure, $|\alpha_L - \alpha_R|$, $|\beta_L - \beta_R|$, and $|\gamma_L - \gamma_R|$, of the elite individuals at the final generation and the standard deviations. Table7 indicates the correlation coefficients between these values and the actual fitness.

Both the averages and the standard deviations are high, so we could not identify distinguishing structural symmetry properties represented by the arrangement of the two visual sensors. However, the fact that the correlation value between $|\beta_L - \beta_R|$ and the actual fitness in Case B was 0.32 suggests that the smaller the difference in the range of the left and right fields of view, the higher the evaluation becomes. Additionally, the correlation value of $|\gamma_L - \gamma_R|$ was 0.48.

Fig. 8 shows the angle distribution of the actual cross point angles $_{cpa}$. Fig. 9 displays the relationship between this angle and the actual fitness at the 30th, 60th, 90th, 120th, and 150th generation in Case A. Fig.9 shows the same for Case B. 37%

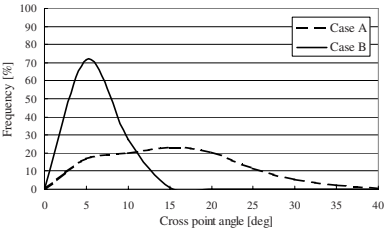


Fig. 8. Comparison of actual cross point angle distributions incase A and case B

of the elite individuals in Case A created actual cross point angles $_{cpa}$ within 10deg, as shown in Fig.9, while 99% did in Case B. This result suggests that in Case B most robotic systems designed the actual cross point in front of the robot and assigned actions based on this configuration. The correlation coefficients between the actual cross point angle $_{cpa}$ and the fitness were 0.31 in Case A and 0.43 in Case B.

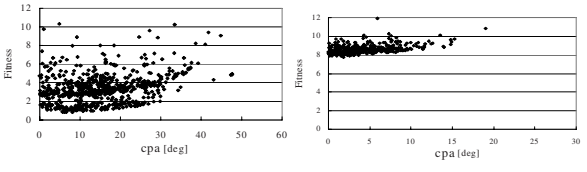


Fig. 9. Correlation between actual cross point angle and actual fitness

This result shows that the smaller the actual cross point angle, the better the fitness. This tendency was more marked in Case B. This was considered to be due to the following. In Case A, the robot can immediately reach the target distance *D*, because the robot velocity is high compared to the object velocity. Once the target distance has been obtained, the Type I robot maintains this distance by wiggling around the actual cross point and the Type II robot maintains it by rotating in place. Thus, the task is not difficult, and the robot does not need to develop a strategy. This is why the individuals with various strategies survived and the distribution of the actual cross point became wide in Case A.

In Case B, the Type III robot allowed the fastest approach by creating an area of MOVE_F in the direction of travel, as shown in State 1 of Fig.7.This type of robot can quickly cope with the random behavior of the object by symmetrically assigning actions based on the actual cross point. We found that, from the viewpoint of physical structure, although the arrangement of the two visual sensors was not symmetric (Fig.7), the state-action space was symmetric about the traveling direction (Fig.7). This result implies that symmetric properties exist in structure and in functionality, thus there are functional symmetry properties about the traveling direction. We found the functional properties emerging from the arrangement of the two visual sensors were one of the important design principles in Case B.

5 Conclusion

We focused on symmetry properties and executed the computational simulations by genetic programming using an evolutionary individual robotic agent with reconfigurable morphology and intelligence. Two kinds of experiment, Case A and Case B, were run. To clarify the mechanism of emergent symmetry properties, we compared and analyzed (1) the characteristics of morphology obtained in each case and (2) the function emerging between morphology and intelligence in each case.

From the viewpoint of physical structure, we could not identify distinguishing structural symmetry properties represented by the arrangement of two visual sensors, since the evolutionary agent robotic arranged them at various positions depending on the simulation. However, there was a trend toward creating structural symmetry properties through evolution, because both the difference of the range of the left and right visual sensors' fields of view in Case B and that of the visual axis in Case A and Case B correlated with the actual fitness.

Furthermore, we defined the angle between the traveling direction and the cross point of the sensor recognition areas as the actual cross point angle. 37% of the robot agent with this angle in Case A and 99% in Case B had an angle within 10 deg. So, functional properties remarkably emerged in Case B where the robot velocity was slow compared to the object's. Additionally, the correlation coefficients between the angle and the actual fitness were 0.31 in Case A and 0.43 in Case B, showing that robotic systems with higher functional symmetry properties got a high evaluation.

For further works, we intend to investigate the relationship between the sensory system and the driving system using an evolutionary robot agent which is able to change the sensor and the motor arrangement, so that find out more about design principles leading to structural symmetry.

Acknowledgments. This research has been studied in Kikuchi Lab. of Chiba Institute of Technology of Japan, and supported by JSPS Grant-in-Aid for Young Scientist (B).

References

1. Wooldridge, M.: Multiagent Systems: A Modern Approach to Distributed Artificial Intelligence, Gerhard Weiss. The MIT press, Massachusetts (1999)
2. Lin, C.H., Song, K.T.: Flexible Real-Time Control of Home Robots Using a Multi-Agent Based Approach. In: Proc. IEEE/RSJ Int. Conf. on Intelligent Robots and Systems, Sendai, Japan, pp. 3092–3097 (2004)
3. Sen, S.: Evolution and Learning in Multiagent Systems. Int. J. Human Computer Studies 48, 1–7 (1998)
4. Mataric, M.J.: New directions: Robotics: Coordination and learning in multirobot systems. IEEE Intelligent Systems 13(2), 6–8 (1998)
5. Maeda, Y.: Behavior Learning and Group Evolution for Autonomous Multi-Agent Robot. In: Proceedings of the IEEE International Conference on FUZZ (1997)
6. Jung, D., Zelinsky, A.: An architecture for distributed cooperative planning in a behaviour-based multi-robot system. Journal of Robots and Autonomous Systems 26(2-3), 149–174 (1999)

7. Earon, E.J.P., Barfoot, T.D.: Development of a Multi-agent Robotic System with Application to Space Exploration. In: International Conference on Advanced Intelligent Mechatronics Proceeding (2001)
8. Earon, E.J.P., Barfoot, T.D.: Development of a Multiagent Robotic System with Application to Space Exploration. In: International Conference on Advanced Intelligent Mechatronics Proceeding (2001)
9. Gerkey, B.P., Mataric, M.J.: Push-watcher: An approach to Fault-Tolerant Tightly-Coupled Robot Coordination. In: Proceedings of the IEEE Int. Conference on Robotics and Automation, IEEE Computer Society Press, Los Alamitos (2002)
10. Gao, Y., et al.: A New Multi-Robot Self-Determination Cooperation Method Based on Immune Agent Network. In: Proceedings of the 2005 IEEE Int. Conf. on Robotics and Automation, Barcelona, Spain, April 2005, pp. 390–395 (2005)
11. Vidal, R., Shakernia, O., Kim, J., Shim, D., Sastry, S.: Probabilistic Pursuit-Evasion Games with Un-manned Ground and Aerial Vehicles: Theory, Implementation and Experimental Evaluation. IEEETransactions on Robotics and Automation 18(5), 662–669 (2002)
12. Arai, T., Pagello, E., Parker, L.E.: Editorial: Advances in Multi-Robot Systems. IEEE Transactions on Robotics and Automation 18(5), 655–661 (2002)
13. Bongard, J.C., Paul, C.: Investigating Morphological Symmetry and Locomotive Efficiency using Virtual Embodied Evolution. From animals to animates 6, 420–429 (2000)
14. Bongard, J.C., Pfeifer, R.: A Method for Isolating Morphological Effects on Evolved Behavior. In: 7th International Conference on the Simulation of Adaptive Behavior, pp. 305–311 (2002)
15. Pascual, A., Huang, K., Neveu, J., Preat, T.: Brain asymmetry and long-term memory. Nature, 427 (2004)
16. Knudsen, E.I.: Instucted learning in the auditory localization pathway of the barn owl. Nature 417, 322–328 (2002)
17. Koza, J.R.: Genetic Programming: On the Programming of Computers by Means of Natural Selection. MIT Press, Cambridge (1992)
18. Koza, J.R.: Genetic Programming II: Automatic Discovery of Reusable Problems. MIT Press, Cambridge (1994)
19. Soule, T.: Exons and Code Growth in Genetic Programming. In: Proc. 5th European Conference on Genetic Programming, EuroGP-2002, pp. 142–151. Springer, Heidelberg (2002)

Design of the Intelligence User Identification Method for the Improved Password Input Method and the Personality Security*

Eun-Ser Lee[1], Haeng-Kon Kim[2], and Sang Ho Lee[3]

[1] Information & Media Technology Institute, Soongsil University
eslee1@ssu.ac.kr
[2] Catholic University of Daegu, 330, Hayangup, Kyungsan, Kyungbuk, Korea
hangkon@cu.ac.kr
[3] College of Information Science School of Computing, Soongsil University
shlee@comp.ssu.ac.kr

Abstract. When executing the various application programs of a computer generally or connecting with a specific site through the Internet A place with many examples into which carry out a user and the user-identification child (ID) and the password are made to enter in order to check a user's identity, If in agreement as compared with the identifier which a user inputs, ID which had the password registered, and a password, the authority to perform an applicable program, and the authority to use the various contents of an applicable site will be granted to the user. This research relates to a user identification by an indirect password, which allows for a password control system to identify user by a received value, which is calculated with a predetermined function known to both user and password control system, and a value of a variable, which is provided form a password control system. According to the research, password cannot be embezzled in the input stage and hacking by intercept on the transmission line is to be in vain.

1 Introduction

When executing the various application programs of a computer generally or connecting with a specific site through the Internet A place with many examples into which carry out a user and the user-identification child (ID) and the password are made to enter in order to check a user's identity, If in agreement as compared with the identifier which a user inputs, ID which had the password registered, and a password, the authority to perform an applicable program, and the authority to use the various contents of an applicable site will be granted to the user[1]. However, it sets to a bank, a card issuer, etc. who the correlation of the aforementioned Prior art of a password is fixed with a user-identification child, and manage many identifiers. The identifier which is in agreement with an applicable password regardless of the count limit of an

* In this paper of data was provided by PassHUB CO.,LTD. This work was supported by Seoul R&BD Program(10581cooperateOrg93112).

B. Apolloni et al. (Eds.): KES 2007/ WIRN 2007, Part II, LNAI 4693, pp. 581–588, 2007.
© Springer-Verlag Berlin Heidelberg 2007

input error if many identifiers to a fixed password are changed and it continues inputting even if it restricts the password input error to a fixed identifier can be discovered. ** -- Since it has the regulation to which anyone can also learn how to enter an identifier and a password, it tries to steal, without being noticed by the user by a user's side, or a camera is installed that it seems that it is easy, Even if it is not him, appearance of the password to an identifier is investigated and carried out. if the password which looks at that a user enters an identifier and a password directly, checks the program which records the contents to input, carries out usurpation and tapping on a communication path, or is kept is stolen -- a user -- There was a problem that a user's authority could be plagiarized now[2].

This research is thought out in order to solve the trouble of the aforementioned conventional technique. The purpose prevents from knowing a password only by a user supervising only entering a password, as a user can express the right to an identifier from the phase of deciding the password input approach oneself and inputting. How to input the result calculated through the count method with which the user is registered in case using a password again will enter a password so that difficultly if it is not him, It is characterized by to enable it to input this paper together with the channel number which classified the exercise approach which inputs only a value and is going to use it for the contents to input to an identifier as a result of a user's coming out by the password input approach which he learns, in order to attain the aforementioned purpose, to constitute the password control system characterized by making it input within the limits of the response time, and to check a user[11][12][13].

The purpose and the advantage of above this paper become still clearer by description of the suitable example of the detailed research with reference to an accompanying drawing.

2 Related Work

2.1 Semantic Web

Semantic Web that Tim Berners-Lee defines is as following.

"The Semantic Web is an extension of the current web in which information is given well-defined meaning, better enabling computers and people to work in cooperation"[5][6]. Ultimate purpose of Semantic Web develop the standard and technology that help computer can understand better information in web, and support semantic search, data integration, navigation and automation of task etc. if describes in detail, it enables following work

- When search information, it bring more correct information.
- Integrates and compares information of different all source.
- Correlates meaningful and descriptive information about certain resources.
- attaches detail information on the web for automation of web service.

Most low floor is consisted of URI(Uniform Resource Identifier) and unicode that is addressing method to access the resource in web protocol. Next floor is XML (eXtensible Markup Language) and namespace that can define concept as modularity. RDF (Resource Description Framework) and RDF schema to describe resource are

located in next floor. Ontology exists in next floor, and technology element for lawyer, logic, proof etc. is located in the above hierarchy.[7][8][9]

2.2 MagPie

MagPie is plug-in program of Internet Explorer. It is a Semantic Web browsing system that shows semantic information using ontology that can speak as most important element of Semantic Web. Basically, It is Active X control program and when it is consented with word that is suitable in domain and word that exist in lexicon, show semantic information of word that user wants[3][4][10].

3 Main Theory

3.1 Brain Key

This research by inputting the value calculated by the variable value offered from the function and password control system which were fixed in advance between the user and the password control system It is a thing about the user symptom which led the indirect password input which enables it to receive a user check. The password control system which controls a password is constituted so that it cannot embezzle, if the password entered even if it supervised the password which a user enters in more detail is led, and it is related with having enabled it to check a user through this.

When executing the various application programs of a computer generally or connecting with a specific site through the Internet A place with many examples into which carry out a user and the user-identification child (ID) and the password are made to enter in order to check a user's identity, If in agreement as compared with the identifier which a user inputs, ID which had the password registered, and a password, the authority to perform an applicable program, and the authority to use the various contents of an applicable site will be granted to the user.

However, it sets to a bank, a card issuer, etc. who the correlation of the aforementioned Prior art of a password is fixed with a user-identification child, and manage many identifiers. The identifier which is in agreement with an applicable password regardless of the count limit of an input error if many identifiers to a fixed password are changed and it continues inputting even if it restricts the password input error to a fixed identifier can be discovered. ** -- Since it has the regulation to which anyone can also learn how to enter an identifier and a password, it tries to steal, without being noticed by the user by a user's side, or a camera is installed that it seems that it is easy, Even if it is not him, appearance of the password to an identifier is investigated and carried out. if the password which looks at that a user enters an identifier and a password directly, checks the program which records the contents to input, carries out usurpation and tapping on a communication path, or is kept is stolen -- a user -- There was a problem that a user's authority could be plagiarized now.

3.2 Environment of the User Identification Method

It is as follows when the definition of the important vocabulary is arranged in describing this invention first.

* A thing besides identifier:, and the thing which has ******. That is, it digitizes so that it may be distinguished from other things by a resident registration certificate, a card, a passbook, USER ID, a social position certificate, a vehicle, digital seal, a document, voice, the iris, the fingerprint, etc., and in the case of a card, it becomes a card number, and, in the case of a fingerprint, becomes fingerprint data and the fingerprint management number which were generated through the fingerprint recognition machine.
* User : those who have a just use right to an identifier
* Channel : that to which the user classified the right execution approach of an identifier into some
* Reference sign : the variable which a system determines and is sent to a user
* Reference variable : it can know without communicating as data which are the variable which the user decided and a system and a user can share, and is a part etc. at a year, the moon, a day, and the time.
* The password input approach : how to determine and input the count approach, analysis, and sequence which were made by the password, the reference sign, the reference variable, the channel, etc. when a user inputted (expression).
* Response time : time amount which it will take before a user calculates as the input approach and inputs a result
* The indirect password input approach : how to enter [inputs the result of having not inputted the contents which are having it registered that the user knows as it was, but having processed them as the approach, to input the answer which inputted and calculated the formula, if a certain formula is registered, and] a password by pushing a specific thing on specific time amount, if a specific thing must be pushed on specific time amount

3.3 Design of the Method

The password storage area 1 where Fig.1 is an example of the general password structure of a system which carried out the introductory configuration of the concept over this paper, and a password [as opposed to an identifier (ID) in the password control system of this research] is saved, The password input approach storage area 2 where the formula which made the password to the same identifier as a basic variable, and a function are saved, The reference variable storage area 3 where the variable referred to in case a user inputs is saved, and the reference-sign management equipment 4 which manages a reference sign, The output unit 5 which displays this, and the channel storage area 6 which defined the class of exercise to an identifier, The input unit 7 into which a user inputs a result, and the reaction-time storage area 8 where the reaction time to input is saved, The central processing unit 9 which carries out the process of an applicable channel which the user demanded and carries out control management after checking whether the password registered by comparing and analyzing the result inputted as the inputted reaction time and the formula are known, It consists of an extraordinary storage area 10 required at the time of execution, an output port 11 connected with the exterior, and a communication wire 12 which exchanges the signal of these mutual.

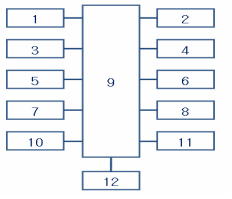


Fig. 1. Structure of a password control system

In case the phase 21 and user who input the identifier for which a user is going to use the password control system of constituted this research like the flow chart of the password control system of Fig.2 input the user-validation approach which led, thus, a password, a reference sign, a reference variable, By the approach of determining and expressing a constant, the count approach made by the channel, analysis, and sequence It judges whether the phase 23 re-calculated according to the approach of classifying an effective thing and an effective channel for the result inputted by the phase 22 of inputting the result which must be inputted, and the user, and calculating the response time by which it was fixed, and the inputted response time are within the response time which the user set. Do in the phase 25 which carries out the process of a channel which regarded it as what checked the user to an identifier when the phase 24 which analyzes with the result which the system calculated and is compared and the result suited, and was demanded together. The phase 26 which carries out the process processed when the result which the user inputted is not in agreement with the result which the system calculated and it is not a user is included.

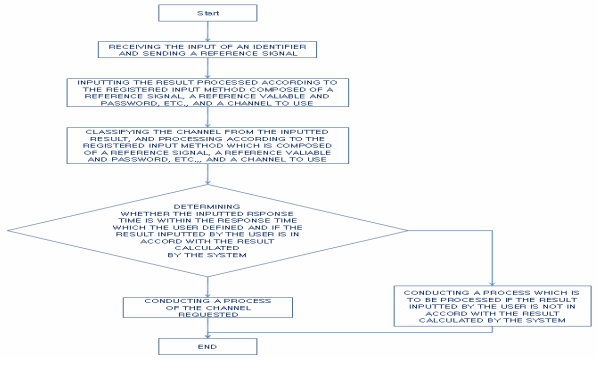


Fig. 2. Method for identification of a user

The user symptom which led the indirect password input of such this invention is as follows when the example which it may be applied to various fields using the approach of checking the user knowing the user validation to an identifier and authentication of industry, in object for **, and was applied to phone banking (Phone Banking) in it is explained in detail. Since the configuration of a phone banking system can be regarded as having added the automatic-answering system (ARS) with the configuration of the password control system of drawing 1 for the user interface, it will be

explained with reference to drawing 1 as it is. In phone banking, an automatic-answering system (ARS) changes according to the class of terminal connected with a password control system in the interface part for communicating with a user and voice with the output unit 5 and input unit 7 of a password control system. Since it is unrelated to the summary of this invention, a detailed technique and a detailed notation decide to omit. The user who is going to receive phone banking service pushes the account number used after connecting with an automatic-answering system (ARS) by telephone. (Phase of receiving the input of an identifier by Fig.2).

Carrying out extraordinary preservation of the time of day which sorted out the figure of arbitration for the reference sign in the reference-sign management equipment 4 of a password control system, and sent one or the figure beyond it to the automatic-answering system (ARS) with the contents of a reference sign one by one in the extraordinary storage area 10 of delivery and a password control system at the target, an automatic-answering system (ARS) tells a reference sign with voice. (Phase of sending a reference sign by Fig.2).

4 Algorithm and Operation of the Method

The password already registered on the other hand is 1234, the password input approach registered is for example, a password +300-reference number *10+ channel number, and if it assumes that the dealings, i.e., less than 100,000won dealings, for which it asks were registered as a channel number 3, 15293 which is a count result value will be inputted. (Phase of receiving the input as a result of Fig.2).

It compares with the contents which the user inputted, after calculating the contents as which the identifier to which the channel number which has grasped the part into which a channel is inputted and was inputted at the end in the personal identification number input approach storage area 2 is saved in the extraordinary storage area 10, a personal identification number, the reference sign sent to the user by the extraordinary storage area 10, and a channel number are substituted for a password control system in the password storage area 1, and the input approach storage area 2 corresponds must be inputted. (Phase which compares whether the result which the password control system calculated is the same as the result which the user inputted).

If it is response within the limits for which calculated the response time if a user input value is compared with the the very thing calculated value of a password control system and it is in agreement in the above, until a user will receive and input a reference sign, and the user opted, the applicable process of an applicable channel will be carried out so that use authority can be used in the authority of the applicable channel which saved in the response-time storage area 8, and was demanded. (Applicable channel execution phase of Fig.2).

It is as follows when the part which was not described to the meaning, the purpose, and means of this invention below, or ran short is again explained to a detail.

How to set up the password input approach first is explained.

If the registered password serves as a decimal number and has become with the N girder, Password P

P=P0x100+P1x101+ -- +PN-1x10N-1+PNx10N Channel number Ch, Ch= {Ch1, Ch2, --} If the reference sign serves as a decimal number N girder, Reference sign S

S=S0x100+S1x101+ -- +SN-1x10N-1+SNx10N When the reference variable V is made only into the date and time amount,

Reference variable V, V={D1, D2 --D30, D31}

The password input approach I

I=f (P0, P1 -- PN-1, PN, Ch1, Ch2, -- ChN, S0, S1, -- SN-1, SN, D1, -- D31, H1, -- H24, 1, 2, -- 9, 0)

that as which can express and a user decides the input approach to be -- it is -- for example

I=P0xS1Mod10, P1xS1Mod10, Chn, Mod(P2+S1) 10, PN-1, PNxS1Mod100

It can set up.

Thus, the result of the digit count of the shoes numerically made when you set up (Sol), Sol={10N+Ch+1}

If the number of the passwords which came out and were registered for a certain reason is four, it will become 7 figures or a digit count beyond it from 4 figures.

Although it will be inputted by 4 figures if it inputs by the superabundance of the number of specification which does not input a channel but comes out by four operations, power, etc. according to each digit.

Since the approach of helping to set up a reference variable etc. by the system so that it may be easy to use a user in this invention could make various instruments in the case of the engineer engaged in this field, but it is not the meaning of this invention, it is omitted.

The variable referred to when entering a password can set many things as the location applied according to a property, when releasing the number of visitors, it can refer to the digit count of the last of a visitor counter when he comes in in the case of a user, when it is a stock account, it can also refer to a broad based stock index, and it can also make it his number of possession stocks. And the replacement password which is used at the time of account replacement in the case of a bank can also make the specific part of the account to send a reference variable. Therefore, a reference variable can be set up anything, if a system and a user can own jointly.

5 Conclusion

The concept of this paper is what may be applied to various fields on industry. Single devices, such as a digital door lock, therefore, of course It may be utilized as a keyword of various encryption, such as the control managerial system of a side gate, a communication link, a file, a digital signature key, a smart card, and a card reader machine, and a decryption. The various user validations and electronic financial dealings which are utilized in the Internet electronic commerce, Small wireless radios, both directions TV, etc., such as a credit card enquiry machine, an ATM terminal and a cellular phone, and PDA. The place which is a thing broadly applicable to the field which exchanges a mutual signal, if this is a person with usual knowledge, it can constitute the password control system suitable for it within limits which are not missed to the summary of this invention easily. The password input approach is what can be decided to like according to liking of a user. The user who desires a password change periodically in addition to the above approaches can also use the date, time amount, etc. as a variable. The formula made by the reference sign, the reference

variable, a personal identification number, etc. can also be used. The approach of inserting a figure only with the figure of a reference sign unrelated to a password in the middle of an input can also be chosen, and the approach of inputting the location of the channel which it is going to use in password order or the middle of a password can also be chosen. Therefore, since it becomes impossible to be sure of the success possibility of hacking even if it submits to risk and spends much time amount and many costs, in order to embezzle, it is effective in losing the attempt to hacking itself.

References

[1] Pressman, R.S.: Software Engineering international edition. Mcgraw-hill, New York (1997)
[2] Luger, G.F.: Artificial intelligence. Addison wesley, Reading (2001)
[3] Dzbor, M., Domingue, J., Motta, E.: Magpie: Towards a Semantic Web Browser. In: Fensel, D., Sycara, K.P., Mylopoulos, J. (eds.) ISWC 2003. LNCS, vol. 2870, Springer, Heidelberg (2003)
[4] Domingue, J., Dzbor, M., Motta, E.: Magpie: Supporting Browsing and Navigation on the Semantic Web. In: Proc. of Intl. Conf. on Intelligent User Interfaces (IUI), Portugal (2004)
[5] The Report of the President's Commission on Critical Infrastructure Protection CCEB (Common Criteria Editorial Board): Common Criteria for Information Technology Security Evaluation, Version 2.0 (May 1998)
[6] Lee, E.-S., Hwang, S.-M.: Definition of Security Requirement Items and Its Process to Security and Progress Management. LNCIS, vol. 344. Springer, Heidelberg (2006)
[7] Lee, E.-S., Hwang, S.-M.: Design Implementation of Web Security Access Control System for Semantic Web Ontoloty. In: Gervasi, O., Gavrilova, M., Kumar, V., Laganà, A., Lee, H.P., Mun, Y., Taniar, D., Tan, C.J.K. (eds.) Computational Science and Its Applications – ICCSA 2005. LNCS, vol. 3481, Springer, Heidelberg (2005)
[8] Lee, E.-S., Lee, M.: Development System Security Process of ISO/IEC TR 15504 and Security Considerations for Software Process Improvement, August 2006. LNCIS, vol. 344. Springer, Heidelberg (2006)
[9] Lee, E.-S., Lee, S.-H.: Design progress management for Security Requirements in Ubiqiiuous computing using COQUALMO. In: Gavrilova, M., Gervasi, O., Kumar, V., Tan, C.J.K., Taniar, D., Laganà, A., Mun, Y., Choo, H. (eds.) ICCSA 2006. LNCS, vol. 3984, Springer, Heidelberg (2006)
[10] Nam-deok, Lee, C.E.-S.: Design and Implementation of Semantic Web Search System using Ontology and Anchor Text. In: Gavrilova, M., Gervasi, O., Kumar, V., Tan, C.J.K., Taniar, D., Laganà, A., Mun, Y., Choo, H. (eds.) ICCSA 2006. LNCS, vol. 3984, Springer, Heidelberg (2006)
[11] ISO/IEC 15408-1:1999 Information technology - Security techniques - Evaluation criteria for IT security - Part 1: Introduction and general model
[12] ISO. ISO/IEC 15408-2:1999 Information technology - Security techniques - Evaluation criteria for IT security - Part 2: Security functional requirements
[13] ISO. ISO/IEC 15408-3:1999 Information technology - Security techniques - Evaluation criteria for IT security - Part 3: Security assurance requirements

Using Tuple Space to Coordinate Multiagent Activities

Wanli Ma, Dat Tran, and Dharmendra Sharma

School of Information Sciences and Engineering
University of Canberra, Australia
{Wanli.Ma,Dat.Tran,Dharmendra.Sharma}@canberra.edu.au

Abstract. Programming languages play vital roles in program development and implementation. On the one hand, a programming language provides a means of algorithm presentation; on the other hand, it reflects the underlying abstract computational model. In this paper, we discuss the extension of the Chemical Abstract Machine with multiple coexisting agents – MACH (Multiagent Extended Chemical Abstract Machine). The aim of MACH is to develop a coordination programming language based on an interactive computational model. The language is simple yet powerful enough to construct complicated multiagent, based software systems. We present MACH as a simple and efficient programming approach with an underlying formal proof system.

Keywords: Multiagent Systems, MAS, Agent Based Software Engineering, Coordination.

1 Introduction

Multiagent systems (MAS) stem from the fields of distributed computing and artificial intelligence [1]. The agent technology originated from artificial intelligence research and can be traced back to the Actor model by Hewitt [2] of 1970'. According to Wooldridge and Jennings [3], if a piece of software has the characteristics of autonomy, social ability, reactivity, and pro-activity, it is an agent in weak notion. On top of these characteristics, if the software has further characteristics of adaptability, mobility, veracity, and rationality, it becomes an agent in strong notion. Nwana suggested that "agent" is an umbrella term, under which many different types of agents exist [4].

Using agents as the basic building blocks to construct complicated software systems was first suggested by Shoham in 1993 [5]. Agent based software development starts with system analysis, which results in agent organizational structure, i.e., agent roles and interactions. Based on the structure, agent types are designed, and then the agents are implemented in a programming language (or languages). From BDI model [6] point of view, the behavior of an agent is driven by its beliefs, desires, and intention. According to Braubach et al [7], in programming term, the beliefs reflect the environment the agent reside, the desires are the goals for the agent, and the intentions are the operational or executable plans of the agents.

A multiagent system can be described at two different levels: the interactions among agents and the internal operations of the agents to achieve their goals.

B. Apolloni et al. (Eds.): KES 2007/ WIRN 2007, Part II, LNAI 4693, pp. 589–596, 2007.
© Springer-Verlag Berlin Heidelberg 2007

Communications among the agents are essential to any multiagent system. KQML [8] and FIPA-ACL [9] are 2 examples of agent communication languages (ACL). Both use asynchronous messages, based on speech act theory [10], for agents to communicate.

Actually, the study of interactions and communications among concurrent activities is not new to MAS. It has been intensively studied in operating systems, database systems, and particularly parallel programming [11]. Milner realized the problem in his CCS [12], where "interaction or communication is the central idea". Ciancarini [13] has the same observation by distinguishing a closed program from an open one. Wegner [14] also strongly advocates this idea by arguing that interaction is more powerful than computing. Orthogonal coordination [15] is the basic idea promoted by Linda [16-18] to coordinate the interactions of the activities which are performed in parallel. Linda uses a global shared tuple space to coordinate these activities. A tuple space is a logically shared memory used for data exchange and synchronization control among the interactive components of a program.

The tuple space can also be used as the (logically) global environment, or global space, for agents to reside and interact, for example, [19-21]. It provides the social environment for the agents. It is also the communication media for the agents to interact with each other.

For any programming language, the three most important aspects are programmmability, execution efficiency, and provability. We propose a new programming paradigm called MACH (Multiagent Extended Chemical Abstract Machine). MACH extends the Chemical Abstract Machine (Cham) [22, 23] with multiple agents. In MACH programming, the pre-condition captures part of the environment. When the condition meets, the agent starts to react. The post-condition describes the goals of the agent. And the intentions of the agent are programmed as the internal behavior of the agent. Cham (Chemical Abstract Machine) is a theoretical refinement, with rigid mathematical definition, of GAMMA [24-26] model. The rigid mathematical definition gives MACH an opportunity to develop an underlying formal proof system.

The rest of the paper is organized as follows: Section 2 briefly explains the syntax of MACH, and then Section 3 illustrates MACH programming by examples. In Section 4, we very briefly introduce the temporal logic verification system for MACH. Section 5 concludes the paper with our future work.

2 MACH Notations

Every MACH program has 3 major sections. A manifest section defines the global names used by the program. The names are just meaningful symbols for human beings to easily remember. This section is just a list of symbolic names and their real values labeled by the keyword `#define`.

A reaction section (`REACTIONS`) describes the chemical reactions. It has several subsections.

- The `tuples` subsection declares all possible tuple types (molecules) which may appear in the tuple space. The declaration only specifies the types of possible tuples. How many of the declared tuples, when, and where they enter

into the tuple space depend on a particular execution and cannot be predicted in advance.

- The `initialization` subsection sets up the initial state of a tuple space. This is where the chemical reaction starts.
- The main subsection is the reaction rule subsection (`reactionrules`), which is a collection of reaction rules. The left hand side of a reaction rule lists the tuples required for this reaction to happen, and the right hand side of the reaction rule gives the impact of this reaction upon being finished, for example,

  ```
  lhs1, lhs2, lhs3 leadsto rhs1, rhs2 by AgentX
  ```

  which means that if there are tuples `lhs1, lhs2,` and `lhs3` in the tuple space, agent `AgentX` will start to operate. As the result of `AgentX` activities, tuples `rhs1` and `rhs2` will be injected back to the tuple space.
- The `termination` subsection is optional. If it exists, it specifies the grand system goal, which the multiple agents are collectively trying to achieve. Upon reaching this condition, all reactions are terminated. In other words, all agents together with their individual goals are aborted.
- The agent goal (`agentgoals`) subsection has a collection of pre- and post-condition pairs for each agent described in the reaction rule subsection. A pre- and post-condition pair specifies that under the context of pre-conditions, an agent will be triggered to perform and achieve the goal described in the post-conditions.

Finally, an AGENT section lists the program code of all agents. These agents do not have to be written in the same program language; however, they have to comply with the predefined interaction standard.

3 Programming with Multiagent and Tuple Space

The Auction Problem: There is 1 item to be auctioned. To simplify the problem for explanation purpose, we assume that there are 1 auctioneer and 3 bidders. The auctioneer is known as A, and the 3 bidders are P, Q, and R. Each bidder has a price range for the item. The goal for a bidder is to acquire the item with minimum possible price within its price range, and the goal for the auctioneer is to achieve the highest possible price for the item. The program is in Figure 1.

The current bidding price of the item is represented by tuple price in the tuple space. It is initialized to the auctioneer's minimum reserve price. Token ticket has to be obtained for an agent to bid. Supposing agent P obtained the token ticket and also the current price, it will check who updated the price last time and also if the current price is still within its acceptable price range. If the price was not updated by itself and is still lower than the maximum acceptable price, it will increase the price and then inject the updated tuple price back to the tuple space, together with a new tuple update. The tuple update has the information on P's action – bid or not. The auctioneer agent A checks the tuple update. If the last bidder agent did increase the price (update.Updated==TURE), agent A will turns the tuple into a ticket tuple and

also preserves the name of the last biding agent. The bidding process continues. If the update tuple indicated no action from last bidding agent, it remains unchanged in the tuple space. If the tuple space accumulates 3 update tuples, it means that none of them can be turned into a ticket tuple. In the other words, all 3 bidding agents have checked the price, and decided no to bid any more. Thus, the bidding process terminates. The winning bidder is price.Bidder and the final price for the item is price.Price.

```
REACTIONS Auction
  tuples
     token price, ticket, update;
  initialization        price.Price = TheReservedPrice;
  reactionrules
    ticket, price leadsto price, update by P;
    ticket, price leadsto price, update by Q;
    ticket, price leadsto price, update by R;
    update leadsto ticket by A when (update.Updated == TURE);
  termination     (|update|==3) do output_price;
end

AGENT P
  P() {
     if (price.Price < P.MaxPrice and ticket.UpdatedBy != P) {
         price.Price += P.Inc; price.Bidder = P;
         update.Updated=TRUE; update.By = P;}
       else { update.Updated=Flase; update.By = P; }
end
/* AGENT Q and AGENT R are similar to AGENT P */
AGENT A
  A() { ticket.UpdatedBy = update.By;}
end
```

Fig. 1. MACH program of the Auction problem

The Predators and Prey Problem: There are 4 predators chasing 1 prey on a grid board, which is a simplified 2 dimensional toriodal world for the predators and the prey. The locations of the predators and the prey are the x and y coordinations (x,y) of the grid. The prey moves randomly, and the goal of the predators are to capture the prey. There are many variations of the problem depending on the different conditions on the predators and the prey. The basic program is in Figure 3.

```
REACTIONS Pursuit
  tuples
    token pt[x1,y1],pt[x2,y2],pt[x3,y3],pt[x4,y4],pr[x,y];
    initialization      random pt, pr;
    reactionrules
     pr leadsto pr by Prey;
     pt leadsto pt by Predator;
     pt, pr leadsto pt, pr by Predator;
    termination   ForAny pt[x,y] and pr[u,v]: x=u and y=v;
end

AGENT Prey
    Prey() { pr = random(pr) }
end

AGENT Predator
  Predator() { pt = CalculatedMove(pt, pr); pr = pr;}
end
```

Fig. 2. MACH program of Predators and Prey problem

This program has a termination condition. There are several ways of defining the capture state, i.e., the termination condition. In Figure 2, the capture state is that a predator occupies the same grid as the prey. Termination can also be defined as the 4 predators immediately surrounding the prey. If this is the case, the termination condition is:

```
pt[x1,y1],  pt[x2,y2],  pt[x3,y3],  pt[x4,y4],  and
pr[u,v]:  x1=u,y1=v+1  and  x2=u,y2=v-1  and  x3=u+1,y3=v
and x4=u-1,y4=v;
```

where pt[x1,y1] and so on are just a place holders and could be any of the pt tuples (predators), and also the + and – operations are on the toriodal world, i.e., moving back to the other sides when at the edges.

The 2 reaction rules for `Predator` agents in the program introduce non-determinism. The first rule, which does not have tuple `pr` at the left hand side, describes the situation when the predator cannot see the prey. It may just move randomly. The second reaction rules describes the situation when the predator sees the prey, it adjusts its movement towards the prey based on its internal algorithm. If we want the `Predator` agents to be aware of the existence of each other, the second reaction can be rewritten as:

```
pt, pt, pt, pt, pr leadsto pt, pt, pt, pt, pr by
Predator;
```

The tuples at the left hand side of the reaction rule are just place holders. They are not locked with the agents. Each time, a `Predator` agent will be associated with a `pt` tuple and becomes the predator at that location. In this case, an agent actually does not have internal memory about the locations of the predators. If we want to always associate an agent with a tuple, we can use 4 different tuple names, say `pt1`, `pt2`, `pt3`, and `pt4`.

4 Temporal Logic Proof System for MACH

MACH has an underlying temporal logic proof system. Although mathematics and logic are the better way to achieve a correct program, most programmers are not so comfortable with the rigid process of mathematical reasoning. People tend to use a natural and intuitive way to express their ideas. For example, people prefer the Venn Diagrams to the mathematical definitions of the set operations. Formal temporal logic semantics provides a means of correctness proof for MACH programs, but the proof system is separated from programming and kept in the background.

This section very briefly explains the idea of the underlying temporal logic proof for MACH programs.

The logic proof system for a program consists of three parts: the uninterpreted logic part, the domain part, and the program part [27]. The uninterpreted logic part is a general temporal logic system [28, 29]. The domain part restricts the proof system to the related domains, for example, the axioms and theorems about integers, strings, trees, etc. The program part of the proof system further restricts the proof system to the acceptable computation sequences of this program.

To represent the program part of the proof system in temporal logic notations, we recast MACH programs to a 4-tuple: *P=(T, S, R, I)*, where

- *T* is the set of all tuples (data) possibly appearing in the tuple space. It is specified in the `tuples` section of a MACH program.
- *S* is the set of all possible tuple space states. Its purpose is twofold. On the one hand, an element of *S* can assign values to tuples or the variables contained in the tuples. In other words, it maps variables to their domains. On the other hand, the element designates the tuples currently in a tuple space. For any tuple $t \in T$, the characteristic function $C(t)$ = `TRUE` if *t* is currently present in the tuple space; otherwise $C(t)$ = `FALSE`. For simplicity, we write just *t* itself for $C(t)$ whenever there is no ambiguity.
- *R* ($R \subseteq S * S$) is the set of reaction rules. The elements of *R* come from the `reactionrules` section of an agent.
- *I* ($I \in S$): the initial state of the tuple space, which is specified by the initialization section of a MACH program, for example, *I* = {`token`, `token`, `token`}.

A MACH deductive system consists of a set of axioms and a set of inference rules. Axioms come from the general axioms of the uninterpreted temporal logic, domain axioms (including fairness properties etc.), and the axioms from a given MACH program. For any reaction rule, say,

$x_1, x_2 \ldots x_n$ `leadsto` $y_1, y_2, \ldots y_m$ `by` *Agent_T* `when` $f(x_1, x_2, \ldots x_n)$

supposing the pre-condition and post-condition of *Agent_T* are *p* and *q*, i.e., *{p} Agent_T {q}* and there are no repetitive elements among x_1, x_2, …, and x_n, we have

$$(\bigwedge_{i=1}^{n} |x_i| \geq 1) \wedge f(x_1, x_2, \ldots x_n) \wedge p \Rightarrow \Diamond [(\bigwedge_{i=1}^{n} |x_i|^{-1}) \wedge (\bigwedge_{j=1}^{m} |y_i|^{+1}) \wedge q]$$

where $|x|$ denotes the number of *x* tuples currently in the tuple space, and $|x|^{-t}$ means the number of tuple *x* has decreased by *t* in comparison to the former state (i.e., the state corresponding to the left hand side of the reaction rules), while $|y|^{-t}$ means that the population of tuple *y* is increased by *t*.

5 Conclusion

MACH (Multiagent Extended Chemical Abstract Machine) is based on the Chemical Abstract Machine model with the extension of multiple agents. MACH keeps the theoretical part of the language in background. It can be ignored, but it is there and ready to use. We expect that MACH can lead us to a new way of separating (i) the logic of a program from its implementation, (ii) correctness from efficiency, and (iii) the rigid formal reasoning aspect from a comfortable intuitive presentation, without heavy penalties on execution efficiency.

Acknowledgments. The research work is supported by The Completion Grant of Division of Business, Law and Information Sciences, University of Canberra and the Multidisciplinary Research Grant of University of Canberra.

References

1. Stone, P., Veloso, M.M.: Multiagent Systems: A Survey from a Machine Learning Perspective. Autonomous Robots 8(3), 345–383 (2000)
2. Hewitt, C.: Viewing Control Structures as Patterns of Passing Messages. Artificial Intelligence 8(3), 323–364 (1977)
3. Wooldridge, M., Jennings, N.: Intelligent Agents: Theory and Practice. The Knowledge Engineering Review 10(2), 115–152 (1995)
4. Nwana, H.S.: Software Agents: An Overview. Knowledge Engineering Review 11(3), 1–40 (1996)
5. Shoham, Y.: Agent-oriented programming. Artificial Intelligence 60(1), 51–92 (1993)
6. Rao, A.S., Georgeff, M.P.: BDI-agents: from theory to practice. In: Proceedings of the First Intnational Conference on Multiagent Systems (ICMAS-95), The MIT Press, San Francisco (1995)
7. Braubach, L., Pokahr, A., Lamersdorf, W.: Jadex: A BDI Agent System Combining Middleware and Reasoning. In: Klusch, M., Unland, R., Calisti, M. (eds.) Software Agent-Based Applications, Platforms and Development Kits (2005)

8. Finin, T., Fritzson, R., et al.: KQML as an Agent Communication Language. In: The 3rd International Conference on Information and Knowledge Management (CIKM'94), ACM Press, Gaithersburg, MD (1994)
9. FIPA: FIPA Specification [cited 2007 February 15, 2007] (2007), Available from http://www.fipa.org/
10. Searle, J.: Speech Acts: An Essay in the Philosophy of Language. Cambridge University Press, Cambridge (1969)
11. Hwang, K.: Advanced Computer Architecture: Parallelism, Scalability, Programmability. McGraw-Hill, Inc., New York (1993)
12. Milner, R.: Communication and Concurrency. Prentice-Hall, Inc., Englewood Cliffs (1989)
13. Ciancarini, P.: Distributed Programming with Logic Tuple Spaces. New Generation Computing 12, 251–283 (1994)
14. Wegner, P.: Why Interaction is More Powerful than Algorithm. Comm. ACM 40(5), 89–91 (1997)
15. Gelernter, D., Carriero, N.: Coordination Languages and Their Significance. Comm. ACM 35, 96–107 (1992)
16. Ahuja, S., Carriero, N., Gelernter, D.: Linda and Friends. IEEE Computer, pp. 26–34 (1986)
17. Carriero, N., Gelernter, D.: How to write Parallel Programs: {A} guide to the Perplexed. ACM Trans. on Programming Languages and Systems 21, 323–357 (1989)
18. Carriero, N., Gelernter, D., et al.: The {Linda} alternative to message-passing systems. Parallel Computing 20, 632–655 (1994)
19. Ciancarini, P., Omicini, A., Zambonelli, F.: Multiagent System Engineering: The Coordination Viewpoint. In: Jennings, N.R. (ed.) ATAL 1999. LNCS, vol. 1757, Springer, Heidelberg (2000)
20. Omicini, A., Denti, E.: From tuple spaces to tuple centres. Science of Computer Programming 41(3), 277–294 (2001)
21. Bergenti, F., Ricci, A.: Three approaches to the coordination of multiagent systems. In: Proceedings of the 2002 ACM symposium on Applied computing, ACM Press, Madrid, Spain (2002)
22. Berry, G., Boudol, G.: The Chemical Abstract Machine. Theoretical Computer Science 96, 217–248 (1992)
23. Boudol, G.: Some Chemical Abstract Machines, in A Decade of Concurrency. {Reflections} and Perspectives. In: Bakker, J.W.D., Roever, W.P.D., Rozenberg, G. (eds.), pp. 92–123. Springer, Heidelberg (1993)
24. Banatre, J.-P., Metayer, D.L.: Programming by Multiset Transformation. Comm. ACM 36(1), 98–111 (1993)
25. Banatre, J.-P.: Parallel Multiset Processing: {From} Explicit Coordination to Chemical Reaction. In: Ciancarini, P., Hankin, C. (eds.) Coordination Languages and Models, {First} International Conference, {COORDINATION'96}, pp. 1–11. Springer, Heidelberg (1996)
26. Banatre, J.-P., Metayer, D.L.: The {GAMMA} model and its discipline of programming. Science of Computer Programming 15, 55–77 (1990)
27. Manna, Z., Pnueli, A.: How to Cook a Temporal Proof System for Your Pet Language. In: Manna, Z., Pnueli, A. (eds.) Proc. 10th Ann. ACM Symp. on Principles of Programming Lang., pp. 141–154. ACM Press, New York (1983)
28. Manna, Z., Pnueli, A.: The Temporal Logic of Reactive and Concurrent Systems: {Specification}. Springer, Heidelberg (1992)
29. Gotzhein, R.: Temporal Logic and Application — {A} tutorial. Computer Networks and ISDN systems 24, 203–218 (1992)

Adaptive Use of Technical Indicators for Predicting the Intra-Day Price Movements

Mieko Tanaka-Yamawaki[1] and Seiji Tokuoka[2]

[1] Department of Information and Knowledge Engineering, Tottori University,
4-101 Koyamacho-Minami, Tottori, 680-8550 Japan
[2] RICOH Software, Inc.,
I-Forefront Tower, 1-12-3 Kachidoki, Chuo-ku, Tokyo, 104-0054 Japan
mieko@ike.tottori-u.ac.jp

Abstract. We propose a system to select the best combination of technical indicators and their parameter values adaptively by learning the patterns from the tick-wise financial data. In this paper, we show that this system gives good predictions on the directions of motion with the hitting rate at 10 ticks ahead of the decision point as high as 70% for foreign exchange rates (FX) in five years from 1996 to 2000 and 8 different stock prices in NYSE market in 1993. Based on this fact, we conclude that the tick-wise price time series carry a long memory of the order of at least a few minutes, which is equivalent to 10 ticks.

Keywords: Technical Indicators, Prediction, Intra-day Price Movements, Foreign Exchange Rates, Stock Prices.

1 Introduction

Financial time series are usually assumed as the random walk in financial engineering. Indeed the auto-correlation function of the price time series quickly diminishes as a function of the time separation. On the other hand, there is enough reason to believe in the existence of long memory in the financial time series, which makes the financial forecast possible. One good reason is the existence of many professional farms whose major function being the financial forecasts. Moreover, financial forecasting is an exciting game to human in general. There are increasing number of day-traders nowadays who spend much time and effort in forecasting intra-day price movements. Many of them use the price charts as well as other information to help their decisions. Still no comprehensive enough way has been established to compile those elements of knowledge into a systematic procedure.

In this paper, we attempt to establish such a systematic procedure by identifying effective combinations of technical indicators applied on the real-world tick data. Since the tick-wise prices have strong trends and features [1,2], unlike the daily close values of stock prices known to be the random walk, we expect such predictions quite useful.

We have earlier observed a long-term stability of the tick-wise price fluctuations in terms of the direction of move of the next tick [3,4]. For example, the conditional

B. Apolloni et al. (Eds.): KES 2007/ WIRN 2007, Part II, LNAI 4693, pp. 597–603, 2007.
© Springer-Verlag Berlin Heidelberg 2007

probability of the up motion after the down motion is almost like a universal constant for several years.

$$P(\text{up}|\text{down}) = 1 - P(\text{down}|\text{down}) = 0.7 \quad (1)$$

Guided by this fact, we have attempted to predict the immediate future price by means of evolutional computations [5,6].

2 High Frequency Financial Data

We use two kinds of financial data in this paper. One is the foreign exchange rates between US Dollar and Japanese Yen, from January 1995 to December 2000, provided by CQG. At each tick, the time (year, month, day, hour, minutes), price and the type (A stands for the ask price and B stands for the bid price). Always the ask price is higher than the bid price. The bank presents those prices at the same moment in most cases, and this price presentation is named as a quote. One tick counts one quote.

We list the total volume of quotes per year and the average intervals between quotes from 1995 to 2000 together with the number of active days in Table 1. From this table, we observe the rapid increase of quotes (decrease of intervals) after 1999. The time duration of 10 ticks ranges from 100 seconds to 200 seconds, which approximately corresponds to 1-3 minutes.

Table 1. Frequencies of quotes and the average intervals for tick-wise currency exchange rate of Japanese Yen/ American Dollar, for six years from 1996 to 2000

Year	Days	Quotes/Yr	Quotes/Day	Interval[s]
1995	260	1,410,321	5,424	16
1996	261	1,198,216	4,590	19
1997	260	1,138,421	4,378	20
1998	260	1,373,849	5,284	16
1999	260	1,722,764	6,626	13
2000	260	2,232,583	8,586	10

The other time series we have examined are from the NYSE market [7]. We have selected 8 stocks from NYSE, by choosing two from each of the four different groups of business area, including retail stores, computer business, oil, automobile. We use bid-prices of those stocks for a year of 1993 (January 1 to December 31). We show the symbols, business types, total ticks in the year of 1993, and the average interval between ticks in Table 2.

Table 2. Tick-wise stock prices of 4 stocks chosen from NYSE in 1993

Symbol	Branch	# of ticks	Tick interval[s]
SMRT	retail	12525	473
BP	Oil	73562	83
IBM	computer	455233	14
GM	automobile	277241	23

3 Pattern Recognition by Technical Indicators

There are many technical indicators in the literature [8,9]. They fall into the following three categories.

(1) Price trends (rising/falling)
(2) Trend turnover (rising/falling to falling/rising)
(3) Market strength (momentum, volume, etc.)

The first group has many kinds of moving averages (abbreviated as MA), obtained by averaging the price time series over a certain period T. They indicate whether the price is in the phase of increasing trend or decreasing trend. The SLMA uses information of relative location of the SMA (MA over shorter T), the LMA (MA over longer T), and the original price time series itself. Likewise, the SLEMA uses relative location of two exponential-weighted moving averages (abbreviated as EMA) over different periods, namely, SEMA (EMA over shorter period) and LEMA (EMA over longer period).

The second group has various signals that are supposed to indicate the approach of trend turnover. Usually indicators in different categories are combined. However, there is no fundamental guideline for the best choice of indicators. The investors use their favorite combinations of those indicators. The period to take MA is also determined ad hoc based on investor's experience and sentiments.

Our purpose is to find the way of choosing the most suitable combination of the indicators and the period T to take MA at moment of concern, and establish the automatic prediction generator on the trend of the price movement at ten ticks ahead. We select 10 popular indicators:

(1) MA: moving average
(2) SLMA: relative magnitude of SMA, LMA, price, where SMA (LMA) is the MA over the shorter (longer) period
(3) SLEMA: relative magnitude of SEMA, LEMA, price, where SEMA (LEMA) is the exponentially weighted MA over the shorter (longer) period
(4) MACD: MA convergence vs. divergence defined as SEMA minus LEMA
(5) BB: Bolinger Band defined as a belt over MA plus (minus) standard deviation*k, where k is taken as an appropriate integer. We take k=3 in this work which gives the eight states, the price is larger than MA + 3 sigma, the price is between MA + 2 sigma and MA + 3 sigma, etc.
(6) MO1: Momentum of 1 step defined as the difference between the current price and the previous price
(7) MO2: Momentum of 2 steps defined as the difference between the current price and the price at 2 ticks before
(8) RSI: Relative Strength Index defined as a percentage of average rises in the last T ticks out of average times of moves in the same period
(9) RCI: Rank Correlation Index defined as a percentage of 1- 6d/(n*(n*n-1)) where d is a sum of the squares of the difference between the price rank and the time rank.
(10) PL: Psychological Line defined as a percentages of rising events in T ticks period

Table 3. Number of states assumed for the indicators

Label	Indicator Symbol	# of States
0	MO1 (Momentum of 1 step)	3
1	MO2 (Momentum of 2 steps)	9
2	MA (Moving Average)	2
3	SLMA (SMA vs. LMA)	6
4	RSI (Relative Strength Index)	6
5	MACD (MA Conv./Div.)	2
6	MAD (MA Deviation Rate)	4
7	RCI (Rank Correlation Index)	5
8	PL (Psychological Line)	6
9	SLEMA (SEMA vs. LEMA)	6

4 Prediction Generator

Our system [10] consists of the following four steps illustrated in Fig. 1.

Step 1: Determine parameters for all the technical indicators. Create Parameter-file.
Step 2: Select indicators using the Parameter-File. Create Indicator-file.
Step 3: Generate Strategy using Indicator-file + Parameter-file. Create Strategy-file.
Step 4: Predict future price level using Strategy-file.

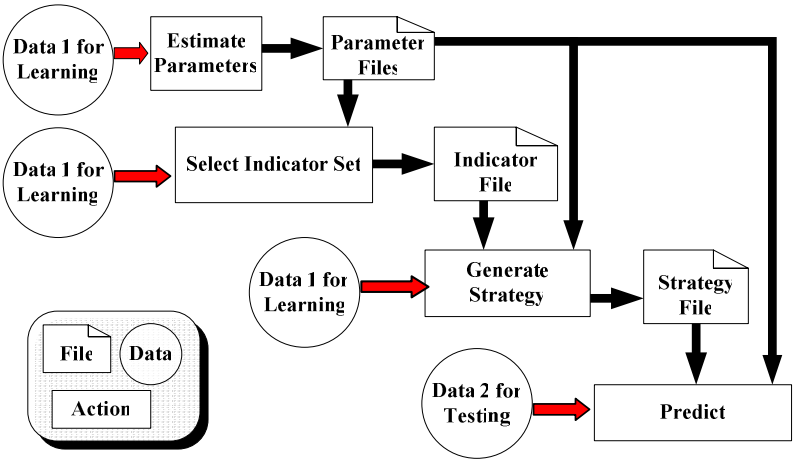


Fig. 1. Job flow of our prediction generator

5 Results and Concluding Remarks

We test those prediction strategies on new data. The last half of the data set is served for this purpose. The prediction referred here is the UP/DOWN motion of the price at

10 ticks ahead compared to the current price. Note that this method has a limitation for the data in which the situation drastically change for the first half and the second half within the data.

The result for FX data is shown in Table 4. The first row shows the years, the second row shows the best performance out of ten different choices of the indicators and the third row shows the average performance of the ten choices of the indicators. We have obtained nearly 70 % of the correct prediction. The lower rate from 1997 to 1998 can be attributed to the rapid fall of Yen against Dollar during this period. Note that the best and the average are almost in the same magnitude.

However, the time variations are not small. Within a year, the correct prediction rates fluctuate vigorously as shown in Fig. 2 where the solid line shows the weekly average, and the dashed line shows the daily average.

The results for 4 stocks in NYSE are shown in Table 5. The highest rates are obtained in IBM, that exceed 80%, and the second highest rates are observed in GM. Much lower hitting rates are obtained for BP and SMRT. The reason for this can be attributed for the difference in the tick intervals listed in Table 2. For IBM and GM, 10 ticks are equivalent to 140 - 230 seconds, which is equal to 2 - 4 minutes, while the same 10 ticks for BP and SMRT are 830 - 4730 seconds, which is equal to 14 – 79 minutes.

Table 4. The best and the average hitting rate of the predictions on the UP/Down direction of the price at 10 ticks ahead for FX data of JPY/USD for five years from 1996 to 2000. [%]

Year	1996	1997	1998	1999	2000
Best	69.9	66.5	64.9	66.0	67.7
Average	69.9	66.5	64.8	66.0	67.7

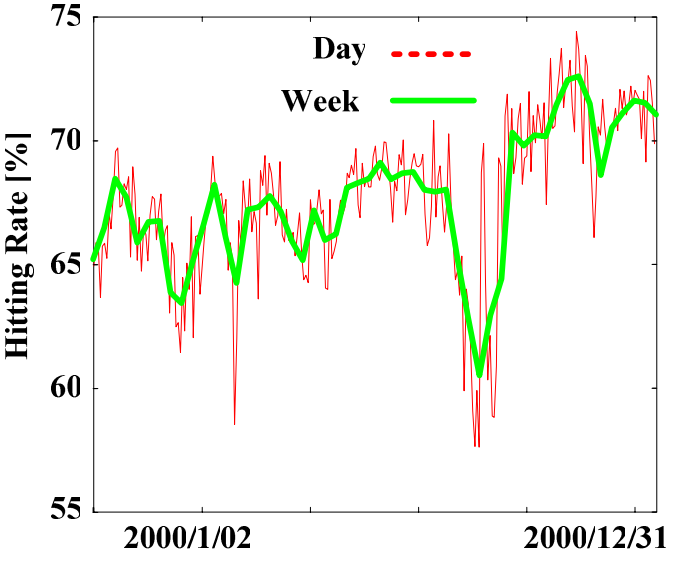


Fig. 2. The hitting rates for FX rates of JPY/USD in 2000

Table 5. The best and the average hitting rate of the predictions on the UP/Down direction of the price at 10 ticks ahead for 4 stocks (BP, SMRT, GM, IBM) in NYSE market. [%]

	BP	SMRT	GM	IBM
Best	65.4	67.2	78.5	82.3
Average	63.4	64.0	77.4	81.4

Finally, we show how the hitting rates decay according to the time, in the case of FX data. The rate is at around 70% after 1 minute, which goes down to 60% after 10 minutes, as shown in Fig.3.

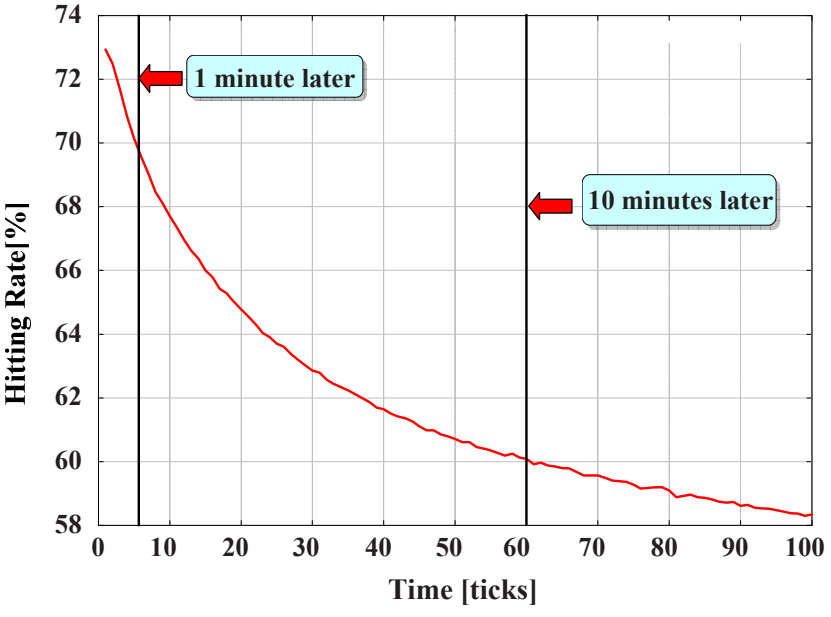


Fig. 3. The hitting rates at X ticks ahead for FX rates of JPY/USD in 2000

References

1. Mantegna, R.N., Stanley, H.E.: An Introduction to Econophysics: Correlations and Complexity in Finance. Cambridge University Press, New York (2000)
2. Takayasu, H. (ed.): Empirical Science of Financial Fluctuantions. Springer, Heidelberg (2001)
3. Tanaka-Yamawaki, M.: Stability of Markovian Structure Observed in High Frequency Foreign Exchange Data. Annals of the Institute of Statistical Mathematics 55, 437–446 (2003)
4. Tanaka-Yamawaki, M.: On the Predictability of High-Frequency Financial Time Series. In: Palade, V., Howlett, R.J., Jain, L. (eds.) KES 2003. LNCS, vol. 2773, pp. 1100–1108. Springer, Heidelberg (2003)

5. Tanaka-Yamawaki, M., Motoyama, T.: Predicting the Tick-wise Price Fluctuations by Means of Evolutionary Computation. In: Proc. IEEE-CEC, pp. 955–958 (2004)
6. Tanaka-Yamawaki, M.: Tickwise Predictions of Foreign Exchange Rates. In: Negoita, M.Gh., Howlett, R.J., Jain, L.C. (eds.) KES 2004. LNCS (LNAI), vol. 3213, pp. 449–454. Springer, Heidelberg (2004)
7. New York Stock Exchange Trade and Quotes: http://www.nysedata.com
8. http://decisionpoint.com
9. Bauer, R.J., Dahlquist, J.R.: Technical Market Indicators, Analysis and Performance. J. Wiley, New York (1999)
10. Tanaka-Yamawaki, M., Tokuoka, S.: Adaptive Use of Technical Indicators for the Prediction of Intra-day Stock Prices, Physica A (2007) (to be published)

Credal Networks for Operational Risk Measurement and Management*

Alessandro Antonucci, Alberto Piatti, and Marco Zaffalon

Istituto "Dalle Molle" di Studi sull'Intelligenza Artificiale
Via Cantonale, Galleria 2, CH-6928 Manno, Switzerland
{alessandro,alberto.piatti,zaffalon}@idsia.ch

Abstract. According to widely accepted guidelines for self-regulation, the capital requirements of a bank should relate to the level of risk with respect to three different categories. Among them, *operational risk* is the more difficult to assess, as it requires merging expert judgments and quantitative information about the functional structure of the bank. A number of approaches to the evaluation of operational risk based on Bayesian networks have been recently considered. In this paper, we propose *credal networks*, which are a generalization of Bayesian networks to imprecise probabilities, as a more appropriate framework for the *measurement* and *management* of operational risk. The reason is the higher flexibility provided by credal networks compared to Bayesian networks in the quantification of the probabilities underlying the model: this makes it possible to represent human expertise required for these evaluations in a credible and robust way. We use a real-world application to demonstrate these features and to show how to measure operational risk by means of algorithms for inference over credal nets. This is shown to be possible, also in the case when the observation of some factor is vague.

Keywords: Credal networks, Operational Risk, Imprecise probabilities.

1 Introduction

In the last two decades, the definition of standards for the stability of the international banking system, aiming at improving consistency of international capital regulations, has been unanimously recognized by the banking institutes worldwide as a keypoint for their self-regulation. The *Basel Committee on Banking Supervision*, which is the main authority in this field, defines the guidelines, accepted by national supervisory institutes worldwide, to measure the capital adequacy of each bank and to guarantee it to undergo a minimum standard. Accordingly, three different categories of risk, associated respectively to market, credit and operational activities, and the corresponding minimal capital

* This research was partially supported by the Swiss NSF grants 200020-109295/1 and 200021-113820/1. We are grateful to Fabio Trojani for introducing us to the problem of operational risk, and to Kwabena Adusei-Poku for providing us with some supplementary information about the model presented in [1].

B. Apolloni et al. (Eds.): KES 2007/ WIRN 2007, Part II, LNAI 4693, pp. 604–611, 2007.
© Springer-Verlag Berlin Heidelberg 2007

requirements hedging each category, should be estimated. While the evaluation of market and credit risk is traditionally supported by (large amounts of) historical data and well-established measurement methods, the evaluation of the latter category, which is called *operational risk* (Sect. 2), is more difficult, requiring the development of models that combine expert opinions with quantitative information. Yet, a standard model was not explicitly specified by the Basel Committee.

To fill this gap, some scholars have proposed *Bayesian networks* (Sect. 3.1) as a possible framework to measure operational risk with respect to the Committee requirements. Bayesian networks are graphical models, whose quantification requires a precise elicitation of the probabilistic relations among the different factors. Yet, this requirement clashes with the kind of uncertainty characterizing qualitative expert judgments about operational risk.

For this reason, we regard *credal networks* (Sect. 3.2), which are a generalization of Bayesian networks to imprecise probabilities, as a more credible model for operational risk. Credal networks allow for the specification of intervals (or, more generally, closed convex sets of mass functions) instead of single values of probability: this appears to be better suited to capture human knowledge. This flexibility regards also the observation of the variables: credal nets can cope with vague observations, where a condition of partial or complete ignorance about the actual state of an observed variable holds (Sect. 3.3).

To demonstrate and clarify these key features, a Bayesian network for operational risk evaluation is considered (Sect. 4.1). This model is first transformed into a credal network by more reliable probabilistic assessments (Sect. 4.2). We then show, by this example, how to employ credal networks for both *measurement* and *management* of operational risk. To solve the first problem, the network is fed with the available evidential information about the bank under consideration. The posterior probability intervals for the risk level can be therefore computed by standard algorithms for credal networks (Sect. 5). These tests are considered also in presence of *soft evidence*, that means observation unable to reveal the state of a variable, which is therefore modeled by (imprecise) probabilistic assessments. A simulation of a management problem, consisting in the evaluation of the configuration of the variables minimizing operational risk, is also considered (Sect. 6).

Overall, we regard this work as a first step towards a systematic approach to operational risk measurements based on credal networks. Conclusions and possible developments in this direction are reported in Sect. 7.

2 Operational Risk

In June 2004, the Basel Committee on Banking Supervision (simply called the *Committee* in the following) released an important document known as *Basel II accord* [9]. This document defines a comprehensive measure and a minimum standard for capital adequacy, which has been implemented by national supervisory authorities worldwide. The aim of these guidelines is to contribute to the stability of the international banking system, by improving the consistency of

capital regulations, making regulatory capital more risk sensitive, and promoting enhanced risk-management practices among financial institutions. Basel II is a revised version of a first accord (*Basel I*), that has been employed since 1988 [7].

Basel II sets standards for measurement and minimal capital requirements with respect to three different types of risk that can be incurred by financial institutions: market, credit, and operational risk. While market risk and credit risk, that were already considered in Basel I, can be evaluated by standard techniques, the evaluation of the *operational risk* (OR), that appears for the first time in Basel II as a separate category of risk, is more difficult.

OR is defined as *"the risk of loss resulting from inadequate or failed internal processes, people and systems or from external events"* [9]. The definition of OR is very general and encompasses several possible events like for example internal fraud, business disruption and system or delivery failures [8]. It is important to note that, unlike factors driving market and credit risk, factors driving OR are mainly internal and specific to the single financial institutions. This feature, together with the scarcity of reliable historical data regarding losses due to OR events, is the main reason for the lack of a standard method in this field.

The Committee has proposed three possible approaches for measuring OR and calculating the corresponding minimal capital requirement: (i) the *Basic Indicator Approach*, (ii) the *Standardized Approach*, and (iii) the *Advanced Measurement Approach* (AMA). Of these approaches, only the latter is risk sensitive, while the others adopt the conservative choice to set the minimal capital requirements as a fixed percentage of the the gross annual income in different business lines, and are therefore useless for OR management. In the AMA, the calculation of minimal capital requirement is based on OR measurement systems developed by the bank itself. Basel II is not prescriptive with respect to the particular models that should be used by financial institutions following the AMA to measure OR, but describes several qualitative and quantitative criteria that should be met by the measurement model in order to be appropriate, e.g., *"use of internal data, external relevant data, scenario analysis and factors reflecting the business environment"* [9]. To meet these requirements, it is therefore necessary to apply models merging expert opinions with data. In Sect. 4, we propose *credal networks* (which are defined in Sect. 3.2) as an appropriate mathematical framework for these evaluations.

3 Mathematical Aspects

In this section we review the basics of Bayesian nets and their extension to *imprecise probabilities* [12], i.e., credal nets. Both the models are based on a collection of random variables $\mathbf{X} \equiv (X_1, \ldots, X_n)$, which take values in finite sets, and a directed acyclic graph $\mathcal{G}$, whose nodes are in one-to-one relationship with the variables of $\mathbf{X}$. For both models, we assume the *Markov condition* to make $\mathcal{G}$ represent probabilistic independence relations between the variables in $\mathbf{X}$: every variable is independent of its non-descendant non-parents conditional on its parents. What makes Bayesian and credal nets different, is a different

notion of independence and a different characterization of the conditional mass functions for each variable given the possible values of the parents, to be detailed next.

Regarding notation, for each $X_i \in \mathbf{X}$, Ω_{X_i} is the possibility space of X_i, x_i a generic element of Ω_{X_i}, $P(X_i)$ a mass function for X_i and $P(x_i)$ the probability that $X_i = x_i$. A similar notation with uppercase subscripts (e.g., X_E) denotes arrays (and sets) of variables in $\mathbf{X}$. The parents of X_i, according to $\mathcal{G}$, are denoted by Π_i and for each $\pi_i \in \Omega_{\Pi_i}$, $P(X_i|\pi_i)$ is the conditional mass function for X_i given the joint value π_i of the parents of X_i.

3.1 Bayesian Networks

In the case of Bayesian networks [10], the modeling phase involves specifying a conditional mass function $P(X_i|\pi_i)$ for each $X_i \in \mathbf{X}$ and $\pi_i \in \Omega_{\Pi_i}$; and the standard notion of probabilistic independence is assumed in the Markov condition. A Bayesian net can therefore be regarded as a joint probability mass function over $\mathbf{X}$ that factorizes as follows: $P(\mathbf{x}) = \prod_{i=1}^{n} P(x_i|\pi_i)$, for each $\mathbf{x} \in \Omega_{\mathbf{X}}$, because of the Markov condition.

A Bayesian network can be regarded as an expert system, from which we can extract probabilistic information about its variables. The updated belief about a queried variable X_q, given some evidence $X_E = x_E$, is given by Bayes rule:

$$P(x_q|x_E) = \frac{P(x_q, x_E)}{P(x_E)} = \frac{\sum_{x_M} \prod_{i=1}^{n} P(x_i|\pi_i)}{\sum_{x_M, x_q} \prod_{i=1}^{n} P(x_i|\pi_i)}, \tag{1}$$

where $X_M \equiv \mathbf{X} \setminus (\{X_q\} \cup X_E)$, the domains of the arguments of the sums are left implicit, and the values x_i and π_i are consistent with (x_q, x_M, x_E). Despite its NP-hardness in general, the problem of computing (1) can be efficiently solved for Bayesian networks defined over singly-connected graphs with standard propagation schemes [10]. Similar techniques apply also for general topologies with increased computational time.

3.2 Credal Networks

Credal nets relax Bayesian nets by allowing for imprecise probability statements. This generalization is based on the fundamental notion of *credal set* [5]. We define a *credal set* $K(X)$ as the convex hull of a collection of mass functions over X. We assume this collection to be finite. $K(X)$ contains therefore an infinite number of mass functions, but only a finite number of *extreme mass functions*, corresponding to the (geometrical) *vertices* of the convex hull. Inference based on a credal set is equivalent to that based only on its extreme mass functions [12], so that the nature of the problem we face is inherently combinatorial.

The specification of a *credal network* over $\mathbf{X}$ consists in the assessment of a conditional credal set $K(X_i|\pi_i)$ for each $X_i \in \mathbf{X}$ and $\pi_i \in \Omega_{\Pi_i}$. We assume that the credal sets of the net are *separately specified*: this implies that selecting a mass function from a credal set does not influence the possible choices in others.

The *strong extension* $K(\mathbf{X})$ of a credal network [3] is defined as the convex hull of the joint mass functions $P(\mathbf{X})$, which, for each $\mathbf{x} \in \Omega_{\mathbf{X}}$, are such that:

$$P(\mathbf{x}) = \prod_{i=1}^{n} P(x_i|\pi_i), \quad \begin{array}{l} P(X_i|\pi_i) \in K(X_i|\pi_i), \\ \text{for each } X_i \in \mathbf{X}, \pi_i \in \Omega_{\Pi_i}. \end{array} \tag{2}$$

Observe that any extreme mass function $P(\mathbf{X})$ of $K(\mathbf{X})$ can be identified with a Bayesian net. In other words, a credal net is equivalent to a set of so-called *compatible* Bayesian nets. Inference over a credal net is intended as the computation of lower and upper expectations over the strong extension $K(\mathbf{X})$. Updating consists therefore in the computation of the lower and upper bounds of (1), denoted respectively as $\underline{P}(x_q|x_E)$ and $\overline{P}(x_q|x_E)$, evaluated over all the compatible Bayesian nets. Credal nets updating is an NP-hard task [4], for which a number of approximate algorithms have been proposed [3].

3.3 Soft Evidence by Credal Networks

In a recent paper [2], credal nets updating has been generalized to situations where the observations of some variables are missing according to an unknown process. Remarkably, we can reproduce this situation by simply augmenting the network with *auxiliary binary child nodes*, whose (imprecise) quantification reproduces the ignorance about the missingness process. This approach can be naturally extended to various forms of *soft evidence* about a variable. If an observation cannot reveal the actual state of a variable, but the knowledge provided by this observation can be modeled as a credal set, we can still fed the credal net with this (soft) information, by an appropriate quantification of the auxiliary child. Section 5 explains this idea by an application to OR.

4 A Credal Network for Operational Risk Measurement

Several authors have proposed Bayesian nets as a possible modeling framework for OR measurement [1,6,11]. These models make it possible to merge qualitative and quantitative factors driving OR measurements in a very natural and easy verifiable way. The most critical and time-consuming task for the realization of a Bayesian net in practical OR measurements is the quantification of the probabilities. This process can be very critical and arbitrary, e.g., what is the probabilistic assessment modeling the fact that the trade type of a bank is more likely to be standard than not? The problem is that these kinds of knowledge cannot be modeled by precise probabilistic statements. On the other hand, a more credible model of such qualitative judgment is the credal set made of *all* the mass functions assigning more than half of the mass to the state *standard*. This is the main reason why we regard credal nets as a model more suited than Bayesian nets for OR measurements. To demonstrate the main features of this approach, a concrete application based on a Bayesian net for OR measurement is considered.

4.1 A Bayesian Network for Operational Risk (Adusei-Poku 2005)

Let us consider the Bayesian network for OR measurement detailed in [1]. As an indicator of the level of OR, the number of SETTLEMENTS FAILURES in foreign exchange operations (e.g., delays, wrong or misdirected payments) in a typical week of a particular bank is considered. For our purposes, the following categorization is sufficient: less than 5 failures (corresponding to *low risk*), from 6 to 10 failures (*medium risk*), more than 10 failures (*high risk*).[1] Thirteen variables, characterizing the internal structure of the bank under consideration, and affecting this number of failures, are indeed considered. Figure 1 depicts the probabilistic dependencies between these variables and the OR indicator (according to the Markov condition). We point the reader to [1, Appendix B] for the quantification of this network.

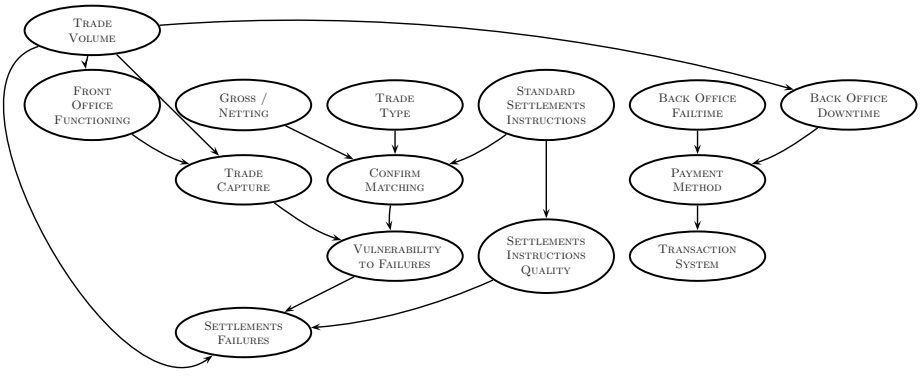


Fig. 1. A directed acyclic graph depicting dependencies between the OR level (represented by the node SETTLEMENTS FAILURES) and its triggering factors

4.2 From Bayesian to Credal Network

The data about the conditional probabilities for the Bayesian net of Section 4.1 attest one of the main limitations of these models for the representation of human expertise. For some variables, the specification in [1] actually reports ranges instead of precise values for the probabilities assessed by the experts. But in the end, a *typical value* in this range has to be adopted for use with Bayesian nets. Thus, the modeling phase requires artificially strong assessments, that can be avoided using credal nets. To show the benefit, we can therefore transform the Bayesian network into a credal network.

The ranges provided for the probabilities of the two states of the node TRADE TYPE, (90–99% for *standard* and the complement for *non standard*) are reproduced by a credal set with two extreme mass functions, and we similarly proceed for the nodes STANDARD SETTLEMENTS INSTRUCTIONS, GROSS/NETTING, and

[1] We use SMALL CAPITAL for the variables and *slanted* for the states.

Settlements Failures, obtaining a credal net with 216 compatible Bayesian nets. This relatively small number makes it possible to solve inferences exhaustively by simply solving each Bayesian net independently.

5 Operational Risk Measurement

The credal net described in the previous section can be clearly employed to measure OR. To this extent, it is sufficient to collect the available information about the variables affecting the OR, and update the probabilities of the node Settlements Failures. These probabilities for a simulated scenario consisting of a bank where the Front Office Functioning is *proper*, the Settlements Instructions Quality is *average* and the Payment System is *manual* are: 86.69–91.94% for *low*, 7.70–12.87% for *medium*, and 0.35–0.45% for *high risk*. This result is largely acceptable from a financial point of view, and based on much more reliable assumptions than the corresponding result obtained by the Bayesian net (90.01% for *low*, 9.61% for *medium*, and 0.38% for *high*).

According to Sect. 3.3, we can also model a situation where only soft evidence about these variables is available. Assume for instance that the observation of the Front Office Functioning certifies that the system is certainly not *down*, but we cannot decide whether it is *malfunctioning* or *properly functioning*, even if our observation suggests that the latter case is more likely. This is a soft evidence we can model by a credal set. Thus, we fed the net with this information by simply adding a binary child to the Front Office Functioning, assuming two possible specifications of the relative conditional probability table:

$$\left\{ \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 1 \end{bmatrix}, \begin{bmatrix} 0.5 & 0.5 & 0 \\ 0.5 & 0.5 & 1 \end{bmatrix} \right\}, \tag{3}$$

where the columns of these tables correspond respectively to the states *down*, *proper*, and *malfunctioning*. The first row of each tables reproduces a vertex of the credal set modeling our soft evidence. As an obvious generalization of the results presented in [2], we conclude that assuming the child node in the state corresponding to the first row of the table, we can compute the OR level of this scenario. The posterior probabilities (85.54–91.94% for *low*, 7.70–14.05% for *medium* and 0.35–0.54% for *high risk*) reveal a relative robustness of the simulated scenario with respect to the partial ignorance about the state of the Front Office Functioning.

6 Operational Risk Management

Finally, let us consider a management problem consisting in the evaluation of the effect of Front Office Functioning, Settlements Instructions Quality, and Payment Method to the level of OR. By simply analyzing the results reported in Tab. 1, we note that the configuration *proper*/*excellent*/*STP* assigns the highest probability to *low risk*. More generally, these kinds of simulations provide important information about the role played by the different variables in the determination of the overall level of OR.

Table 1. Posterior probabilities of OR for 18 simulated scenarios

Front Office functioning	Quality of SI	Payment System = *STP* P(Failures Settlements) [%] *low*	*medium*	high	Payment System = *manual* P(Failures Settlements) [%] *low*	*medium*	*high*
proper	*excellent*	[97.97,98.23]	[01.77,02.02]	[00.00,00.00]	[97.63,97.91]	[02.09,02.36]	[00.00,00.01]
proper	*average*	[88.40,93.36]	[06.38,11.27]	[00.25,00.33]	[86.69,91.94]	[07.70,12.87]	[00.35,00.45]
proper	*bad*	[46.54,54.35]	[42.49,47.58]	[03.15,05.89]	[44.40,52.66]	[43.39,48.76]	[03.95,06.85]
mal func.	*excellent*	[96.65,96.83]	[03.15,03.33]	[00.02,00.02]	[96.06,96.11]	[03.87,03.92]	[00.02,00.02]
mal func.	*average*	[81.35,86.64]	[12.63,17.84]	[00.73,00.81]	[79.33,84.13]	[14.90,19.66]	[00.97,01.01]
mal func.	*bad*	[37.69,46.17]	[46.48,51.96]	[07.31,10.43]	[35.55,43.48]	[47.18,52.94]	[08.75,11.59]
down	*excellent*	[95.61,96.03]	[03.92,04.34]	[00.04,00.05]	[94.86,95.12]	[04.82,05.08]	[00.06,00.06]
down	*average*	[78.67,84.28]	[14.66,20.06]	[01.06,01.26]	[76.46,81.49]	[17.12,22.01]	[01.38,01.53]
down	*bad*	[35.74,44.13]	[46.09,51.00]	[09.67,13.46]	[33.55,41.32]	[47.19,51.86]	[11.36,14.81]

7 Conclusions and Further Work

The main goal of this paper is to point out as credal networks can be regarded as a natural framework for OR measurement problem, with the same desirable features of Bayesian networks, but allowing for more freedom and robustness in the elicitation of the underlying probabilities and in the treatment of the available evidence. This argument is demonstrated by a concrete application, which represents a starting point for similar and more complex applications.

References

1. Adusei-Poku, K.: Operational Risk Management – Implementing a Bayesian Network for Foreign Exchange and Money Market Settlement. PhD thesis, University of Göttingen, Faculty of Economics and Business Administration (2005)
2. Antonucci, A., Zaffalon, M.: Equivalence between Bayesian and credal nets on an updating problem. In: Lawry, J., Miranda, E., Bugarin, A., Li, S., Gil, M.A., Grzegorzewski, P., Hryniewicz, O. (eds.) Soft Methods for Integrated Uncertainty Modeling, pp. 223–230. Springer, Heidelberg (2006)
3. Cozman, F.G.: Graphical models for imprecise probabilities. Int. J. Approx. Reasoning 39(2-3), 167–184 (2005)
4. de Campos, C.P., Cozman, F.G.: The inferential complexity of Bayesian and credal networks. In: Proceedings of the International Joint Conference on Artificial Intelligence, Edinburgh, pp. 1313–1318 (2005)
5. Levi, I.: The Enterprise of Knowledge. MIT Press, London (1980)
6. Neil, M., Fenton, N.E., Tailor, M.: Using bayesian networks to model expected and unexpected operational losses. Risk Analysis 25(4), 963–972 (2005)
7. Basel Committee on Banking Supervision: International Convergence of Capital Measurement and Capital Standards. Basel (1988)
8. Basel Committee on Banking Supervision: Sound Practices for the Management and Supervision of Operational Risk. Basel (2002)
9. Basel Committee on Banking Supervision. International Convergence of Capital Measurement and Capital Standards, a Revised Framework. Basel (2004)
10. Pearl, J.: Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference. Morgan Kaufmann, San Mateo (1988)
11. Ramamurthy, S., Arora, H., Ghosh, A.: Operational risk and probabilistic networks – an application to corporate actions processing. Infosys White Paper (2005)
12. Walley, P.: Statistical Reasoning with Imprecise Probabilities. Chapman and Hall, New York (1991)

Fair Consistency Evaluation in Fuzzy Preference Relations and in AHP

Matteo Brunelli and Michele Fedrizzi

Dipartimento di Informatica e Studi Aziendali, Università di Trento
Via Inama 5, TN 38100 Trento, Italy
michele.fedrizzi@unitn.it

Abstract. In this paper we first highlight what we consider a drawback of most consistency indices for pairwise comparison matrices. In our opinion they do not take into account whether the elicited judgements are close to indifference or, on the contrary, they express strong preferences in comparing the alternatives, with the result that the latter case is unfairly penalized. We then introduce a consistency preserving transformation. By means of this transformation we define an equivalence relation in the set of pairwise comparison matrices and propose a new method for a more fair evaluation of the consistency. Finally, we extend the new method to fuzzy preference relations.

Keywords: Pairwise Comparison Matrices, Consistency Indices, Fuzzy Preference Relations.

Introduction

Consistency of judgements is considered an important issue in decision making problems and the achievement of a satisfactory consistency level is viewed as a desirable property. In [5], for example, an aggregation process for group decision making problems is proposed where the more consistent the preferences of the various decision makers, the more importance is given to those preferences. Therefore, it is crucial that the consistency evaluation is carried out in a fair way. In this paper we argue that almost all known consistency indices suffer from the same drawback: they unfairly penalize the pairwise comparison matrices with strong preferences and favor, on the contrary, those with preferences close to indifference, even if they are contradictory. The paper is organized as follows. In section 1 we introduce the problem of pairwise comparison with some necessary notation. In section 2 we consider a consistency preserving transformation of a Pairwise Comparison Matrix, *PCM* in the following, and we describe what, in our opinion, is a drawback of almost all known consistency indices. In order to overcome the previously described drawback, we propose in section 3 a new method for evaluating the consistency of a *PCM*. Finally, in section 4, we extend our approach to the fuzzy preference relations framework by means of a simple transformation function.

B. Apolloni et al. (Eds.): KES 2007/ WIRN 2007, Part II, LNAI 4693, pp. 612–618, 2007.
© Springer-Verlag Berlin Heidelberg 2007

1 Pairwise Comparisons and Consistency Indices

Let $A = \{A_1, A_2, ..., A_n\}$ be a set of n alternatives and let the judgements of a decision maker be expressed by pairwise comparisons of the alternatives. We assume that the reader is familiar with Saaty's approach [14], where the decision maker's judgement a_{ij} estimates the ratio w_i/w_j of the preference intensities of alternative A_i over alternative A_j and the 1– 9 ratio scale is used. A *PCM* $A = [a_{ij}]$ is then positive and reciprocal, $a_{ji} = 1/a_{ij}, \quad \forall i, j$. We denote by Ω_n the set of the $n \times n$ pairwise comparison matrices.

A *PCM* $A = [a_{ij}]$ is called consistent if and only if

$$a_{ij} = a_{ih}a_{hj} \quad \forall i, j, h. \tag{1}$$

If (1) is not satisfied, it is a relevant issue to evaluate, by means of a consistency index, how 'far' is $A = [a_{ij}]$ from full consistency condition (1). In his seminal paper on the Analytical Hierarchy Process [13], T. Saaty proposes the consistency index *CI*,

$$CI(A) = \frac{\lambda_{max} - n}{n - 1}, \tag{2}$$

where λ_{max} is the maximum eigenvalue of A. By means of (2), Saaty defines the Consistency Ratio $CR(A) = CI(A)/RI$, where the Random Index RI is the mean value of the CI of random $n \times n$ *PCM*s. After that, several other authors proposed different consistency indices [1] [3] [5] [8] [12] [15]. For brevity, we do not describe here the various proposals and refer the interested reader to the cited papers.

2 A Consistency Preserving Transformation

Assuming that a *PCM* A is consistent, i.e. that (1) is satisfied, a transformation $f(\cdot)$ preserves consistency property (1) if and only if

$$f(a_{ij}) = f(a_{ih})f(a_{hj}). \tag{3}$$

From (1) and (3) the following Cauchy's functional equation is obtained,

$$f(a_{ih}a_{hj}) = f(a_{ih})f(a_{hj}). \tag{4}$$

Excluding the trivial solution $f(a_{ij}) = 0$, for positive arguments and assuming continuity [2], the general solution of (4) is

$$f(a_{ij}) = (a_{ij})^k, \quad k \in \Re. \tag{5}$$

As a consequence, if a *PCM* $A = [a_{ij}]$ is consistent, then it is consistent also every *PCM* $A' = [a'_{ij}]$ obtained from A by means of (5), for every real value of k,

$$a'_{ij} = f(a_{ij}) = (a_{ij})^k. \tag{6}$$

The general result stated above clearly requires that the scale $\frac{1}{9}, \ldots, 9$ originally proposed by Saaty is extended to the whole set of positive real numbers. If it is required that the entries of the *PCM* remain in the interval $[\frac{1}{9}, 9]$, it is sufficient to conveniently bound the value of k in (5).

Transformation (5) is used in [9] in order to rescale in Saaty's interval $[\frac{1}{9}, 9]$ a consistent *PCM* with some entries not belonging to this interval. Note that (5) also preserves properties weaker than (1) introduced and studied in [4] [6].

At this point we can better explain why we consider most of the known consistency indices to be unfair. The first argument is the following. A *PCM* with all entries equal to one is evaluated as fully consistent by all the known consistency indices. Note that this is the case where the decision maker expresses indifference in every pairwise comparison between alternatives. A small change in the values of the entries causes a small change in the value of the consistency index, except for the index proposed in [3]. Therefore, this means that the *PCM* remains very close to consistency, and this happens whatever are the changes, even if cycles arise stressing contradictory preferences. The following example should clarify this fact. Let $A \in \Omega_n,\ A_{modif} \in \Omega_n$, with

$$A = \begin{pmatrix} 1 & 1 & 1 \\ 1 & 1 & 1 \\ 1 & 1 & 1 \end{pmatrix}, \quad A_{modif} = \begin{pmatrix} 1 & \frac{5}{4} & \frac{4}{5} \\ \frac{4}{5} & 1 & \frac{5}{4} \\ \frac{5}{4} & \frac{4}{5} & 1 \end{pmatrix}.$$

Matrix A_{modif} contains a cycle: $A_1 \succ A_2 \succ A_3 \succ A_1$ and therefore it is contradictory. Nevertheless, since all the judgements are still close to indifference, its consistency index (2) is quite good: $CI = \frac{1}{40}$. To summarize: in our opinion, not all *PCM*s with entries close to one should be considered as satisfactorily consistent.

The second argument is the following. Given that transformation (6) preserves consistency for a consistent *PCM*, it is justified to assign the same consistency level to both *PCM*s A and A' (non necessarily consistent) if A' is obtained by applying (6) to A. In the following section we will formalize this idea.

3 A New Proposal for Consistency Evaluation

In the previous section we have argued that an unbiased method for consistency evaluation should take into account only the mutual coherence of the judgements and thus should be, in a suitable way, independent from the size of the entries of the corresponding *PCM*. Starting from consistency preserving transformation (6), let us define an equivalence relation in Ω_n.

Definition 1 (Consistency–Equivalence for *PCM*). *Let $A \in \Omega_n$ and $B \in \Omega_n$, $A = [a_{ij}]$, $B = [b_{ij}]$. A is said consistency–equivalent to B, $A \sim B$, If and only if $\exists k \neq 0$ s.t. $a_{ij} = b_{ij}^k\ \forall i, j$.*

Proposition 1. *Consistency–Equivalence $\sim$ is an equivalence relation.*

Proof

1. *Reflexivity.* Clearly, with $k = 1$, it is $A \sim A$.
2. *Symmetry.* From $a_{ij} = b_{ij}^k$, it is $b_{ij} = a_{ij}^{\frac{1}{k}}$; then $A \sim B \Rightarrow B \sim A$.
3. *Transitivity.* Let $C \in \Omega_n$, $C = [c_{ij}]$. If $a_{ij} = b_{ij}^k$ and $b_{ij} = c_{ij}^h$, then $a_{ij} = c_{ij}^{kh}$. Therefore $(A \sim B \text{ and } B \sim C) \Rightarrow A \sim C$. □

As a consequence of Proposition 1, set Ω_n is partitioned by $\sim$ into equivalence classes. Let $\Omega_n/\sim$ be the quotient set. To overcame the above mentioned drawback, we propose to consider equivalent, from the point of view of consistency, all the *PCM*s in the same equivalence class $\Theta \in \Omega_n/\sim$. Therefore, we assign to all the *PCM*s $A \in \Theta$ the same numerical value to quantify their consistency. We denote this value by $C(A)$ and we will call it *inconsistency level* of A. Since all the *PCM*s in Θ share the same inconsistency level, we can denote it by C_Θ. Using this notation, it is $A \in \Theta \Rightarrow C(A) = C_\Theta$.

We propose to define C_Θ in the following way. We chose a particular *PCM* $\hat{A} \in \Theta$ as representing the whole equivalence class, we compute its Consistency Ratio $CR(\hat{A})$ and we assign to all the matrices in Θ the inconsistency level

$$C_\Theta = CR(\hat{A}). \tag{7}$$

The choice of $\hat{A}$ can be made on the basis of different criteria. One possibility is to define $\hat{A}$ as the (unique) *PCM* in Θ with maximum entry equal to 9, i.e. the maximum feasible value in Saaty's scale. A second possibility is to define $\hat{A}$ as the (unique) *PCM* in Θ with 'average size entries'. More precisely, referring to 1– 9 Saaty's scale, which has mean value 5, it is possible to define $\hat{A}$ as the *PCM* having the mean value of the $\frac{n(n-1)}{2}$ comparisons with $a_{ij} \geq 1$ equal to 5.

The previous definition of C_Θ doesn't apply to the equivalence class containing the single matrix with all the entries equal to one. Clearly, the inconsistency level of this particular *PCM* is defined to be equal to zero.

Let us now apply (5), for $k \in]0, 10]$, to the entries of the matrix A_{modif} introduced in section 2. For every obtained matrix A_{modif}^k we compute Saaty's Consistency Index (2). Results are shown in Fig.1, where it is clear that the CI increases with respect to k.

According to our proposal, on the contrary, since all the matrices A_{modif}^k are in the same equivalence class, they share the same inconsistency level (7).

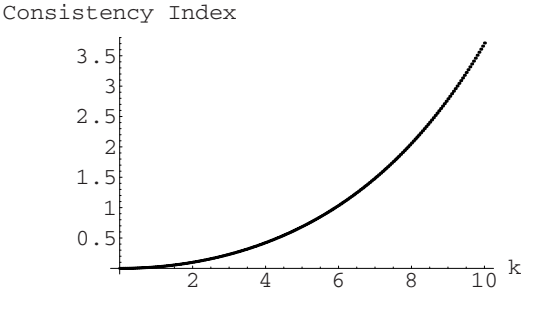


Fig. 1. Consistency Index of A_{modif}^k

Note that, instead of Saaty's Consistency Ratio, any other known consistency index can be used in (7), since what we propose is not a new index, but a new way of evaluating consistency, based on equivalence relation $\sim$. Analogously, any other ratio scale different from Saaty's one can be used. It is sufficient to substitute, in what described above, the number 9 with the maximum value of the considered scale. To our knowledge, the consistency index $RC(A)$ proposed in [3] by Barzilai is the only invariant one with respect to (5). Note that in [3] unbounded scales are used. In future research we will go deeper into the comparison between our proposal and Barzilai's index.

4 Extension to Fuzzy Preference Relations

The results reported in the previous sections can be easily extended to Fuzzy Preference Relations, *FPR* in the following, [10] [11] defined on the same set of alternatives $\Lambda = \{A_1, A_2, ..., A_n\}$. In fact, each *PCM* $A = [a_{ij}]$ can be transformed into a matrix $R = [r_{ij}]$ associated with a *FPR* and vice versa using the following function $g : [\frac{1}{9}, 9] \rightarrow [0, 1]$, introduced in [7],

$$r_{ij} = g(a_{ij}) = \frac{1}{2}(1 + \log_9 a_{ij}). \tag{8}$$

Function (8) transforms the a_{ij} values into the r_{ij} values in such a way that all the relevant properties of $A = [a_{ij}]$ are transformed into the corresponding properties for $R = [r_{ij}]$. In particular, multiplicative reciprocity $a_{ji} = 1/a_{ij}$ is transformed into additive reciprocity $r_{ij} + r_{ji} = 1$ [16] and multiplicative consistency (1) is transformed into additive consistency [16],

$$(r_{ih} - 0.5) + (r_{hj} - 0.5) = (r_{ij} - 0.5), \quad i, j, h = 1, ..., n \ . \tag{9}$$

As a consequence, every result obtained for *PCM*s can immediately be reformulated into the corresponding one for *FPR*s and vice versa.

Thus, let $R = [r_{ij}]$ be a consistent *FPR*. The Consistency preserving transformation for R can be obtained from (6) through (8), and it is

$$r'_{ij} - 0.5 = k(r_{ij} - 0.5). \tag{10}$$

Clearly, the same result could be obtained following a method analogous to that described in section 2 for *PCM*s. In this case, instead of (4), the following Cauchy's functional equation should have been solved,

$$f((r_{ih} - 0.5) + (r_{hj} - 0.5)) = f(r_{ih} - 0.5) + f(r_{hj} - 0.5), \tag{11}$$

obtaining again (10) as a solution.

The extension to *FPR*s of the results exposed in section 3 is briefly presented in the following, having denoted by Ψ_n the set of the n–dimensional *FPR*s.

Definition 2 (Consistency–Equivalence for *FPR*). *Let $R \in \Psi_n$ and $S \in \Psi_n$, $R = [r_{ij}]$, $S = [s_{ij}]$. R is said consistency–equivalent to S, $R \sim S$, If and only if $\exists k \neq 0$ s.t. $r_{ij} - 0.5 = k(s_{ij} - 0.5)$ $\forall i, j$.*

Proposition 2. *Consistency–Equivalence for FPR* $\sim$ *is an equivalence relation.*

We skip the proof of Proposition 2 for brevity, as it is similar to the proof of Proposition 1.

The set Ψ_n is partitioned by $\sim$ in equivalence classes, $\Psi_n/\sim$ being the quotient set. All the *FPR*s in the same equivalence class Φ share the same inconsistency level C_Φ.

As observed for (5), the general validity of the results presented above requires the use of an open scale, as it is assumed in [3]. Nevertheless, if it is required that the entries of R remain in the interval $[0, 1]$, it is sufficient to conveniently bound in (10) the value of k.

In [9] a related problem is addressed: the function

$$f(x) = \frac{1}{1+2a} \cdot x + \frac{a}{1+2a} \tag{12}$$

has been used to rescale into the interval $[0, 1]$ a consistent *FPR* with entries in the interval $[-a, 1+a]$. Note that function (12) is a special case of (10), obtained for $k = \frac{1}{2a+1}$.

Acknowledgements. The authors would like to thank the referees for their useful suggestions.

References

1. Aguarón, J., Moreno–Jiménez, J.M.: The geometric consistency index: Approximated thresholds. European Journal of Operational Research 147, 137–145 (2003)
2. Aczel, J.: Lectures on Functional Equations and their Applications. Academic Press, New York (1966)
3. Barzilai, J.: Consistency measures for pairwise comparison matrices. J. Multi–Crit. Decis. Anal. 7, 123–132 (1998)
4. Basile, L., D'Apuzzo, L.: Transitive Matrices, Strict Preference and Intensity Operators. Mathematical Methods in Economics and Finance 1, 21–36 (2006)
5. Chiclana, F., Herrera–Viedma, E., Herrera, F., Alonso, S.: Some induced ordered weighted averaging operators and their use for solving group decision–making problems based on fuzzy perference relations. European Journal of Operational Research (2006)
6. D'Apuzzo, L., Marcarelli, G., Squillante, M.: Generalized consistency and intensity vector for Comparison Matrices. International Journal of Intelligent System (to appear)
7. Fedrizzi Michele: On a consensus measure in a group MCDM problem. In: Multiperson Decision Making Models using Fuzzy Sets and Possibility Theory Theory and Decision Library, series B: Mathematical and Statistical Methods, vol. 18, Kluwer Academic Publishers, Dordrecht, The Netherlands (1990)
8. Fedrizzi Mario, Fedrizzi Mario, Marques Pereira, R.A.: On the issue of consistency in dynamical consensual aggregation. In: Bouchon Meunier, B., Gutierrez Rios, J., Magdalena, L., Yager, R.R (eds.) Technologies for Constructing Intelligent Systems, Physica, Studies in Fuzziness and Soft Computing, vol. 1, 89, pp. 129–137, Springer, Heidelberg (2002)

9. Herrera–Viedma, E., Herrera, F., Chiclana, F., Luque, M.: Some issues on consistency of fuzzy preference relations. European Journal of Operational Research 154, 98–109 (2004)
10. Kacprzyk, J., Roubens, M.: Non–Conventional Preference Relations in Decision–Making. Springer, Berlin (1988)
11. Orlovsky, S.A.: Decision–making with a fuzzy preference relation. Fuzzy Sets and Systems 1, 155–167 (1978)
12. Peláez, J.I., Lamata, M.T.: A new measure of consistency for positive reciprocal matrices. Computers and Mathematics with Applications 46, 1839–1845 (2003)
13. Saaty, T.L.: A scaling method for priorities in hierarchical structures. J. Math. Psychology 15, 234–281 (1977)
14. Saaty, T.L.: The Analytic Hierarchy Process. McGraw-Hill, New York (1980)
15. Shiraishi, S., Obata, T., Daigo, M.: Properties of a positive reciprocal matrix and their application to AHP. J. Oper. Res. Soc. Japan 41, 404–414 (1998)
16. Tanino, T.: Fuzzy preference orderings in group decision making. Fuzzy Sets and Systems 12, 117–131 (1984)

Making Financial Trading by Recurrent Reinforcement Learning

Francesco Bertoluzzo[1] and Marco Corazza[2,3]

[1] Department of Statistics, University of Padua
Via Cesare Battisti 241/243, 35121 Padua, Italy
fbertoluzzo@stat.unipd.it
[2] Department of Applied Mathematics, University Ca' Foscari of Venice
Dorsoduro 3825/E, 30123 Venice, Italy
corazza@unive.it
[3] School for Advanced Studies in Venice Foundation
Dorsoduro 3488/U, 30123 Venice, Italy

Abstract. In this paper we propose a financial trading system whose strategy is developed by means of an artificial neural network approach based on a recurrent reinforcement learning algorithm. In general terms, this kind of approach consists in specifying a trading policy based on some predetermined investor's measure of profitability, and in setting the financial trading system while using it. In particular, with respect to the prominent literature, in this contribution: first, we take into account as measure of profitability the reciprocal of the returns weighted direction symmetry index instead of the wide-spread Sharpe ratio; second, we obtain the differential version of this measure of profitability and obtain all the related learning relationships; third, we propose a procedure for the management of drawdown-like phenomena; finally, we apply our financial trading approach to some of the major world financial market indices.

Keywords: Financial trading system, recurrent reinforcement learning, no-hidden-layer perceptron model, returns weighted directional symmetry measure, gradient ascent technique, world financial market indices.

1 Introduction

When an economic agent invests in financial markets, she/he has to make decisions under uncertainty. In such a context, her/his task consists in coping with financial risk in order to maximize some predetermined measure of profitability. A wide-spread class of tools which are used to support such risky decisions is the one of the financial trading systems (FTSs).

A standard approach which is usually followed to specify a FTS consists:

- in identifying one or more variables (asset prices, ...) related to the time-behaviour of one or more quantities of interest (trading signals, ...);
- in utilizing the current and the past values of these variables to extract information concerning with the future values of the quantities of interest;

B. Apolloni et al. (Eds.): KES 2007/ WIRN 2007, Part II, LNAI 4693, pp. 619–626, 2007.
© Springer-Verlag Berlin Heidelberg 2007

– in using these predictions/information to implement a trading strategy by which to make effective trades.

The distinctly operative valence of the FTSs has made them popular from long time among professional investors and practitioners. A lot of book has been devoted to the working utilization of these tools (see, among the various ones, [8]). Nevertheless, in recent years also the academic world has began to recognize the soundness of some of the features related to the FTSs (see, for example, [5]). Moreover, the really enormous – and continuously increasing – mass of collected financial data has led to the development of FTSs whose information extraction processes are based on data mining methodologies (see, for example, [3]).

Alternative approaches to the standard building of FTSs have been proposed both in the operative literature and in the academic ones. Among the approaches belonging to the latest category, in this paper we consider the ones which exploit artificial neural network (ANN) methodologies based on recurrent reinforcement learnings. In general terms, these approaches consist:

– in specifying a trading policy based on some predetermined investor's measure of profitability (in such a manner one avoids to have to identify the quantities of interest, and avoids to have to perform the information extraction concerning with such quantities);
– in setting the frame and the parameters of the FTS while using it (in such a way one can avoid to carry out the off-line setting of the trading system).

Among the first contributions in this research field we recall [7], [6] and [4]. In general, they show that such strategies perform better than the ones based on supervised learning methodologies when market frictions (transaction costs, ...) are considered.

In this paper, with respect to the cited contributions:

– we take into account as measure of profitability the reciprocal of the returns weighted direction symmetry index (see the third section) reported in [1] instead of the wide-spread Sharpe ratio, which is the only one used in the quoted literature;
– we obtain the differential version of the considered measure of profitability and obtain all the new related learning relationships (see the third section);
– we propose a simple procedure for the management of drawdown-like phenomena by which to integrate the considered FTS (see the fourth section);
– we apply our financial trading approach (FTA) (i.e. FTS + drawdown-like phenomenon management) to some of the major world financial market indices (see the fourth section).

Finally, in the last section we provide some concluding remarks.

2 Recurrent Reinforcement Learning: A Short Recall

In this section we give a short qualitative introduction of the recurrent reinforcement learning.

In general terms, this kind of learning concerns an agent (in our case the FTS) dynamically interacting with an environment (in our case a financial market). During this interaction, the agent perceives the state of the environment and undertakes a related action. In its turn, the environment, on the basis of this action, provides a negative or positive reward (in our case the investor's loss or gain). The recurrent reinforcement learning consists in the on-line detection of a policy (in our case a trading strategy) which permits the maximization over the time of a predetermined cumulative reward (see, for technical details, [9]).

3 The Financial Trading System

In this section, we describe our discrete-time trading strategy and we obtain all the new learning relationships related to the considered measure of profitability.

3.1 The Trading Strategy

Let we start by considering a discrete-time frame $t = 0, \ldots, T$. Our trading strategy at time t, F_t, is based on the sign of the output, y_t, of a suitable ANN:

- if $y_t < 0$ then $F_t = -1$, and one short-sells the considered portfolio;
- if $y_t = 0$ then $F_t = F_{t-1}$, and one does nothing;
- if $y_t > 0$ then $F_t = 1$, and one buys the considered portfolio.[1]

We assume that this strategy depends on the current and past values of the logarithmic rate of return of the portfolio to trade, and on the previous value of the trading strategy itself.

The ANN we consider is a simple no-hidden-layer perceptron model (both the architectural structure and the squashing function are the ones commonly used in the relevant literature):

$$y_t = \tanh\left(\sum_{i=0}^{M} w_{i,t} x_{t-i} + w_{M+1,t} F_{t-1} + w_{M+2,t}\right),$$

where $w_{0,\tau}, \ldots, w_{M+1,\tau}$ are the weights of the ANN at time τ; $x_\tau, \ldots, x_{\tau-N}$ are the current and past values of the logarithmic rate of return of the portfolio to trade at time τ; and $w_{M+2,\tau}$ is the threshold of the ANN at time τ.

As net reward at the generic time period $(t-1,\ t]$ we take into account the following quantity:

$$R_t = \mu\left[F_{t-1} r_t - \delta\left|F_t - F_{t-1}\right|\right],$$

where μ is the amount of capital to invest; r_τ is the geometric rate of return at time τ of the portfolio to trade; and δ is the per cent transaction cost related to the portfolio quota to trade.

[1] F_t plays the role of the action.

It is to notice that, with respect to the cited contributions, in this paper we consider a net reward formulated in terms of rate of return instead of price.

Given the net reward of period, it is easy to define the net cumulative reward from time 1 to time t as $CR_t = \sum_{i=1}^{t} R_i$. This expression for the net cumulative reward is a particularization of the expression $CR_t = \sum_{i=1}^{t} R_i \left(1 + i_f\right)^{t-i}$, where i_f is the free-risk rate of return of period.

Finally, we give the new investor's gain index at time t whose utilization permits the determination – *via* recurrent reinforcement learning – of the optimal values of the weights of the considered ANN:

$$I_t = \frac{\sum_{i=1}^{t} g_i \left|R_i\right|}{\sum_{i=1}^{t} b_i \left|R_i\right|}, \text{ with } \sum_{i=1}^{t} b_i \left|R_i\right| \neq 0, \tag{1}$$

where $g_\tau = \begin{cases} 0 \text{ if } R_\tau \leq 0 \\ 1 \text{ if } R_\tau > 0 \end{cases}$; and $b_\tau = \begin{cases} 1 \text{ if } R_\tau \leq 0 \\ 0 \text{ if } R_\tau > 0 \end{cases}$.

This index, which is the reciprocal of the returns weighted directional symmetry measure reported in [1], is given at each time t by the ratio between the cumulative "good" (i.e. positive) rewards and the cumulative "bad" (i.e. not positive) rewards.

3.2 The Recurrent Reinforcement Learning

The considered ANN is characterized by $M+3$ parameters: $w_{0,t}, \ldots, w_{M+2,t}$. For determining their optimal values we perform an economically founded approach consisting in the maximization of an investor's utility function depending, at time t, on $R_1, \ldots, R_t$. In particular, as (non-standard) utility function we take into account (1). At this point, we determine the optimal values of the considered parameters by using an usual weight updating method based on the following gradient ascent technique:

$$w_{i,t} = w_{i,t-1} + \rho_t \frac{dU_t}{dw_{i,t}}, \text{ with } i = 0, \ldots, M+2, \tag{2}$$

where U_τ is the investor's utility function at time τ; and ρ_τ is a suitable learning rate at time τ.

It is to notice that, in each generic time period $(t-1,\ t]$, the investor is (obviously) interested in the marginal variation of her/his utility function, i.e. in $D_t = U_t - U_{t-1} = I_t - I_{t-1}$. Now, in order to provide the expression for $dU_t / dw_{i,t} = dI_t / dw_{i,t}$, it is to notice that computing I_t becomes as harder as t increases. So, we resort to the following exponential moving formulation of (1):

$$\widetilde{I}_t = \frac{A_t}{B_t} \tag{3}$$

where $A_\tau = \begin{cases} A_{\tau-1} \text{ if } R_\tau \leq 0 \\ \eta R_\tau + (1-\eta)\, A_{\tau-1} \text{ if } R_\tau > 0 \end{cases}$ is the exponential moving estimates of the numerator of (1) at time τ; $B_\tau = \begin{cases} -\eta R_\tau + (1-\eta)\, B_{\tau-1} \text{ if } R_\tau \leq 0 \\ B_{\tau-1} \text{ if } R_\tau > 0 \end{cases}$ is the

exponential moving estimates of the denominator of (1) at time τ; and η is an adaptation coefficient.

So, $U_t = \widetilde{I}_t$.

Then, in order to provide an expression for D_t, we act in a way similar to the one used in [7], [6] and [4], i.e.:

- firstly, we consider the expansion of (3) in Taylor's series about $\eta = 0$;
- secondly, we utilize $d\widetilde{I}_t/d\eta\Big|_{\eta=0}$ as approximation of D_t, i.e., after some arrangements:

$$D_t \cong \left.\frac{d\widetilde{I}_t}{d\eta}\right|_{\eta=0} = \begin{cases} -A_{t-1}\left(R_t + B_{t-1}\right)/\, B_{t-1}^2 \text{ if } R_t \leq 0 \\ \left(R_t - A_{t-1}\right)/\, B_{t-1} \text{ if } R_t > 0 \end{cases}.$$

At this point, it is possible to prove that (see, for more details, [7], [6] and [4]):

$$\frac{dU_t}{dw_{i,t}} = \sum_{j=1}^{t} \frac{dU_j}{dR_j}\left(\frac{dR_j}{dF_j}\frac{dF_j}{dw_{i,t}} + \frac{dR_j}{dF_{j-1}}\frac{dF_{j-1}}{dw_{i,t-1}}\right)$$

where $dU_\tau/\, dR_\tau = d\left(U_\tau - U_{\tau-1}\right)/\, dR_\tau = \begin{cases} \left(A_{\tau-1}B_{\tau-1}\right)/\, B_{\tau-1}^3 \text{ if } R_t \leq 0 \\ 1/\, B_{\tau-1} \text{ if } R_t > 0 \end{cases}$ (it is to notice that $dU_\tau/\, dR_\tau = d\left(U_\tau - U_{\tau-1}\right)/\, dR_\tau$ as $U_{\tau-1}$ does not depend on R_τ); $dR_\tau/\, dF_\tau = -\mu\, \delta\, \mathrm{sign}\left(F_\tau - F_{\tau-1}\right)$; $dR_\tau/\, dF_{\tau-1} = P_\tau - P_{\tau-1} - \mu\, \delta\, \mathrm{sign}\left(F_\tau - -F_{\tau-1}\right)$; and $dF_\tau/\, dw_{i,\tau}$, which depends on the chosen squashing function, is easily obtainable (see, for instance, [2]).

Finally, in order to implement (2), we approximate the previous exact relationship, which holds for batch learnings, in the following one, which holds for on-line learnings:

$$\frac{dU_t}{dw_{i,t}} \cong \frac{dU_t}{dR_t}\left(\frac{dR_t}{dF_t}\frac{dF_t}{dw_{i,t}} + \frac{dR_t}{dF_{t-1}}\frac{dF_{t-1}}{dw_{i,t-1}}\right).$$

4 Applications

In this section, at first we present the plan of our experimentation, then we propose a procedure for the management of drawdown-like phenomena, finally we give the results coming from the applications of our FTA.

We apply the considered FTA to 9 of the major world financial market indices, i.e. to some of the ones that at March 31, 2007 were quoted from at least 15 years: Dow Jones Industrials index (D), FTSE 100 index (F), Jakarta SE Composite index (J), Madrid SE General index (M), NASDAQ Composite index (NA), NIKKEI 225 Stock Average index (NI), Shanghai SE Composite index (Sh), Swiss Market index (Sw), and Tel Aviv SE General index (T). In detail, we consider close daily data from April 1, 1992, to March 31, 2007.

In order to make operative our FTA, we need to determine the optimal values/time-evolution of the parameters M, δ, ρ_t, η and of the one related to the management of drawdown-like phenomena. For carrying out this optimal setting, we perform in a manner similar to the one utilized in [7], [6] and [4], i.e. we articulate the whole trading time period (from 0 to T) in the sequence of overlapping time sub-period $[0,\ \Delta t_{off} + \Delta t_{on}]$, $[\Delta t_{on},\ \Delta t_{off} + 2\Delta t_{on}]$, ..., $[T - \Delta t_{off} - 2\Delta t_{on},\ T - \Delta t_{on}]$, $[T - \Delta t_{off} - \Delta t_{on},\ T]$, where Δt_{off} is the length of the initial part of each time sub-period during which the FTA works in an off-line modality for performing the optimal setting of the considered parameters; and Δt_{on} is the length of the final part of each time sub-period during which the FTA works in an on-line modality for making the financial trading.[2]

By so acting, our FTA performs uninterruptedly from Δt_{off} to T, and in the meantime the parameters are periodically updated.

4.1 The Drawdown-Like Phenomenon Management

Any well-working FTS should be able to minimize large losses during its running since these losses could reduce so much the capital at investor's disposal to make impossible the continuation of the trading itself. So, also our FTS should be able to minimize large losses, and in case they occur it should be able to guarantee the continuation of the trading.

In order to minimize large losses, the financial trading is not performed (and is definitively interrupted) at first instant time $t \in \{1, \ldots, T-1\}$ in which the net cumulative reward is negative. In such a case, the loss can be at most $-\mu$.

In order to (attempt to) guarantee the continuation of the financial trading in case a large loss occurs, we utilize as amount of capital to invest $\mu + \mu_0$, where $\mu_0 = \left|\min\left\{\min_{0<t\leq\Delta t_{off}}\left\{R_t :\ R_t < 0\ \wedge\ CR_t < 0\ \wedge\ F_{t-1}F_t = -1\right\};\ 0\right\}\right|$, i.e. the absolute value of the largest loss (associated to the occurrence that the net cumulative reward is negative and the trading strategy is changed) happened during the initial part of the first time sub-period. Of course, given such an amount of capital to invest, the loss can be at most $-(\mu + \mu_0)$.

4.2 The Results

In all the 9 applications we have used an amount of capital to invest, μ, equal to 1. As far as regards the optimal setting of the parameters, we have determined that, in general, $M \in \{2, 3, \ldots, 13, 14\}$, $\delta \in \{0.0050, 0.0075, \ldots, 0.0725,\ 0.0750\}$, $\rho_t = 0.01$, $\eta = 0.01$, $\Delta t_{off} = 500$, and $\Delta t_{on} = 65$; of course, μ_0 varies as the investigated financial market index varies. Moreover, in each application we suitably initialize the numerator and the denominator of (1).

In Table 1 we present the results of our applications. In particular, with respect to each investigated financial market index: in the first column we report the identifier; in the second column we report the CR_T, with $T =$ March 31,

[2] Of course, we need also to determine the optimal values of Δt_{off} and Δt_{on}.

2007, obtained from the application of our FTA; in the third column we report μ_0; in the fourth column we report the largest loss, i.e. $\min_{\Delta t_{off} \leq t \leq T} \{R_t : R_t < 0 \wedge F_{t-1}F_t = -1\}$; in the fifth column we report the CR_T, with $T =$ March 31, 2007, obtained from the application of our FTS (i.e. FTA − drawdown-like phenomenon management);[3] in the sixth column we report the number of performed trades; in the seventh/eighth/ninth column we report the percentages of the performed trades for which $R_t = 0/R_t = 0/R_t > 0$.

Table 1. The results of our applications

ID	FTA-based CR_T	μ_0	Largest loss	FTS-based CR_T	Number of trades	% $R_t < 0$	% $R_t = 0$	% $R_t = 0$
D	0.070	0.000	−0.172	0.043	22	0.455	0.000	0.545
F	0.482	0.003	−0.139	0.545	19	0.579	0.000	0.421
J	0.480	0.000	−0.431	0.480	101	0.436	0.109	0.455
M	0.980	0.000	−0.137	0.782	79	0.430	0.076	0.494
NA	0.229	0.016	−0.402	0.406	86	0.488	0.012	0.500
NI	−1.036	0.036	−0.352	−1.264	86	0.570	0.023	0.407
Sh	0.967	0.048	−0.521	1.295	23	0.478	0.000	0.522
Sw	0.922	0.000	−0.297	0.597	19	0.368	0.000	0.632
T	0.658	0.129	−0.179	0.099	88	0.477	0.091	0.432

As far as concerns these results, it is to notice:

- that only 1 of the 9 investigated financial market indices shows a negative FTA-based CR_T, and that the summation of all the FTA-based CR_Ts is positive and equal to 3.752. It can be interpreted as the fact that, under a suitable diversification in the investments, our FTA is well-working in the long-run term;
- that the magnitude of this summation is not satisfying. It depends both on the choice to utilize in the expression for CR_T a free-risk rate of return, i_f, equal to 0, and on the need to effect the setting of the parameters in some more refined way;
- that our drawdown-like phenomenon management is enough well-performing. In fact, as the summation of the FTA-based CR_Ts is greater than the summation of the FTS-based CR_Ts, our FTA appears able to reduce the losses which occur during its running (although it is not able to guarantee the continuation of the financial trading of the only critical case, the one related to the NIKKEI 225 Stock Average index).

[3] The FTS-based CR_T has to be considered as a comparison term for the FTA-based CR_T. It is to notice that the FTS-based CR_T is sometimes unrealistic because, being lacking the management of the drawdown-like phenomena, the associated financial trading could continue to T even if $CR_t < 0$, with $t \in \{1, \ldots T-1\}$.

5 Concluding Remarks

In this section we provide some concluding remarks. In particular:

- in order to reduce the percentage of performed trades for which $R_t < 0$, we conjecture that could be fruitful to utilize the following so-modified trading strategy:
  - if $y_t < -\varepsilon^-$, with $\varepsilon^- > 0$, then $F_t = -1$;
  - if $\varepsilon^- \leq y_t \leq \varepsilon^+$, with $\varepsilon^+ > 0$, then $F_t = F_{t-1}$;
  - if $y_t > \varepsilon^+$ then $F_t = 1$;[4]
- in order to improve the learning capabilities of the ANN, we guess that could be profitable to use a multi-layer perceptron model (it is to notice that such a check is carried out in [4], but without meaningful results);
- in order to make more informative the set of the variables related to the time-evolution of the quantities of interest, we conjecture that could be fruitful to consider, beyond the logarithmic rate of return, at least the transaction volume;
- finally, in order to explicitly take into account the risk in our FTA, we guess that could be profitable to utilize some risk-adjusted version of (1).

References

1. Abecasis, S.M., Lapenta, E.S., Pedreira, C.E.: Performance metrics for financial time series forecasting. Journal of Computational Intelligence in Finance, 5–23 (July/August 1999)
2. Bishop, C.: Neural networks for pattern recognition. Oxford University Press, Oxford (1995)
3. Corazza, M., Vanni, P., Loschi, U.: Hybrid automatic trading system: technical analysis & group method of data handling. In: Marinaro, M., Tagliaferro, R. (eds.) Neural nets. 13th Italian workshop on neural nets, WIRN Vietri 2002, Vietri sul Mare, Italy, May/June 2002, pp. 47–55. Springer, Berlin (2002) (revised paper)
4. Gold, C.: FX trading via recurrent reinforcement learning. In: Proceedings of IEEE international conference on computational intelligence in financial engineering, pp. 363–370. IEEE Computer Society Press, Los Alamitos (2003)
5. Lo, W.A., Mamaysky, H., Wang, J.: Foundations of technical analysis: computational algorithms, statistical inference, and empirical implementation. Journal of Finance LV, 1705–1769 (2000)
6. Moody, J., Saffell, M.: Learning to trade via direct reinforcement. IEEE Transactions on Neural Networks 12, 875–889 (2001)
7. Moody, J., Wu, L., Liao, Y., Saffell, M.: Performance functions and reinforcement learning for trading systems and portfolios. Journal of Forecasting 17, 441–470 (1998)
8. Murphy, J.J.: Study guide to technical analysis of the financial markets. Prentice Hall Press, New York (1999)
9. Sutton, R.S., Barto, A.G.: An introduction to reinforcement learning. MIT Press, Cambridge (1997)

[4] Of course, in such a case we should need to determine the optimal values of two other parameters, ε^- and ε^+.

Multiple Ant Colony Optimization for a Rich Vehicle Routing Problem: A Case Study

Paola Pellegrini, Daniela Favaretto, and Elena Moretti

Department of Applied Mathematics, University Ca' Foscari of Venice
Dorsoduro 3825/E, I–30123 Venice, Italy
paolap@pellegrini.it, favaret@unive.it, emoretti@unive.it

Abstract. Starting from a case study, a rich vehicle routing problem is analyzed. It is characterized by multiple time windows, heterogeneous fleet, maximum duration, and multiple visits. Two variants of Ant Colony Optimization are proposed in a multiple colonies framework. Two algorithms are tested, giving results that appear satisfactory with respect to the ones achieved by the firm.

Keywords: ant colony optimization, vehicle routing, case study.

1 Introduction

The analysis at the basis of this paper is motivated by the need of solving a case study, which can be modeled as a rich vehicle routing problem.

The vehicle routing problem (VRP) consists in the determination of the optimal set of routes to serve a given set of customers using a fixed fleet of vehicles. In recent years, thanks to the increasing efficiency of heuristic and metaheuristic approaches and the availability of a larger computing power, the interest has been shifted to variants which combine various types of constraints, such as: time windows in which the service must be performed, the requirement of inserting in the same tours both pickup and delivery of goods, the availability of different types of vehicles, the presence of more than one depot. These variants are grouped under the denomination of rich VRP.

The rich vehicle routing problem studied in this paper is characterized by: the presence of multiple time windows, the availability of different types of vehicles, the requirement of multiple visits to some customers and the limitation of the duration of each subtour. Moreover, two objectives are to be considered in hierarchic order. The first one is the minimization of the number of vehicles used, and the second one is the minimization of the total time required by the subtours.

We use the metaheuristic approach known as Ant Colony Optimization [1] to tackle this problem. We consider two of its most successful variants: Ant Colony System [2] and $\mathcal{MAX}$–$\mathcal{MIN}$ Ant System [3]. They are both studied in a multiple colonies framework. The respective results are compared on a set of instances generated on the basis of real ones. The Randomized Nearest Neighbor heuristic [4] and a Tabu Search algorithm are used as reference elements.

B. Apolloni et al. (Eds.): KES 2007/ WIRN 2007, Part II, LNAI 4693, pp. 627–634, 2007.
© Springer-Verlag Berlin Heidelberg 2007

The algorithm which performs the best on the generated instances, is then used for a comparison with the solutions proposed by the firm considered.

First of all, the case study is presented. Secondly, the resulting rich vehicle routing problem is proposed. Then the two Ant Colony Optimization algorithms considered and Tabu Search approach are described. A following section describes the instances used for the analysis and another one reports the computational results obtained. Finally some conclusions are drawn.

2 The Case Study

The case study at the basis of this analysis is concerned to an Italian firm. Its major activity consists in the delivery of a wide number of food products to restaurants and retailers in the North-East region of the country. This task is accomplished by external suppliers of vehicles. The cost for completing the services consists in a fixed amount related to each vehicle used and in a variable one due to the working hours of the drivers. The firm decides the set of customers to be assigned to each vehicle and the sequence according to which the customers must be visited. The fee owed to a driver is much lower for the first eight hours than for the following ones, and then the firm imposes tours of limited duration.

This structure allows to consider as virtually unlimited the dimension of the fleet available. The usable vehicles are of two kinds. They differ for what concerns the external dimensions and the load and unload system. The use of the a specific type may be necessary for serving a precise customer.

As it is widespread practise, the firm allows customers to choose when to receive the delivery. To this aim, they can indicate at most three time intervals for each of the five working days of a week. This gives a total of at most fifteen time windows per customer. Moreover, each customer can require to be served more than once a week. When this is the case, the firm must take care of placing the visits in non-consecutive days unless the customer himself explicitly requires consecutive visits.

The aim of the firm is the minimization of the cost of completing all the deliveries. The fix cost, related to the utilization of the vehicles, has a stronger impact than the variable one concerning the total time needed. This makes a set of few tours requiring a certain time preferable to another set including more tours and implying a smaller amount of hours. The general aim then can be split in two objectives which have to be pursued hierarchically: the first is the minimization of the number of vehicles required to complete the services, the second is the minimization of the total time needed.

The procedure currently applied by the firm to establish the tours consists first of all in splitting the customers in groups of seventy to eighty, and then in deciding the assignments to the different vehicles according to a weekly time horizon. Tours are scheduled on the basis of the experience and the knowledge of the territory of the logistic manager. In agreement with the firm, the aim of the research presented in this paper is proposing an algorithm for solving this second step, considering the first clusterization as non-modifiable.

3 The Rich Vehicle Routing Problem

To the best of our knowledge, the rich vehicle routing problem that is object of this study has not been considered in the literature. The element which differentiates the most the current variant from the ones that can be found in the literature, is the objective function. Often, only one objective is considered. Moreover, the quantity which is typically minimized is the total distance (or traveling time), while in this study we must take into account also the waiting time.

Let us focus on the most particular constraints of our problem. A first type of constraint is related to the requirement of multiple visits to a single customer. These visits must be scheduled at a suitable time distance, as discussed in the following. The time horizon considered is split in disjoint subperiods, which represent different days. For each day, customers may indicate one or more time windows in which to be served. Each subtour must belong to only one of the subperiods and its total time length must not exceed a fixed duration. Furthermore, an heterogeneous fleet is available, and each customer may require to be served by a specific type of vehicle. A specific type of vehicle, then, must be associated to each subtour.

A mathematical formulation of a very similar model can be found in [5].

4 Multiple Ant Colony System and Multiple $\mathcal{MAX}$–$\mathcal{MIN}$ Ant System

Ant Colony Optimization (ACO) is a metaheuristic based on the foraging behavior of ants. A complete analysis of this metaheuristic can be found in [1]. Many ACO algorithms have been proposed in the literature. The two most successful variants, when dealing with routing problems are recognized to be Ant Colony System (ACS) first proposed in [6] and $\mathcal{MAX}$–$\mathcal{MIN}$ Ant System ($\mathcal{MM}$AS) first introduced in [3]. In this study, the *swap* local search is combined with the algorithm. This procedure is well known in the literature and it was first proposed by Lin in 1965. It offers the advantage of not being very expensive in terms of computational time although it offers satisfactory results. In our case this element is crucial since the addition of many constraints could make other typical local search procedures very time consuming.

In order to apply the two algorithms to the rich VRP considered, some elements must be clarified.

For taking into account the differences among vehicles, the following procedure is applied. First of all, one type of vehicle $\hat{k}$ is fixed as default. Then the insertion of the customers begins considering the capacity constraint imposed by type $\hat{k}$, and without taking into account the type of vehicle possibly required by the customers. When the first customer requiring a specific type of vehicle is inserted, this type is considered for the subtour. The following customers will be feasible either if they do not require a specific class of vehicles or if they require the class already established. When considering customers for the insertion while

the vehicle is still the default one, if the cumulated demand is larger than the capacity of some type of vehicle ($\tilde{k}$), then all the customers requiring $\tilde{k}$ are set unfeasible.

Periodic constraints must be satisfied. Before starting all the procedures, the set of customers is analyzed. Each of those requiring more than one visit is duplicated once for each visit required. To each dummy customer the edges and some time windows of the original one are assigned. If the partition of the time windows allows a suitable separation between the services, it is recorded that the visits must happen with at least a certain distance. This distance depends from the number of visits required. For a formal description of this procedure we refer the reader to [5].

As far as the multiple time windows are concerned, the procedure used is quite straightforward. It consists in choosing dynamically the one to be considered. This time window corresponds to the first one closing after the instant in which the customer can be reached, if such time window exists. The number and duration of the not yet closed time windows is nonetheless taken into account. In particular, the choice is biased in favor of customers that would be the hardest to serve in the future, considering all the time windows they indicate. This is done by introducing a heuristic measure proposed in [5] in the random proportional rule (pseudo-random proportional rule for ant colony system) [1].

The constraint on the duration of each subtour is treated by stating that, if visiting customer j and going back to the depot implies that the current subtour lasts more than the available time, then j is unfeasible.

Finally, the algorithms must operate considering two objectives in hierarchical order. They represent the minimization of the number of vehicles used, and the minimization of the time required to complete the tour. To this aim, we follow the approach presented in [7]: two kinds of colonies with different specializations are exploited. This system has been developed to tackle a multi-objective vehicle routing problem with time windows using Ant Colony System and has never been implemented for other ACO variants, despite reaching very encouraging results.

The two algorithms obtained following these steps will be referred as Multiple Ant Colony System (M-ACS) and Multiple $\mathcal{MAX}$–$\mathcal{MIN}$ Ant System (M-$\mathcal{MM}$AS).

5 Tabu Search

For evaluating the performance of the M-ACS and M-$\mathcal{MM}$AS, we consider as reference element a Tabu Search algorithm. Tabu Search [8] is one of the most successful metaheuristics for the vehicle routing problem with single time windows [9]. For this reason it is considered as a reference point for the algorithms proposed here. In the algorithm, the starting solution is generated with the Nearest Neighbor heuristic and its neighborhood is explored via the *swap* local search procedure. Following [10], a *variable neighborhood* is considered: for any current solution, instead of considering the whole neighborhood, we choose a subset of it according to a given probability distribution.

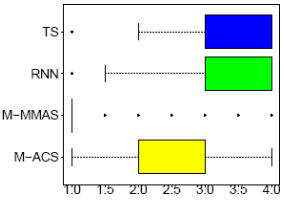


Fig. 1. Ranking of the performance of the algorithms

6 Experimental Results

In order to choose which algorithm to suggest to the firm, we compared the performance of Multiple Ant Colony System, Multiple $\mathcal{MAX}$–$\mathcal{MIN}$ Ant System, Tabu Search and Randomized Nearest Neighbor (RNN) [4] on a set of instances generated on the basis of the real ones.

As reported when the case study has been presented, we are interested in sets of customers with cardinality of the order of 70 to 80 elements, as for the subsets of customers considered by the firm. The area actually served by the firm is used as a reference for the localization of the customers. We use for the analysis a set of instances generated in order to mimic the real instances proposed by the firm [4].

Experiments on Generated Instances

The experiments are run on a cluster of 32 AMD Opteron^{TM} 244. The executables have been generated from C++ source. The code is publicly available on the web page www.paola.pellegrini.it. The common practice used for dealing with metaheuristics is followed [11,12,13]. A set of 400 instances is used for the experiments.

The behavior of the Deterministic Nearest Neighbor (DNN) is considered as reference. The element that first emerges is the dominance of the M-$\mathcal{MM}$AS over the other algorithms. In Figure 1 the ranking of the different approaches is reported. As it appears evident in the strong majority of the cases the M-$\mathcal{MM}$AS outperforms the other approaches. The difference appears statistically significant according to the Friedman two-way analysis of variance by ranks. This better behavior is observable in Table 1, in which the number of instances with an improvement with respect to DNN, and the average percentage of these improvement, are presented for each algorithm. The improvement with respect to the first objective (minimum number of vehicles) is reported in the first line. The one concerning the minimum traveling and waiting time is given in the second line. Owing to the priority of the minimization of the number of vehicles, if a solution is preferable in this sense the travel time is not considered. The second objective is checked when no difference is detectable about the first. As it can be seen, in general M-$\mathcal{MM}$AS and M-ACS behave in a comparable way. They improve significantly the number of vehicles used in the best solution

Table 1. Number of instances in which any improvement with respect to the DNN has been found and average percentage improvement

M-ACS		M-$\mathcal{MM}$AS		RNN		TS	
veh.	%	veh.	%	veh.	%	veh.	%
395	21.36	375	25.06	213	12.66	321	12.84
time	%	time	%	time	%	time	%
0	0.00	0	0.00	97	11.91	42	8.84

found and do not make a real difference with respect to the total travel time when they are not able to make a difference in terms of the first objective. It can be observed, moreover, that even if the number of instances in which the number of vehicles decreases with respect to the DNN is very similar for the two algorithms, the average value of this difference is much higher when considering M-$\mathcal{MM}$AS. A completely different behavior is observable for RNN, where the number of vehicles is quite seldom improved (in about 50% of the cases). When this happens the difference is less stressed than for the ACO algorithms. On the other hand the total travel time is diminished much more often. Finally, TS places itself between the ACO algorithms and RNN. The number of instances in which the number of vehicles is improved is close to the one detected for ACO algorithms, and once it is even bigger. Nonetheless, the average improvement is always quite lower. The value referred to the total travel time in this case is comparable to the one concerning RNN. The differences reported in Table 1 concerning the number of vehicles, are significant according to the Wilcoxon test.

Experiments on Real Instances

Let us try to compare the approaches presented with the one currently applied by the firm object of the case study. It is not very easy to analyze the differences, since following the tours provided by the firm about 20% of the time windows are violated. In any case, the ACO algorithms find feasible solutions which imply in average the same number of vehicles currently used, even if the total travel time needed is bigger. TS is the third algorithm, followed by RNN and DNN, all supplying tours requiring more vehicles than those currently used but clearly not violating any constraint. For giving a clearer idea of the different solutions, let us consider a representative instance. It is graphically shown in Figure 2 and the characteristics of the customers follow the trends reported in the description of the instances generated. The tours proposed by the firm and the ones found by $\mathcal{MAX}$–$\mathcal{MIN}$ Ant System are graphically shown in Figure 3. The customers whose time windows are violated are represented by black bullets. As it can be seen, in the tour proposed by the firm there is a sort of equilibrium in the number of customers served by each vehicle. In the solution found by $\mathcal{MAX}$–$\mathcal{MIN}$ Ant System this element is not detectable. The disequilibrium may appear as a weakness of the set of tours proposed by the ACO algorithm. On the other hand, the fact that only a few customers are inserted in the last two tours, implies that relaxing only a few constraints, such as some time windows, it might be

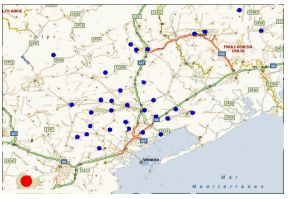
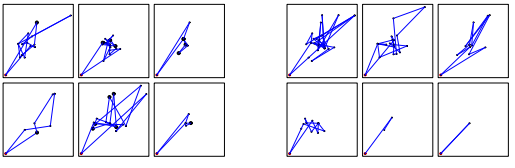


Fig. 2. Graphical representation of the instance considered



(a) (b)

Fig. 3. Solution proposed by the firm and by $\mathcal{MAX}$–$\mathcal{MIN}$ Ant System

possible to eliminate them, and then to complete the services with two vehicles less. Moreover, by slightly changing the clusters which currently are considered fixed under suggestion of the firm – as previously explained – it would probably be possible to find more efficient solutions.

7 Conclusion

Two Multiple-ACO algorithms are presented for dealing with a case study which can be modeled as a rich vehicle routing problem. Its characteristics are: multiple hierarchically ordered objectives, multiple time windows, heterogeneous fleet of vehicles, periodic constraints and limited duration of the subtours. As reference elements the Deterministic and Randomized Nearest Neighbor heuristics and a Tabu Search algorithms are considered. The ACO algorithms proposed perform significantly better than these alternatives, both on instances generated for the theoretical analysis and on those available for the case study. The results of this study are currently being exploited by the firm which proposed the analysis.

Acknowledgements. The authors would like to thank Mauro Birattari for the useful advice during the preparation of this paper. They also thanks IRIDIA, Université Libre de Bruxelles, Belgium, for allowing the use of computing power.

References

1. Dorigo, M., Stützle, T.: Ant Colony Optimization. MIT Press, Cambridge (2004)
2. Dorigo, M., Gambardella, L.M.: Ant colonies for the traveling salesman problem. ByoSystems 43, 73–81 (1997)

3. Stützle, T., Hoos, H.: The $\mathcal{MAX}$–$\mathcal{MIN}$ Ant System ant system and local search for the traveling salesman problem. In: Angeline, P. (ed.) Proceedings of the IEEE International Conference on Evolutionary Computation, Indianapolis, USA, pp. 308–313. Springer, Heidelberg (1997)
4. Pellegrini, P.: Application of two nearest neighbor approaches to a rich vehicle routing problem. Technical Report TR/IRIDIA/2005-15, Iridia, Université Libre de Bruxelles, Brussels, Belgium (2005)
5. Favaretto, D., Moretti, E., Pellegrini, P.: Ant colony system for a vrp with multiple time windows and multiple visits. Journal of Interdisciplinary Mathematics (to appear, 2006), Available on `http://www.paola.pellegrini.it`
6. Dorigo, M., Gambardella, L.M.: Ant colony system: A cooperative learning approach to the traveling salesman problem. IEEE Transaction on Evolutionary Computation 1, 53–66 (1997)
7. Gambardella, L.M., Taillard, E., Agazzi, G.: MACS-VRPTW: A multiple ant colony system for vehicle routing problem with time windows. In: Corne, D., Dorigo, M., Glover, F. (eds.) New Ideas in Optimization, pp. 63–76. McGraw-Hill, London (1999)
8. Glover, F.: Future paths for integer programming and links to artificial intelligence. Computers & Operations Research 13, 533–549 (1986)
9. Cordeau, J.F., Laporte, G., Mercier, A.: A unified tabu search heuristic for vehicle routing problem with time windows. Journal of the Operational Research Society 59, 928–936 (2001)
10. Bianchi, L., Birattari, M., Chiarandini, M., Manfrin, M., Mastrolilli, M., Paquete, L., Rossi-Doria, O., Schiavinotto, T.: Hybrid metaheuristics for the vehicle routing problem with stochastic demand. Journal of Mathematical Modelling and Algorithms 5(1), 91–110 (2006)
11. Birattari, M.: On the estimation of the expected performance of a metaheuristic on a class of instances. How many instances, how many runs? Technical Report TR/IRIDIA/2004-01, Université Libre de Bruxelles (2004)
12. Birattari, M., Stützle, T., Paquete, L., Varrentrapp, K.: A racing algorithm for configuring metaheuristics. In: Langdon, W. (ed.) GECCO 2002: Proceedings of the Genetic and Evolutionary Computation Conference, pp. 11–18. Morgan Kaufmann Publishers, San Francisco, CA (2002)
13. Birattari, M., Zlochin, M., Dorigo, M.: Towards a theory of practice in metaheuristics design: A machine learning perspective. Theoretical Informatics and Applications 40(2), 353–369 (2006)

Portfolio Optimization Through Elastic Maps: Some Evidence from the Italian Stock Exchange

Marina Resta

Department of Applied Economics and Quantitative Methods, University of Genova
Via Vivaldi 2, 16126, Genova, Italy
mresta@unige.it

Abstract. In this paper we discuss the use of elastic maps as support tool in the decision process underlying the selection, optimization, and management of financial portfolios. In particular, we suggest an allocation scheme which is interely driven by neural networks, in contrast to the traditional model where investors distribute their money among assets chosen according to the mean and variance of their returns. Our optimization procedure is based on the selection of assets from clusters originated by the nets, according to their proximity to the nodes of the map; this, in turn, is the criterion thanks to which we assign the proper weight to each asset into the portfolio. In order to check the profitability of the approach, we have empirically tested the method with stocks from the Italian Stock Exchange; market reference index has been then used to build proper performance benchmarks. Our main results may be summarised as follows: (*i*) our approach has revealed to be generally more informative than classical mean−variance method, since it allows to take into account additional variables in the selection procedure; (*ii*) our procedure can work both in a static framework (i.e. for one time choice), and into a dynamic context (i.e. to the purpose of re−calibration of original decisions). The overall performances appear to be superior to the benchmark in both the static and dynamic case.

Keywords: Elastic Maps, Portfolio Optimization, Italian Stock Exchange.

1 Motivation of the Study

The Markowitz model for portfolio selection and optimization is considered a milestone of modern financial theory. Markowitz [3] suggested to view at investors stocks selection procedure as a diversification task: within the final objective of risk minimization, each market operator makes his choice among the available stocks, using as discriminating factor their average return. In a more formal view, being $A = \{S_1, S_2, \ldots, S_n\}$ the set of all possible assets in the market[1], each asset i $(i = 1, \ldots, n)$, may be associated to a history, i.e. a timeseries

[1] Although asset is a generic expression that can be used to indicate every kind of market financial instrument, we assume to refer to stocks.

B. Apolloni et al. (Eds.): KES 2007/ WIRN 2007, Part II, LNAI 4693, pp. 635–641, 2007.
© Springer-Verlag Berlin Heidelberg 2007

of k observable price levels: $\{p_i(t)\}_{t=1,\ldots,k}$, and then to the corresponding series of returns $r_i(t) = [p_i(t) - p_i(t-1)]/p_i(t-1)$, $t = 2, \ldots, k$. If x_i $(i = 1, \ldots, n)$ is the (unknown) quantity that the trader can invest into each stock, then we will have:

$$\begin{aligned} & \min_{\mathbf{x}} \sum_{i=1}^{n} \sum_{j=1}^{n} x_i x_j \sigma_{i,j}^2 \\ s.t. & \\ & \sum_{i=1}^{n} x_i \quad = 1 \\ & \sum_{i=1}^{n} x_i \mu_i = \mu_\pi \end{aligned} \tag{1}$$

where $\mathbf{x} = \{x_i\}, i = 1, \ldots, n$, $\sigma_{i,j}^2$ is the covariance between assets i and j, μ_i is the mean return value on asset i, and $\mu_\pi \geq 0$ is the so called *portfolio mean return*, and indicates the average return threshold value that the investor expects *at least* to obtain by speculating in the market.

Criticisms have been moved to the model described above: the main point is that labeling the assets according to the mean and variance of the series of their returns can be proficient only within a Gaussian framework, i.e. assuming that stocks returns are *iid* (independent and identically distributed) normal random variables. This, in turn, means that the procedure is not able to gain advantage of additional information that may come by observing higher empirical moments, or by considering ad−hoc variables that may be built through proper transformations of the original input. However, the Markowitz model still remains widely used in practical financial applications, offering to practitioners a friendly and easy to use selection procedure.

The main point of our work is to verify if it is possible to implement a procedure of asset selection that conjugates a set of highly desiderable features: ease of use, strong readability of results, and procedural robustness. In our opinion, such desiderable requirements can be well fit by the method we are going to describe, which is based on elastic maps. Our optimization procedure selects assets from clusters originated by the nets, according to the significance of clusters and to proximity of assets to the nodes of the map; this, in turn, is the criterion thanks to which we assign the proper weight to each asset into the portfolio.

The structure of the paper is as follows. Section II briefly introduces elastic maps, presenting main features of the algorithm. Section III discusses the proposed methodology, and deals with a practical application with stocks quoted in the Italian Stock Exchange. Finally, Section IV concludes.

2 Elastic Maps: The General Framework

Formally, an elastic map [1] is an undirected graph $G(Y, E)$, where $Y = \{Y^{(j)}, j = 1, \ldots, N\}$ denotes the collection of graph nodes, and $E = \{E^{(j)}, j = 1, \ldots, s\}$ is the collection of graph edges, each of them with a beginning node $E^{(j)}(0)$ and

final node $E^{(j)}(1)$. Additionally, each pair of adjacent edges $R^{(j)} = \{E^{(j)}, E^{(k)}\}$ is an elementary rib, with a beginning node $R^{(j)}(1)$, end node $R^{(j)}(2)$ and the central node $R^{(j)}(0)$. The algorithm merges a finite input set $\Xi = \{\xi_r\}_{r<\infty}$ into the neural space through an iterative procedure whose main steps are here recalled.

1. The first step consists in the division of input samples into subcollections satisfying the following criterion respect to map nodes:

$$K_j = \{\xi^{(j)} : \min_Y ||\xi^{(j)} - Y^{(j)}||\} \tag{2}$$

2. The second step moves to define the graph G energy function U, that summarizes energies of every node, edge and rib: $U = U^{(Y)} + U^{(E)} + U^{(R)}$ with:

$$\begin{aligned} U^{(Y)} &= \tfrac{1}{N}\sum_{j=1}^{N} \sum_{\xi^{(r)} \in K^j} ||\xi^{(r)} - Y^{(j)}||^2 \\ U^{(E)} &= \sum_{j=1}^{s} \lambda_j ||E^{(j)}(1) - E^{(j)}(0)||^2 \\ U^{(R)} &= \sum_{j=1}^{s} \mu_j ||R^{(j)}(1) + R^{(j)}(2) - 2R^{(j)}(0)||^2 \end{aligned} \tag{3}$$

   where values λ_j and μ_j are, respectively, coefficients of stretching elasticity of every edge $E^{(j)}$, and coefficients of bending elasticity of every rib $R^{(j)}$. We assume: $\lambda_1 = \lambda_2 = \ldots = \lambda_s = \lambda(s)$, and $\mu_1 = \mu_2 = \ldots = \mu_r = \mu(r)$. While $U^{(Y)}$ is somewhat self−explaining, $U^{(E)}$ may be correctly thought as the analogue of summary energy of elastic stretching, and $U^{(R)}$ as the summary energy of elastic deformation of the net. This makes every node connected by elastic bonds to the closest input data points, and simultaneously to the adjacent nodes.
3. The final step consists in the minimization of graph energy U, calculating new positions of nodes accordingly, remembering that:

$$\begin{aligned} \tfrac{1}{2}\tfrac{\partial U^{(Y)}}{\partial Y^{(j)}} &= N_j Y^{(j)} - \sum_{\xi^{(r)} \in K^{(j)}} \xi^{(r)} \\ \tfrac{1}{2}\tfrac{\partial U^{(E)}}{\partial Y^{(j)}} &= \sum_{k=1}^{N} Y^{(k)} \sum_{i=1}^{s} \lambda_i \Delta E^{ij} \Delta E^{ik} \\ \tfrac{1}{2}\tfrac{\partial U^{(R)}}{\partial Y^{(j)}} &= \sum_{k=1}^{N} Y^{(k)} \sum_{i=1}^{r} \mu_i \Delta R^{ij} \Delta R^{ik} \end{aligned} \tag{4}$$

   where N_j is the number of points in $K^{(j)}$.
4. If changes in U values become less than a small (arbitrary) ε, the procedure stops, otherwise steps 1.−3. are repeated until such threshold value is reached.

3 Data and Methodology

3.1 The Experimental Procedure

Figure 1 gives an overview about the procedure we have implemented. The idea is quite simple. We assume that each asset is a d−dimensional point into the input space. Those points are mapped into the neural space, according to the procedure explained in the previous section. This originates a projection manifold, where the points (assets) are clustered according to the minimization of energy graph U. The so obtained clusters are variously significant, since they can contain a greater or lesser number of nodes of the map: the greater is the number of nodes belonging to the cluster, the higher is the significance of the cluster itself. In the same way, we can assess the representativity of the nodes within each cluster. However, since the representativity of the node depends on the distance among the assets and their referring node, in order to manage homogeneous data we have rescaled within the range $[0, 1]$. We have then used the so obtained values to generate a rank for both clusters, nodes and assets. This ranking, in turn, makes possible to define the final portfolio weights x_i, up to their total sum of one:

$$x_i = max\left(0, x_i^{(init)} + \frac{1}{cr_i - nrc - 1} + \left(1 - \frac{asr_i}{nrn}\right) \cdot resc_i\right) \tag{5}$$

where: nrc is the number of identified clusters, cr_i and asr_i are the cluster and asset i ranking respectively, nrn is the number of nodes within the cluster i belongs to, $x_i^{(init)} = 1/(nrc+1)$, and, finally, $resc_i$ is the rescaled distance between asset i and its corresponding node. Clearly, Eq.(5) punishes/rewards according to the previously introduced concepts of significance and representativity.

The main difficulty at this stage relies in the choice of inputs dimension, i.e. in the choice of the number of explicative factors needed to give a good representation of market stocks: the assets manifold is non−linear separable, but also there isn't any *a priori* knowledge about its dimensions. To this aim, we have run a heuristics, that spans over the largest number of candidate input components, and eliminates those satisfying the following criteria: (a) exhibition of strong ($> |0.75|$) correlation with at least 60% of remaining variables; (b) low significance of induced clustering. In this way, we have analyzed 24 perspective input components that evaluate over 200,100, 50 and 25 days the following variables: mean, standard deviation, skewness, intensity: $(r_i(k) - r_i(1))/r_i(1)$, return to risk ratio: μ_i/σ_i, and downside risk [5]:$\sum_{i=1}^{k}\sum_{j=1}^{k}(\sigma_{i,j}^{-})^2$, where: $\sigma_{i,j}^{-} = \min(0, \sigma_{i,j})$. In order to discriminate according to criterion (a), we have examined the 24×24 covariance matrix associated to such variables, while the analysis of the significance of induced clusters led us to choose the elements to be included into the input vectors.

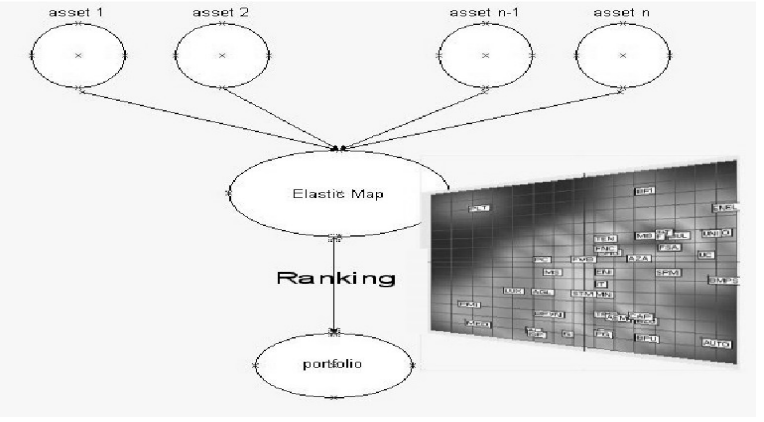


Fig. 1. A sketch about the implemented procedure

3.2 Case Study

We refer to the case of Italian Stock Exchange. Our data sample consists of the 40 assets included into the Italian S&P MIB index that since 20 September 2004 has replaced the MIB30 index as the new benchmark for the Italian stock market. The resulting 40 timeseries are made up by 293 daily observations, taken in the period 30 September 2005−21 November 2006. Observations from 22 November 2006 to 15 January 2007 have been used as test set. According to the criteria (a) and (b) given in the previous section, the following variables have been associated to each asset:

1. mean return over 25 days;
2. standard deviation over 50 days;
3. standard deviation over 25 days;
4. skewness over 25 days;
5. return to risk ratio at 200 days;
6. return to risk ratio at 25 days;
7. downside risk at 50 days;
8. downside risk at 25 days.

Such 8−dimensional arrays (one for each asset) have been employed to run simulations in a static environment (i.e. for one−time choice of portfolio composition). Also a dynamic hypothesis has been considered, according to which every portfolio has been re−calibrated every 15 days: in this latter case, each asset was presented to the map four times, through four different arrays of the variables given above, evaluated by sliding windows along the overall length of timeseries.

We have monitored various maps, with different degrees of elasticity, thus obtaining different portfolios. Related performances have been evaluated by analyzing, for each portfolio, the corresponding average return (AR), and two

Table 1. A comparison among performances of elastic net driven portfolios and standard portfolios. In the second column, R,S, and SS are used as abbreviations for Rigid, Soft, and Super Soft, and refer to the kind of connections between ribs and nodes. NR ASSETS in the fifth column shows the number of assets per portfolio with non−zero positive weight.

ID	Type	Dyn/Static	Nr Clusters	Nr Assets	AR	SR	L
M1010a	10x10 R	static	3	6	-0.005	-0.090	0.91
M1010b	10x10 S	static	3	9	-0.003	-0.034	3.00
M1010c	10x10 SS	static	5	15	-0.001	-0.070	2.11
M1212a	12x12 R	static	9	13	0.020	0.030	2.93
M1212b	12x12 S	static	5	9	0.022	0.051	1.78
M1212c	12x12 SS	static	5	14	0.012	0.070	2.41
M1616a	16x16 R	static	6	11	0.017	0.051	3.12
M1616b	16x16 S	static	6	15	0.015	0.055	2.27
M1616c	16x16 SS	static	6	12	0.016	0.096	1.85
M2525a	25x25 R	static	5	15	0.005	0.000	1.15
M2525b	25x25 S	static	6	14	0.009	0.036	2.00
M2525c	25x25 SS	static	6	15	-0.005	0.050	1.32
DM1010b	10x10 S	dyn	4	9	0.001	-0.023	2.79
DM1212b	12x12 S	dyn	4	11	0.019	0.112	1.13
DM1616b	16x16 S	dyn	4	11	0.021	0.134	1.98
DM2525b	25x25 S	dyn	6	15	-0.003	-0.373	1.44
ptf1	–	–	–	9	-0.002	-0.358	2.92
ptf2	–	–	–	9	0.020	0.030	1.14
ptf3	–	–	–	15	0.013	0.039	2.95
ptf4	–	–	–	15	0.003	-0.100	1.22

well−known conventional indexes: the Sharpe ratio [4] SR, and the leverage factor L, being:

$$SR = \frac{\mu_{ptf} - rf}{\sigma_{ptf}};\ L = \frac{\sigma_{mkt}}{\sigma_{ptf}} \tag{6}$$

where μ_{ptf}, σ_{ptf} are, respectively, the portfolio mean return and standard deviation, rf is the risk−free rate[2], and, finally, σ_{mkt} is the reference index (S&P MIB) standard deviation. The results are given in Table 1, where we have reported performances of elastic maps both in the static and in the dynamic case. In this latter case, we have evidenced results only for maps with soft strength connections, i.e. for those that have proved to work at the best also in the static hypothesis; additionally, performances refer to the overall period. The scores produced by running classical Markowitz portfolios have been also considered and reported at the bottom of Table 1. In general, elastic maps have given better results than standard technique. This is particularly true for 12×12 and 16×16 maps. An interesting point is then related to the leverage effect that suggest the particular profitability of almost all the examined portfolios.

[2] We have set $rf = 0.005$, considering it equivalent on bi−monthly basis, to the annual rate $rf_y = 0.03$.

4 Conclusions

We have presented a technique that allows to manage financial portfolios in a way completely driven by neural networks. Our scheme is based on elastic maps that in turn are strongly related to the more known Kohonen's Self Organizing Maps[2], but even more reliable for a number of features. In particular, this is due to the possibility to control the strength of each node, edge and rib of the map, thus producing a more *plastic* representation of input manifold. Starting from this point, we have developed a method that clusters assets into the neural space, and uses their proximity to the reference nodes to select portfolio stocks, and to weight them into the portfolio. We have applied this method to data observable on the Italian Stock Exchange, using 293 daily observations to train the net, and 60 observations to test its performance, opposed to that of portfolios weighted according to the standard Markowitz procedure. The results, in our opinion, are interesting from different points of views. Firstly, our optimization scheme is generally more informative, because it admits the selection of stocks through a wider set of significative factors. Looking deepest to the results, we have evidenced that plastic connections can fit well the input manifold, thus assuring good overall performances. Additionally, better performances gained by nets of "intermediate"dimensions (12×12, and 16×16, rather than 10×10 or 25×25) suggest the rule of thumb of putting worst at the extremes, in financial forecasting tasks. However, almost three questions remain open. The first question is related to the real potential of improving portfolios results by adding/changing or refining the variables used as explicative factors. Additionally, those problems need to be deeply investigated into the dynamic framework, where a great deal of work both from the theoretical and the practical point of view is still awaiting. Finally, since the technique presented belongs to the group of data driven optimization methods, the big deal will be that of comparing the results of our scheme with those of most known data driven optimization models. At the present stage, we are working toward such direction, by opposing the technique relying on elastic maps to optimization models based on genetic algorithms, swarm optimization and support vector machines. The earlier (and very preliminary) results appear to be promising, and we expect to detail them in a deepest fashion sooner in a further work.

References

1. Gorban, A.N., Zinovyev, A.Y.: Visualization of Data by Method of Elastic Maps and its Applications in Genomics, Economics, and Sociology. Institut des Hautes Etudes Scientifiques Preprint, IHES/M/01/34 (2001), available at `http://www.ihes.fr/PREPRINTS/M01/Resu/resu-M01-34.html`
2. Kohonen, T.: Self-Organizing Maps, 2nd edn. Springer, Heidelberg (1997)
3. Markowitz, H.M.: Portfolio Selection. Journal of Finance 7, 77–91 (1952)
4. Sharpe, F.: Mutual Fund Performance. Journal of Business, 119–138 (January 1966)
5. Sortino, F.A., van der Meer, R.: Downside risk. Journal of Portfolio Management, 27–31 (1991)

Prediction of Golden Cross and Dead Cross by Neural Networks and Its Utilization

Norio Baba and Kou Nin

Information Science, Osaka Kyoiku University,
Asahiga-Oka, 4-698-1, Kashiwara City, Osaka Prefecture, 582-8582, Japan

Abstract. In this paper, we propose a new decision support system (DSS) for dealing stocks which utilizes both the prediction (obtained by neural networks) concerning the occurrence of the "Golden Cross (GC) and Dead Cross (DC)" and that concerning the increase (decrease) rate of the future stock price several weeks ahead. Computer simulation results concerning the dealings of (randomly chosen) several individual stocks confirm the effectiveness of our approach.

Keywords: neural networks, traditional technical analysis, golden cross, dead cross, improved DSS for dealing stocks.

1 Introduction

In recent years, soft computing techniques such as Neural Networks (NNs), Evolutionary Computation, Fuzzy Engineering, and etc. have been studied quite extensively by many researchers and have been applied to various real problems. They have been successfully utilized for constructing a large number of intelligent systems [1]-[6].

In this paper, we shall propose a new DSS for dealing stocks which utilizes intelligently both the outputs from NNs which give a prediction concerning the occurrence of GC & DC and those which predict increase (decrease) rate of a stock price several weeks ahead. The outline of this paper is as follows. In Section 2, we shall briefly touch upon the traditional technical analysis which predicts future tendency of the stock price by taking the relative relationship between "long term moving average" and "short term moving average" into account. In Section 3, we first refer to our trial having been done so far for predicting GC & DC several weeks before they occur. Then, we mention that a DSS relying upon only the prediction concerning the occurrence of GC & DC becomes sometimes unreliable. In the last part of Section 3, we propose a new DSS for dealing stocks which utilizes intelligently both the prediction concerning the occurrence of GC & DC and that concerning the increase (decrease) rate of future stock price several weeks ahead. In Section 4, we give computer simulation results concerning the dealings of the (randomly chosen) several individual stocks in the Tokyo Stock Market which confirm the effectiveness of the proposed (new) DSS. Finally, the paper concludes with a brief summary and future perspectives.

B. Apolloni et al. (Eds.): KES 2007/ WIRN 2007, Part II, LNAI 4693, pp. 642–648, 2007.
© Springer-Verlag Berlin Heidelberg 2007

2 A Traditional Technical Analysis in the Stock Market

For getting an information concerning the current tendency of the stock market, many stock traders have often utilized the traditional technical analysis which takes the relative relationship between "Long Term Moving Average (LTMA)" and "Short Term Moving Average (STMA)" into account. Particularly, they have paid their attention to the crossings of LTMA & STMA. They have believed that GC (DC) is a reliable indicator that predicts the upward (downward) moving of the future stock price.

Remark 2.1. *Fig.1 illustrates a typical example of the GC and the DC which appeared in the two moving averages of the Nissan Motors in the Tokyo Stock Market.*

The point A (B) denotes DC (GC) where STMA cuts LTMA downwards (upwards). Many stock traders have so far believed that GC (DC) is a reliable indicator that suggests the upward (downward) moving of the future stock price.

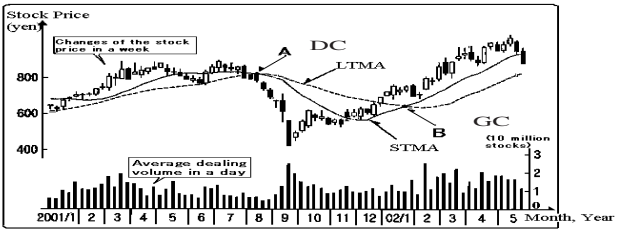


Fig. 1. Golden Cross and Dead Cross of the LTMA & STMA (Nissan Motors)

3 Utilization of Two Kinds of NNs Systems for Further Improving the Traditional Technical Analysis

3.1 Our Trial Having Been Tried So Far for Improving the Traditional Technical Analysis

During the past several years, we have been involved in the research for improving the traditional technical analysis. We have suggested that the use of NNs for predicting the GC & DC several weeks before they occur is quite helpful in order to construct an intelligent decision support system (DSS) for dealing stocks[5],[6]. A number of computer simulation results concerning dealings in the TOPIX, Nikkei-225, and individual stocks (in the Tokyo Stock Market) have confirmed the effectiveness of our approach.

3.2 A New Trial for Further Improving Our Approach

Although several computer simulation results have confirmed the efficacy of our approach, we have recently noticed that our DSS could be further improved by adding conditions concerning the predictions of the increase (decrease) rate of the future

stock price several weeks ahead to the DSS having been used so far[5],[6]. In the following subsection, we shall propose our new DSS "NDSS".

Remark 3.1. *Although GC & DC are powerful indicators for detecting the current tendency of the changes of a stock price, they are not completely reliable for this purpose. Their reliability depends strongly on the relative changes of the LTMA & STMA near the crosses.*

Look at the Fig.2.1 – 2.3 which show the relative changes of the LTMA & the STMA near the Golden Cross. Roughly speaking, in the Fig.2.1, both the LTMA and the STMA are going up. In the Fig. 2.2, the LTMA is increasing, and the STMA is almost constant. In Fig.2.3, the LTMA is going down while the STMA is going up. Among many traders, it has been empirically believed that the relative relationship between the two moving averages as shown in the Fig.2.1 gives a strong sign that suggests the upward tendency of a stock, but those of the other two figures don't give a definite sign.

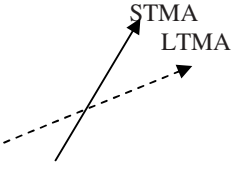


Fig. 2.1. A rough sketch of the STMA and the LTMA where both moving averages are going up

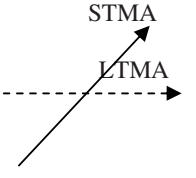


Fig. 2.2. A rough sketch of the STMA and the LTMA where the STMA is going up and the LTMA is almost constant

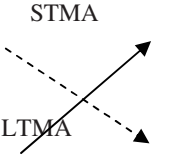


Fig. 2.3. A rough sketch of the STMA and LTMA where the STMA is going up and the LTMA is going down

3.3 Proposal of a New DSS Which Contains an Additional System That Makes a Prediction of the Increase (Decrease) Rate of the Future Stock Price Several Weeks Ahead

In the above section, we mentioned that the DSS relying upon only the prediction (by NNs) concerning the occurrence of GC & DC becomes sometimes unreliable.

In the following, we shall propose a new DSS which utilizes additional predictions (given by the NNs) concerning the increase (decrease) rate of the future stock price several weeks ahead.

NDSS

Carry out dealing only when the following two conditions A) and B) are satisfied.

Condition A)
One of the following two conditions ((1) & (2)) is at least satisfied.

(1) All of the outputs from n NNs are positive (negative) and the average of their absolute values is above 0.5.
(2) All of the following three conditions (2-a), (2-b), and (2-c) are satisfied.

(2-a) The greater part of the outputs from n NNs is positive (negative).
(2-b) The rate of the outputs having values over 0.7 (below - 0.7) exceeds 50 % of the number of the outputs having the same sign.
(2-c) The average of the outputs from n NNs is above 0.5 (below – 0.5)

Condition B)
The number of the NNs whose outputs have positive (negative) sign among the m NNs which are prepared for making predictions concerning the increase (decrease) rate of future stock price is larger than s.

Remark 3.2. *The n NNs in the Condition A) are prepared for making a prediction of GC (DC) several weeks before it occurs.*

Remark 3.3. *Signs "positive" and "negative" (Signs "increase" and "decrease") in Condition A) and Condition B) correspond to buying and selling, respectively.*

4 Computer Simulations

We have carried out computer simulations concerning dealings of the 5(randomly chosen) individual stocks (in the Tokyo Stock Market) from 1994 to 2004. Table 1 shows the average of the changes of the initial money (10 billion yen) during each year together with average return during 11 years by utilizing the Buy-and-Hold method (BHM), the traditional technical analysis (TTA), and the DSS (DSSL) proposed last year [6]. Table 2 shows the average of the changes of the initial money (10 billion yen) during each year together with average return during 11 years by utilizing the proposed decision support system NDSS. The computer simulation results presented in the two tables confirm the effectiveness of the proposed NDSS.

Table 1. Dealing Result by the Three Methods (BHM, TTA, DSSL) (Billion yen; Numbers below 10 million yen are rounded.)

	M		L		M		
	. 9	.4	.	7	.	.	9.4
	. 7	.	4. 9		9.	.	.
	.	9.4	.	9	7.	.	.9
4	9.74	.9	.		.	9.	9.99
	.4	9.4	4.		7. 7	.	.
	.	. 7	.				
Average Return: BHM: +1.04; TTA: +0.56; DSSL: + 1.66							

Table 2. Dealing Result by the Proposed Method (NDSS) (Billion yen; Numbers below 10 million yen are rounded.)

	DSS s)	DSS s)	DSS s)	DSS s)	DSS s 6)		DSS s)	DSS s)	DSS s)	DSS s)	DSS s 6)
							0 0	0	0 6	0	6
	0	0	0	0	0			0			
		6	6	6	6					0 6	0 6
	0 6	0			0	0					
	0	0	0				00				
6	0	6		6	6						
Average Return: NDSS(s=12): +1.52; NDSS(s=13): +2.08; NDSS(s=14): +2.02 NDSS(s=15): +2.18; NDSS(s=16): +2.19											

Remark 4.1. *In our simulations, we have set n = 7 and m = 21. We have used 7 k×k×1 (k≤N) neural network models (where k denotes the number of input variables having been chosen by the sensitivity analysis [7],[8]) for checking whether GC (DC) will occur in several weeks. We have also used 3×7 = 21 neural network models for making predictions concerning the increase (decrease) rate of changes of each individual stock in the Tokyo Stock Market 3 weeks, 4 weeks, and 5 weeks in the future.*

Remark 4.2. *We have carried out neural network training by using the past data for three years. Table 3 shows the learning periods and the prediction periods.*

Table 3. Learning Periods and Prediction Periods

	Learning Period	Prediction Period
A1	January 1991 - December 1993	January 1994 - December 1994
A2	January 1992 - December 1994	January 1995 - December 1995
.	.	.
.	.	.
.	.	.
A11	January 2001 - December 2003	January 2004 - December 2004

Remark 4.3. *In the above simulations, we have taken the charge for dealing into account by subtracting (0.001* (total money used for dealing) + 250,000) yen from the total fund for dealing.*

Remark 4.4. *In the above simulations, we have used the rule which allows "dealing on credit". Due to space, we don't go into details. Interested readers are kindly asked to attend our presentation.*

Remark 4.5. *5 individual stocks randomly chosen from our available data set for carrying out computer simulations are the followings: Nomura Holdings, SHIONOGI & CO., Advantest, Shin-Etsu Chemical, and AEON.*

5 Concluding Remarks

A decision support system for dealing stocks which improves the traditional technical analysis by utilizing the two types of the NNs has been proposed. Computer simulation results suggest the effectiveness of the proposed DSS. However, these simulations have been done only for the randomly chosen 5 individual stocks in the Tokyo Stock Market. In order to execute full confirmation concerning the developed NDSS, we need to check whether it can be successfully applied for dealing in the other individual stocks in various markets. For this purpose, we also need to carry out computer simulations concerning various indexes such as Nikkei-225, S&P 500, DAX, and etc.

Acknowledgements

The authors would like to express their thanks to the Grant – In –Aid for Scientific Research (C) by Ministry of Education, Science, Sports and Culture, Japan and the Foundation for Fusion of Science & Technology (FOST) who have given them partial financial support. They would also like to express their thanks to Prof. Y. Kai (Kwanseigakuin University) and QUICK Corporation for their kind support in giving them various financial data.

References

1. Rumelhart, D.E., et al.: Parallel Distributed Processing. MIT Press, Cambridge (1986)
2. Azoff, E.M.: Neural Network Time Series Forecasting of Financial Markets. Wiley, Chichester (1994)
3. Baba, N., Kozaki, M.: An intelligent forecasting system of stock price using neural network. In: Proceedings of IJCNN 1992, pp. 371–377 (1992)
4. Baba, N., Suto, H.: Utilization of artificial neural networks and TD-Learning method for constructing intelligent decision support systems. European Journal of Operational Research 122, 501–508 (2000)
5. Baba, N., Nomura, T.: An Intelligent Utilization of Neural Networks for Improving the Traditional Technical Analysis in the Stock Markets. In: Khosla, R., Howlett, R.J., Jain, L.C. (eds.) KES 2005. LNCS (LNAI), vol. 3681, pp. 8–14. Springer, Heidelberg (2005)

6. Baba, N.: Utilization of NNs for Improving the Traditional Technical Analysis in the Financial Markets. In: Gabrys, B., Howlett, R.J., Jain, L.C. (eds.) KES 2006. LNCS (LNAI), vol. 4251, pp. 663–669. Springer, Heidelberg (2006)
7. Zurada, J.M., et al.: Sensitivity analysis for minimization of joint data dimension for feedforward neural network. In: Proceedings of the IEEE International Symposium on Circuits and Systems, pp. 447–450. IEEE Computer Society Press, Los Alamitos (1994)
8. Viktor, H.L., et al.: Reduction of symbolic rules from artificial neural network using sensitivity analysis. In: Proceedings of the IEEE International Symposium on Circuits and Systems, pp. 1788–1793. IEEE Computer Society Press, Los Alamitos (1995)

An Agent-Based Early Manufacturability Assessment for Collaborative Design in Coating Process

Chii-Ruey Lin[1], Pei-Shu Fan[2], Yea-Jou Shiau[3], and MuDer Jeng[4]

[1,2] Graduate Institute of Mechanical and Electrical Engineering, National Taipei University of Technology, Taiwan
crlin@ntut.edu.tw, fanps202@yahoo.com.tw
[3] The 202nd Arsenal. MND, Taiwan
joeshiau@yahoo.com
[4] Department of Electrical Engineering, National Taiwan Ocean University
jeng@mail.ntou.edu.tw

Abstract. In this paper, an approach for developing a new theoretical platform and pilot environment for rapid distributed product design in extended enterprises using collaborative, autonomous agents. The approach is based on exchange of requests and information between design, engineering and facility agents. The resulting decisions are based on design criteria which is then optimised by an intelligent adaptive design moderator which interact with the different agents to generate new solutions. This research aims to provide engineers with a generic prototype tool for collaborative design evaluation and rapid assessment of the manufacturing feasibility and resource availability at different stages of the product development in film coating product.

Keywords: Collaborative design, Agent system, Coating process.

1 Introduction

The increased globalisation of competition, more demanding customer requirements, and rapid technological developments have recently resulted in previously unknown levels of complexity and dynamism in the manufacturing process. Early design assessment activities have a significant impact on reducing the cost of manufacture [1]. The efficient utilisation of the product development time and the level of coordination of the early design activities are becoming key factors for success in manufacturing today.

Coating Technology is a cross-sectional technology, which is needed by almost any producing industry for the outline or optimization of their products. Coatings add critical functionality and can be a key factor in determining the performance of the product. Conventional coatings which are formed by their natural properties, can address a number of issues ranging from surface hardness and toughness, corrosion and anti-sticking to lubricity, thermal stability and decoration. This is necessary in particular for the precision engineering and electronics industry due to the increasingly stringent requirements in component specifications and sophisticated

B. Apolloni et al. (Eds.): KES 2007/ WIRN 2007, Part II, LNAI 4693, pp. 649–655, 2007.
© Springer-Verlag Berlin Heidelberg 2007

application conditions. Without applying proper coatings to tools or products, they would have a short life span, poor quality, low productivity, poor performance or, in some cases, they do not perform at all. However, past research into the design of coating was always through trial and error, which is a time-consuming task and could lead to the divergence of the system so that the agent-based early manufacturability assessment is in appropriate.

Early design assessment is an effective method for cutting cost throughout the product development phase. In order to dramatically reduce the cost at the early stages, the planners require methodologies and tools that would allow them to judge the implications of the design specification on the required manufacturing processes and facilities requirements. However, the complexity of the specialised feedback of film coating knowledge to the designer is often not sufficiently assimilated or utilised efficiently.

The design tasks become even more complicated when the new product design process is carried out in various enterprise environment which the capabilities of different geographically distributed plant entities have to be taken into account. A distinctive feature of the early conceptual design assessment in virtual enterprises is that it depends on successful process knowledge management between closely interrelated domains of products, manufacturing processes, and resources information.

The system presented below attempts to address some of these issues by considering a multi-agent framework where the responses from the various distributed agents are combined using neural network techniques to compare the responses with past design scenarios.

2 System Architecture

In constructing the design moderator neural network techniques are used as a key component in the structure. This offers an alternative approach for efficiently integrating the domain-specific solutions generated by the individual agents and produces a collective solution from the responses it receives. Indeed, the design agent's knowledge is developed following the assumption that reasoning and searching for solutions on presently complex design tasks is an inductive process directed by examining past experiences, drawing analogies, reshaping patterns in the data corpus and adapting past solutions to newly experienced problems [2].

To achieve the necessary flexibility, the knowledge of the design agent is split into two components: a problem domain database, and an adaptive decision tool based on the neural network algorithms is shown in Fig. 1.

As opposed to the traditional multi-agent architectures, where the design agent receives isolated responses from each agent to the queries related to a specific design, in this system the design moderator integrates the responses generated by the individual agents with parameters from the initial design problem. The problem domain database thus constructed is a collection of the design specifications, designer queries, agent solutions, and design outcome of all previously encountered design cases, and constitutes the starting cognitive environment.

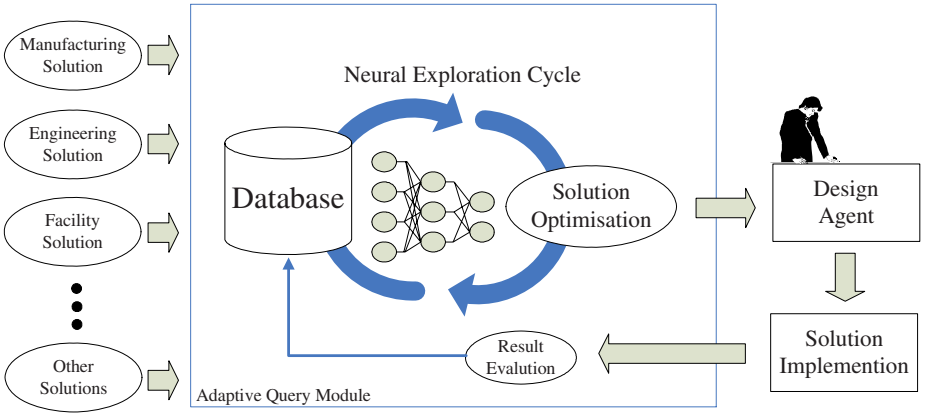


Fig. 1. The intelligent adaptive design moderator

The database is used by the adaptive neural network tool to answer specific design questions, by finding the closest matching design situations from past successful design projects. This process is aided by the ability of the adaptive tool to learn from past data as well as generate new design scenarios that closely match past situations from the existing data. The advantages of using a neural network implementation in contrast to other approaches, such as case based reasoning [3], is that in the situation of an incompletely described design query, or when no similar past cases exist, the system has the ability to internally manipulate the existing data to produce a new scenario. The neural network is thus providing an effective possibility to perform the adaptive task in case based reasoning.

The adaptive decision tool provides the design agent with a more comprehensive and integrated view of the information stored in the database, as well as consider new parameters coming from the design scenario.

Understanding the problem, building a design strategy, and identifying the possible obstacles of the project are accomplished simultaneously. In a single computational effort the neural decision tool is able to integrate the individual specific domain solutions and recommend a successful design strategy by extrapolating from past experiences.

In order to enhance the design agent knowledge for future decision-making, the measurement of the performance of the presently endorsed design solution is also entered into the database. This provides the system with a feedback mechanism, which constantly updates the design agent's problem domain database.

3 Multi-agent Approaches to Collaborative Design

A key characteristic of a concurrent engineering environment is that design is performed in a coordinated and collaborative manner. Coordination is the key to maintain the consistency of concurrent decision-making and the integrity of information in the product development lifecycle [4]. Coordination in collaborative

design takes time and requires knowledge, and if the proper tools are available it could reduce the lead-time and let designers focus on the essential problems of design [5].

A concurrent engineering framework based on a multi-agent planning environment is reported by Balasubramanian et al [6] and is further developed by Gu et al [7], where the proposed methodology assumes an agile manufacturing environment based on intelligent machines capable of making autonomous planning decisions.

Another multi-agent framework is proposed by Tan et al [8]. The framework draws on design-team behaviour in different domains with the goal of providing the infrastructure to enable designers, planners, and managers from different functional areas to collaborate using conflict resolution techniques in improving the initial product design.

In an extended manufacturing environment the decision making is based on collaboration between different geographically distributed facilities, each with a high level of autonomy in deciding on how best to utilise resources. Such autonomy can be effectively described by using a multi-agent representation of the main decision-making modules within a distributed design and planning environment (see Fig. 2). In the context of the reported research, an agent is considered to be an autonomous collection of objectoriented decision-making algorithms, software tools, and data structures that support a human-centred decision making process in a well defined domain and that provides communication to other agents.

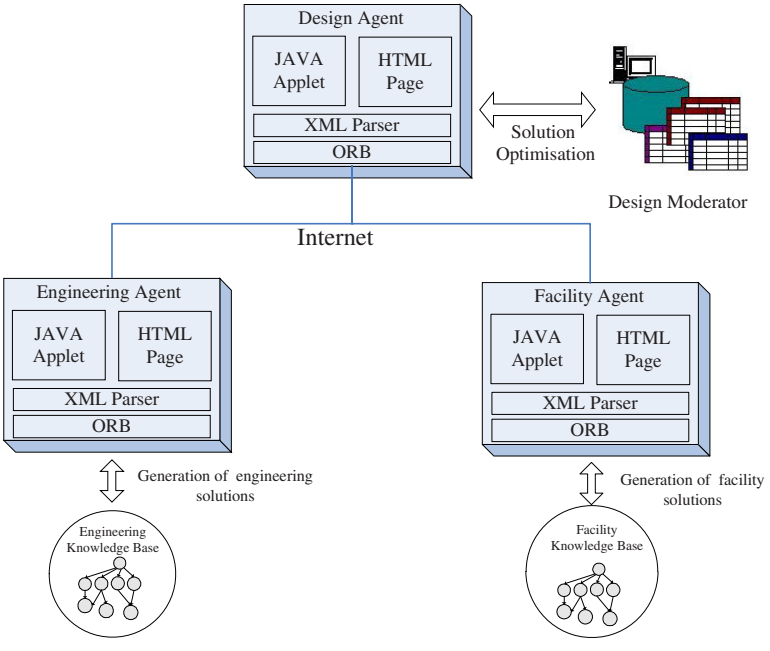


Fig. 2. Multi-agent system architecture

The agent's functionality is defined by the specific need for the integration of the decision-making steps into a common manufacturing decision-making environment. The design agent provides support to the design activities and initiates queries to the other agents on manufacturability assessment. The engineering agent is responsible for assessing the design and generating process solution. The facility agent uses the design and processing information as an input to select the most appropriate resources and facilities for manufacturing the product.

3.1 The Design Agent

The solid model of the part design is developed by a designer and stored in form-feature formatted table and should also be possible to basic knowledge of coating process, test methods of coating and standardization in the local design system. The WWW sharable CAD model of the part is generated in 3D CAD format by the designer. The 3D models of the part and associated product data model elements can be directly and simultaneously fetched by the designer using the interface of the design agent. The designer can then interactively specify part information such as feature, surface finishes, tolerances, coating requirement, etc., on the feature based product data model of the part. The design agent broadcasts the details of the part to the engineering and facility agent for process solutions and adapted facilities. The design agent interface provides access to the product data model and handles the communication with other agents.

3.2 The Engineering Agent

Once the types of the features requiring machining and their relationships are established, the data model is broadcasted by the design agent to the engineering agent. The selection of the manufacturing process is provided by the product data model based on form features and coating requirements. The engineering agent generates processing alternatives for each feature. The process is supported by the process data model generically describing the available capabilities in the machining and coating process. As a result, a set of alternative feature processing solutions is attached to each feature on the answer window in the engineering agent. According to the agent's conversation, the engineering agent could decide to send the result to the design agent, or the facility agent for checking the available facilities of using the process result.

3.3 The Facility Agent

The facility agent receives the part data model with the attached feature and coating processing alternatives from the engineering agent. As an output of the feature process selection, the part processing alternatives are described as sequences of form generating schemas representing the required machining and coating operations in a generalised format. The optimisation process at extended facility level is designed to decide upon the number and the distribution of operation required for selection the constraints combination of design criteria and manufacturing resources in the enterprise. Depending on the variety of manufacturing resources available in the enterprise' several planning alternatives can be generated at the facility agent.

4 Agent Communication

This section provides a detailed description of the languages and protocols utilised by the multi-agent systems in communicating within the suggested framework.

The agent abstraction is utilised as a mechanism for encapsulating and exchanging the distributed knowledge. The agent knowledge can be acquired implicitly by having the agents make assumptions independently or by jointly utilising support from other agents for coordinated activities by explicit communication.

Recently, researchers in the ARPA Knowledge Sharing Effort have defined the components for communication, called the Agent Communication Language (ACL) [9], which satisfies these needs. However, in many cases XML (eXtensible Markup Language) could be used to substitute ACL as a more broadly understood - albeit less expressive form of representation [10].

The Internet is increasingly becoming a popular, distributed, and open environment and hence attracts more people to investigate agent related knowledge and techniques on it. Web pages are essentially constructed using Hypertext Markup Language (HTML). However, by considering XML to encode the necessary information and services using content specific structures and semantics, it could easily be understood by agents as well. Other researchers have created XOL (XML-based Ontology exchange Language) which is designed to provide a format for exchange ontology definitions that is easily published on the WWW for exchange among a set of application developers [11].

In the agent communication, a method is proposed which involves isolating the design domain, engineering domain and facility domain, and modelling each domain using an object-relational data model. The structured data provided by the XML meta tags are then stored and linked to the domain knowledge. The communication scenario between the agents using XML is as follows:

A design agent sends a query to the engineering agent for a particular manufacturing solution. This message is delivered to the XML parser/generator where the message is then parsed by the design application module to generate the code which includes details such as the location of other agents and the requirement string.

This request is sent to the receiving agent where the needed operations are executed according to the content of the message from the sending agent, using its domain knowledge base.

A reply message using similar steps detailed above, will be sent to the design agent. The result is a response from the domain specific agents to a query from the design agent within a distributed environment.

5 Conclusion

In this paper, a framework for intelligent multi-agent decision making is presented. The individual autonomous agents with their domain specific knowledge are interrogated by the design agent for a domain specific solution. The response from these individual distributed agents are collectively filtered by the intelligent adaptive design moderator where comparisons are made with outcomes of previous design scenarios and a feasibility decision is suggested by the system to the design agent.

The collaborative nature of the decisions which considers past design scenario's allows a more responsive, informed decisions to be made that would encourage quicker design times and shorter time to market.

The reported methodology facilitates the decision-making process by providing greater flexibility in suggesting processing solutions, and allocating resources to new designs and provides higher responsiveness in supporting a more dynamic early product design. It is a step towards the development of a generic prototype tool for collaborative design evaluation and rapid assessment of the manufacturing feasibility and resource availability at different stages of the product development.

References

1. Daniel, S.: Innovation and manufacturability analysis in an integrated design context. Computer Industry 43(2), 111–121 (2000)
2. Im, K.H., Park, S.C.: Case-based reasoning and neural network based expert system for personalization. Expert Systems with Applications. 32(1), 77–85 (2007)
3. Lottaz, C., Smith, I.: Collaborative Design Using Constraint Solving. Lausanne, Swiss Federal Institute of Technology, Switzerland (EPFL) (1997)
4. Shiau, J., Ratchev, S., Valtchanov, G.: Distributed collaborative design and manufacturability assessment for extend enterprise in XML-based agent system. In: Proceeding of the IEEE 9th international workshop on enabling technologies: infrastructure for collaborative enterprise. Gaithersburg, MD, USA, pp. 260–265 (1999)
5. Karp, P.: Pangea Systems Inc.: XOL: An XML-Based Ontology Language for Bioinformatics. Objects in Bioinformatics'99. Silicon Valley, San Jose (1999)
6. Balasubramanian, S., Maturana, F.P., Norrie, D.H.: Multi-Agent Planning and Coordination for Distributed Concurrent Engineering. International Journal of Co-operative Information System. 5, 153–179 (1996)
7. Gu, P., Balasubramanian, S., Norrie, D.H.: Bidding-based Process Planning and Scheduling in a Multi-Agent System. Computers and Industrial Engineering. 32(1), 477–496 (1997)
8. Tan, G.W., Hayes, C.C., Shaw, M.: An Intelligent Agent Framework for Concurrent Product Design and Planning. IEEE Transactions on Engineering Management. 43, 297–306 (1996)
9. Genesereth, M.R., Ketchpel, S.P.: Software Agents. Communications of the ACM 37(7), 49–53 (1994)
10. Lassila, O.: RDF Metadata and Agent Architectures. In: Workshop on Compositional Software Architectures; Monterey, California (1998)
11. Murthy, S., Akkiraiu, R., Rachin, J., Wu, F.: Agent-based Co-operative Scheduling. Yorktown Heights, IBM T.J. Wrson Research Center, NY, USA (1998)

Design of a Simulated Environment for Flexible Manufacturing Systems

Yi-Sheng Huang[1,*] and Pin-June Su[2]

[1] Department of Electrical and Electronic Engineering, Chung Cheng Institute of Technology, National Defense University, 335 Taiwan, R.O.C.
yshuang@ccit.edu.tw
[2] School of Defense Science, Chung Cheng Institute of Technology, National Defense University, 335 Taiwan, R.O.C.

Abstract. A new simulated environment of flexible manufacturing systems is proposed in this article. The simulated environment can help the engineers reduce their efforts and better deal with some frustrating situations in the beginning of research and development. A modeling tool, called scenario, is employed in this paper. The scenario is able to depict the actions between equipment and its controller. This article starts with scenario to model the operation sequences of the system's equipment. In addition, this article demonstrates how a scenario is converted into a Petri net. To our knowledge, it is the first time that scenarios are used to model the operations of the equipment's handshaking. Finally, two emulators are given for the simulated environment.

Keywords: Petri nets, flexible manufacturing systems, scenario, emulator.

1 Introduction

In a flexible manufacturing system (FMS), different types of raw parts enter the system at discrete points of time and are executed simultaneously, sharing a limited number of resources, such as robots, AGVs, machine tools, and buffers. More preciously, an FMS is a computer controlled configuration where different operations can be processed. Generally it consists of various kinds of general purpose workstations, a palletized and programming material handing system, and other types of facilities such as fixtures and butters. The increasing importance of FMS requires a big effort in design and synthesis of computer software control to manage facility and schedule jobs. Hence, a cell controller plays a role to coordinate the facilities of the FMS. A cell controller is a unit in a facility that accepts commands from a parent (controller), and implements these commands by instructing children units under its control. These children may also be cell controller, or may be equipment controllers with sufficient intelligence to interpret the commands sent from a parent cell controller [1]. This article focuses on how to design the equipment emulators and their controllers.

* Corresponding author.

B. Apolloni et al. (Eds.): KES 2007/ WIRN 2007, Part II, LNAI 4693, pp. 656 – 663, 2007.
© Springer-Verlag Berlin Heidelberg 2007

Below are some contributions that are concerning this research field. A novelty of modern production systems is their dynamics, i.e., production facility, tools or fixtures and layout may be changed [7]. And some characteristics of FMSs are frequent jobs change, dynamic production scheduling, and frequent reconfigurations of production resources. As a result, the software control must be flexible and reconfigurable to satisfy the demand of a changeable production [1]. In order to develop and validate system control software for managing FMSs, most of laboratories need actual manufacturing equipment to make experimental test jobs. Typically, those experimental systems are necessary to occupy a large amount of laboratory space. In [3] proposed a physical emulator for a flexible manufacturing system. However, this paper focuses only on AGV equipment. In [4] (Huang and Chang, 1992) proposed a development environment based on Colour Time Petri nets (CTPN). Unfortunately, each different system requires a different programming. Recently, Huang *et al.* [5] proposed a Petri nets-based development environment for a cell controller. This advantage is that one can defined and/or change operation flow directly on the modeled Petri nets, without programming efforts. However, the cell controller must be supported by the real-world manufacturing system. Some approaches are closed this work as follow. One proposed a framework for a knowledge-based cell controller for FMS [10], another one presented a simulation-based controller for a FMS using a physical emulator [2], the other one submitted a virtual factory based on colored Petri net [6]. In this contribution, a solution to this dilemma is given. An emulator environment has been designed and implemented. And it is almost completely composed of software whose is based on the tool of scenarios. The proposed emulator development environment has several advantages as follows. First, it is inexpensive and can be replicated easily at other laboratories. Second, it can be reconfigured for the alternative manufacturing processes. Third, it can be considered as a personal computer rather than complex actual equipment. Finally, it is totally reliable and safe. And also it requires only a computer with a few RS-232 serial ports. After describing the system description in the next section, a scenario models and their simulated environment are present in section 3. Section 4 presents the analysis of the simulated equipment. Conclusion is summarized in Section 5.

2 The Description of the Actual Equipment Functions

A real-world FMS is introduced in [5]. This actual FMS manufactures motorcycle's linkers and pistons from two raw parts. The FMS consists of an automated storage/retrieval system (AS/RS), an automated guided vehicle (AGV), a CNC lathe, a CNC machine center (MC), a coordinate measuring machine (CMM), a robot, a flexible assembler system (FAS) and several buffer stations. The robot is dedicated to loading and unloading tasks of CNC lathe and CMM. The AGV transports pallets with raw parts, pallets with working pieces, and empty pallets between AS/RS and the buffer stations. A raw part of linker starting as a solid block is milled and drilled on the CNC MC, to become a geometrically shaped part with a linker hole on its side. A raw part of piston starting as a cylindrical solid peg is machined by the CNC lathe into a cylindrical shaped peg whose base fits into the hole of the finished linker.

Functions and operation flows for each of the major components are summarized below:

- AS/RS is a buffer station where pallets with raw parts and/or pallets with intermediate parts can be stored. The pallets are removed into and out from the AS/RS through buffer station_11 and buffer station_12, respectively.
- AGV is used to transport pallets with raw parts and empty pallets between AS/RS and the buffer stations.
- CNC MC is a milling and drilling machine.
- CMM is used to inspect the finished parts.
- The robot is used to load and unload the parts between CNC lathe and the buffer station_5 and station_6. And the robot is also used between CMM and the buffer station_3 and station_4.
- The buffer stations are used to load and unload a pallet.
- The FAS is used to assemble the linker and piston.

Based on the preceding system description, this article should be more detail discussing the actions of the real-world equipment. In the following subsection, how to design and implement the simulation environment should be introduced.

2.1 The AGV Communication Data Format

The communication data of the equipment may help one to understand the equipment actions. One can catch the communication data information from the controller (i.e., a PLC) of AGV. In this case, the authors apply ComWatch software to catch the data form the PLC. The date format consists of 109 bites and it is shown in table 1. The date format includes four different type formats which are command, reply, process end, and ACK reply formats. The detailed information of the communication data format is listed below.

Table 1. The communication data of AGV

Byte	0	1	2-3	4	5	6-7	8-107	108
Code	05 06 15	01 02	00	01	03	00	AGV path	0D

For convenience, the explanation of the command format is explained in a compact way as follows.

Byte 0. Three cases 05H, 06H, and15H are possible presented in this byte. It represents three actions ENQ, ACK, and NAK, respectively.
Byte 1. The operation code is represented by this byte. “01H” means AGV is testing on-line and “02H” means AGV is sending a command.
Byte 2-3. The bytes are represented what the working serial number is. It can be assigned to (00 00H) - (FF FFH) by the engineer.
Byte 4. The byte is represented the number of AGV. In this case, there are only one AGV in the system.

Byte 5. The byte is represented the state of AVG (i.e., auto model (03H) or manual model (02H)).
Byte 6-7. The bytes are reserved (00H).
Byte 8-107. The bytes are represented the path of AGV.
Byte 108. The byte is represented the end code (0DH).

2.2 The Scenario Realization

A scenario is able to describe the handshaking of the system components. Recently, scenarios have been widely used and documented in object-oriented design and user interface design as a technique to elicit and validate requirements, particularly for the operator of the system [9]. In addition, programmers often use them to understand an already-built system, by asking how the system responds to a particular input or operational situation. A simple scenario is shown in Fig. 1(a). The system components are represented as vertical arrows called instances. Instances model a system components (i.e., controller and equipment) participating in a scenario. Interactions between instances are shown as horizontal arrows called messages. The direction of a message indicates which instance the interaction. Messages labels (i.e., start and end) indicate the type of interaction that is occurring. The points on instances where and arrow starts and finishes are called (send and receive) events. An event can be considered the phenomena observed by an instance because of an interaction. Message direction also provides information on the order in which events occur; a send-event can never occur after its corresponding receive-event. In addition, messages can be considered to represent synchronous (handshaking) or asynchronous communication [9]. In this paper, a scenario-based is used to develop the emulator environment for the actual equipment.

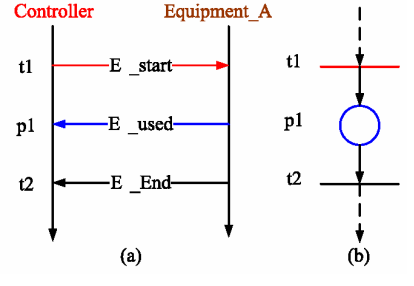


Fig. 1. The relationships with scenario and Petri net

In view of these observations, the author proposes a new modeling scheme that is based on scenarios. The scenarios can be characterized as follows.

1. Every arc has a label to represent an action that can be an operation start, operation end or a resource available.
2. From the arc definition, the designer understands which resource is required and the designer understands its dynamic situation.

3 Design of a Simulated Environment

From those characteristics of scenarios, the Petri net models will become easy to construct by mapping scenarios to Petri net models. In summary, there are close relationships with a Petri net and scenario. Fig. 1 can explains the relationships; a transition *t1* in Petri net (Fig. 1(b)) represents an event "*E_start*" in a scenario (Fig. 1(a)). In the following section, a scenario is converted into a Petri net is demonstrated.

The AGV driver and emulator. In order to design the AGV driver and emulator, the designer has to follow the data format of the AGV controller. First, break the actions down into smaller steps. And then the steps can be modeled by scenarios. Based on the AGV commands, one can construct its scenario step by step. Fig 2 shows AGV and ASRS scenarios. Due to the space limitation, the communication data format of ASRS is omitted. In the beginning of the explanation, the author assumes the readers understand any equipment work which is required by its cell controller. It states the AGV works when the cell controller sent a requirement for a service of AGV.

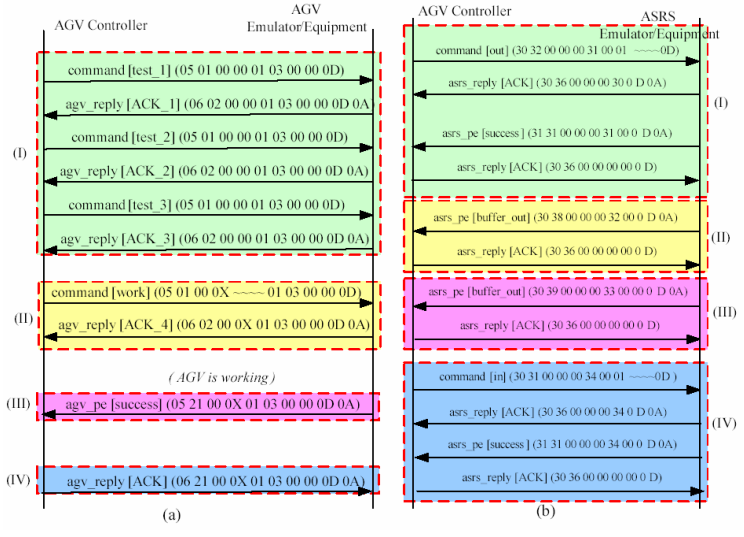


Fig. 2. (a) The AGV scenario. (b) The ASRS scenario.

In this case, the AGV can work normally if the AGV must be passed in the on line test. It states the AGV can send "command" and receive "reply" three times by itself successfully. The detail information of the test command and the reply are shown below.

Command Format

Command[test]□ "05 01 00 00 01 03 00 00 0D"

Agv_reply[ACK]□ "06 02 00 00 01 03 00 00 0D 0A"

Once the AGV passed in the online tests, the controller sends the work command to the AGV. At this moment, the AGV should send a reply to the controller. The detailed information is shown as follows.

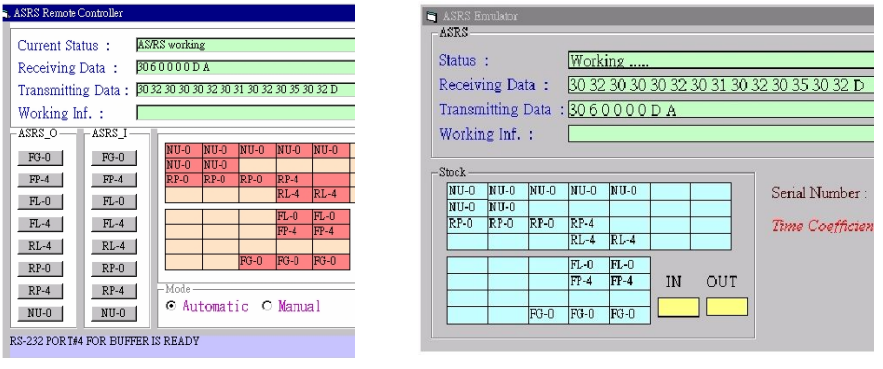
Reply Format
Command[work]□ "05 02 00 02 01 03 00 00 10 12 19 05 15 19 5119 01 19 12 19 11 17 24 17 11 12 19 12 19 01 19 5119 05 17 23 17 0D"

Agv_reply[ACK]□ "06 02 00 02 01 03 00 00 0D 0A"

Once the work is completed by AGV, AGV should send a process end event to the controller automatically. And then the controller sends AGV a reply. The detailed information is listed below.

Response format
Agv_pe[success]□ "05 21 00 02 01 03 00 00 0D"

ACK reply format
Agv_reply[ACK]□ "06 21 00 02 01 03 00 00 0D"



(a) (b)

Fig. 3. (a) The ASRS driver. (c) The ASRS emulator.

Based on the AGV specifications, the AGV scenario can be depicted in Fig. 2(a). Due to the space limitation, the detailed information about the AS/RS communication data is omitted. One can construct the AS/RS scenario (i.e., Fig. 2(b)) as the same as the preceding example. Moreover, the driver and emulator are easily to be implemented by Visual Basic. This article shows the simulated environment of the ASRS driver (i.e., Fig. 3(a)) and ASRS emulator (i.e., Fig. 3(b)). From the drive and emulator, one can easily check all events are listed in the scenario. It is worthy to notice that the emulator can completely replace the actual ASRS equipment functions.

By the same way, the other drivers and emulators of the FMS could be obtained. In fact, this simulation environment of the FMS is totally reliable and safe, and it requires only a computer and its related interface equipment. The main advantage is that the actual equipment can be replaced by the emulator. Using the simulation development environment, the designers have a virtual FMS in a computer.

4 Analysis of the Simulated Equipment

The scenario is able to model the communications between the drivers and equipment. However, it has not the ability to analyze the performance by itself. Petri nets have been recognized as one of the most power tool for analysis of the system. Hence, we shall construct the Petri nets model for the scenarios. Using the Petri net, the simulation environment's performance can be derived. For the above AGV and ASRS scenarios, one can convert the two scenarios into two Petri net models directly. And also the two Petri net models are shown in Fig. 4 (a) and Fig. 4 (b), respectively. Note that the two buffer station_11 and station_12 are included in the AS/RS. All pallets are moved into the AS/RS by the buffer station_11. And all the pallets are moved out the AS/RS by buffer station_12. Hence, the AS/RS, buffer station_11 and buffer station_12 are considered as a group in the scenario model.

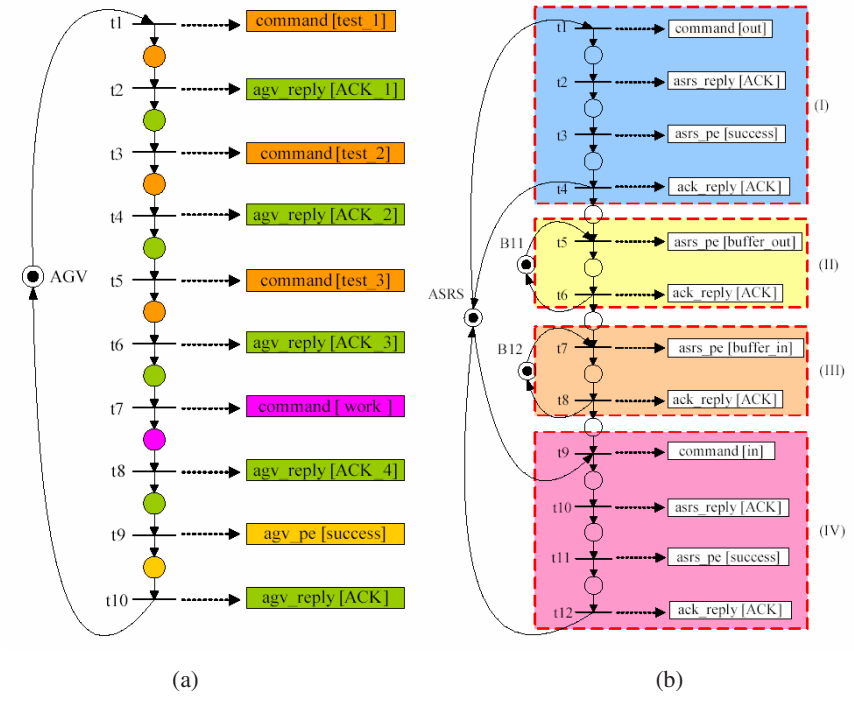


Fig. 4. (a) A Petri net model of the AGV. (b) A Petri net model of the ASRS.

Based on the marked graph definition [8], the two Petri nets (i.e., Fig. 4) belong to *marked graph*. Because the two Petri nets are ordinary Petri nets and each place has exactly one input transition and exactly one output transition. As a result, we can conclude that the two nets are liveness and reversibility.

5 Conclusion

In this paper we have been showed a simulation development environment for a real-world FMS. The author starts with a scenario tool to model the communication of the equipment. And the Petri nets have been constructed. Based on the basic definition of Petri net, the Petri net models have good performances. In addition, this simulation environment of the FMS is totally reliable and safe, and it requires only a computer and its related interface equipment. The main advantage is that the emulator that does not need the actual equipment might achieve the experimental tests. In addition, further research may develop the virtual shop for the FMS based on the simulation development environment.

Acknowledgments. This article was supported in part by the National Science Council of Taiwan, R.O.C., under Grant NSC94-2218-E-014-001.

References

1. Fanti, M.P., Maione, B., Piscitelli, G., Turchiano, B.: System approach to design generic software for real-time control of flexible manufacturing systems. In: IEEE Trans. Syst., Man, Cyber. A, vol. 26, pp. 190–202. IEEE Piscataway, New Jersey (1996)
2. Gonzalez, F.G., Davis, W.J.: A simulation-based controller for a flexible manufacturing cell. In: Proceedings of IEEE Syst., Man, Cyber. Cyber., vol. 4, pp. 3642–3647. IEEE Piscataway, New Jersey (1997)
3. Gonzalez, F.G., Davis, W.J.: Developing a physical emulator for a flexible manufacturing system. In: Proceedings of IEEE Syst., Man, Cyber., pp. 710–715. IEEE Piscataway, New Jersey (1998)
4. Huang, H.P., Chang, P.C.: Specification, modeling and control of a flexible manufacturing cell. International Journal of Prod. Res. 30, 2515–2543 (1992)
5. Huang, Y.S., Jeng, M.D., Chung, S.L.: Design, analysis and implementation of a real-world manufacturing cell controller based on Petri nets. International Journal of Prod. Res. 14, 304–318 (2001)
6. Jianhua, Y., Fujimoto, Y.: Modeling, Implementation and simulation of virtual factory based on colored timed Petri net. In: Proceeding of the Emerging Technologies and Factory Automation. vol. 1, pp. 574–579 (2003)
7. Lawley, M.A.: Deadlock avoidance for production systems with flexible routing. In: IEEE Trans. Robotic and Automation, vol. 15, pp. 497–509. IEEE Piscataway, New Jersey (1999)
8. Murata, T.: Petri nets: properties, analysis and applications. In: Proceeding of the IEEE, vol. 77, pp. 541–580. IEEE Piscataway, New Jersey (1989)
9. Uchitel, S., Kramer, J., Magee, J.: Synthesis of behavior models from scenarios. In: IEEE Trans. Sofeware Enginerring, vol. 29, pp. 99–115. IEEE Piscataway, New Jersey (2003)
10. Yep, C.K., Boey, S.H., Goh, J.: A framework for a knowledge-based cell controller for flexible manufacturing systems. In: Proceeding of the Emerging Technologies and Factory Automation, pp. 257–264 (1993)

MyHome: A Residential Server for Smart Homes

Sheng-Luen Chung and Wen-Yuan Chen

Department of Electriacl Engineering,
National Taiwan niversity of Science and Technology,
Taipei, Taiwan 106
{slchung, D9307002}@mail.ntust.edu.tw

Abstract. To take advantage of the recent advances in computer science, control, biomedical, and communication, an experimental project of MyHome is conducted. The objective is to explore innovation technology for future design and implementation of smart homes. This paper reports on current status of the MyHome project, which comprises of a local residential server, MyServer, and remote access interface, MyIcon. By clicking the icon, a user has information about his or her home as a whole at the finger tip. Information related to: alarm security, inventory tracking, facility management, and family members are all to be available through internet access. Design philosophy and proposed architecture are presented. Current technology related to the proposed architecture is surveyed before the ongoing status is reported. Scenarios of full function are also sketched[1].

Keywords: residential server, smart home, system integration.

1 Introduction

Technology evolves quite quickly nowadays. One particular application that takes advantage of the recent advances in computer science, control, biomedical [5], and communication [8,3,5]is the new idea in design and materialization of smart homes [1]. A feasible functional list for smart home now includes at least: multimedia and entertainment, networking and communication, security and firing, health care, and energy and power, etc. Study [2] also shows much room for Digital Home or home automation to grow in coming years.

From infrastructure point of view, there are hardware devices to be integrated in the smart home system, and all devices are linked through versatile networking available at home. Besides, in order to maintain interoperability and sufficient quality in communication among device, many standards have been proposed and used. In addition to these conventional requirements, new requirement and new peripherals stemming from renewable energy from the extended notion of green life is also included.

To take advantage of the recent advances in computer science, control, biomedical, and communication, an experimental project of MyHome is conducted. The objective is to explore innovation technology for the advancement of smart homes. In particular, one focus of the project is to design and implement a “My Home” icon

B. Apolloni et al. (Eds.): KES 2007/ WIRN 2007, Part II, LNAI 4693, pp. 664–670, 2007.
© Springer-Verlag Berlin Heidelberg 2007

displayed in computer. By clicking the icon, a user can see information related to the idea of smart home proposed in this project at the user's finger tip. Information related to: alarm security, inventory tracking, facility management, and family members are all to be available through internet access.

For the rest of this paper, Section II presents the idea of MyHome project. Section III is on the design and implementation of residential server MyServer and remote access interface MyIcon. Current technology related to peripheral and data integration is reported in Section IV before the conclusion that includes a tentative working scenario in Section V.

2 The Idea of HyHome Icon

The idea of a MyHome icon is an integrated information and command gateway, presented on the desktop of an internet accessible personal computer. Just as the "My Computer" icon in Windows operating system is to summarize all relevant hardware and software configuration of a computer, the proposed MyHome icon is to summarize all information related to a smart home, which may be at a distance from the computer viewing the MyHome icon.

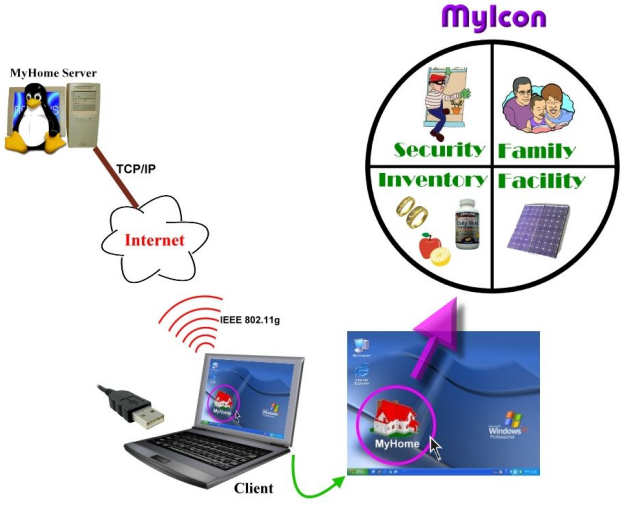


Fig. 1. MyHome icon

As shown in Fig. 1, the left hand side denotes a residential server that integrates information in a smart home; the right hand side depicts a remote access to the smart home. The MyHome icon, to be displayed on an internet accessible computer, is to display information about, as shown counter-clockwise of the quadrant of the MyIcon circle: alarm security, inventory tracking, facility maintenance, and family caring. For illustration, alarm security concerns safety monitoring in the house, in the garage, and to the entrance of a building; inventory tracking concerns tracking on kitchen

grocery, warehouse items, and valuables tagged by RFID tags; Facility maintenance concerns more with the notion of *green life*: environment monitoring, utility usage, and recycled resource; family caring concerns for each family member: where-about, e-health statistics, and daily schedule.

3 Residential Server MyServer and MyHome Icon

All the information mentioned above are to be provided by external collaborative sub-systems and to be integrated in the proposed residential server, called MyServer. Then, the MyHome icon, serving as a user graphic interface to view and manipulate the information can be a display icon in the desktop of a computer. In addition to viewing the information just mentioned, our design foresees the possibility of allowing users to configure operating rules: for instance, when a pre-defined condition is met, a pre-defined handling routine is activated to send out a voice mail to a designated mobile number to signify such an event occurrence. It is expected, the persons to be notified are also subject to rules defined by event emergency, and availability of notification.

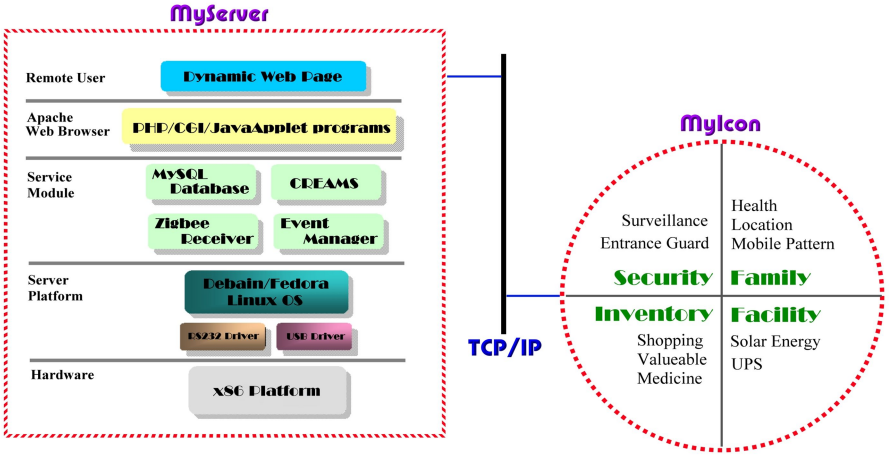


Fig. 2. Integration of MyHome icon to other supplementary subsystem

Fig. 2 shows the proposed architecture of the MyHome icon. With reference to the technology availability, we adopt client/server architecture for the final implementation of the proposed system. In our design, located in each home is the MyServer as sketched in Fig. 3 that collects all the local residential information. Meanwhile, at the access of internet is the MyIcon, which is mainly a friendly user interface to view and manipulate the home as a whole.

Left hand side shows the making of the residential server, MyServer. Built on top of Linux operating system through API interface, MyServer takes in monitored sensors, and is posed to be triggered by scheduled or conditioned driven events. In addition, through external interface to other sub-systems, the MyServer is to collect

information that keeps track of the daily activity of family members; the module of e-health of family member's health related statistics and medical profile as well as medical cabinet; the module of FRID tracking for both valuables and grocery; and finally the module of smart media to integrate TV, cable, internet as well as wireless gateway.

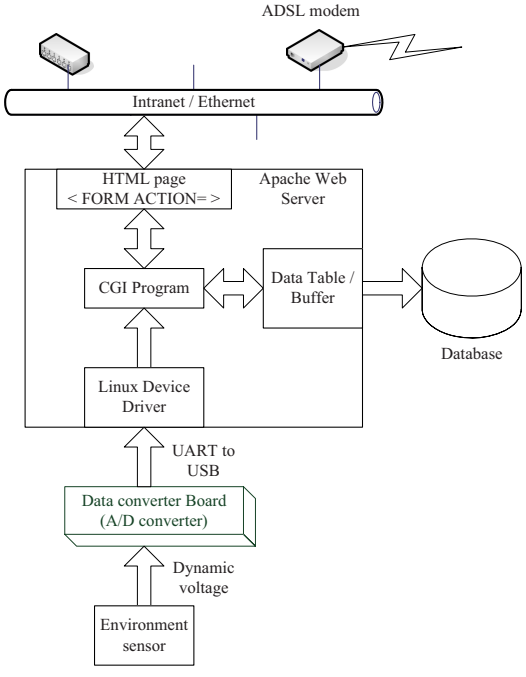


Fig. 3. Local implementation of MyHome server in MyServer

Right hand side of Fig. 2 shows the information to be displayed by the icon, MyIcon. By taking our proposed system as an interactive system, we adopt UML method as modeling tool. Particular emphases are on the interactive nature between sub-systems; data content required for each module; database schema design; inter-process communication arrangement; and the user interface presentation method.

4 Peripheral and Data Integration

Behind the unified presentation of a smart home requires successful design of peripheral and data integration in the smart home. Many studies indicate that a smart home system may consist of residential gateway (RG) [7] [6] or router [8], home server with some integrated function [2] [6]. Then, miscellaneous facilities and devices are connected to the home server through home networking system for different purposes: door sensors, fire detectors, security sensors, entertainment equipment, home health care telemedicine devices, etc.

We limit our discussion to peripheral integration. Design approach of remote configuration mechanism with standardization for communication and services is suggested in [7], which proposed standard-based methodologies and tool, XML, for the purpose: remote configuration, management, life-cycle support and device testing. TR-069 was proposed as an open standard for configuration mechanism. Based on this standard, one tool is developed with XML technology. With the proposed standard and method, home automation system is classified as Customer Premises Equipment (CPE), which includes residential gateways (RG) and all UPnP-based electrical facilities in home, and Automatic Configuration Server (ACS). The ACS interfaces the LAN based in home components, which are connected to RG, and the remote controller in public network through wide area network (WAN). The interface between ACS and RG is named as ACS southbound interface, whereas, the interface between remote configuration controller and ACS is called ACS northbound interface. With XML structure, all conceivable CPE are included in it, and CPE can be instantiated with "Device Descriptor" and "Device Creator". Meanwhile, CPE WAN management protocol provides the function of set/get for description, control, event of CPE devices. As a result, TR-069 compliant devices can be developed with Java-based tool suite for the purpose of building and testing.

Table 1 summarizes current implementation of peripherals for the proposed MyHome server. Under the four major categories to be presented by *MyHome*, listed items indicates communication channels utilized and the information collected.

Table 1. Applications and associated technologies

Item	Application	Utilized Technology	
		Communication	In/out
Family caring			
1	Where-about	GSM	GPS, Display, LED
2	Health Information	Zigbee	
Facility maintenance			
1	Recycled garbage		
2	Utilities	Zigbee	Solar energy conversion, Lux meter,
3	Environment	Zigbee	Sensors: temperature, humidity, noise, air quality, sunlight
Inventory tracking			
1	Kitchen	Zigbee	RFID, sensors: fire,
2	Warehouse	Zigbee	RFID, sensors: fire, door,
Alarm security			
1	Outdoor	Zigbee, WiFi, LAN	Camera, Sensor, Card Reader
2	In house		

MySQL is adopted for database engine. Frame format between communication gate way and end point device are defined. Dataflow between the MyHome server and MyHome client is illustrated in Fig. 4.

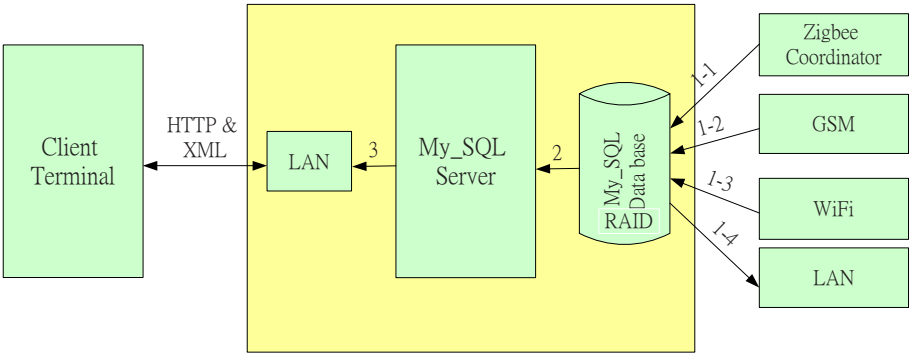


Fig. 4. Tentative concept for the data flows from end point device (EPD) to MySQL server database, and Client terminal

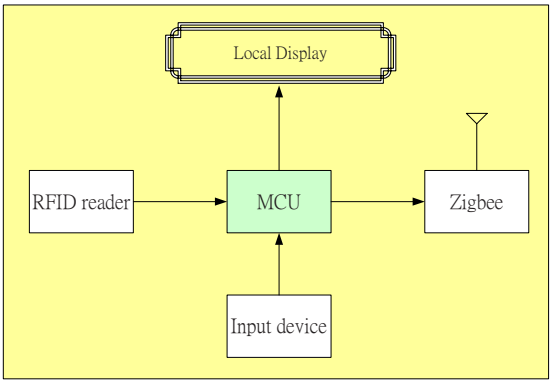


Fig. 5. End point device – RFID reader with Zigbee communication capability

RFID technique is employed to read out data from tracking items. Fig. 5 depicts an end point device with RFID reader, which is to be installed near refrigerator, warehouse or storage room. Then the change status of goods can be transmitted out through Zigbee. In addition, there is an input device for local display and simple manual setting. MCU is used to get the data from RFID reader, and convert its data into format compliant with the protocol between this device and the coordinator connected to MySQL server.

5 Conclusion and Future Work

The goal of the proposed MyHome icon is to provide a comprehensive view of a smart home in terms of the following four categories of information at a user's finger tips:

(1) alarm security of the whole residential area;
(2) RFID tag-based tracking and positioning system;
(3) In-house facility management system; and
(4) family care with possible extension to e-Health profile and daily scheduler.

Long term objective of this project is to support a complete MyHome system as illustrated by the following tentative scenario. No matter where a person is, he or she can always access an overview of his or her house through either a browser on the internet, or in the near future a third generation mobile. At the click of thumb, the person can know where a family member is, and foresees this family member's next scheduled activity is, at the permission of the concerned family member. He or she can even check medical appointment and intake notice of medication. The user can check the surrounding environmental information of his or her home, utilization of renewable energy consumption, and the status of recycled resource in the house, information traffic and index of content flowing through the house. Video image and history record, for instance, of gate access, are instantly available for home security check. The person in charge of shopping can also check inventory at the refrigerator,

References

1. Alheraish, A.: Design and Implementation of Home Automation System. IEEE Transactions on Consumer Electronics, 50(4) (November 2004)
2. Cheng, M.K.: Digital Home Main driver for Asian tech in 2006. CLSA Research & sales offices (2006)
3. Jostschulte, K., Kays, R., Endemann, W.: Enhancement of Wireless LAN for Multimedia Home Networking. IEEE Transactions on Consumer Electronics, 51(1) (February 2005)
4. Kao, W.C., Chen, W.H., Yu, C.K., Hong, C.M., Lin, S.Y.: Portable Real-Time Homecare System Design with Digital Camera Platform. IEEE Transactions on Consumer Electronics, 51(4) (November 2005)
5. Katsuno, Y., Aihara, T.: Autonomic Network Configuration for Networkable Digital Appliances. IEEE Transactions on Consumer Electronics, 51(2) (May 2005)
6. Kuwahara, M., Yoneyama, K.: A Portable Camcorder/Server for Wireless Video Transmission. IEEE Transactions on Consumer Electronics, 51(2) (May 2005)
7. Nikolaidis, A.E., Papastefanos, S.S., Stassinopoulos, G.I., Drakos, M.P.K., Doumenis, G.A.: Automating Remote Configuration Mechanisms for Home Devices. IEEE Transactions on Consumer Electronics, 52(2) (May 2006)
8. Sakamoto, N., Muguruma, K., Koshino, N., Chiba, S., Sakurai, M.: A Digital HDTV Receiver with Home Networking Function and Digital Content Storage. Transactions on Consumer Electronics, 51(3) (August 2005)

Semi-automatic Production Testing of Spark Plugs

S.D. Walters, P.A. Howson, and R.J. Howlett

Engineering Research Centre, School of Engineering
The University of Brighton, Brighton, U.K.
s.d.walters@brighton.ac.uk
http://www.brighton.ac.uk

Abstract. The spark plug has been in existence for nearly 150 years, with few significant design changes over that time. Similarly, spark plug production testing systems have evolved slowly – mainly because there have been simple, yet ingenious, tests which offered adequate 'Go / No go' testing. This paper describes a new method of spark plug testing, featuring elementary detection and diagnosis of faults. Spark voltage waveforms are captured and classified using a neural network. This paper follows up initial work reported in some other recent publications by the Author, presenting the prototype testing system and the method and results of an initial factory-based test.

Keywords: Neural network, spark plug.

1 Introduction

Spark plugs, and the means by which they are production tested, have changed very little since they were invented [1,2]. In recent years, shortcomings have emerged in respect of existing functional tests for spark plugs in the reliability and range of faults detectable; health and safety legislation has also played a part [3]. As a result of successful demonstrations of the air- and fluid-quenched tests outlined in [2,4,5], a factory-ready, prototype, on-line test system was designed and constructed by UoB and Champion Spark Plug UK. The need for a safe, sensitive data capture system to be present in the factory environment was correctly foreseen to be a potential problem, which consumed a significant proportion of the system design time.

Initially, a working requirements specification was formulated for the prototype on-line spark plug testing system, which became known as 'Brighton On-line Tester' or 'BOLT'. The air-quenched-gap and the fluid-quenched-gap tests were later named the 'A-test' and 'F-test' respectively. The objectives for the system were to:

(1) Initially equal the Pass/Fail function of the existing Champion Oil Spark Test (OST) dielectric tester [1,3]; offer potential for future improvement of fault detection.
(2) Later be able to distinguish between fault classes, as an off-line facility.
(3) If possible operate automatically, allowing operation by a factory worker.

B. Apolloni et al. (Eds.): KES 2007/ WIRN 2007, Part II, LNAI 4693, pp. 671–680, 2007.
© Springer-Verlag Berlin Heidelberg 2007

(4) Be safer than the OST.
(5) Use a minimum of different tests and software programs to cover as wide a range of different types and heat-ranges of spark plug as possible.

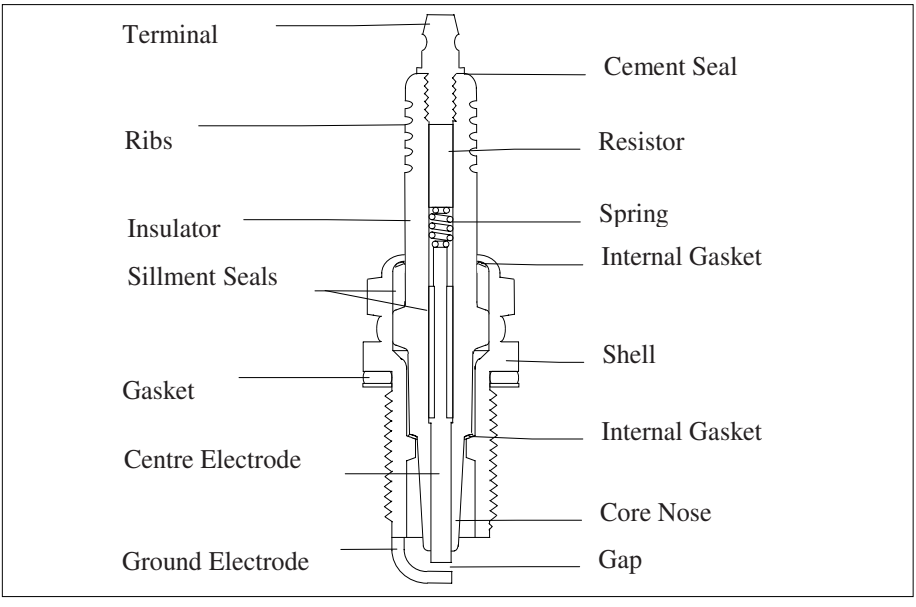


Fig. 1. Cross-section View of a Typical Spark Plug (Not to scale)

2 Hardware Development

In view of the experimental nature of such a prototype test system, the hardware comprised a P.C.-controlled digital storage oscilloscope as data capture, processing and output system and a specially constructed spark plug test-rig. It had been discovered earlier in the investigation that two types of test were required to detect all faults on the Champion list [2-5]. It was decided that two separate interchangeable jig fittings would be used for the plug handling of the A- and F-tests. The complete hardware solution was jointly produced by Champion and University of Brighton; the central equipment layout for both the A- and F-tests is shown in Fig.2. Both tests were undertaken using the same basic procedure, the main differences being in parameters in the new system software, the loading of the jig and the testing chamber conditions. The typical voltage waveforms resulting from the A- and F-tests were very similar to those shown previously in [2,5], respectively.

3 Software Development

New software was written in the C-language to perform: waveform data capture, data pre-processing, editing of input files and neural network functions; this contrasted

with the cumbersome use of four separate pieces of software in the first laboratory tests [2]. The data capture functions were specially written by combining Gould oscilloscope remote control instructions with National Instruments IEEE488 control instruction shells and instructions. Permission was granted for the use of part of the original C-language code of the BKP Back-propagation Neural Network Software [6] to provide the neural network functions. Modifications were made to suit the new program shell and to fix bugs. The new control, data pre-processing and formatting, menus and output display functions were all specially written.

Automatic training and recall programs were written. This speeded up the training process significantly and facilitated easier and quicker testing of modifications; however, care was required to ensure the reliability of training data captured on-line.

Different training/recall pairs of programs were required for the A- and F-tests, the main differences being in the waveform pre-processing and the post-processing, optimising the neural network's responses to the particular types of waveforms being tested. The A-test and the F-test produced significantly different waveshapes, owing to the intended presence and absence, respectively, of a spark.

Special versions of the programs were also written for various extra purposes, e.g. storing of training or recall waveform data and/or results, as necessary.

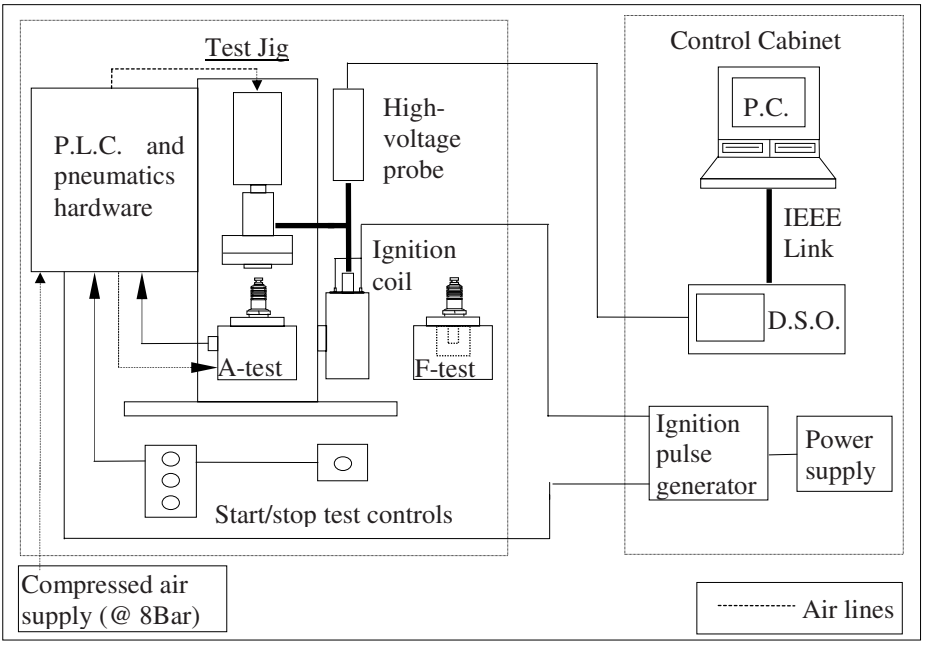


Fig. 2. The BOLT Hardware

The training program of each pair facilitated the following, two important functions:

(1) capture and pre-processing of training waveforms directly from the oscilloscope;
(2) training of the neural network using the resulting pre-processed data.

The recall program of each pair facilitated the following two important functions:

(1) capture and pre-processing of recall waveform data directly from the oscilloscope;
(2) "on-line testing" of the pre-processed data using the neural network.

Data capture. The data capture process was controlled automatically by the software, using the National Instruments IEEE bus to link the PC and the oscilloscope. Before each test the oscilloscope was initialised by sending all relevant set-up commands via the software and the PC; this was included to improve reliability under adverse operating conditions such as electrical-noise and possible accidental misuse or abuse by factory staff. Each pulse from the ignition pulse generator triggered the oscilloscope which then informed the PC that data was ready to be read back. Two consecutive waveforms of differing time-bases were captured for each sample test.

Data Pre-processing and Neural Network Input Format. The data pre-processing methods used in BOLT were specially designed to try to maximise the probability of correct responses from the neural network under adverse conditions such as: inconsistent data, electrically noisy environment and what was subsequently discovered to be non-ideal hardware. Two consecutive waveforms were ‘edited and joined’ to make each "spark" or "pulse" waveform for input to the neural network. The first waveform featured a sufficient time duration to represent in detail the breakdown and arc phases of a spark (or first peak of a F-test pulse) and the other featured a sufficient time duration to represent the complete ignition waveform breakdown, arc and glow discharge phases (or full damped-sinewave F-test pulse). The data point spacings and scalings were based on the features of good waveforms being ideally represented. In the case of the A-test the "breakdown" waveform was captured using a time/division of 2μs/division compared with the "full" waveform's 500μs/division. Similarly the F-test used time/division values of 20μs and 100μs for covering the initial negative peak of the waveform and the "full" waveform respectively. This had the benefit of offering the neural network a complete detailed representation of the waveform profile, whether a spark occurred anywhere around the spark plug or not. The aforementioned method of data pre-processing was found to be of particular benefit where the neural network was being used in fault classifier mode i.e. not Boolean output, (Pass/Fail / fault detection mode).

The Neural Network. The design of the neural network architecture was undertaken by experimentation and using guidelines [6,7]. The final network configurations are

shown in Table 1. The output format of the neural network software was a confidence level (0 - 100) and a class, where class 0 represented good. Any other class value represented a fault. The confidence level was inherently non-linear due to the sigmoidal activation function of the network's neurons.

Table 1. BOLT Neural Network Parameters

Parameter	Value		Parameter	Value (A, F)
Input Nodes	100		Learning Rate (Hidden))	0.80, 0.08
Hidden Nodes	25		Learning Rate (Output)	0.20, 0.02
Output Nodes	6		Momentum	0.80
Training Sets	30		Error Limit	0.10
Training Mode	Sequential		Initial Weight Size	0.10
			Error Interval	1

4 Testing the New Hardware and Software Together

Objectives. To undertake a pilot test of Brighton On-line Spark Plug Tester (BOLT). Also, to prove BOLT's abilities as an integrated Pass/Fail tester and spark waveform classifying system, for one build specification of spark plug.

Procedure. The spark plug testing system was set up as shown in Fig. 2. The required test (air or fluid quenched) was selected. The remainder of the Procedure was common to both tests.

The training mode was selected. The system was trained using suitable examples of RC9YCCs. A total of 11 different classes of RC9YCC spark plug were involved, and these are shown in Table 2. The resulting weight file was stored.

The recall mode was selected and the aforementioned weight file loaded. Samples of RC9YCCs to be tested were placed in the tester jig, as for training, but on recall ten pairs of pulses were used for each sample. The results of each test were printed out.

Table 2. Classification Tests of the A- and F-tests

A-test Fault	T-test Fault
GOOD – No fault	GOOD - No fault
NR – No resistor fitted	PTI - Puncture through insulator
NS – No spring fitted	CFE - Chipped firing end
2S – Spring substituted for resistor	CGS - Cracked gasket seat
NSW - No side wire	CK1 - Internal insulator crack 1
CG - Closed gap	CK2 - Internal insulator crack 2

Results. Tables 3 and 4 show results for runs of the A-test and F-test, respectively, with a summary in Table 5. Note that the recall testing of some sample spark plugs,

Table 3. A-Test Results

Ref.	Type	Training Sample?	Results Obtained			Correct?	
			C	P/F	Class	P/F	Class
1	Good	No	91	P	Good	Yes	Yes
2	Good	No	89	P	Good	Yes	Yes
3	Good	No	90	P	Good	Yes	Yes
4	Good	No	91	P	Good	Yes	Yes
5	Good	No	89	P	Good	Yes	Yes
6	Good	Yes	90	P	Good	Yes	Yes
7	Good	Yes	91	P	Good	Yes	Yes
8	Good	Yes	88	P	Good	Yes	Yes
9	Good	Yes	90	P	Good	Yes	Yes
10	Good	Yes	90	P	Good	Yes	Yes
11	NR	No	65	F	NR	Yes	Yes
12	NR	No	84	F	NR	Yes	Yes
13	NR	No	74	F	NR	Yes	Yes
14	NR	No	75	F	NR	Yes	Yes
15	NR	No	65	F	NSW	Yes	No
16	NR	Yes	54	F	NR	Yes	Yes
17	NR	Yes	73	F	NR	Yes	Yes
18	NR	Yes	52	F	NR	Yes	Yes
19	NR	Yes	58	F	NR	Yes	Yes
20	NR	Yes	49	F	NR	Yes	Yes
21	NS	No	91	F	NS	Yes	Yes
22	NS	No	92	F	NS	Yes	Yes
23	NS	No	91	F	NS	Yes	Yes
24	NS	No	90	F	NS	Yes	Yes
25	NS	No	81	F	NS	Yes	Yes
26	NS	Yes	90	F	NS	Yes	Yes
27	NS	Yes	92	F	NS	Yes	Yes
28	NS	Yes	91	F	NS	Yes	Yes
29	NS	Yes	92	F	NS	Yes	Yes
30	NS	Yes	92	F	NS	Yes	Yes
31	2S	No	95	F	2S	Yes	Yes
32	2S	No	96	F	2S	Yes	Yes
33	2S	No	95	F	2S	Yes	Yes
34	2S	No	95	F	2S	Yes	Yes
35	2S	No	95	F	2S	Yes	Yes
36	2S	Yes	96	F	2S	Yes	Yes
37	2S	Yes	96	F	2S	Yes	Yes
38	2S	Yes	96	F	2S	Yes	Yes
39	2S	Yes	96	F	2S	Yes	Yes

Continued overleaf…

Table 3. (*continued*)

Ref.	Type	Training Sample?	Results Obtained			Correct?	
			C	P/F	Class	P/F	Class
40	2S	Yes	96	F	2S	Yes	Yes
41	NSW	No	95	F	NSW	Yes	Yes
42	NSW	No	95	F	NSW	Yes	Yes
43	NSW	No	94	F	NSW	Yes	Yes
44	NSW	No	95	F	NSW	Yes	Yes
45	NSW	No	95	F	NSW	Yes	Yes
46	NSW	Yes	70	F	NR	Yes	No
47	NSW	Yes	58	F	NSW	Yes	Yes
48	NSW	Yes	77	F	NSW	Yes	Yes
49	NSW	Yes	59	F	NSW	Yes	Yes
50	NSW	Yes	95	F	NSW	Yes	Yes
51	CG	No	95	F	CG	Yes	Yes
52	CG	No	95	F	CG	Yes	Yes
53	CG	No	95	F	CG	Yes	Yes
54	CG	No	95	F	CG	Yes	Yes
55	CG	No	91	F	CG	Yes	Yes
56	CG	Yes	95	F	CG	Yes	Yes
57	CG	Yes	95	F	CG	Yes	Yes
58	CG	Yes	95	F	CG	Yes	Yes
59	CG	Yes	95	F	CG	Yes	Yes
60	CG	Yes	95	F	CG	Yes	Yes
61*	NSNR	No	73	F	NR	Yes	N/A
62*	NSNR	No	92	F	NSW	Yes	N/A
63*	NSNR	No	92	F	NR	Yes	N/A
64*	NSNR	No	86	F	NR	Yes	N/A
65*	NSNR	No	56	F	NR	Yes	N/A
66*	NSNR	No	76	F	NR	Yes	N/A
67*	NSNR	No	73	F	NR	Yes	N/A
68*	NSNR	No	80	F	NSW	Yes	N/A
69*	NSNR	No	75	F	NR	Yes	N/A
70*	NSNR	No	75	F	NSW	Yes	N/A

* GENERALISATION TEST; network was not trained to categorise this fault.

originally used to provide training data, was deemed necessary and acceptable for some F-test samples due to the serious shortage of samples and because the waveforms actually displayed variations each time.

The good results attained in this test were a repeat of those obtained in the feasibility tests of [2,4,5]. The F-test was not easy to optimise given the lack of suitable faulty-insulator samples. A solution to the sample problem is presented in another paper [5].

Table 4. F-Test Results

Ref.	Type	Training Sample?	Results Obtained			Correct?	
			C	P/F	Class	P/F	Class
1	Good	No	86	P	Good	Yes	Yes
2	Good	No	97	P	Good	Yes	Yes
3	Good	No	99	P	Good	Yes	Yes
4	Good	No	99	P	Good	Yes	Yes
5	Good	No	99	P	Good	Yes	Yes
6	Good	Yes	94	P	Good	Yes	Yes
7	Good	Yes	98	P	Good	Yes	Yes
8	Good	Yes	99	P	Good	Yes	Yes
9	Good	Yes	100	P	Good	Yes	Yes
10	Good	Yes	97	P	Good	Yes	Yes
11	PTI	No	91	F	PTI	Yes	Yes
12	PTI	Yes	62	F	PTI	Yes	Yes
13	PTI	Yes	99	F	PTI	Yes	Yes
14	CFE	No	80	F	CFE	Yes	Yes
15	CFE	No	94	F	Good	No	No
16	CFE	No	55	F	CGS	Yes	No
17	CFE	No	79	F	CFE	Yes	Yes
18	CFE	No	98	F	CFE	Yes	Yes
19	CFE	Yes	52	F	CFE	Yes	Yes
20	CFE	Yes	62	F	CFE	Yes	Yes
21	CFE	Yes	40	F	CFE	Yes	Yes
22	CFE	Yes	60	F	CFE	Yes	Yes
23	CFE	Yes	86	F	CFE	Yes	Yes
24	CGS	No	62	F	CGS	Yes	Yes
25	CGS	No	81	F	CGS	Yes	Yes
26	CGS	No	75	F	CGS	Yes	Yes
27	CGS	No	100	F	CGS	Yes	Yes
28	CGS	Yes	93	F	CGS	Yes	Yes
29	CGS	Yes	100	F	CGS	Yes	Yes
30	CGS	Yes	98	F	CGS	Yes	Yes
31	CGS	Yes	97	F	CGS	Yes	Yes
32	CGS	Yes	89	F	CGS	Yes	Yes
33	CK1	Yes	100	F	CK1	Yes	Yes
34	CK2	Yes	88	F	CK2	Yes	Yes
35*	CGSU	No	60	F	CFE	Yes	N/A

* GENERALISATION TEST; network was not trained to categorise this fault.

Note: CGSU was a CGS sample with the broken section of core-nose not displaced downwards.

Table 5. Final Results of the Classification Tests

	A-test Results	F-test Results
Go/No-go (Pass/Fail) Test	70/70	34/35
Correct Fault Diagnoses	58/60	32/34

Conclusions. The efficacy of the BOLT system as a Pass/Fail tester, of the specified samples, was proven in this relatively small-scale evaluation. It was therefore decided that the system would work well in the required capacity in the forthcoming major factory test. The only sample which gave an incorrect response was a Chipped Firing End sample, so marginal that it would almost certainly not have been detected by the OST either. Discrimination between categories could have been improved, but this was a low priority. (A Pass/Fail tester was all that was required in the short-term).

The reason for the shortage of results for some faults (related to punctured or cracked insulator) was simply a shortage of samples, in this plug type, that were available for testing at the time. This was due to an inability to manufacture accurate replicas of some faults at the time.

Acknowledgment. The Author is grateful to Champion Spark Plug U.K. for their financial and technical support which was offered during this investigation into spark plug testing.

References

1. Walters, S.D.: Characterization and Analysis of Kettering-type Automotive Ignition Systems and Electrical Spark Profiles, Ph.D Thesis, University of Brighton, in Association with Champion Spark Plug (Europe) (1998)
2. Walters, S.D., Howson, P.A., Howlett, R.J.: Production Testing of Spark Plugs using a Neural Network, Paper No. kes05-159. In: Proceedings, International Conference on Knowledge-based Intelligent Information and Engineering Systems, Melbourne, vol. 4(4), pp. 74–80. Springer, Heidelberg (2005)
3. Walters, S.D., Howson, P.A., Howlett, R.J.: A Survey of Spark Plug Testing Tools and Techniques, Accepted for: International Journal of Condition Monitoring and Diagnostic Engineering Management, Pub. COMADEM International (2007) ISSN1363-7681
4. Walters, S.D., Howson, P.A., Howlett, R.J.: Functional Testing of Spark Plugs using Ignition Sparks. In: Paper No. 403, 40th Universities' Power Engineering Conference, Cork, Proceedings, Full text CD, Abstracts book (2005) ISBN0-9502440-4-X
5. Walters, S.D., Howson, P.A., Howlett, R.J.: Dielectric Testing of Spark Plugs using Neural Networks. In: Accepted for: 42nd Universities' Power Engineering Conference, Brighton, Proceedings (2007)
6. Howlett, R.J.: A Distributed Neural Network For Machine Vision (Ph.D. Thesis), University of Brighton (1994)
7. Haykin, S.: Neural Networks. Macmillan College Pub. Co. Inc., NYC (1999)
8. Champion: Private Communication, Champion Spark Plug (U.K.) (1994)
9. SAE ARP159A Standard: Dielectric Testing of Spark Plugs, SAE (1994)

10. Walters, S.D., Howson, P.A., Howlett, R.J., Ryder, D.M., Miller, R.: Inverse Analysis of Electrical Discharge Phenomena by Neural Network. In: 30th Universities' Power Engineering Conference, Procs. University of Greenwich, vol. 2, pp. 850–853 (1995)
11. Champion: Straight Talk about Spark Plugs, Champion Spark Plug (1987)
12. NGK: Engineering Manual for Spark Plugs, NGK Spark Plug CoLtd, OP-00769105 (1991)
13. Bosch: Spark Plugs, Robert Bosch GmbH, Delta Press (1985 and 1999)
14. Lippman, R.P.: An Introduction to Computing with Neural Networks. IEEE ASSP Magazine 4, 4–22 (1987)
15. Hush, D.R., Horne, B.G.: Progress in Supervised Neural Networks, IEEE Signal Processing Magazine, 8–39 (1993)
16. Spark Plug Defects and Tests: National Advisory Committee for Aeronautics, Report No. 51, Washington Govt. Printing Office (1920)

A Gait Based Approach to Detect Directional Bias of Four-Legged Robots' Direct Walking Utilizing Acceleration Sensors

Ding-Jie Huang and Wei-Chung Teng

National Taiwan University of Science and Technology,
43 Keelung Rd. Sec. 4, Taipei 10607, Taiwan
{M9315048, weichung}@mail.ntust.edu.tw

Abstract. In this paper, we propose an effective approach to detect the directional bias of walking four-legged robots. We found that light weight four-legged robots tend to have unstable directional bias even when walking on flat ground, thus this approach is designed to be able to dynamically and continuously detect the directional bias. By observing and analyzing data from acceleration sensors of AIBO robots, we found the relational model between the acceleration data and the walking gaits by which we may utilize to develop an ideal gait based approach to detect directional bias. In order to make sure the correctness of our approach, several experiments were designed and executed. As a result, we can correctly detect the direction of bias within five gaits, and the correct rate of single gait can be up to 90%.

Keywords: four-legged robot, gait, directional bias, accelerometer.

1 Introduction

In recent years we found more and more robots designed to be sold as household products. One famous example is AIBO robot. The AIBO robot is designed to be light weighted and to be soled plastic hemisphere on its feet such that it doesn't scratch the walking plane. However, this design also makes the robot not able to step firmly, and thus produces directional bias even when walking straight on flat ground. Since the directional bias of some robots is obvious, and the directional bias accumulates as long as the robot is walking, we started trying to develop an algorithm to automatically detect and correct the walking directional bias.

The observed directional bias during walking varies for each step, and the ranges of bias may be quite different between two AIBO robots, thus the detection algorithm should be based on sensor's feedback, and be able to dynamically detect the directional bias. The most popular sensors used to detect the heading direction are video camera, gyroscope and accelerometer. Since the image processing algorithms to detect directional bias are much time consuming and require more computing power than accelerometer based approach, and AIBO doesn't have gyroscope equipped, we choose to develop our algorithm from accelerometer data. In theory, the distance of bias can be

B. Apolloni et al. (Eds.): KES 2007/ WIRN 2007, Part II, LNAI 4693, pp. 681–688, 2007.
© Springer-Verlag Berlin Heidelberg 2007

calculated by integrating the acceleration twice, but the acceleration data we get from AIBO is not accurate enough to generate trustworthy data. In the remainder of the paper, we will briefly review related work before providing a detailed description of our bias detection algorithm. Then, we will describe the AIBO robot and the process how to analyze the acceleration data. By utilizing the statistical data, we present an ideal detection algorithm. Finally, we will examine our algorithm to look for the correct rate of our algorithm.

1.1 The AIBO Robot

Our main experiment platform is AIBO ERS-220A made by SONY. The basic hardware specification of AIBO is shown in Table 1. By using OPEN-R SDK, we may record the data from three-axial acceleration sensors in AIBO [1][2][4][5]. The positive direction of X, Y and Z axis in the acceleration sensor are mapped to right, front and up of the robot respectively. According to the result of measurement, each axis has different unit length [11]. The unit length of X axis is 0.086784, Y axis is 0.169080 and Z axis is 0.172046 meter. The sampling rate of a frame is 125Hz, and AIBO takes 140 frames (1.12 seconds) to complete one gait cycle. Besides, in order to get more information about the status of AIBO, we also recorded the data from touch sensor under each foot. With the help of touch sensor's feedback, we can analyze and determine the order of steps of AIBO's walking pattern.

Table 1. The hardware specification of AIBO

Processor	64bit RISC Processor. 32 MB Main memory.
Actuators	16 degrees of freedom with a continuous range of motion. Three for each of the legs, one from head light and three from the head.
Sensors	Temperature sensor. Infrared distance sensor. Acceleration sensor. Touch sensors(head,face,back,legs,tail). Vibration sensor. 100,000 CMOS image sensor

1.2 Related Works

H. Liu and G. Pang had proposed that the position of the robot can be detected by using acceleration sensors [7]. So if the position is known, then the direction bias is easy to be computed. D. Vail and M. Veloso have used acceleration data from AIBO to identify the surface under the robot as it walks[3]. In addition, they also used the acceleration data to predict gait velocity by using a k-nearest neighbor approach. In order to find out the directional bias, walking distance is a good starting point. Seong Yun Cho et al [10] described how to use low-cost acceleration sensors to measure the system of walking distance. Some previous research showed that the acceleration sensors were very

shaky, so it is too hard to calculate the speed or traveled distance of the dog [12]. Fortunately, it is practicable to determent the relation between actual walking and acceleration by analyzing and observing the walking robot, because acceleration has some particular characteristics which would be affected by walking condition.

2 The Algorithm to Detect Directional Bias

As has been noted above, the integrated data are different from the real distance. Maybe the original data is the critical issue. Consider the characteristic of each acceleration sensor data, a series of study was designed to check the validity of the data. So, the flowchart of the main research method appears in Fig.1. For the first step, we want to confirm the reliability of three-axial acceleration sensors by using Matlab software package, so a simple test was designed. Then we analyze each axis acceleration data and find out the relation between them. Finally we summarize the result and propose our method to detect directional bias.

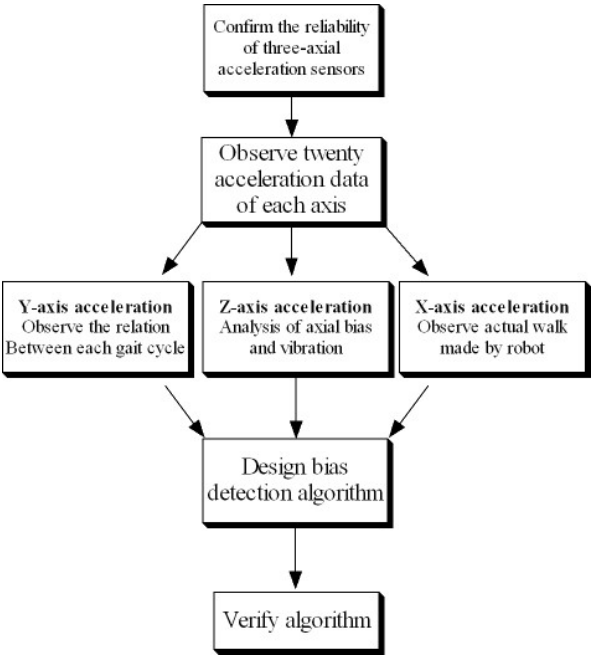


Fig. 1. Flowchart of research method

The first experiment is to let AIBO walks 5 gait cycles on the top of pasteboard for 20 times. We recorded the acceleration data of each axis and drew the data on one graph, as Fig.2 has shown. In Fig.2(a), we find that the X-axis acceleration data is unstable, but each gait cycle has similar feature. For example, there are fixed10 pulses during 5 gait cycles, and they almost happen in the same time. So we believe that the three-axial acceleration sensor has specific repeatability. Especially in first gait cycle,

because the vibrations are so few that the data are very similar. But after second gait cycle the vibrations are increasing, so each data are getting different. Since the three-axial acceleration sensor on AIBO has general reliability, then we will analyze each axis data respectively.

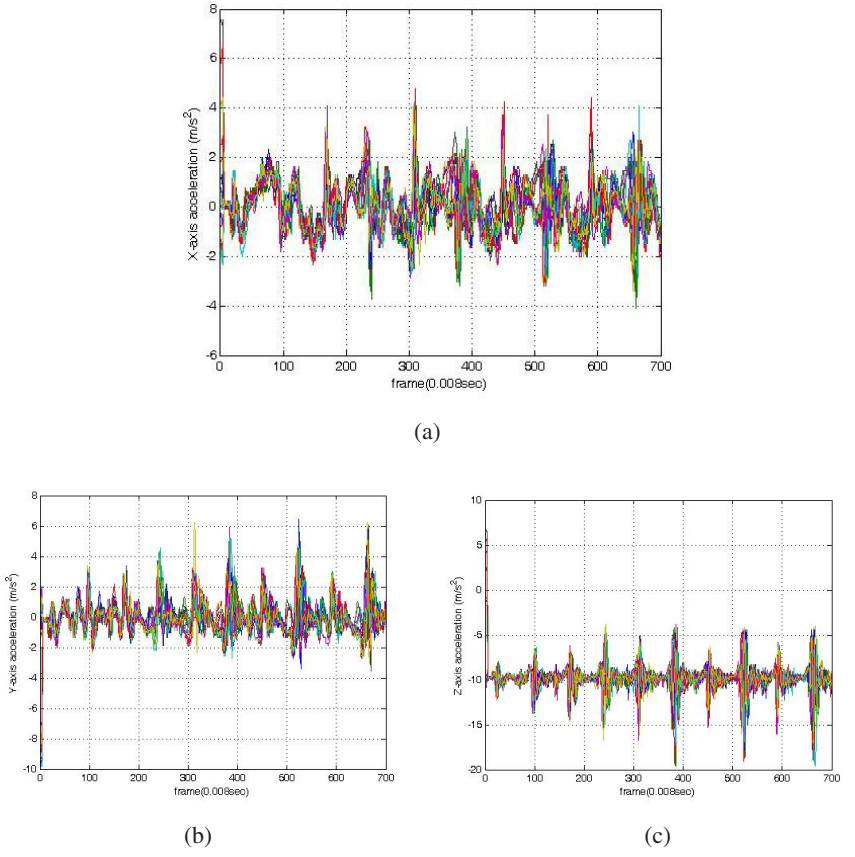


Fig. 2. Each axis acceleration data during 5 gait cycles measured 20 times: (a) X-axis (b) Y-axis (c) Z-axis

2.1 The Characteristics of Acceleration Sensor

After confirmed the reliability of three-axial acceleration sensor on AIBO, we started to analyze each axis. For the first step, we analyze Y-axis acceleration data by integrating the acceleration twice directly. As mentioned above, the integrated data are far from the real distance measured manually. So we consider that there must be some kinds of factors that affect the Y-axis acceleration data. So far we can not compute the distance directly by using acceleration data, but the real distance has particular relation with the acceleration data. As the result of observation, we think that the vibration and the axial bias may be the main factors that cause the difference between integrated data and the

real distance. It seems that the factors appear when AIBO change its own gait. It is because AIBO is on unstable state when it change gait. For example, the vibration occurs in the moment that the left front leg is raising up and the left back leg is putting down. The vibration causes the axial bias and a pulse. In order to find when the vibration occurs, we start to analyze Z-axis acceleration data.

In the second step, we analyze Z-axis acceleration data by comparing with the data from the touch sensor on bottom of each leg. If the reading of the touch sensor is 0, it means that the AIBO is during swing phase. If the reading is 1, it means that the AIBO is during stance phase. The reading from touch sensor is 0 or 1, but in order to compare the touch sensor data with acceleration data we add a fixed value on each reading of touch sensor. As Fig.3has shown, "○" means the reading from right front touch sensor; "□" means the reading from left front touch sensor; "▷" means the reading from right back touch sensor; "◁" means the reading from left back touch sensor. In addition, we find that the time points when the pulses occur during 5 gait cycle are steady. It begins from 0.2 second (25 frame), and there is a pulse every 0.56 second (70 frame). Comparing with each leg phase, the start time point 1 in Fig.3 of these pulses is the time when right front leg is raising up and the right back leg is putting down. The time point 2 is the time when left front leg is raising up and the left back leg is putting down. So the pulse is caused by the changes between two legs on the same side. This kind of vibrations would affect each axial data extremely. So the data seem to be unreliable during the unstable period and can not be integrated directly.

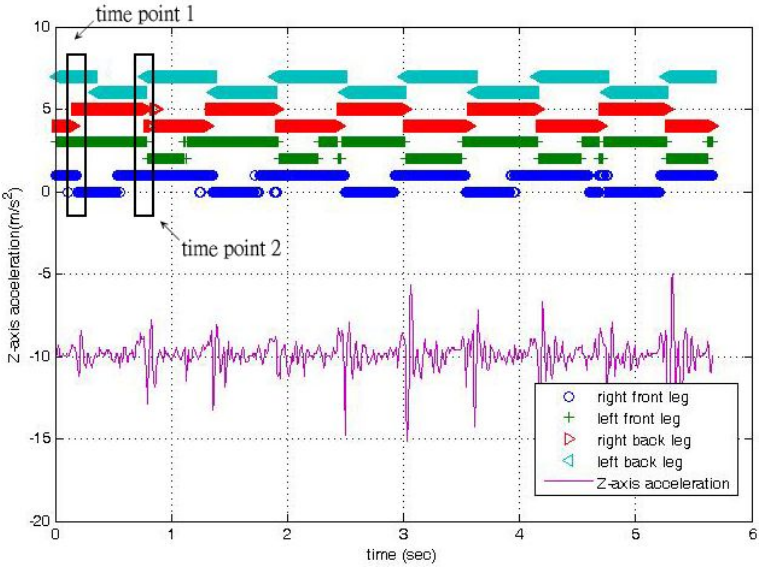


Fig. 3. Z-axis acceleration data during 5 gait cycles compare with each leg phases

Besides, from Fig.3 we can figure out that AIBO starts its gait in the order of right front, left back, left front and right back. This gait [8][9] is shown in Fig.4 and is the same with the notion that Hirose and Kunieda proposed [6]. Via observing walking

AIBO, we find that its walking direction is straight when it raising right front leg. When it is putting right front leg down and raising left back leg up, it gets the most right position during a single gait. Then it raises left front leg and its walking direction is left forward. Eventually when it is putting left front leg down and raising right back leg up, it gets the most left position during a single gait.

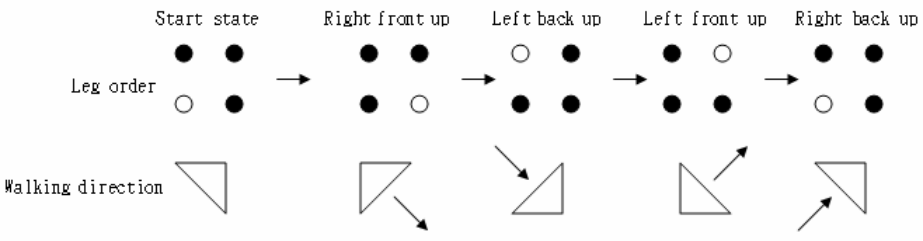


Fig. 4. Start gait order of the robot with direction

In the third part, we want to know at what frame the vibrations start. We start to analyze X-axis acceleration comparing with the trajectory of AIBO. As Fig.5 shown, the gap between dotted lines is the distance during one single gait cycle. The arrows point out the most right positions or the most left positions. We assume the velocity is 0 when AIBO reach the most right positions or the most left positions. It is because that the instant velocity should be 0 at the point when the robot changes its direction. So we can integrate data from the next arrow partially.

2.2 Bias Detection Algorithm

As has been discussed, it seems reasonable to believe the acceleration sensors on AIBO are useful in some degree of accuracy. But we find that AIBO vibrate largely at some fixed time point, so it is important to take out the acceleration data while left leg or right leg in the swing phase individually. As Fig.6 has shown, by observing the phase of each leg we decided to collect data from 23th frame to 63th frame which represented the acceleration data that the right front leg in the swing phases. In addition, we collect data from 97th frame to 137th which represented the acceleration data that the left front leg in the swing phases. By integrating this two data twice individually and then summing these two values up, we get a value that we called it as partial integrating value (PIV). If the PIV is positive, the center position of AIBO is going to right position. If the PIV is negative, the center position of AIBO is going to left position. So, this method is our "single gaited bias detection algorithm".

3 Experiment Results

To test the "single gaited bias detection algorithm", we gathered a single cycle of each sensor reading while AIBO walk on the pasteboard and record the center position of AIBO, and repeated this experiment for 30 times. We list the first 10 records to explain the result of our experiment, as shown in Table 2. The X1 field is the X-axial position of

left front leg, X2 is the X-axial position of left back leg, Xc is the center position of AIBO, or the average of X1 and X2, DIV is the value directly integrated twice from acceleration, and θ is the bias angle which is computed by X1 subtracting X2 and then divided by the distance (23cm) between front leg and back leg. If the value in Table.2 is positive, it means that the directional bias is right; otherwise, if the value in Table.2 is negative, it means that the directional bias is left.

Comparing Xc with DIV, they are different from each other from exp1 to exp10. The values of DIV are not only too large but also upside-down in positive-negative sign. So the directional bias can not be simply computed by integrating X-axial acceleration. Comparing Xc with PIV, there are only two data (exp2 and exp7) in wrong judgment. If the value of Xc equal to 0, θ would help to explain the directional bias. Take exp3 for example, the value of θ is negative, so the directional bias is left, and the method determine correctly. The other data show that the proposed method can determine the directional bias. The correct rate of "single gaited bias detection algorithm" is 93.33%, because 28/30 data are correct.

Table 2. Experiment results (in unit of cm)

NO.	X1	X2	Xc	DIV	PIV	θ
exp1	-1.3	0.7	-0.3	7.542884	-1.9959	-4.96974
exp2	-0.4	0.6	0.1	15.72617	-0.84257	-2.48955
exp3	-0.5	0.5	0	14.12934	-0.49655	-2.48955
exp4	-0.8	0.6	-0.1	13.84468	-0.64595	-3.48327
exp5	-1.2	0.6	-0.3	15.4393	-0.72316	-4.4749
exp6	-0.5	0.4	-0.05	8.680384	-0.54403	-2.24087
exp7	-0.6	0.8	0.1	8.871726	-1.71764	-3.48327
exp8	-0.5	0.4	-0.05	6.40316	-0.49071	-2.24087
exp9	-0.3	0.2	-0.05	17.49574	-0.83063	-1.24536
exp10	-1.2	0.6	-0.3	9.290792	-1.34884	-4.4749

4 Conclusion

We have proposed an effective approach to detect the directional bias of walking four-legged robots by using acceleration data. It is because that sensor on AIBO has poor precision, so during the process of acceleration analysis we must compare the actual walking gait with acceleration data. We adjust our algorithm by the try-and-error method continuously, and with lots of attempts the relation between them had been found out. Besides, the maximum bias relative to center position of AIBO in one gait cycle is ±0.6cm, but the factor error is up to ±5.443cm so that we can not precisely get the actual bias. Although there are many possible factors to the directional bias, such as material and gradient of ground, lack of calibration, hardware malfunction, etc…, but the proposed algorithm is considered effective to all kinds of situation as long as the accelerometers function correctly.

Acknowledgements. This work is supported by the National Science Council of the Republic of China, under the grant NSC 95-2221-E-011-127.

References

1. Wijbenga, A., van de Sanden, M.: How To Make AIBO Do Tricks, University of Groningen,Netherlands (2004)
2. Hengst, B., Ibbotson, D., Pham, S.B., Sammut, C.: The UNSW United 2000 Sony Legged Robot Software System, University of New South Wales?Australia (2000)
3. Vail, D., Veloso, M.: Learning from Accelerometer Data on a Legged Robot, Computer Science Department, Carnegie Mellon University (2004)
4. Serra, F., Baillie, J.-C.: Aibo programming Using OPEN-R SDK. Tutorial, ENSTA, France (2003)
5. Rico, F.M., González-Careaga, R., María, J., Plaza, C., Olivera, V.M.: Programming Model Based on Concurrent Objects for the AIBO Robot, Universidad Rey Juan Carlos (2004)
6. Bekey, G.A.: Autonomous Robots From Biological Inspiration fo Implementation and Control. The MIT Press, Cambridge, Massachusetts (2005)
7. Liu, H., Pang, G.: Accelerometer for Mobile Robot Positioning. In: Proc. Of the 1999 IEEE Industry Applications Society Conference (October 3-7, 1999)
8. Schmiedeler, J.P., Waldron, K.J.: The Mechanics of Quadrupedal Galloping and the Future of Legged Vehicles. The International Journal of Robotics Research 18(12), 1224–1234 (1999)
9. Toyomasu, M., Shinohara, A.: DEVELOPING DYNAMIC GAITS FOR FOUR LEGGED ROBOTS. In: The International Symposium on Information Science and Electrical Engineering, pp. 577–580 (2003)
10. Cho, S.Y., Park, C.G., Jee, G.I.: Measurement System of Walking Distance Using Low-cost Accelerometers. In: The 4th Asian Control Conference,Singapore, pp. 1799–1803 (September 2002)
11. Sony Corporation, OPEN-R SDK (2004)
12. Sven Westermark, Programming tools for AIBO (2005)

An Intelligent Algorithm for Scheduling Jobs on a Single Machine with a Common Due Date

Zne-Jung Lee[1], Chen-Chia Chuang[2], and Kuo-Ching Ying[3]

[1] Deptartment of Information Management, Huafan University, No. 1, Huafan Rd. Shihding Township, Taipei County, 22301, Taiwan, R.O.C.
johnlee@hfu.edu.tw
[2] Department of Electrical Engineering, National Ilan University, Taiwan, R.O.C.
[3] Deptartment of Industrial Engineering and Management Information, Huafan University, No. 1, Huafan Rd. Shihding Township, Taipei County, 22301, Taiwan, R.O.C.

Abstract. In this paper, an intelligent algorithm for scheduling jobs on a single machine with a common due date is introduced. This problem of scheduling jobs on a single machine, known as NP-complete problem, is against the common due date with respect to earliness and tardiness penalties. In the proposed algorithm, the properties of scheduling jobs on a single machine with a common due date are well derived with ant colony optimization (ACO). Furthermore, the heuristic is embedded into ACO as local search to escape from local optima. Simulation results show that the proposed intelligent algorithm yields the best performance among existing algorithms.

Keywords: Intelligent Algorithm; Common Due Day; Ant Colony Optimization.

1 Introduction

Scheduling jobs on a single machine with a common due date, just one job can be completed exactly on the due date, and all other jobs must be completed either before or after this common due date. To meet this requirement, an earliness penalty will be incurred if jobs have been completed before this common due date. On the other hand, a tardiness penalty will also be incurred if jobs have been completed after this common due date [1]-[5]. The problem of scheduling on a single machine with a common due date, known as a NP-complete problem, has been extensively investigated [2]. The objective for scheduling on a single machine with a common due date is to find an optimal schedule that can minimize the sum of earliness and tardiness costs for all jobs.

Many exact and heuristics algorithms have been proposed to solve the problem of scheduling jobs on a single machine with a common due date [6]-[9]. Due to the complexity of this problem, it is difficult for above approaches to obtain the optimal solution when the problem size is large [10], [11]. In the literature, meta-heuristic approaches such as simulated annealing (SA), genetic algorithms (GAs) and tabu search (TS) have also been proposed to find the near optimal solutions for this

B. Apolloni et al. (Eds.): KES 2007/ WIRN 2007, Part II, LNAI 4693, pp. 689–695, 2007.
© Springer-Verlag Berlin Heidelberg 2007

problem [12]-[17]. Recently, many search activities have been devoted to ACO and shown premising results in many applications [18], [19]. ACO has been widely performed as a cooperative search algorithm for solving optimization problems. It needs to incorporate the heuristic into its search process to ameliorate search performance. In this paper, a heuristic is embedded into ACO as local search for solving the problem of scheduling on a single machine with a common due date. The proposed algorithm can demonstrate excellent performance in later simulations.

The rest of this paper is presented as follows. Section 2 describes the formulation of problem for scheduling jobs on a single machine with a common due date. The basic concept of ACO is introduced in Section 3. Section 4, the structure of the proposed algorithm is discussed in detail. Next, simulation results are compared with those of existing approaches in Section 5. Concluding remarks are finally drawn in Section 6.

2 Problem Formulation

The problem of scheduling jobs on a single machine with a common due date is to minimize the total earliness and tardiness penalties. There is only one operation for each job to be scheduled on a single machine with the common due date d. For each job i, the processing time λ_i, the penalty per unit time of earliness α_i, and the penalty per unit time of tardiness β_i are given. Let S denote a feasible schedule of the jobs. A penalty of $\alpha_i * E_i$ is incurred when job i is completed E_i time units earlier than the common due date, whereas a penalty of $\beta_i * T_i$ is incurred when it is completed T_i time units later than the common due date. The minimization of earliness and tardiness penalties of single-machine scheduling problems with a common due date can be formulated as follows [20]-[24].

$$s_i + \lambda_i + E_i - T_i = d,\ i = 1, \cdots, n \tag{1}$$

$$s_i + \lambda_i - s_j - \gamma(1 - x_{ij}) \le 0,\ i = 1, \cdots, n-1;\ j = i+1, \cdots, n \tag{2}$$

$$s_j + \lambda_j - s_i - \gamma x_{ij} \le 0,\ i = 1, \cdots, n-1;\ j = i+1, \cdots, n \tag{3}$$

$$s_i, E_i, T_i \ge 0,\ i = 1, \cdots, n \tag{4}$$

where $x_{ij} \in \{0,1\}$, x_{ij}=1 if job i precedes job j and otherwise x_{ij}=0, n is the number of jobs, γdenotes a sufficiently large number and s_i denotes the starting time of job i. Eq. (1) indicates each job is early or tardy. Eq. (2) represents that the starting time plus processing time of job i is earlier than or equal to the starting time of job j if job i precedes job j. Eq. (3) represents that the starting time plus processing time of job j is far ahead of the starting time of job j if job i lags job j. Eq. (4) ensures that the staring time, tardiness and earliness of jobs must be exceeding or equal to zero. Then, the objective function for scheduling jobs on a single machine with a common due date is presented as follow.

$$F(S) = \sum_{i=1}^{n} (\alpha_i * E_i + \beta_i * T_i) \tag{5}$$

To efficiently solve the problem of scheduling jobs on a single machine with a common due date, three well known properties will be used in the proposed intelligent algorithm. These three properties are proofed in the literature [1], [7], [9], [17], [21].

Property 1. An optimal schedule exists if no idle time between any consecutive jobs for scheduling jobs on a single machine with common due dates.

Property 2. For the optimal schedule, V-shaped property exists around the common due date. This means that jobs completed before or on the common due date d are scheduled in non-increasing order of the ratios λ_i / α_i, and jobs starting on or after d are scheduled in non-decreasing order of the ratios λ_i / β_i in an optimal schedule.

Property 3. In an optimal schedule, either the first job starts at time zero or the completion time of one job coincides with the common due date d.

3 The Concept of ACO

The proposed intelligent algorithm is based on ant colony optimization. In this section, the basic concept of ACO is introduced. ACO uses a set of ants in a cooperative way to build solutions by exchanging pheromone information. The pseudo code of ACO algorithm is stated as [18]:

```
Procedure: ACO algorithm
Begin
   While (ACO has not been stopped) do
     ConstructAntsSolution;
     UpdatePheromone;
     DaemonAction          % optional
   End;
End;
```

In the algorithm, ants find solutions starting from a start node and moving to feasible neighbor nodes in the process of ConstructAntsSolution. ConstructAntsSolution decides a colony of ants that cooperatively and interactively visit next states by choosing form feasible neighbor nodes. They move by applying a stochastic local ant-decision policy that consists of pheromone information and heuristic information. UpdatePheromone is a process of updating the intensities of pheromone information over time. In this process, ants can release pheromone while building the solution (online step-by-step) or while the solution is built (online delayed). It is used to avoid locally convergence and to explore more search space. DaemonActions is used to implement centralized actions such as local optimization procedure or the collection of global information that decides whether to deposit additional pheromone or not. It is optional for ACO. In this paper, ACO is used to specify the interaction between above processes for scheduling jobs on a single machine with a common due dates.

4 The Proposed Algorithm

The problem of scheduling jobs on a single machine against the common due date with respect to earliness and tardiness penalties is known as NP-complete problem. It seemed that ACO applied to solve this problem is promising. In the proposed intelligent algorithm, ants successively choose feasible jobs into subsequence to construct feasible solutions until all jobs are scheduled. For constructing feasible solution, the ant-decision policy for the *l*-th ant positioned on job *r* successively selects the next job *s* into subsequence at iteration *t* is governed by

$$z = \begin{cases} \arg\{\max_{u=allowed_l(t)} [\tau_{ru}(t)\ \eta_{ru}^{\beta}]\} & when(\varsigma \le \varsigma_0) \\ Z & otherwise \end{cases} \tag{6}$$

where $\tau_{ru}(t)$ is pheromone information, $_{ru}$ is called heuristic information, β is a parameter representing the importance of heuristic information, ς is a random number uniformly distributed in [0,1], and $allowed_l(t)$ is the set of feasible jobs currently not assigned by ant l at iteration t. In Eq.(6), ς_0 is the probability of exploiting the learned knowledge when $\varsigma \le \varsigma_0$. It indicates that ants will directly select next jobs by the product of learned pheromone information and heuristic information. It performs a biased exploration for the next job for $\varsigma > \varsigma_0$, and Z is an index of node selected from $allowed_l(t)$ according to the probability distribution given by

$$P_{rz}^{k}(t) = \begin{cases} \dfrac{\tau_{rz}(t)\eta_{rz}^{\beta}}{\sum_{u \in allowed_l(t)} \tau_{ru}(t)\eta_{ru}^{\beta}} & \text{if } z \in allowed_l(t) \\ 0, & \text{otherwise;} \end{cases} \tag{7}$$

For $_{ru}$, it is separately decided when the next job positions in the earliness or tardiness subset of the V-shaped. According the Property 2, let $\eta_{ru} = \frac{\lambda_r}{\alpha_r} + \frac{\lambda_u}{\alpha_u}$ if next job positions in the earliness subset, otherwise $\eta_{ru} = (\frac{\lambda_r}{\beta_r} + \frac{\lambda_u}{\beta_u})^{-1}$.

In finding feasible solutions, ants perform online step-by-step pheromone updates as:

$$\tau_{ij}(t+1) = (1- \psi)\ \tau_{ij}(t) + \psi \tau_0 \tag{8}$$

where $0<\psi \le 1$ is a constant, $\tau_0 = (m * \sum_{i=1}^{n} \lambda_i)^{-1}$ is the initial value of pheromone information and m is the number of ants. After all ants have constructed complete solutions, the global update rule is performed as follow.

$$\tau_{ij}(t+1) = (1- \rho)\ \tau_{ij}(t) + \rho \Delta\tau_{ij}(t) \tag{9}$$

where $0<\rho\leq 1$ is a parameter governing the pheromone decay process, $\Delta\tau_{ij}(t) = 1/F(S)^{elitist}$, and $F(S)^{elitist}$ is the objective function of elitist solution obtained from the beginning of the search process.

After obtained the elitist solution, the heuristic is performed to escape local optima and hope to find the global optima. In the proposed algorithm, the idea of heuristic is to greedily swap jobs between the earliness subset (S_{EA}) and tardiness subset (S_{TA}) in the V-shaped of the elitist solution. There are two phases in the heuristic of the proposed algorithm. In Phase I, the jobs in S_{EA} are successively selected from the first job to swap with all jobs in S_{TA} that could obtain better objective function than that of elitist solution. In this phase, the i-th job in S_{EA} and j-th job in S_{TA} are swapped when this swapping causes the objective function improvement. Furthermore, the elitist solution is replaced by the swapped solution and then the global update rule of Eq. (9) is performed. On the contrary, the elitist solution is not changed if this swapping does not cause any objective function improvement. This phase continues until all jobs in S_{EA} have been examined. In Phase II, the jobs in S_{TA} are successively selected from the first job to swap with the jobs in S_{EA} that could obtain better objective function than that of elitist solution. This phase continues until all jobs in S_{TA} have been examined, and the swapping process is similar to above Phase I if a better objective function is obtained.

5 Simulation Results

In simulation, we need to identify a set of parameters. We performed the simulation with various values, and the results are all similar. Experiments were conducted on PCs with P4 3GHz processor. In the following simulations, we keep the following values as default: ρ=0.1, ψ=0.1, q_0=0.8, β=2, and the number of ants m=20. For fair comparisons, these compared approaches are performed to see which approach can find the best solution after a fixed period of running without improving objective function [17]. To verify the effectiveness of the proposed algorithm, the problem sets are taken from [2], [7], [17]. The number of jobs n=200 is selected, and four restrictive factors h=(0.2, 0.4, 0.6, 0.8) are used to determine the common due date $d = \lfloor h\sum\lambda_i \rfloor$. Therefore, each test problem was solved for 3 runs. To evaluate the results, the percentage improvement of the obtained solution (F_{ob}) was compared with regard to the benchmark values provided by Biskup and Feldman (F_{bf}) [2], [17].

$$Improvement\ rate = \frac{F_{bf} - F_{ob}}{F_{bf}} * 100\% \tag{10}$$

Table 1 tabulates all simulation results for various approaches. From Table 1, it shows that proposed algorithm has the same solutions as Lin [17] for h=(0.2, 0.4, 0.6). The proposed intelligent algorithm has the best solution among these compared approaches for h=0.8. It is clearly evident that the proposed algorithm indeed has admirable results.

Table 1. Average improvement rates (%) of various approaches

n	*h*	Meta-heuristic [2]	TS [25]	GA [25]	HTG [25]	HGT [25]	Lin [17]	The proposed algorithm
200	0.2	5.73	5.76	5.74	5.76	5.76	5.78	5.78
	0.4	3.79	3.74	3.75	3.75	3.75	3.75	3.75
	0.6	0	0.01	0.13	-0.37	-0.07	0.15	0.15
	0.8	0	0.04	0.14	-0.26	-0.07	0.15	**0.07**

6 Conclusions

In this paper, an intelligent algorithm for scheduling jobs on a single machine with a common due date is proposed. In the proposed algorithm, ACO with heuristic can perform global search and avoid trapping in local optima. Simulation results show that the proposed intelligent algorithm outperforms other existing approaches. The proposed algorithm is able to derive new upper bounds for test case.

Acknowledgements

Authors appreciate Dr. Shih-Wei Lin for providing the data and discussions.

Reference

1. Baker, K.R., Scudder, G.D.: Sequencing with Earliness and Tardiness Penalties: A Review. Operations Research 38, 22–36 (1990)
2. Biskup, D., Feldmann, M.: Benchmarks for Scheduling on A Single Machine Against Restrictive and Unrestrictive Common Due Dates. Computers and Operations Research 28, 787–801 (2001)
3. Bose, R.: Information Theory, Coding, and Cryptography. McGraw Hill, New York (2002)
4. Gordon, V., Proth, J.-M., Chu, C.: A Survey of the State-of-art of Common Due Date Assignment and Scheduling Research. European Journal of Operational Research 139, 1–25 (2002)
5. Gupta, J.N.D., Lauff, V., Werner, F.: Two-machine Flow Shop Problems with Nonregular Criteria. Journal of Mathematical Modelling and Algorithms 3, 123–151 (2004)
6. Liaw, C.F.: A Branch-and-Bound Algorithm for the Single Machine Earliness and Tardiness Scheduling Problem. Computers and Operations Research 26, 679–693 (1999)
7. Lin, S.W., Chou, S.Y., Ying, K.C.: A Sequential Exchange Approach for Minimizing Earliness-Tardiness Penalties of Single-machine Scheduling with a Common Due Date. European Journal of Operational Search 177, 1294–1301 (2007)
8. Jaynes, E.T.: On The Rationale of The Maximum Entropy Methods. Proceedings of the IEEE 70, 939–952 (1982)
9. Hoogeveen, J.A., Van De Velde, S.L.: Scheduling Around a Small Common Due Date. European Journal and Operational Research 55, 237–242 (1991)

10. Ibarraki, T., Katoh, N.: Resource Allocation Problems. The MIT Press, Cambridge, Massachusetts (1988)
11. Lee, Z.-J., Lee, C.-Y.: A Hybrid Search Algorithm with Heuristics for Resource Allocation Problem. Information sciences 173, 155–167 (2005)
12. Feldmann, M., Biskup, D.: Single-machine Scheduling for Minimizing Earliness and Tardiness Penalties by Meta-heuristic Approaches. Computers and Industrial Engineering 44, 307–323 (2003)
13. James, R.J.W.: Using Tabu Search to Solve the Common Due Date Early/Tardy Machine Scheduling Problem. Computers and Operations Research 24, 199–208 (1997)
14. Lee, C.Y., Kim, S.J.: Parallel genetic algorithms for the earliness-tardiness job scheduling problem with general penalty weights. Computers and Industrial Engineering 28, 231–243 (1995)
15. Liu, M., Wu, M.: Genetic Algorithm for the Optimal Common Due Date Assignment and The Optimal Policy in Parallel Machine Earliness/Tardiness Scheduling Problems. Robotics and Computer-Integrated Manufacturing 22, 279–287 (2006)
16. Mittenthal, J., Raghavachari, M., Rana, A.I.: A Hybrid Simulated Annealing Approach for Single Machine Scheduling Problems with Non-regular Penalty Functions. Computers and Operations Research 20, 103–111 (1993)
17. Lin, S.W., Chou, S.Y., Ying, K.C.: Meta-heuristic Approaches for Minimizing Total Earliness and Tardiness Penalties of Single-Machine Scheduling with a Common Due Date, Journal of Heuristics (2007)
18. Dorigo, M., Stützle, T.: Ant Colony Optimization. The MIT Press, Cambridge (2004)
19. Lee, C.-Y., Lee, Z.-J., Su, S.-F.: Ant Colonies With Cooperative Process Applied To Resource Allocation Problem. Journal of the Chinese Institute of Engineers 28, 879–885 (2005)
20. Hall, N.G., Kubiak, W., Sethi, S.P.: Earliness-Tardiness Scheduling Problem, II: Deviation of Completion Times About a Restrictive Common Due Date. Operations Research 39, 847–856 (1991)
21. Kahlbacher, H.G.: Scheduling with Monotonous Earliness and Tardiness Penalties. European Journal of Operational Research 64, 258–277 (1993)
22. Mondal, S.A., Sen, A.K.: Single Machine Weighted Earliness-Tardiness Penalty Problem with a Common Due Date. Computers and Operations Research 28, 649–669 (2001)
23. Raghavachari, M.: Scheduling Problems with Non-regular Penalty Functions: A Review. Operations Research 25, 144–164 (1988)
24. Szwarc, W.: Single-machine Scheduling to Minimize Absolute Deviation of Completion Times from a Common Due Date. Naval Research Logistics 36, 663–673 (1989)
25. Hino, C.M., Ronconi, D.P., Mendes, A.B.: Minimizing Earliness and Tardiness Penalties in a Single-machine Problem with a Common Due Date. European Journal of Operational Research 160, 190–201 (2005)

Application of Intelligent Neural Networks to Prediction of Micro-electroforming for U-Type Micro-cavity

Sheau-Wen Shiah[1], Pai-Yu Chang[2], Tzeng-Yuan Heh[1], Po-Hung Lin[1], and Fu-Cheng Yang[1]

[1] Department of Power Vehicle and System Engineering, Chung Cheng Institute of Technology, National Defense University
[2] Department of Product Design, Fortune Institute of Technology
{Sheau-Wen Shiah, Tzeng-Yuan Heh, Fu-Cheng Yang, Po-Hung Lin, g970403}@ccit.edu.tw

Abstract. The physical transport mechanisms of micro-electroforming are innately complex and generally entail a great effort to comprehend the nature through either experimental measurements or theoretical simulations. In this study, the approach of using neural networks is implemented for demonstrating its effectiveness in the real-time determination of an electrolyte to achieve on U-type micro-cavities. Three intelligent back-propagation neural networks are established via the training process with the Kondo's database to predict the distributions for cathode various parameters *Pe* and for various cavity widths are determined. Comparisons of the predictions with the test target vectors indicate that the averaged root-mean-squared errors from three back-propagation neural networks are well. This study also examines in detail the effects of various network parameters, including number of hidden nodes, transfer function type, number of training pairs, learning rate-increasing ratio, learning rate-decreasing ratio, and momentum value, on the performance of neural networks.

Keywords: micro-electroforming, neural networks, back-propagation.

1 Introduction

The development of technology has advanced to the point where new products are constantly being developed to achieve convenient features, leading to the second generation micro-system technology. Nano-engineering and micro-system technology will be the next major developments in manufacturing technology. These technologies can be integrated with high precision manufacturing and semiconductor technologies and collocated with technologies of synchronous radiation using new materials, modulus and elements. Micro-electroforming technology has been successfully applied in various fields, for example, information, communication, optical electronics and biomedicine. First, the effect of internal stress, see relevant research by Yang and Kang [1], indicated that the cladding material induces the phenomena of twist and deformation, also used the second cathode to reduce the internal stress in the

B. Apolloni et al. (Eds.): KES 2007/ WIRN 2007, Part II, LNAI 4693, pp. 696–703, 2007.
© Springer-Verlag Berlin Heidelberg 2007

Ni-cladding process and increase the uniformity of the Ni-cladding material. Savinell et al. [2] examined micro-electroforming and used pulse current to increase the tension and yielding stress for the alloy of Ni-P. Moreover, Tan and Chan [3] used reverse pulse current technology and reduced the internal stress in micro-electroforming for Ni material. Furthermore, the effect of Current density, DC source is frequently used for micro-electroforming. However, the larger internal stress in the cladding layer causes the surface to be uneven. Lee et al. [4] and Yin et al. [5] used the pulse current for the micro-electroforming of Ni-W alloy, and determined the degree to which the wolfram material is increased by the pulse current. To determine the stirring effect, Andricacos et al. [6] used the ring disk technology for depositing Ni-Fe film in the cathode electrode, increased the convection flow effect, obtained much alloy ion mass diffusion and reduced the proliferation thickness. Yamasaki et al. [7] noted that chloride ammonium concentration in electrolyte seriously influenced the Ni-W alloys of the cladding layer. When chloride ammonium 0.5 mol/l was taken, the optimum wolfram content in cladding layer was obtained, while an excessive chloride ammonium concentration limited the wolfram content of the cladding layer. Gould and Lopez [8] conducted sandwich structure electroforming on SiO_x, and found that the gold electrolyte is more stable than aluminum electrolyte.

For the numerical analysis and the discussion of the physical aspects of micro-electroforming are sourced in the last 20 years. Hessami and Tobias [9] simulated the co-deposit of Ni-Fe alloys, and established diffusion and convection equations. Pesco and Cheh [10] demonstrated that electrolyte mass was only affected by diffusion flow for cavities with a high aspect ratio. However, both diffusion mass flow effect and convection mass flow effect existed when the electrolyte flow was laminar. Kondo et al. [11] described the width of the cavity more 200 μm, the effect of convection mass transfer governs the behavior of the electrolyte for micro-electroforming. However, when cavity width is less 100 μm, the effect of diffusion mass transfer is to dominate.

2 Micro-electroforming and Neural Network

The electrochemistry reaction of micro-electroforming is to deposit the metal ion or alloy ion for the microstructure. The micro-electroforming process to satisfy presented in the electrolyte flow, electrolyte mass flow and electrolyte voltage distribution. The working fluid is electrolyte 0.6 mol/liter $CuSO_4$ + 1.85 mol/liter H_2SO_4, which the electrolyte with constant properties. The temperature of copper electro-deposition bath is 20° C. A uniform chemical reaction rate is processing for the electrolyte in the U-type micro-cavity, and the process is steady state and laminar flow. To determine the domain of evaluation, the fully developed velocity boundary layer thickness δ, can be represented approximately by Schmidt number *Sc* following from Incropera and DeWitt [12]:

$$\delta / \delta_c \approx Sc^{1/3} = (\nu / D)^{1/3} \tag{1}$$

where δ, δ_c respectively present the velocity boundary layer thickness and the concentration diffusion boundary layer thickness, υ is the kinematic viscosity and *D* is the mass diffusion coefficient. The value of $Sc = \upsilon / D \approx 64$ was calculated by $\upsilon =$

1.07×10^{-6} m^2/s, D= 1.67×10^{-8} m^2/s. Because the concentration diffusion boundary layer thickness δ_c is very thin and let the concentration is C_b at the inlet face, so the concentration diffusion boundary layer thickness δ_c represents the cavity depth b. Therefore, the velocity boundary layer thickness δ = 4b= h could be evaluated to satisfy Schmidt number *Sc*, i.e., the height of the free concentration electrolyte surface and the height of the free velocity electrolyte surface both could be determined. The values of inlet and outlet velocity gradient in the calculated domain are proportional to *Pe* as shown by Kondo et al. [11]. The values of inlet and outlet velocity gradient could be calculated by velocity boundary layer thickness $\delta = 4\times10^{-5}$ m and mass diffusion coefficient D= 1.67×10^{-8} m^2/s.

$$Pe = [\delta^2 (du/dy)_{y=0}]/D \tag{2}$$

Table 1. Test Conditions for Examining the Effects of Network Parameters on Training Quality

Test Condition	Input Layer	Hidden Layer			Overall Parameters		
	Training Pairs	Hidden Layer	Hidden Nodes	Transfer Function	*Lr-inc*	*Lr-dec*	Momentum
A1	39	1	1	Sigmoid	1.06	0.7	0.9
A2	39	1	2	Sigmoid	1.06	0.7	0.9
A3	39	1	4	Sigmoid	1.06	0.7	0.9
A4	39	1	8	Sigmoid	1.06	0.7	0.9
A5	39	1	16	Sigmoid	1.06	0.7	0.9
B1	39	1	16	Hyper-tang	1.06	0.7	0.9
B2	39	1	16	Pure-linear	1.06	0.7	0.9
C1	29	1	4	Sigmoid	1.06	0.7	0.9
C2	19	1	8	Sigmoid	1.06	0.7	0.9
C3	9	1	16	Sigmoid	1.06	0.7	0.9
D1	39	1	16	Sigmoid	1.2	0.7	0.9
D2	39	1	16	Sigmoid	1.5	0.7	0.9
D3	39	1	16	Sigmoid	2.0	0.7	0.9
E1	39	1	16	Sigmoid	1.06	0.6	0.9
E2	39	1	16	Sigmoid	1.06	0.8	0.9
E3	39	1	16	Sigmoid	1.06	0.9	0.9
F1	39	1	16	Sigmoid	1.06	0.7	0.6
F2	39	1	16	Sigmoid	1.06	0.7	0.7
F3	39	1	16	Sigmoid	1.06	0.7	0.8

A neural network is a computational system, which employs artificial neurons linked mutually to simulate a biological neural network. An artificial neuron, considering as a simple imitation of a physical neuron, can acquire information from the environment or neurons in the upstream layer and dispatch its calculation result back to the environment or to neurons in the downstream layer [13,14]. Among a variety of neural network models, back-propagation networks are most widely used in

the areas of diagnosis and prediction mechanisms. As a supervised learning network, the back-propagation network is first trained with the training sets, consisting of known input and output vectors. It learns the nonlinear relationship between input and output vectors based on an existing database from theoretical computations and/or experimental measurements. This nonlinear relationship can be expressed in the form of connected weights between nodes and the bias of each node. In this work, the back-propagation-learning rule is used to adjust the weights and biases of networks for minimizing the sum squared error of the network. When certain inexperienced input vectors are presented, trained back-propagation networks can readily predict reasonable output vectors for those new cases [15].

In this study, three intelligent back-propagation neural networks are established via the training process with the Kondo's database to predict the distributions of the mass fractions Sh/Sh_{max} on the surface of the cathode for various parameters Pe = 1.31~ 41.6, and for various cavity widths a= 15~ 150 μm are determined. The training sets are obtained from Kondo's results introduced above. The training quality can be then determined using sum-squared error (SEE) or averaged root-mean-squared error (RMS), defined as follows.

$$SSE = \sum_{j=1}^{m} \sum_{i=1}^{n} \left(T_{ij} - P_{ij} \right)^2 \tag{3}$$

$$RMS = \frac{1}{m} \cdot \frac{1}{n} \cdot \sqrt{SSE} \tag{4}$$

Where m, n, T, P stand for the number of training pairs, the number of output nodes, the target value, and the predicted value by the associated neural network. This study also examines the effects of different network parameters, including number of hidden nodes, type of transfer function, number of training pairs, learning rate-increasing ratio, learning rate-decreasing ratio, and momentum value, on the training quality. The test conditions are listed as an analysis matrix in Table 1.

3 Results and Discussion

When a=100 μm and Pe=1.31~41.6, the mass fraction Sh/Sh_{max} on the cathode surface is evaluated to be as illustrated in Fig. 1.The obtained results for Pe=1.31 and 41.6 agree excellently with the numerical data of Kondo et al. The root-mean-squared deviations were under 2.13 %. Therefore, the numerical data of Kondo et al. were obtained to ensure the accuracy of the results herein. At Pe=1.31, the copper is transported to the cathode by diffusion flow because more copper ions on the upstream side. The vortex captures more copper ions, the local mass transfer to the cathode is clearly increased, causing the flux to maximal upstream of X= 0. When a high copper ion concentration causes the penetration flow, the fresh bulk electrolyte first passes through the upstream side, and then moves along the surface of the cathode to the downstream side. During this process, the copper ions are consumed. When they have reached the downstream side, the local mass transfer to the cathode is again increased to an extent determined by diffusion from the vortex. Therefore, the increase in the flux is at downstream of X=0.9~1.0. For Pe= 41.6, the hump is slightly shifted and appeared at X=0.16; this process is mainly governed by both convection

of the penetration flow and the effect of the vortices at the corners. Accordingly, the flux gradually decrease toward the downstream side and is flattened by the local resistance at X=0.95~1.0. The Sh/Sh_{max} in cases Pe=20 is similar to that at Pe =41.6.

Fig. 2 presents the mass fraction for Pe=1.31 and a=15~150 μm. The Sh/Sh_{max} values on the surface of the cathode are evaluated, and those at a= 15 and 30 μm agree closely with the numerical data obtained by Kondo et al. The root-mean-squared deviations were under 1.37 %. The flux shows that the 30 μm cavity width peaks at about X=0. The flux also increases slightly at X= 0.75~1, in which range the copper ions are captured by the recirculation vortices. The profile at a cavity width of a= 30 μm resembles those at a= 100 and 150 μm. However, for a cavity width a= 15 μm, the profile becomes almost flat and little difference is observed between the upstream and downstream sides.

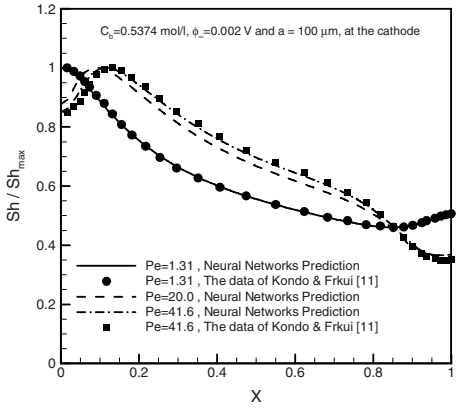


Fig. 1. Mass fraction Sh/Sh_{max} versus X for a=100 μm and various Pe=1.31~ 41.6

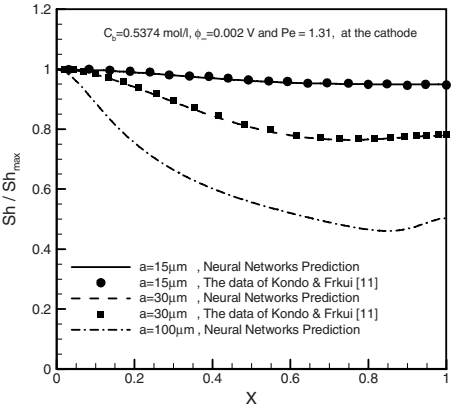


Fig. 2. Mass fraction Sh/Sh_{max} versus X for Pe=1.31 and various a=15~100 μm

In the present study, the effects of various neural network parameters on the training quality were also examined extensively. Fig. 3 illustrates the effect of number of hidden nodes on the training quality. Referring to the test conditions A1 to A5, the averaged RMS errors are plotted with respect to the number of hidden nodes changed from 1 to 16. In most cases, the increase in the number of hidden nodes tends to reduce the averaged RMS errors. However, for all three networks, addition of hidden nodes to a certain critical number could lead to excessive complexity of network structure and in turn result in larger training errors along with larger storage space and longer computation time for training needs. In the current cases, the predicted results based on 16 hidden nodes display negligible improvement of training quality over the predictions based on 8 hidden nodes.

Different transfer functions can substantially affect the training expediency. This investigation considers three types of transfer functions, comprising sigmoid, hyperbolic tangent, and pure linear function, defined as below:

Sigmoid: $f(x)=1/(1+e^{-x})$□Hyperbolic tangent: $f(x)=tanh(x)$□Pure linear: $f(x)=x$.

Fig. 4 shows the effect of transfer function on the training quality. The setting values of network parameters are given in test cases of A5, B1, and B2 in Table 1.

The averaged RMS errors based on the sigmoid function are obviously much lower than those based on the hyperbolic tangent and the pure linear functions. As a result, it can be inferred that the sigmoid function is the most appropriate transfer function for all three neural networks.

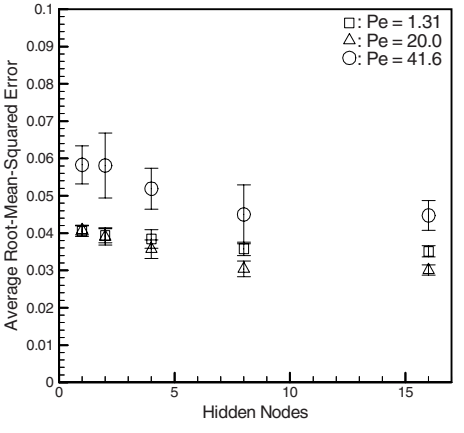


Fig. 3. Effect of Number of Hidden Nodes on Averaged RMS Error

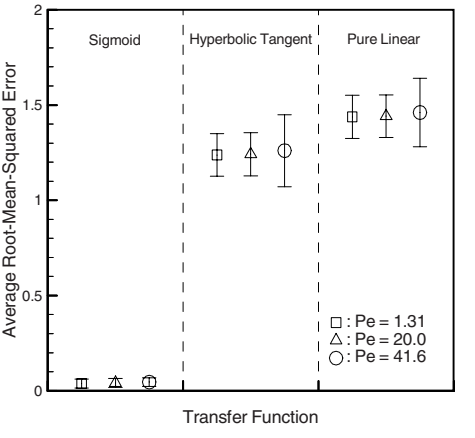


Fig. 4. Effect of Transfer Function r on Averaged RMS Error

As shown in Fig. 5, the effect of the number of training pairs on the training quality is studied in accordance with the test conditions A5 and C1 to 3 of Table 1. The averaged RMS errors are charted versus training pairs ranging from 9 to 39. For all three networks, the training quality is enhanced somewhat when the number of training pairs is increased up to 29. As the number of training pairs is greater than 29, it shows insignificant effect on the averaged RMS errors.

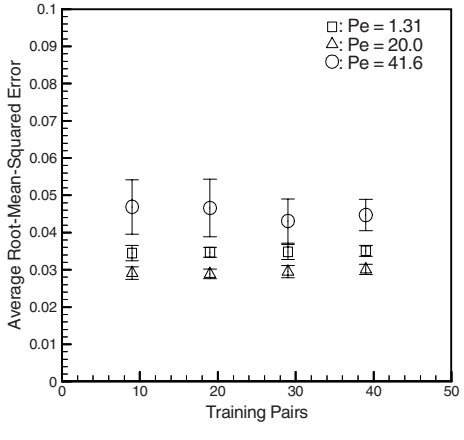


Fig. 5. Effect of Number of Training Pairs on Averaged RMS Error

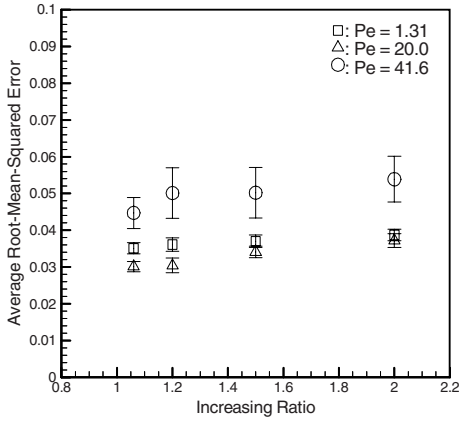


Fig. 6. Effect of Increasing Ratio of Learning Rate on Averaged RMS Error

Fig. 6 reveals the effect of increasing ratio of learning rate on the averaged RMS error. The neural network parameters for acquiring the results can be referred to the test conditions of A5 and D1 to D3 in Table 1. The initial learning rate was set to be 0.001. Following the typical expression, the *Lr-inc* is defined as the ratio of the new learning rate to the old learning rate as the ratio of SSE between iterations is greater than a certain prescribed error ratio, here set at 1.04. Similarly, when the ratio of SSE between iterations is less than the prescribed error ratio, the ratio of the new to the old learning rates is named as decreasing ratio of learning rate. As studying Fig. 6, the lowest averaged RMS errors are attained at *Lr-inc*=1.06 for all three back-propagation networks. Therefore, an increasing ratio of learning rate on the brink of unity tends to be favorable for achieving a better training quality. In addition, Fig. 7 divulges the effect of decreasing ratio of learning rate on the averaged RMS error with network parameters specified in test conditions A5 and E1 to E3 of Table 1.

By changing the momentum value from 0.6 to 0.9, the effect of momentum on the training quality is presented in Fig. 8 with respect to test conditions A5 and F1 to F3 of Table 1. For the void fraction network, the averaged RMS error is minimized as the momentum value is over 0.8. Further, the training errors are relatively not influenced by the momentum effect for the liquid velocity and gas velocity networks. Among four test cases, the best training quality could be achieved at the momentum value of 0.9 for all three neural networks.

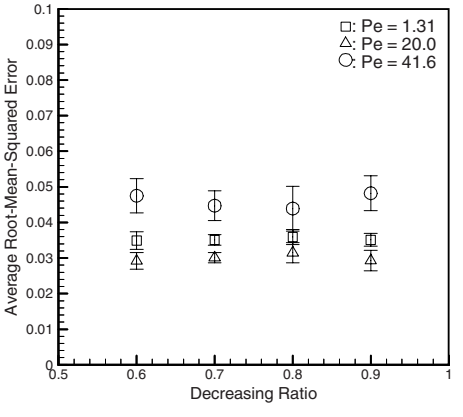


Fig. 7. Effect of Decreasing Ratio of Learning Rate on Averaged RMS Error

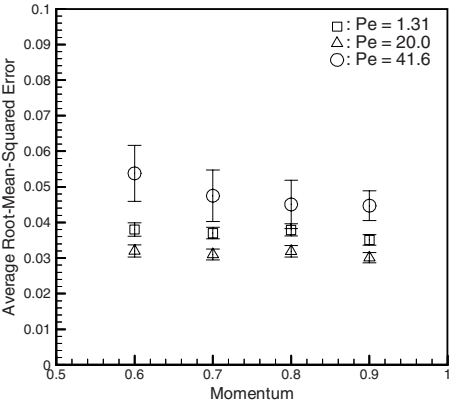


Fig. 8. Effect of Momentum on Averaged RMS Error

4 Conclusion

In this study, Effects of various network parameters on the training quality are also explored thoroughly. It is found that the sigmoid function is the most suitable transfer function for all three back-propagation networks. Within the range of 16 hidden nodes and 39 training pairs, the best training quality can be attained when the numbers of hidden nodes and training pairs are greater than 8 and 29. The averaged RMS errors can be minimized as the increasing ratio of learning rate and the momentum value are

set to be 1.06 and 0.9, correspondingly. The approach of using neural networks is implemented for demonstrating its effectiveness in the real-time determination of an electrolyte to achieve on U-type micro-cavities. Three intelligent back-propagation neural networks are established via the training process with the Kondo's database to predict the distributions for cathode various parameters *Pe* and for various cavity widths are determined. Comparisons of the predictions with the test target vectors indicate that the averaged root-mean-squared errors from three back-propagation neural networks are well.

References

1. Yang, H., Kang, S.W.: Improvement of Thickness Uniformity in Nickel Electroforming for the LIGA Process. J. Machine Tools and Manufacture 40, 1065–1072 (2000)
2. Savinell, R.F., Zeller, R.L., Adams, J.A.: Electrochemically Active Surface Area: Voltammetric Charge Correlation for Ruthenium and Iridium Dioxide Electrodes. J. Electrochem. Soc. 137(2), 489–494 (1990)
3. Tan, H.J., Chan, K.C.: Numerical Analysis of an Inside-out Tube Inversion Process. J. of Materials Processing Technology 66, 130–136 (1997)
4. Lee, S.L., Lee, Y.F., Chang, M.H., Lin, J.C.: Pulse Plating Effects During Ni-W Electrodeposition. Corrosion Prevention & Control, 71–76 (1999)
5. Yin, K.M., Jan, S.L., Lee, C.C.: Current Pulse with Reverse Plating of Nickel-Iron Alloys in a Sulphate Bath. Surface and Coatings Technology 88, 219–225 (1996)
6. Andricacos, P.C., Tabib, J., Romankiw, L.T.: Stripping Voltammetry of Nickel-Iron Film Electrodeposited on Platinum Using a Rotating Ring-Disk Electrode. J. Electrochem. Soc. 135(5), 1172–1174 (1988)
7. Yamasaki, T., Schlobmacher, P., Ehrlich, K., Ogino, Y.: Formation of Amorphous Electrodeposited Ni-W Alloys and Their Nanocrystallization. Nanostructured Materials 10(3), 375–388 (1998)
8. Gould, R.D., Lopez, M.G.: Electrical Conductivity and Dynamics of Electroforming in Al-SiOx-Al Thin Film Sandwich Structure. Thin Solid Films 433, 315–320 (2003)
9. Hessami, S., Tobias, C.W.: A Mathematical Model for Anomalous Codeposition of Nickel-Iron on a Rotating Disk Electrode. J. Electrochem. Soc. 136, 3611–3616 (1989)
10. Pesco, A.M., Cheh, H.Y.: The Current Distribution within Plated Through-Holes: II. The Effect of Periodic Electrolysis. J. Electrochem. Soc. 136(2), 408–414 (1989)
11. Kondo, K., Fukui, K., Uno, K., Shinohara, K.: Shape Evolution of Electrodeposited Copper Bumps. J. Electrochem. Soc. 143(6), 1880–1886 (1996)
12. Incropera, F.P., DeWitt, D.P.: Fundamentals of Heat and Mass Transfer, 5th edn., pp. 348–354. John Wiley and Sons, New York (2002)
13. Hertz, J., Krogh, A., Palmer, R.G.: Introduction to the Theory of Neural Computation. Addison-Wesley, New York (1991)
14. Demuth, H., Beale, M.: Neural Network Toolbox User's Guidance. The Math Works Inc., Natick, Massachusetts (1993)
15. Rumelhart, D.E., Hinton, G.E., Williams, R.J.: Learning Internal Representation by Error Propagation, Parallel Distributed Processing. MIT Press, Cambridge, MA (1986)

Design of a Supervisor for Traffic Light Systems

Ta-Hsiang Chung[1] and Yi-Sheng Huang[2,*]

[1] School of Defense Science, Chung Cheng Institute of Technology,
National Defense University, Taiwan, R.O.C.
[2] Department of Electrical and Electronic Engineering, Chung Cheng Institute of Technology,
National Defense University, Taiwan, R.O.C.
yshuang@ccit.edu.tw

Abstract. Timed Coloured Petri net (TCPN) has been utilized as a visual formalism for the modeling of complex systems. A module of basic traffic system of TCPN model is successfully constructed in this paper. Especially, a supervisor of an urban traffic light system, which consists of five intersections with multiple phases, is implemented by the new methodology. And all the traffic operations are ruled by the control logic of TCPN. In addition, a new design tool is proposed and it is efficiently to state and solve the problem of traffic signal timing plans. Finally, a supervisor of the traffic light system is completed whose operation flow is modeled by the TCPN models.

Keywords: timed coloured Petri nets, traffic control, signal timing plan.

1 Introduction

With the growing number of vehicles, the traffic congestion and transportation delay on urban arterials are increasingly worldwide. Therefore it is of practical importance to develop, verify and validate simple, yet powerful models that help in design and improve the safety and efficiency of transportation. It is a significant issue to control traffic lights in road-vehicle systems. The main reasons to the traffic signals are used to manage conflicting requirements for the use of road space – often at road junctions – by allocating right of way to different sets of mutually compatible traffic movements during distinct time intervals.

The problems of traffic light plan optimization gathers all the control strategies that act with the aim of reducing traffic congestions, minimize queues at the intersections, which are provided by the model [1]. In [1] the relations between states can be observed, however the procedure and the concurrence can not be described explicitly. Petri nets (PN) have been proven to be a powerful modeling tool for various kinds of discrete event systems [2-3], and its formalism provides a clear means for presenting simulation and control logic. Hence, the PN is availably applied in traffic control. In [4] chooses Deterministic and Stochastic Petri Nets (DSPNs) as the modeling tools. The behavior of the pre-timed two-phase signal is depicted**.** The green periods, and

* Corresponding author.

B. Apolloni et al. (Eds.): KES 2007/ WIRN 2007, Part II, LNAI 4693, pp. 704–711, 2007.
© Springer-Verlag Berlin Heidelberg 2007

the cycle lengths are predetermined and of fixed duration. The approach via programmable logic control (PLC) and Petri nets synthesis is proposed in [5]. Its traffic lights contain three color lights, and each lane has a series light whose signal will be changed by the regular time intervals. And the model of traffic light is used by Petri nets. In [6] a signal timing plan is proposed by timed Petri net (TPN). The streams allowed us to proceed with the eight-phase of the signal timing plan. Recently, an urban traffic light controller using statecharts is proposed [7] which includes eight-phase, six-phase and two-phase.

As mentioned above, the models of traffic light only have one set of phase duration. But just we know the timing plan of traffic light should take into account a given time of day and determine the optimal splits; and optimal cycle time, based on the historical constant demands over the considered signalized urban area. It means that there should have many sets of phase durations which vary with the traffic flow in a day. As a result, the authors propose a new modeling methodology to tackle this difficult problem in this paper.

2 The Description of the Basic Traffic System

The vehicle movement that goes straight and turns right is usually to see in an intersection with two-way roads. Hence, such a system will be introduced in this section, is called a basic traffic system in this paper. The green lights are changed back and forth between the two directions (*NS* and *EW*). There are two phase in the basic traffic system. And the two phases consist of phase_*NS* and phase_*EW*. Their operations are described as follows. In phase_*NS*, the northbound and southbound vehicles are allowed to go straight and turn right. In phase_*EW*, the westbound and eastbound vehicles are allowed to go straight and turn right. Table 1 shows the schedule of the traffic light signals. There are a few definitions have to introduce in order to obtain TCPN model. The definition of "period cycle" represents one cycle of a period. And the definition of "day cycle" states one cycle of a day. As a result, there are several periods in a day cycle. In addition, the phase_*N*S duration means the duration in *NS* direction of the basic traffic system. Similarly, the phase_*EW* duration represents the duration in *EW* direction of the basic traffic system. The duration of the phase includes both the duration of green and the duration of yellow. A cycle time is defined as the summation of the two durations. Note that there are filled with numerous repetition of cycle time in a period. For example, a period of a traffic system is 60 minutes and the traffic system's cycle time is 120 seconds. Hence, the numbers of repletion are 30.

Table 1. A schedule of traffic light signals

Period	t_1	t_2	…	t_i
Execution time	0-h_1	h_1-h_2	…	h_{i-1}-h_i
Phase_*NS* duration	τ_{11}	τ_{12}	…	τ_{1i}
Phase_*EW* duration	τ_{21}	τ_{22}	…	τ_{2i}
Cycle time	ct_1	ct_2	…	ct_j

3 A TCPN Model

We first introduce the new methodology which is based on a global clock. The global clock values represent the system model time where they may either be discrete time (i.e., all clock values are integer numbers) or continuous time (i.e., all clock values are real numbers). More precisely, each token has a time stamp on it. The time stamp describes the earliest model time at which the token can be moved by a binding element [8-9].

In this section, the paper focuses on how to model the traffic lights using TCPN. It is an interesting work for the traffic light system model. Fig. 1 shows a model of TCPN which is obtained from the basic traffic system. The model consists of two main parts: the left part, called *NS* traffic lights, describes the states of both the northern and southern traffic lights; the right part, called *EW* traffic lights, depicts the states of both the eastern and western traffic lights. More precisely, each part can be divided into two sub-models. One model is for signal indication; the other is for the signals' timing. It is important to point that the latter model is applied in the signal indication model's running. The yellow duration is assumed as 3 seconds and the lost time are neglected in this paper. The more detail information is described as follows.

3.1 Signal Indication Model

Usually, the green light means the vehicles can go straight and turn right. And the vehicles do not permit turning left. In the former papers [4-5], this situation is hardly presented in detail. However, this case is easily modeled by the new methodology. A signal indication model can show the traffic light in red, yellow or green. The traffic lights are modeled as places *EW_G*, *EW_Y*, *EW_R*, *NS_G*, *NS_Y*, and *NS_R* (i.e., Fig. 1). The places of the traffic lights are triggered by tokens in the CTPN model. Based on the rules of CTPN, the tokens are defined as 4-tuple (d, n, t, ct) with time stamp. The first element d represents the traffic lights in *NS/EW*. Thus it can be written as $d \in \{ns, ew\}$. The second one n (i.e., $n \in I^+$) represents the numbers of the repetition. The element t stands for the period time, where $t \in I^+$. The final element ct represents the cycle time, where $ct \in I^+$. Here, I^+ is non-negative integer.

Places *NS_G*, *NS_Y*, *NS_R*, *EW_G*, *EW_Y*, *EW_R*, *CP1* and *CP2* belong to this sub-model. A token in place *NS_G* means the green lights turn on in *NS* traffic lights. And then place *NS_Y* (*NS_R*) with a token represents the yellow (red) lights turn on in *NS* traffic lights. Similarly, the places *EW_G*, *EW_Y* and *EW_R* with tokens means the green lights, yellow lights and red lights turn on in *EW* traffic lights, respectively. Places *CP1* and *CP2* play the guard roles, which guarantee only one direction is allowed to proceed for the vehicle streams. For example, when the green light turns on in *NS*, the red light turns on in *EW*. More specifically, the two places ensure the safety of the traffic. The transitions of the system model will be introduced in the following paragraph.

The transitions *T5* and *T9* stand for the green durations. Because the phase duration involves green duration and yellow duration (i.e., 3 seconds), the time inscription τ-3 that deducts yellow duration from phase duration is a green duration. The transitions *T6* and *T10* represent the yellow duration (i.e., 3 seconds). The transition *T7* stands

for the red durations in *NS*. And then the red duration in *EW* is controlled by the token of place *CP1*. The physical means the red light of the traffic lights in *EW* is going on till the enabled binding element (*T8*, *e*) is taken. Obviously, the red duration of traffic lights in *EW* is the τ_{1i} seconds. It reveals that the vehicle streams only proceed in either *NS* or *EW*.

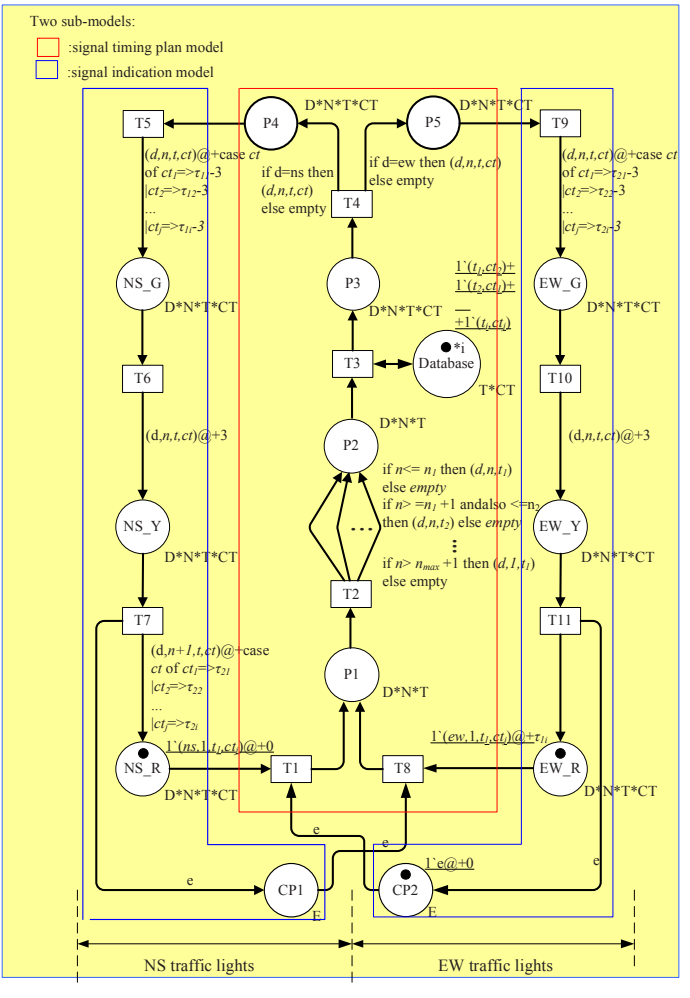


Fig. 1. A simple traffic light TCPN model

3.2 The Signal Timing Plan Model

The sub-model of the signal timing plan model is the kernel of the TCPN model (i.e., Fig. 1). It is the model's duty to assign the cycle times to the different periods. Especially, there are no any durations and delay times in the sub-model transition.

Indeed, it uses to estimate what is the cycle time of a period. In the sub-model, we must count the numbers (i.e., n_i) of the repetition of the cycle time. Once the numbers of repetition in the running system meet the numbers n_i, the current period t_i will enter next period t_{i+1}. And then a new cycle time will be given in terns of period t_{i+1}.

The definitions of the elements of tokens in the sub-model are the same as the signal indication model. It is worthy to notice that the component element of token in this sub-model consists of two (i.e., (t, ct)), three (i.e., (d, n, t)), or four (i.e., (d, n, t, ct)) elements. And the signal indication model consists of four elements (i.e., (d, n, t, ct)). The more explicit expression describing as follows.

There are several arc expressions between transition *T2* and place *P2*. The arc expressions determine when the current period goes to next period. The numbers of repetition of the cycle time in a "period cycle" is represented by r_i. And a values n_i is determined from the formula, where $n_i = \sum_{i=1}^{n} r_i, n \in I^+$. For example, $n_2=r_1+r_2$. Once the numbers of repetition in the running system are equal to n_i. The current period (t_i) will go to the next period (t_{i+1}). And then the current token (d, n, t_i) will change to (d, n, t_{i+1}). In addition, a maximum numbers of repetition $n_{\max}$ for the "day cycle" is required. Because the system model needs a mechanic to judge what is the day off. At the new day, the number n_{max} is enforced to reset. It means the state is coming back to the initial state.

In summary, when a token is removed from *P1* to *P2*, the token is with a t_i. And the token is removed from *P2* to *P3*, the token is now with a ct_i. Note that the ct_i is offered by the Place *Database*. The function of Place *Database* is that the model can show both of the cycle time and the current period together. Therefore, we have to create a database of (t_i, ct_j) by tokens in *Database*. Meanwhile, the output token is changed and it is with four elements like as (d, n, t_i, ct_i).

4 Modeling and Implementation of the Urban Traffic Model

For the analysis purpose, a real-world urban traffic net is used to verify the model. The real traffic net consists of five intersections which are shown in Fig.2. Table 2 shows the information of the phase transition in this real case. And also table 3 shows the signal timing plan of traffic light model. The urban traffic net model will be constructed based on the basic traffic model.

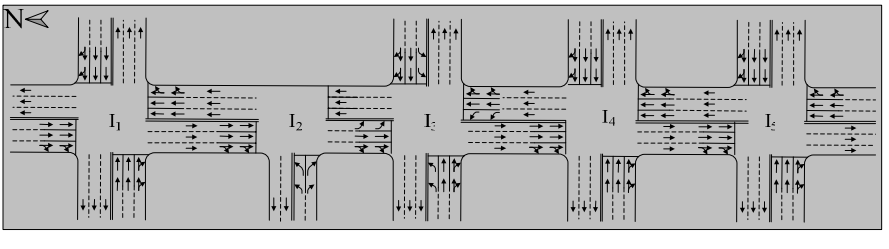


Fig. 2. The intersection of the case study

Table 2. The phase transition of the intersections

	Phase1	Phase2	Phase3	Phase4
I1,I4,I5				
I2				
I3				

Table 3. The schedule of the signal timing plan

	Period	t_1	t_2	t_3	t_4	t_5	t_6	t_7	t_8	t_9
	Execution time	00:00-01:00	01:00-05:00	05:00-07:00	07:00-09:00	09:00-13:00	13:00-16:30	16:30-19:00	19:00-23:00	23:00-24:00
I1	Phase 1	60	46	60	110	110	110	110	95	60
	Phase 2	60	44	60	90	90	90	90	85	60
I2	Phase 1	80	60	80	140	140	140	140	120	80
	Phase 2	40	30	40	60	60	60	60	60	40
I3	Phase 1	55	30	55	85	85	85	85	80	55
	Phase 2	10	10	10	15	15	15	15	15	10
	Phase 3	45	40	45	85	85	85	85	70	45
	Phase 4	10	10	10	15	15	15	15	15	10
I4	Phase 1	60	45	60	100	100	100	100	90	60
	Phase 2	60	45	60	100	100	100	100	90	60
I5	Phase 1	90	60	90	160	160	160	160	140	90
	Phase 2	30	30	30	40	40	40	40	40	30

The models of the intersection I_1, I_4 and I_5 are the same as basic traffic system's model. Hence, the module of basic traffic system could be applied directly to this type intersection. Form the schedule of the signal timing plan, there are four types of cycle time are given. For fitting the basic models, the notations $ct_1 = 90$, $ct_2 = 120$, $ct_3 = 180$ and $ct_4 = 200$ are assigned. And then the values of r_i and n_i have to be determined. Because the type intersections have the same cycle time in the same period, the values of r_i and n_i should be the same. In this group, an example I_1 is used to illustrate how to obtain the values r_i and n_i. Table 3 shows the execution time and the cycle time for every period of the intersection I_1. From the table, the numbers of the repetition (i.e., r_i) is easily to count. For example, the numbers of the repetition of r_1 is 30 for the period t_1 in the intersection I_1. As a result, all the numbers of the repetition (i.e., r_i) are obtained and are shown in table 4. Based on table 4, a formula is formed by the values n_i (i.e., given in table 4). The formula is given as $n_i = \sum_{i=1}^{n} r_i$, $n = 9$. For example, $n_2 = r_1 + r_2 = 30 + 160 = 190$. It is worthy to notice that the values of n_i in table 4 can be applied to intersections I_2 and I_3. In fact, the cycle times are the same between with the intersections I_1, I_2 and I_3.

Finally, three TCPN models should be derived if the values n_i, τ_{1i}, τ_{2i}, and ct_j are put into the original model (i.e., Fig. 1). Similarly, the models of the intersection I_2, I_3 will be obtained. The TCPN model of the urban traffic light system is shown in Fig. 3. Due to the space limitation, we only show the structure of TCPN model. It is worthy to notice that the intersection I_2 and I_3 involve a left turn arrow on green. As a result, the additional places, represented the left turn green light, are added on the original model.

Table 4. The values of r_i and n_i

i	1	2	3	4	5	6	7	8	9
r_i	30	160	60	36	72	63	45	80	30
n_i	30	190	250	286	358	421	466	546	576

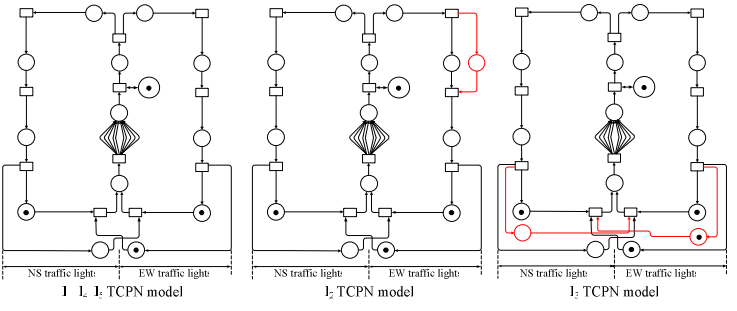


Fig. 3. The structure of the urban traffic light TCPN model

The Simulation Results. The TCPN model is implemented and simulated by the CPN tools. Aim of the simulation is to observe the relation between cycle time and



Fig. 4. A snapshot of the urban traffic light system

execution time in each period. In order to design and implement the traffic light control system, the object-orient programming technique is used. The demonstration program is written with Visual Basic 6.0 taking about 1.63 M bytes, whose operation flow is modeled by TCPN. Fig. 4 shows a snapshot of the urban traffic light system. In summary, the simulation results are the same as the predetermining time (i.e., table 3). And the cycle time of each period is consistent with table 3.

5 Conclusion

This article presents the modeling, analysis and implementation of an urban traffic lights system using TCPN models. The advantage of the proposed approach is the clear presentation of the system behavior and readiness for implementation. To summarize, this paper has the following contributions: 1) This paper has been demonstrated how to use TCPN to model the traffic lights for the urban traffic lights. And then the applications of TCPN to urban traffic lights have been realized. 2) The schedule of the signal timing plan of the traffic system has been converted into TCPN models. The examples are helpful us to obtain a TCPN model for a complex urban traffic lights system. 3) A module of basic traffic system of TCPN model has been constructed in this paper. The intended advantage of the proposed module is that designers can extend the module to meet complex traffic light systems.

The authors believe that our research using TCPN to model traffic light systems will become more important in this field due to the increasing demands in many features of the traffic light systems.

Acknowledgements. This work was supported by the National Science Council of Taiwan, R.O.C. under Grant NSC 94-2218-E-014-001.

References

1. Lei, J., Ozguner, U.: Decentralized Hybrid intersection control. In: proc. 40th IEEE Int. Conf. Decision and Control, vol. 2, pp. 1237–1242. IEEE Computer Society Press, Los Alamitos (2001)
2. Murata, T.: Petri Nets: Properties, Analysis and Applications. Proc. IEEE 77, 541–580 (1989)
3. Peterson, J.L.: Petri net theory and the modeling of systems. Prentice-Hall, Englewood Cliffs (1981)
4. Wang, H., List, G.F., DiCesare, F.: Modeling and evaluation of traffic signal control using timed Petri nets. Proc. Systems, Man and Cybernetics 2, 180–185 (1993)
5. Li, L., Tang, N., Mu, X., Shi, F.: Implementation of Traffic Lights Control Based on Petri Nets. Proc. IEEE Intelligent Transportation Systems 2, 1087–1090 (2003)
6. Dotoli, M., Fanti, M.P., Lacobellis, G.: Validation of an Urban Traffic Network Model using Colored Timed Petri nets. In: IEEE Int. Conf. on. Systems, Man and Cybernetics, vol. 2, pp. 1347–1352 (2005)
7. Huang, Y.S.: Design of Traffic Light Control System Using Statecharts. Computer Journal 49, 634–649 (2006)
8. Jensen, K.: Coloured Petri Nets, vol. 1. Springer, New York (1995)
9. Jensen, K.: Coloured Petri Nets, vol. 2. Springer, New York (1997)

A Knowledge-Based Approach for the Analysis of Abnormal Situations

Tetsuo Fuchino[1], Rafael Batres[2], and Yukiyasu Shimada[3]

[1] Department of Chemical Engineering, Tokyo Institute of Technology,
2-12-1, O-okayama, Meguro-ku, Tokyo 152-8552, Japan
fuchino@chemeng.titech.ac.jp
[2] Department of Production Systems Engineering, Toyohashi University of Technology,
1-1, Hibarigaoka, Tempaku, Toyohashi, Aichi 441-8580, Japan
rbp@pse.tut.ac.jp
[3] Chemical Safety Research Gr., National Institute of Occupational Safety and Health,
1-4-6, Umezono, Kiyose, Tokyo 204-0024, Japan
shimada@s.jniosh.go.jp

Abstract. This paper presents an approach to assist engineers in modeling and capturing abnormal situations. In these days, safety becomes a social strict requirement in the process industry. An incident influences the competitiveness and/or sustainability. Therefore, PHA has been carried out through the lifecycle. However, incidents and/or disasters occur as ever, and many incident investigation reports point out incompleteness of PHA, and reputation of past experienced similar incidents. In this study, a generic exchangeable representation of incident information is proposed. Our approach is based on the knowledge representation of both the plant structure and the abnormal events. A computer implementation is presented to demonstrate the efficacy of the representation method. Finally, a case study is included to illustrate the application of the proposed approach.

Keywords: abnormal situations, process hazard analysis, system safety.

1 Introduction

Recently, the safety regulations and the social safety requirements have become a strict and serious constraint for production, and realization of the plant life cycle safety is very much important issue not only to maintain competitive power, but also to accomplish the sustainability in production. Thus, PHA (Process Hazard Analysis) such as HAZOP (Hazard and Operability study)[1] plays an important role in safety management. The PHA is carried out in general through the cycle of 1) hazard identification, 2) hazard evaluation and 3) mitigation. However, many incident reports [2] point out the incompleteness of hazard identification in PHA, and reputation of past experienced similar incidents. If the potential hazard had identified from past incidents, then some incidents and/or disasters might have been prevented.

Although, there exist various databases which access incident reports using keywords, the incident reports are telling us the case dependent facts from the initial event and sequential events to the incident. It is heavily dependent on the person whether

B. Apolloni et al. (Eds.): KES 2007/ WIRN 2007, Part II, LNAI 4693, pp. 712–719, 2007.
© Springer-Verlag Berlin Heidelberg 2007

similar potential hazard can be captured analogically from such case dependent incident information. In order to utilize the past incident in the PHA effectively without personal experiences and/or knowledge, the case dependent information should be represented into the case independent generic exchangeable knowledge.

In the most of incidents in the process industry, an initial event causes a process abnormality, and the abnormality is propagated along the process structure. However, the process structures differ from plant to plant and process to process, and the propagation path through a plant and/or a process is depending on the incident case. It is the primary reason why the incident information could not but be expressed as a case dependent way in the incident reports.

On the other hand, HAZOP study is basically equivalent to tracing causality from possible initial events to possible hazardous condition on flow diagrams, and various flow diagrams to represent processes are used according to the planning and/or design stages, in chemical engineering. The experienced engineers, especially the chairmen of HAZOP study would generalize their experiences and be able to consider the initial causes and causalities analogously in the different process and/or process structure.

In this study, the activity of HAZOP study from the view point of experienced engineers is analyzed, and on the basis of this analysis, a generic representation of past experienced incident applicable for any process of flow diagram is proposed. A computer implementation is presented to demonstrate the efficacy of the representation method. Finally, a case study is included to illustrate the application of the proposed approach.

2 Activity Model of Performing HAZOP Study

In this study, IDEF0 (Integration DEfinition for Functional model) [3] is adopted to make the activity model of performing the HAZOP study under the corporation of experienced engineers. In IDEF0 activity model, the rectangles represent activities, and the arrows describe information. Each activity is decomposed into sub-activities, and the model has hierarchical nature. The information is classified into four categories; i.e., 'Input' to be changed by the activity, 'Control' to constraint the activity, 'Output' to be results of the activity and 'Mechanism' to be resources for the activity.

The top activity of HAZOP Study; “A-0: Perform Process Hazard Analysis” is decomposed into five sub-activities; “A1: Manage Process Hazard Analysis”, “A2: Identify Trigger Event”, “A3: Identify Propagation to Hazard”, “A4: Design Mitigation Plan” and “A5: Provide Resources for Perform Process Hazard Analysis”. The A1 to A5 activities are further decomposed into sub-activities. Although, only the decomposition of A2 and A3 activities are shown in Figures 1 and 2 from the limitation of space here, it is possible to know the experienced engineers’ way of thinking and/or their view point by focusing attention to the 'Mechanism' information.

From the 'Mechanism' information shown in Figs. 1 and 2 for A23, A32 and A34, which are sub-activities of A2 and A3 activities, it is clarified that the experienced engineers generalize their experiences as knowledge on equipment unit classes. In identifying trigger event, an equipment unit is specified on the process flow diagram, and a trigger event is provided from the knowledge on equipment unit class of the specified unit. The process status change by the trigger event is also considered using the

knowledge on unit class of the abnormal unit, and the abnormity propagation is traced on the process flow diagram. Consequently, the experienced engineers should utilize their knowledge on equipment unit classes and the process structure information from process flow diagram independently, and associate them each other in considering process abnormity propagation through the process. That is why the experienced engineers can apply their knowledge flexibly to any process and/or any level of process diagram.

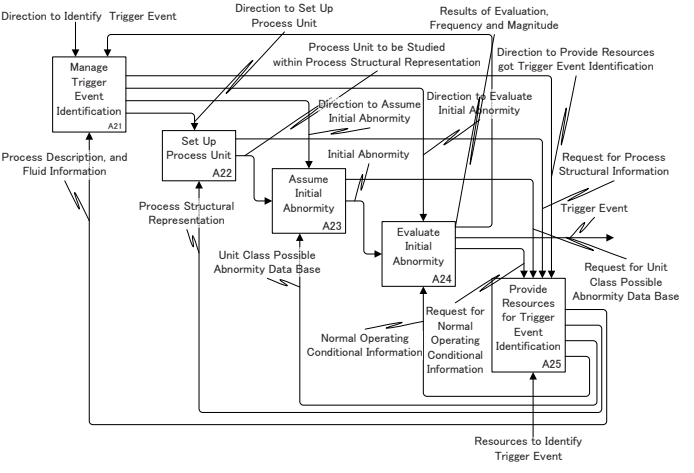


Fig. 1. Decomposition of "A2: Identify Trigger Event" activity

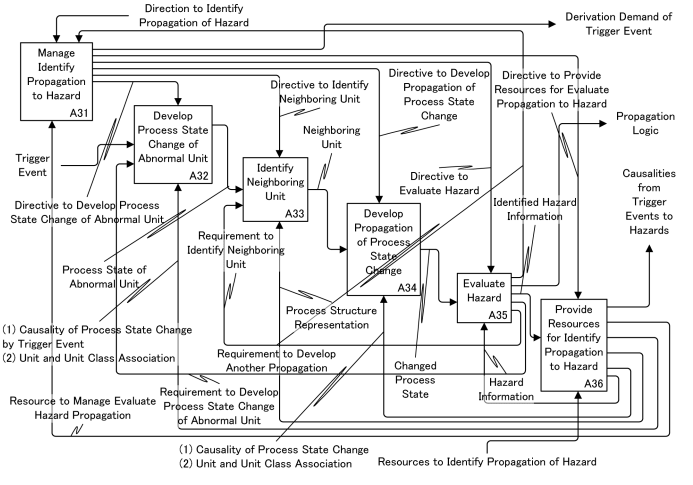


Fig. 2. Decomposition of "A3: Identify Propagation to Hazard" activity

In the previous study [4], we expressed the process abnormality with process state of process functional equipment unit class, and proposed the FT (Failure Tree) like logic diagram to represent a hazard scenario. However, the FT like logic diagram was

dependent on the process structure information, and the representation of the abnormal events was dependent on specific data about the plant equipment. Therefore, its application was very much limited.

In this study, the process structure and causality of process abnormity are represented independently [5], and the causality is considered for every equipment unit class.

3 Representation of Abnormal Situations

In order to explain the representation of abnormal situations resulting incidents, let us consider a case in which engineers are interested in the analysis of an abnormal situation around the feed surge drum shown in Figure 3. The feed of this process is diesel which is a flammable liquid. In this example, we assume that pump PU-201 fails-off causing an increase of liquid in feed surge drum D-201 which leads to overflow and liquid filling the vapor line. A situation like this can result in fires and explosion as documented at U.S. Chemical Safety and Hazard Investigation Board web site [2].

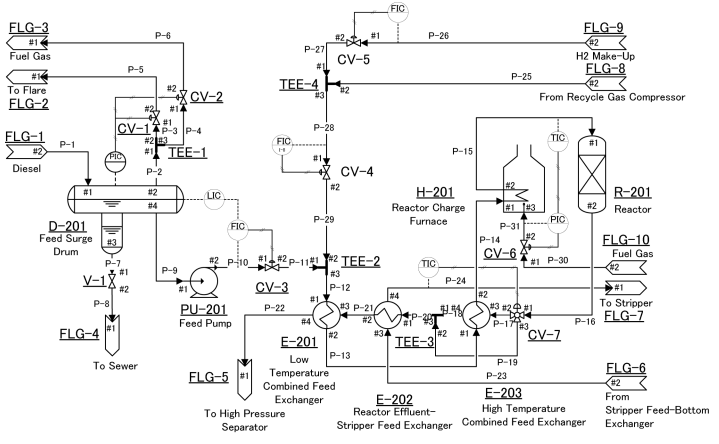


Fig. 3. A part of process flow diagram of HDS (Hydro De-Sulphurization) process

3.1 Process Structure Representation

In this study, representation of the process structure is on the basis of the ISO STEP (ISO-10303) [6], and GPM (General Plant Model) [7]. The knowledge is encoded in Prolog. The process topology is described by three kinds of prolog predicates; i.e. "topology", "obj_unit" and "port" as shown in Equations (1) to (3).

$$\text{topology('Connection_ID','Object_ID}_1\text{','Port_ID}_1\text{','Object_ID}_2\text{','Port_ID}_2\text{')}. \quad (1)$$

$$\text{obj_unit('Object_ID','Object_Class','Object_Description')}. \quad (2)$$

$$\text{port('Object_ID','Port_ID','Port_Type')}. \quad (3)$$

The “topology” predicate means 'Port_ID1' of 'Object_ID1' is topologically connected with 'Port_ID2' of 'Object_ID2' through 'Connection_ID', and “obj_unit” means 'Object_ID' classified in 'Object_Class' class is as 'Object_Description'. For example, the equipment pump PU-201, feed surge drum D-201 and control valve CV-1 are classified as 'Pump', 'Vapor_Liquid_Separator' and 'Control_Valve', respectively. The “port” predicate means 'Port_ID' of 'Object_ID' is 'Port_Type' class port. To consider the internal state of equipment unit and/or unit class, the hypothetical port is introduced in this study. Therefore, the port types are categorized into three types; i.e. 'In', 'Internal' and 'Out' ports.

To define process state, vapor-liquid phase is important for this example, so that “has_phase” predicate as shown in Equation (4) describes phase ('Liquid', 'Vapor' and 'Mixture') of all the port of all equipment units is defined.

has_phase ('Object_ID', 'Port_ID', 'Phase'). (4)

A part of the representation from 'D-201' to 'PU-201' is as below.

```
topology('P-9','D-201','#4','PU-201','#1').
obj_unit('D-201','Vapor_Liquid_Separator',
'Feed_Surge_Drum').
obj_unit('PU-201','Pump','Feed_Pump').
port('PU-201','#1','In').
port('PU-201','#0','Internal').
port('PU-201','#2','Out').
has_phase ('PU-201','#1','Liquid').
has_phase ('PU-201','#2','Liquid').
has_phase ('PU-201','#0','Liquid').
```

3.2 Abnormal Events

Here we focus on abnormal situations of processes that are caused by chains of events (abnormal events). Let us assume an abnormal situation described in terms of the following events.

1. PU-201 fails-off.
2. No flow through PU-201.
3. No flow at inlet-port #1 of PU-201.
4. No flow at outlet-port #4 of D-201.
5. Level of liquid in D-201 passes the high-level limit.
6. Liquid flows through gas port #2 of D-201.

Each step describes an elementary propagation, and they can be represented by “transition” predicate, which means 'Event1' at 'Port_Type1' having phase of 'Phase1' results in the presence of material with 'Phase2' in 'Port_Type2' giving place to 'Event2', as shown in Equation (5) generally using equipment unit class and port type.

transition ('Object_Class', 'Port_Type$_1$', 'Phase$_1$', 'Event$_1$', 'Port_Type$_2$', 'Phase$_2$', 'Event$_2$'). (5)

Some equipment unit class has two or more ports for the same type of port. The vapor liquid separator has at least two 'Out' ports; one is for vapor and the other is for liquid. To limit the ports concerning to "transition" predicate, 'Phase1' and 'Phase2' are introduced. Phase at the port of unit and unit class can be related by "has_phase" predicate as shown in Equation (4). A part of prolog fact code for the elementary propagation logic and phase definition are as following.

```
transition ('Pump','Internal','liquid','Failure_Stop',
'In','liquid','No_Flow').

transition ('Vapor_Liquid_Separator','Out','Liquid',
'No_Flow','Internal','Mixture','Level_High').

transition ('Vapor_Liquid_Separator','Internal',
'Mixture','Level_High','Out','Gas','Liquid_Flow').

transition ('Manual_Valve','Out','Liquid','No_Flow',
'Internal','Liquid','No_Flow').

logic('Manual_Valve','Internal','Liquid','No_Flow','In'
,'Liquid','No_Flow').
```

In this study, the elementary propagation concerning to the abnormal situation should be defined for all equipment unit classes to enable to apply the proposed knowledge representation to any plant.

The chain of events is initiated by means of the "trigger" predicate as shown in Equation (6), which means that the trigger 'Event' at 'Port_Type1' with phase of 'Phase1' changes the process state at 'Port_Type2' with phase of 'Phase2' to 'Abnormal_Event' state in the 'Unit_Class', and is coded for the example abnormal situation as shown below.

trigger('Unit_Class', 'Port_Type$_1$', 'Phase$_1$', 'Event', 'Port_Type$_2$', 'Phase$_2$', 'Abnormal_Event'). (6)

```
trigger('Pump', 'Internal', 'Liquid', 'Failure_Stop',
'In','Liquid','No_Flow').
```

The abnormal event at 'In' or 'Out' port of the trigger unit propagates through the plant units and finally a hazard event occurs. There are almost four patterns for abnormity propagation through a unit; 1) 'In' (Inlet) port to 'Internal' port and 'Internal' port to 'Out' (outlet) port, 2) 'In' port to 'Internal' port and 'Internal' port to 'In' port, 3) 'Out' port to 'Internal' port and 'Internal' port to 'In' port, 4) 'Out' port to 'Internal' port and 'Internal' port to 'Out' port. Which propagation pattern to be accommodated is depending on type and location of the unit, so that "causality" predicate to describe abnormity propagation through an unit is defined using Prolog rule as shown in Equation(7).

causality('Unit_Class', P_1, PC_1, A_1, P_2 ,PC_2, A_2):-
P_1='In', P_2='Out', P_0='Internal', transition('Unit_Class', P_1, PC_1, A_1, P_0, PC_0, A_0), (7)
transition ('Unit_Class', P_0, PC_0, A_0, P_2, PC_2, A_2);

P_1='Out', P_2='In', P_0='Internal', transition ('Unit_Class', P_1, PC_1, A_1, P_0, PC_0, A_0), transition ('Unit_Class', P_0, PC_0, A_0, P_2, PC_2, A_2);
P_1='In', P_2='In', P_0='Internal', transition ('Unit_Class', P_1, PC_1, A_1, P_0, PC_0, A_0), transition ('Unit_Class', P_0, PC_0, A_0, P_2, PC_2, A_2);
P_1='Out', P_2='Out', P_0='Internal', transition ('Unit_Class', P_1, PC_1, A_1, P_0, PC_0, A_0), transition ('Unit_Class', P_0, PC_0, A_0, P_2, PC_2, A_2).

where P is port type, PC is port condition, A is abnormal state and/or event. The "causality" predicate means event A_1 at port type P_1 having phase of PC_1 changes the process condition at port type P_2 with phase of PC_2 to A_1 state in the 'Unit_Class'. In case of shell and tube heat exchanger class, the chain of events is a little bit more complicated, but "causality" predicate is common to every abnormal situation, and can be prepared for every equipment unit before hand.

In tracing the abnormal events through the process, the direction of propagation is not always same as the direction of process flow. To consider just reachable direct from 'In' or 'Out' port of certain equipment unit, "reachable" predicate describing connection between two units regardless of process flow direction by using Prolog rule generally as shown in Equation (8) with "up_stream" and "down_stream" predicates as shown in Equations (9) and (10).

$$\text{reachable}(PP,OB_1,PN_1,OB_2,PN_2){:-}\ \text{up_stream}(PP,OB_1,PN_1,OB_2,PN_2);\ \text{down_stream}(PP,OB_1,PN_1,OB_2,PN_2). \tag{8}$$

$$\text{up_stream}(PP,OB_1,PN_1,OB_2,PN_2){:-}\text{topology}(PP,OB_1,PN_1,OB_2,PN_2). \tag{9}$$

$$\text{down_stream}(PP,OB_1,PN_1,OB_2,PN_2){:-}\text{topology}(PP,OB_2,PN_2,OB_1,PN_1). \tag{10}$$

where PP is pipe connection, OB is object unit, PN is port number.

Consequently, the process abnormity propagation can be represented independent form the process structure specific information. The tracing the abnormal situation through the process can be realized by performing query on the representation of process structure and the process abnormal event propagation. In this study, it is not the purpose to reproduce a abnormal situation automatically on computer, so that the propagation is just trace one by one basis in the next section.

4 Tracing Abnormal Situation

In this section, forgoing example abnormal situation is traced on the HDS process. To demonstrate the effect of the proposed representation of abnormal situation, two cases for process structure are considered. The process flow diagram as shown in Fig.3 is applied, for the first case. On the basis of the first case, a manual valve (V-2) is added between D-201 and PU-201 as shown in Fig. 4, for the second case.

For the first case, PU-201 internal ('Failure_Stop'), port #1 of PU-201 ('No_Flow'), port #4 of D-201 ('No_Flow'), D-201 internal ('Level_High') and port #2 of D-201 ('Liquid_Over_Flow') are traced in this order. On the other hand, PU-201 internal ('Failure_Stop'), port #1 of PU-201 ('No_Flow'), port #2 of V-2 ('No_Flow'), V-2 internal ('No_Flow'), port #1 of V-2 ('No_Flow'), port #4 of D-201 ('No_Flow'), D-201 internal ('Level_High') and port #2 of D-201 ('Liquid_Over_Flow') can be traced for

the second case. It is obvious that it is possible to trace a process abnormal event propagation for different level of process flow diagrams. The proposed representation of abnormal situation can be used to learn from the past experienced incident of any processes.

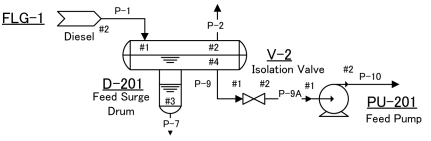


Fig. 4. Modified Process Flow Diagram

5 Conclusion

A novel representation of abnormal situation is proposed. Based on the experienced engineers' activity model of performing HAZOP study, abnormal situation representation policy is clarified. In this study, process structure and the process propagation are independently represented, and propagation from the trigger event to hazardous condition is traced recursively on these independent representations. The proposed representation can be applied for any process and/or any level of process flow diagram. The effectiveness of the proposed representation has been demonstrated by applying an abnormal situation for different process structures.

References

1. A Guide to Hazard and Operability Studies: Chemical Industry Safety and Health Council of the Chemical Industries Association (1981)
2. U.S. Chemical Safety and Hazard Investigation Board Web Site: http://www.csb.gov/
3. Federal Information Processing Standards Publications: Integration Definition for Function Modeling (IDEF0), http://www.itl.nist.gov/fipspubs/
4. Fuchino, T., Shimada, Y., Tanaka, T.: Knowledge Based Incident Database for Lifecycle Process Hazard Analysis. In: Proc. of 11th International Symposium on Loss Prevention, PG2475-PG2480 (2004)
5. Batres, R., Naka, Y., Lu, M.L.: A Multidimensional Design Framework and its Implementation in an Engineering Design Environment. In: Proc. of The Advances in Concurrent Engineering, Tokyo, Japan, pp. 73–82 (1998)
6. ISO 10303-221: Industrial automation systems and integration - Product data representation and exchange - Part 221, Functional Data and Schematics for Process Plants
7. Shibao, K.: Technology and Knowledge Infrastructure. J. of the Japan Society for Artificial Intelligence (Japanese) 18(2), 138–143 (2003)

A Multi-agent Approach to Process Design

Rafael Batres[1], Hikaru Takashima, and Tetsuo Fuchino[2]

[1] Department of Production Systems Engineering
Toyohashi University of Technology, Toyohashi, Japan
rbp@pse.tut.ac.jp
[2] Department of Chemical Engineering
Tokyo Institute of Technology, Tokyo, Japan
fuchino@chemeng.titech.ac.jp

Abstract. In this paper, a two-layer methodology is proposed, consisting of a process design layer and a facility design layer. The objective is the decision maker in finding technologies that can be combined to meet a set of requirements. Specifically, the approach consists of a technology selection agent (TSA) that submits design requests to technology agents. If a technology agent determines that the design is possible then proceeds to design the technology and submits the design to the TSA. Although the TSA is responsible to bundle the designed technologies into a solution, technology agents help to find the best solution by assisting the TSA in pruning away inefficient partial solutions.

1 Introduction

To compete in the ever-expanding global market, as well as to meet increasingly tighter safety and environmental constraints companies are being compelled to develop manufacturing systems in shorter time and less cost.

Manufacturing process design deals with the selection and order of the operations starting with raw materials and ending with the product. It is the stage between product design and production. Process design considers functional requirements of the product, quantities, tools and equipment, and eventually the costs for manufacture.

Much research has been done in the area of chemical process design [3]. Chemical process design methods include artificial intelligence methods (knowledge-bases and expert systems to generate and configure solutions) and mathematical approaches (optimization algorithms for problems defined as MINLP). Few research papers have been published in the area of automatic process planning of discrete manufacturing products. Nau et al. use AI planning and integer programming to design the process plan for the production of microwave modules [4]. Nau's approach is based on the *a priori* construction of a template that contains all the possible equipment items grouped by function which is similar to a superstructure used in process synthesis problems. This implies that the larger the number of equipment and operations the more complex the process template. In current methods, the designer requires the design and cost models

B. Apolloni et al. (Eds.): KES 2007/ WIRN 2007, Part II, LNAI 4693, pp. 720–727, 2007.
© Springer-Verlag Berlin Heidelberg 2007

to be available locally in the computer encoded in the format used by the specific method. Unfortunately, this is not always possible because many equipment and machine suppliers keep their design and cost calculations confidential.

This paper presents a framework for process design of discrete manufacturing systems, based on distributed cooperative agents. In this framework, each agent represents a piece of equipment and is assumed to be owned by the equipment manufacturer. In other words, the process design is carried out with equipment models that are distributed over the network (Internet). Agents communicate through the network by exchanging messages using a shared knowledge representation encoded in the Web Ontology Language [5]. Firstly the methodology is described then the multi-agent environment is explained. Finally, a case study is presented to illustrate the proposed approach.

2 Problem Statement

We formulate the problem as $\Psi = \langle r, p, A, s \rangle$, where

- R represents raw materials
- P represents desired products
- A is a set of agents $a \in A$ each of which has the ability to design a technology
- S is the solution that is composed of technologies each of which consume a feed and produce an output. The solution S is a set of technologies $s_0 ... s_n$ such that $r \subseteq R$, $r \subseteq in(s_0)$, $in(s_j) \subseteq \bigcup_{i=0}^{n} out(s_i)$, and $p \subseteq P$, $p \subseteq \bigcup_{i=0}^{n} out(s_i)$.

3 Semantics

In order for agents to communicate information about their technologies, a common data representation is needed. In this paper, information about an individual technology is represented as a tuple $T = \langle K, I, O, Q, C \rangle$, where

- K represents the classification of the technology.
- I represents a set of inputs that are to be converted into outputs.
- O represents a set of outputs produced by the technology.
- Q constraints related to the inputs in I. For example, a constraint requiring the input to be an instance of a printed wiring board.
- C represents the cost associated to the transformation of the inputs in I into the outputs in O.

Information about the classification of the technology, the kinds of objects in the inputs and outputs, as well as the physical quantities in the constraints and cost is represented using classes of objects and their relationships defined by shared ontologies implemented in the OWL language [5]. Ontologies define classes and properties of things such as substances, and processes using a modeling approach which is application-independent. Domain-independent concepts such as physical objects, activities, mereological and topological relations, classes and relations for physical quantities are defined in an upper ontology described elsewhere [1].

4 Methodology

A two-layer methodology is proposed, consisting of a process design layer and a facility design layer.

4.1 Process Design Layer

This is the layer where the manufacturing process is generated. The Technology Selection Agent (TSA) is the agent responsible of this layer. TSA works by specifying an input and then finding agents in the equipment layer that can process that input.

4.2 Equipment Layer

We assume that there is a set of equipment agents each of which has the ability to design an equipment item that converts an input into one or more outputs. An equipment agent receives a potential feed from the Technology Selection Agent. Then the equipment agent determines whether it can process the feed based on a series of constraints. If the feed can be processed then the equipment is designed, generating information about products, cycle time, operation and capital cost.

Fig. 1 shows the message sequence that takes place between Technology Selection Agent (TSA) and Equipment Agents $\bigcup_{i=1}^{n} E_i$. Since E_1 and E_2 can process raw materials R, the equipment design is simultaneously carried out by these two agents. Other agents who cannot process R wait for further opportunities.

After a design request, agents E_1 and E_2 send their design information to TSA, $EA1D1$ and $E2D1$ respectively. Subsequently, TSA asks the equipment agents whether they can process the output of $EA1D1$, namely $out(EA1D1)$ and the output of $EA2D1$ namely, $(out(EA2D1))$. After checking the feed against their constraints, TSA is informed that equipment agent E_3 can process both works in process. At this point, E_3 generates two instances of equipment: one instance that converts $out(E1D1)$ and another that converts $out(EA2D1)$. E_3 being aware of completing two successful designs, compares the cost of the two partial processes that result at this point: $EA1D1 \prec EA3D1$ and $E2D1 \prec E3D1$. As partial process $(EA1D1 \prec EA3D1)$ has the lowest cost, the design of $(EA1D1 \prec EA3D1)$ continues while the design of $(EA2D1 \prec EA3D1)$ stops. This approach is introduced to avoid combinatorial explosion.

The methodology is shown in Algorithm 1.

5 Implementation

Information about the classification of the technology, the kinds of objects in the inputs and outputs, as well as the physical quantities in the constraints and cost is represented using classes and relationships defined by shared ontologies implemented in the OWL language. The class hierarchy of the classes used in the case study of Section 6 are shown in Fig. 2. Classes play an important role

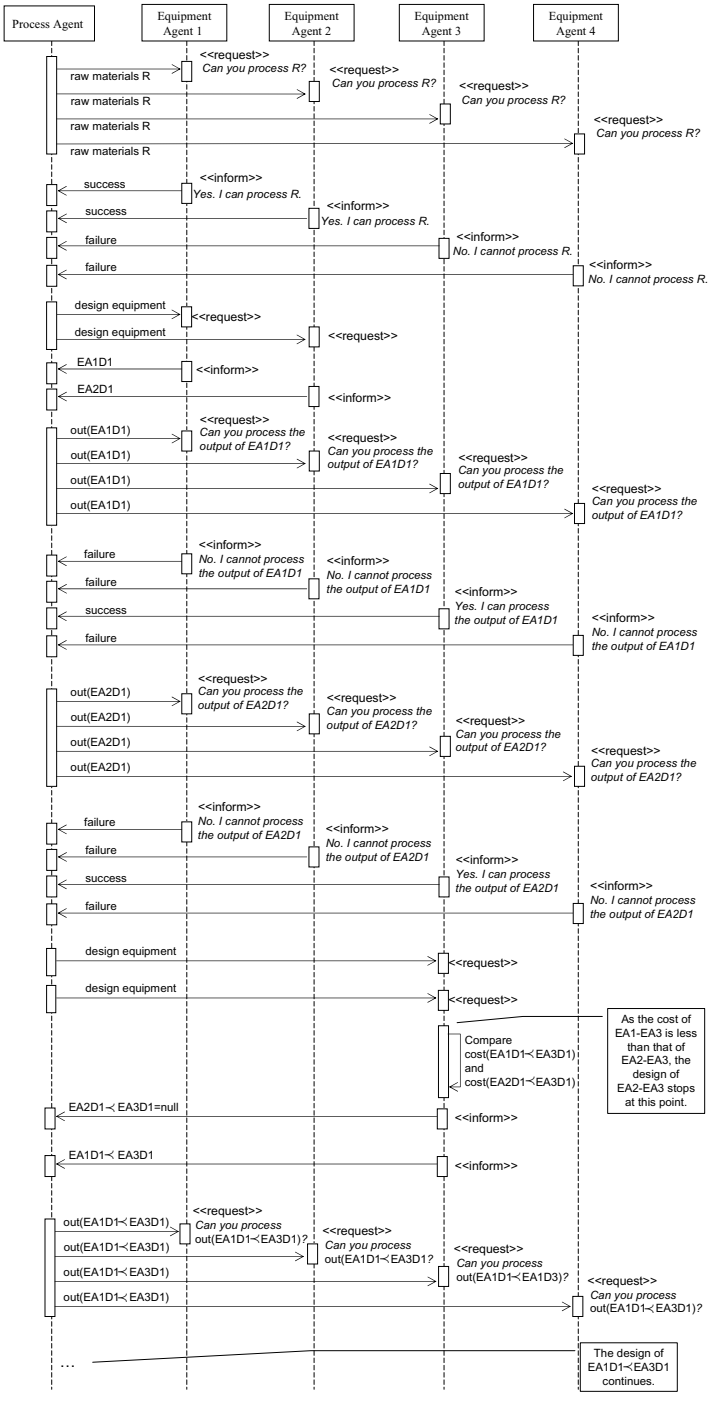


Fig. 1. A typical message exchange

Algorithm 1 Technology composition algorithm

```
1: procedure initiate():
2: TSA ← thisAgent()
3: r ← getRawMaterial()
4: newElement ← ⟨null, rawMaterialSourceID, TSA, r, 0⟩
5: create a solution with newElement
6: insert solution in solutionTable
7: for A_i ∈ A do
8:    send(EVALUATE_FEED, r, TSA) to A_i
9: end for

10: procedure when received(REPLY, prevElementID, E, canProcess, feed):
11: TSA ← thisAgent()
12: find a solution in solutionTable whose output is feed
13: cost ← solution.getCost()
14: if canProcess is true then
15:    send(DESIGN_EQUIPMENT, cost, feed, TSA) to E
16: end if

17: procedure when received(DESIGN_SUCCESSFUL, newElementID, E, feed,
    output, cost):
18: TSA ← thisAgent()
19: find solution in solutionTable such that solution.output = feed
20: copy solution and insert copy in solutionTable to store partial designs
21: newElement ← ⟨feed, newElementID, E, output, cost⟩
22: solution.connectElement(newElement)
23: for A_i ∈ A do
24:    send(EVALUATE_FEED, output, TSA) to A_i
25: end for

26: procedure when received(EVALUATE_FEED, feed, sender):
27: E ← thisAgent()
28: c ← loadConstraints()
29: canProcess ← true
30: for c_i ∈ c do
31:    result ← check(c_i, feed)
32:    if result is false then
33:       canProcess ← false
34:    end if
35: end for
36: send(REPLY, E, canProcess, feed) to sender

37: procedure when received(DESIGN_EQUIPMENT, currentCost, feed, requester):
38: d ← design(feed)
39: solutionCost ← currentCost + d.getCost();
40: taskEntry ← (requester, d, feed, solutionCost)
41: add taskEntry to designTaskTable

42: procedure compareDesigns():
43: E ← thisAgent()
44: best ← findBestDesign(designTaskTable)
45: d ← best.getDesign()
46: cost ← best.getSolutionCost()
47: ID ← d.getID()
48: output ← d.getOutput()
49: feed ← best.getFeed()
50: send(DESIGN_SUCCESSFUL, ID, E, feed, output, cost)
```

in the constraints of the technology. For example, if *add_adhesive* has a precedence constraint requiring *solder_paste* to be carried out before, any subclass of *solder_paste* such as *automatic_solder_paste* will satisfy this constraint.

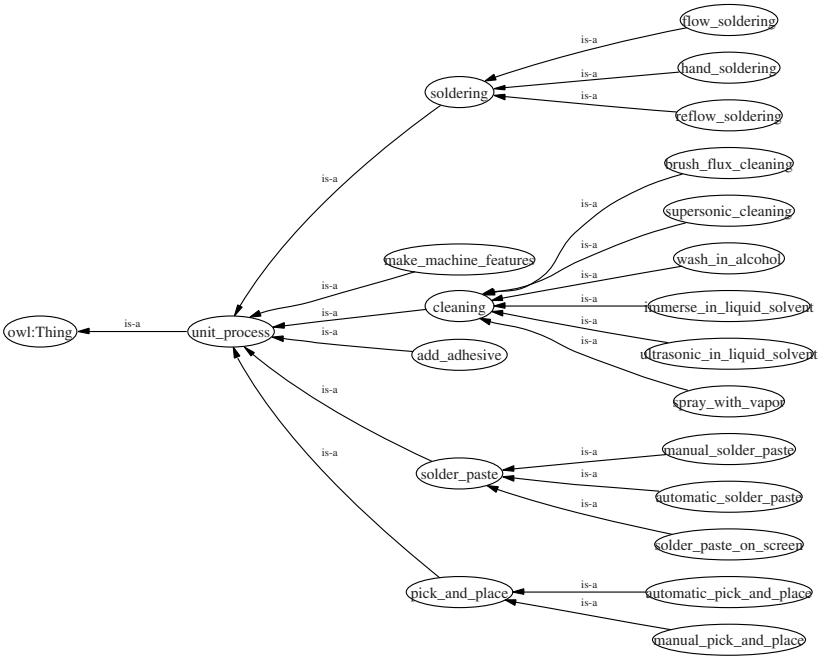


Fig. 2. Classes and subclasses of technologies for PCB assembly

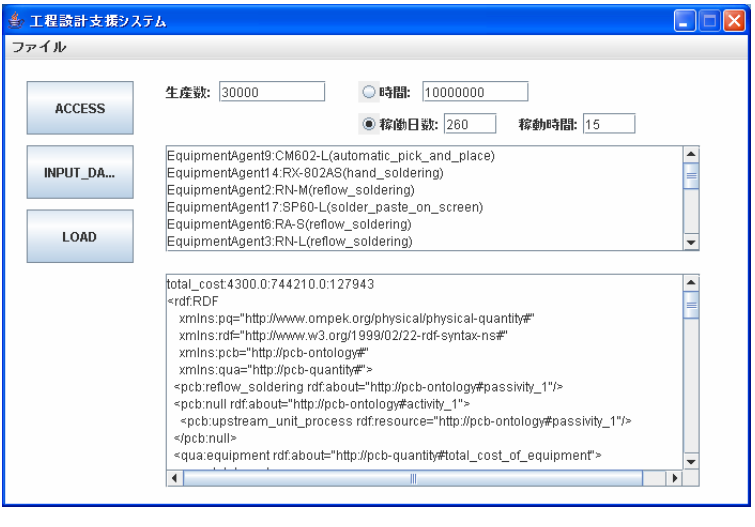
The constraints of the technologies are encoded in CLIF (Common Logic Interchange Format) which is a simplified and updated version of the knowledge representation language known as KIF (Knowledge Interchange Format). Evaluation of the constraints is done by means of inferences using first-order logic reasoners. The actual content of the model is encoded in a predefined XML format. An example of constraint is shown below that specifies that the length of the PWB should be less than 250 mm.

```
<constraint_code>
  (and (rdf:type ?x qua:board) (qua:length ?x ?l)
       (qua:millimeter ?l ?mm) (pq:content ?mm ?value)
       (less_than ?value 250))
</constraint_code>
```

The technology selection environment was programmed in Java using the JADE (Java Agent DEvelopment Framework) library for distributed agent applications [2] and the JTP (Java Theorem Prover) inference system [6]. JADE

provides a Java computer library that can be used to implement multi-agent systems that comply with the FIPA specifications [7]. The agent platform can be distributed across computers and other Java-enabled machines. It also provides tools for monitoring and configuration of the agents. Messages are encoded in FIPA ACL (Agent Communication Language). An ACL message contains a number of parameters such as performative, sender, receiver, content, language and ontology. The performative defines the intention of the message so that an agent knows what to do when a message is received. Specifically, the technology selection environment implements the request, query-ref and inform performatives.

Constraints are evaluated using JTP (Java Theorem Prover) which is a reasoning system that can derive inferences from knowledge encoded in the OWL language. JTP is composed of a number of reasoners that implement algorithms such as generalized modus ponens, backward-chaining, and forward chaining and unification [6]. JTP translates each OWL statement into a CLIF sentence of the form (*PropertyValue Value Predicate Subject Object*). Subsequently, JTP simplifies the CLIF sentences using a series of axioms based on the OWL semantics. OWL statements are finally converted to the form (*(Predicate Subject Object)*). All agents have copies of upper and domain ontologies that can be retrieved from an Internet agent. This means that agent code remains unchanged even when new classes are added to the ontology file. The inference engine is also used for determining the degree of matching between the requirements (feedstock and product specifications) and the constraints of the process. When the TSA agent sends a message to an equipment agent to evaluate an input, the equipment agent asserts the feedstock and product information as facts in the knowledge base of JTP. Subsequently, the equipment agent checks the requirements against the technology constraints by performing queries to the knowledge base.


工程設計支援システム
ファイル
ACCESS
INPUT_DA...
LOAD
生産数: 30000
時間: 10000000
稼働日数: 260
稼動時間: 15
EquipmentAgent9:CM602-L(automatic_pick_and_place)
EquipmentAgent14:RX-802AS(hand_soldering)
EquipmentAgent2:RN-M(reflow_soldering)
EquipmentAgent17:SP60-L(solder_paste_on_screen)
EquipmentAgent6:RA-S(reflow_soldering)
EquipmentAgent3:RN-L(reflow_soldering)
total_cost:4300.0:744210.0:127943
<rdf:RDF
xmlns:pq="http://www.ompek.org/physical/physical-quantity#"
xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
xmlns:pcb="http://pcb-ontology#"
xmlns:qua="http://pcb-quantity#">
<pcb:reflow_soldering rdf:about="http://pcb-ontology#passivity_1"/>
<pcb:null rdf:about="http://pcb-ontology#activity_1">
<pcb:upstream_unit_process rdf:resource="http://pcb-ontology#passivity_1"/>
</pcb:null>
<qua:equipment rdf:about="http://pcb-quantity#total_cost_of_equipment">

6 Case Study

As an application of the process synthesis system, we consider a design problem from the electronics industry. The objective is to design an assembly line that will produce printed circuit boards (PCBs). Printed circuit board assembly consists of attaching electronic components to the Printed wiring boards (PWB). A printed wiring board is the physical structure that holds together the various electronic components of a printed circuit board.

The targeted annual volume of product is 30000 units with 15 equipment agents. For this example, we assume that there are 260 working days per year, each day having two eight-hour shifts.

As shown in Fig. 3 The resulting manufacturing system is composed of three machines: SP18P-L (solder paste on screen), DT401-F (automatic pick and place machine) and RA-M (a reflow soldering machine).

7 Conclusions

In this paper we have described an agent-based approach for process design. Technologies are managed by distributed agents that have design capabilities. Local evaluation of partial designs ensures that the design is completed in polynomial time. This approach has been tested for a PCB production problem. Nevertheless, there are several issues that will require additional research and development. The current approach produces linear processes so future work will be needed in order to manage material recycles.

References

1. Batres, R., West, M., Leal, D., Price, D., Katsube, M., Shimada, Y., Fuchino, T.: An upper ontology based on ISO 15926. Computers & Chemical Engineering 31(5-6), 519–534 (2007)
2. Java Agent DEvelopment Framework [Online], Available at `http://www.cselt.it`
3. Li, X., Kraslawski, A.: Conceptual process synthesis: past and current trends. Chemical Engineering and Processing 43(5), 583–594 (2004)
4. Nau, D.S., Ball, M., Baras, J., Chowdhury, A., Lin, E., Meyer, J., Rajamani, R., Splain, J., Trichur, V.: Generating and evaluating designs and plans for microwave modules. AI in Engineering Design and Manufacturing 14, 289–304 (2000)
5. W3C.OWL Web Ontology Language Overview: W3C Recommendation (2004), [Online] Available `http://www.w3.org/TR/owl-features/`
6. Fikes, R., Jenkins, J., Gleb, F.: JTP: A System Architecture and Component Library for Hybrid Reasoning. In: Proceedings of the Seventh World Multiconference on Systemics, Cybernetics, and Informatics. Orlando, Florida, USA (July 27-30, 2003)
7. FIPA: FIPA specifications (2005), [Online] Available at `http://www.fipa.org/specifications/index.html`

A Sliding-Mode Control Approach in Ultra-Precision Positioning System

Insung Choi, Seungok Choi, Gunhaeng Heo, Kiheon Park, and Kwanho You

Sungkyunkwan University, Suwon, 440-746, Korea
{david2044, seungoki, idealgo, khpark, khyou}@ece.skku.ac.kr

Abstract. It is known that the time optimal control seeks a target position faster than sliding-mode control. However, the time optimal control is vulnerable to chattering effect more than sliding-mode control. Especially for the nanometric scale positioning system under the parameter uncertainty and external disturbances, this can cause an actuator saturation. To solve this problem, we present the design of a robust optimal controller for an ultra-precision positioning stage. The control algorithm is a sliding-mode approach based on linear quadratic optimal control. Some simulation results are examined to demonstrate the effectiveness of the proposed controller.

1 Introduction

Recently, with the development of microprocessor and information industry, nano technology has been developing rapidly. Most of the measurements and implementations are realized through ultra-precision positioning process. The application of the nanometric scale positioning system can be included in various industries such as semiconductor lithography devices, chip mount machines, high density hard disk drivers and so on. Even though the resolution of an actuator is considered as important, generally it is influenced by various external conditions such as the actuator installation, vibration, temperature, air refractive index, and air turbulence. Also the precision capability can be restricted by nonlinearity noise generated in a control system.

To control the nanometric scale positioning system, many approaches have been provided [1-4]. Topcu [1] proposed a method based on a heterodyne Michelson's interferometer and a homemade high frequency electronic circuit. Control of the repetitiveness of each steps with a high-stability quartz oscillator enables the translation stage to get an uniform speed with a same accuracy. A precise control of a linear pulse motor (LPM) using neural network was suggested [2]. To improve the position control of LPM, the neural network with conventional feedback controller is presented to overcome the nonlinearity caused by model uncertainties. This neural network of feedback error learning type scheme modifies the current commands to increase the position accuracy.

A motor-piezo actuator for nanometric scale positioning system was proposed by Liu [3]. The system design consists of a coarse stage and a fine stage. To

B. Apolloni et al. (Eds.): KES 2007/ WIRN 2007, Part II, LNAI 4693, pp. 728–735, 2007.
© Springer-Verlag Berlin Heidelberg 2007

compensate the hysteresis and nonlinearity of fine stage, and simultaneously to increase the linearity of micro control, exact model matching (EMM) control law was proposed. With the improved repeatability of the fine stage's motion, it is possible to follow the same path in the positive and negative movement. The ultra-precision positioning using the interferometer is presented by Yi [4]. It detects the fringe movement of interferometer on a linear CCD to find the fringe peak position and tracks the fringe movement to get the displacement of the target mirror. As the number of the interference fringe increases, the effect of wavefront distortion becomes less significant.

Meanwhile an optimal control method was tried for synchronous reluctance and induction motor position control [5]. The proposed control scheme is composed of classical state feedback control and variable structure control. For the motor system modelled as a 2nd order state equation, the optimal control strategy achieved the performance requirement regardless of the parameter uncertainties and disturbances.

In this paper, an ultra-precision positioning method base on sliding-mode control is suggested. We use the piezoelectric actuator to drive the nano stage and the capacitive displacement sensor to measure the reference displacement for accuracy comparison. The dSPACE controller board is used to acquire data between the controller and nano stage. By applying the system identification method, the system transfer function is approximated as a 4th order state model. As a classical controller design, the time optimal control can be used. It has the advantage of reaching the target position in minimal time. However the chattering effect occurs near the target position. Especially, it is vulnerable to parameter uncertainty and external noise. To solve this problem, we suggest a modified sliding-mode control based on linear quadratic optimal control.

2 Modelling of Ultra-Precision Positioning System

As an experimental apparatus in this paper, the ultra-precision positioning system is established on an optical table as shown in Fig. 1. The nano stage is mounted upon a micro stage of a linear motor and is connected to a PZT (piezoelectric actuator) controller. The model of nano stage to be implemented is a linear piezoelectric transducer (PI: P-621.CL).

The configuration of a nano stage is designed with a PZT controller to actuate the stage and a capacitive displacement sensor (PI: D-100) installed inside the nano stage. The PZT controller consists of three parts; LVPZT-amplifier, PZT-servo controller to correct the hysteresis and nonlinearity of PZT, and display/interface device. As a first step of system modelling, we have tried to derive the transfer function of this nano stage using the system identification (SI) method. The SI method is known to be appropriate for a system in which system parameters or dynamics are difficult to be expressed exactly in mathematical relations.

Moreover, the dSPACE controller board (DS1104) is used as a real-time interface in control system. We have adopted a pseudo-random binary signal (PRBS)

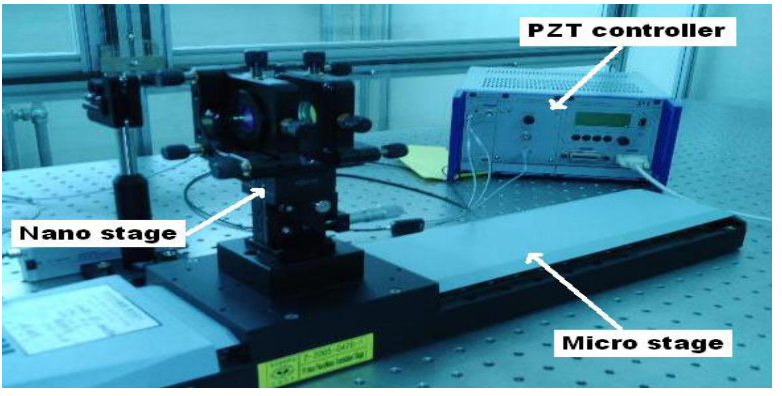


Fig. 1. Configuration of an ultra-precision positioning system

which has a characteristic similar to a white noise as an input signal to the nano stage. This signal is generated by using the MATLAB program and is provided to an input terminal of PZT controller through a dSPACE board. The SI can be processed by using a graphic user interface (GUI) in system identification toolbox (SITB). The proper model construction and system order are selected by the rule of trial and error, and the 4th order is appropriate as a system order for the nano positioning system in Fig. 1. The following describes the transfer function found by the SI method.

$$G(s) = \frac{\beta_4 s^4 + \beta_3 s^3 + \beta_2 s^2 + \beta_1 s + \beta_0}{s^4 + \alpha_3 s^3 + \alpha_2 s^2 + \alpha_1 s + \alpha_0} \tag{1}$$

where the parameters are $\alpha_3 = 100.8$, $\alpha_2 = 2.192 \times 10^5$, $\alpha_1 = 2.075 \times 10^7$, $\alpha_0 = 7.589 \times 10^8$, $\beta_4 = 0.008531$, $\beta_3 = -5.957$, $\beta_2 = 4050$, $\beta_1 = -1.343 \times 10^6$, $\beta_0 = 4.708 \times 10^8$.

The transfer function in (1) is now expressed as a state-space equation through realization in controllable canonical form

$$\begin{aligned} \dot{x}(t) &= Ax(t) + Bu(t) \\ y(t) &= Cx(t) + Du(t) \\ A &= \begin{bmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ -\alpha_0 & -\alpha_1 & -\alpha_2 & -\alpha_3 \end{bmatrix}, \quad B = \begin{bmatrix} 0 \\ 0 \\ 0 \\ 1 \end{bmatrix} \\ C &= \begin{bmatrix} \gamma_3 & \gamma_2 & \gamma_1 & \gamma_0 \end{bmatrix}, \quad D = \beta_4 \end{aligned} \tag{2}$$

where the parameters are $\gamma_3 = -6.8269$, $\gamma_2 = 2180$, $\gamma_1 = -1.520 \times 10^6$, $\gamma_0 = 4.643 \times 10^8$, and $x(t)=[x_1, x_2, x_3, x_4]^T$.

3 LQ Sliding-Mode Control

In this section, we design a modified sliding-mode controller based on linear quadratic (LQ) optimal control. Since the very small fluctuation in ultra-precision positioning system can cause a problem, the controller is designed to reduce the chattering and eventually makes the states to be maintained as stable around the target position. To do so, the objective function in LQ optimal control is now set as the minimization of the state changes and control input variations around the target position in spite of model uncertainties.

To design a robust optimal controller, let's suppose that the state equation is modified by adding the parameter uncertainty as following:

$$\begin{aligned} \dot{x}(t) &= (A + \Delta A)x(t) + (B + \Delta B)u(t) \\ y(t) &= (C + \Delta C)x(t) + (D + \Delta D)u(t) \end{aligned} \tag{3}$$

where ΔA, ΔB, ΔC, and ΔD are parameter uncertainties and are supposed to be bounded by δ. In equation (3), the general solution of the nominal system without parameter uncertainty is

$$x(t) = e^{At}x(0) + \int_0^t e^{A(t-\tau)}Bu(\tau)d\tau \tag{4}$$

With the solution of (4), we define a new sliding surface for the perturbed system model of (3) as

$$\sigma(x) = h\left[x(t) - e^{At}x(0) + \int_0^t e^{A(t-\tau)}BKx(\tau)d\tau\right] \tag{5}$$

where h is chosen as a positive constant and K is a linear feedback gain set from LQ optimal control problem. The feedback controller using LQ theory is obtained by minimizing the following objective function.

$$J = \frac{1}{2}\int_0^\infty \left[x^T(t)Qx(t) + u^T(t)Ru(t)\right]dt \tag{6}$$

where Q and R are real cost weights and are supposed to be the the positive definite real matrices selected by the designer. Following the solution of LQ optimal control problem, we can find a state feedback controller $u(t) = -Kx(t)$. The feedback gain can be obtained from

$$K = (R + B^TSB)^{-1}B^TSA \tag{7}$$

where S is a unique stabilizing solution of the following algebraic Riccati equation (ARE).

$$A^TSA - S + Q - A^TSB(R + B^TSB)^{-1}B^TSA = 0 \tag{8}$$

The sliding-mode control law is derived based on the reaching condition as $\sigma(x)\dot{\sigma}(x) < 0$. Since the sliding-mode controller satisfies the reaching condition, we can obtain a hybrid sliding-mode controller as following:

$$u(t) = -Kx(t) - \mu\text{sat}\left[\sigma(x)\right] \tag{9}$$

where sat$(\cdot)$ means a saturation function.

4 Simulation Results

In this section we do some simulations to confirm the robustness of the proposed algorithm. The modified sliding-mode control is applied to the nano stage in Fig. 1. Moreover we add some noises to the actual system in order to test the stability against parameter uncertainty, noise and disturbance. We compare two different control methods (time optimal control and LQ sliding-mode control).

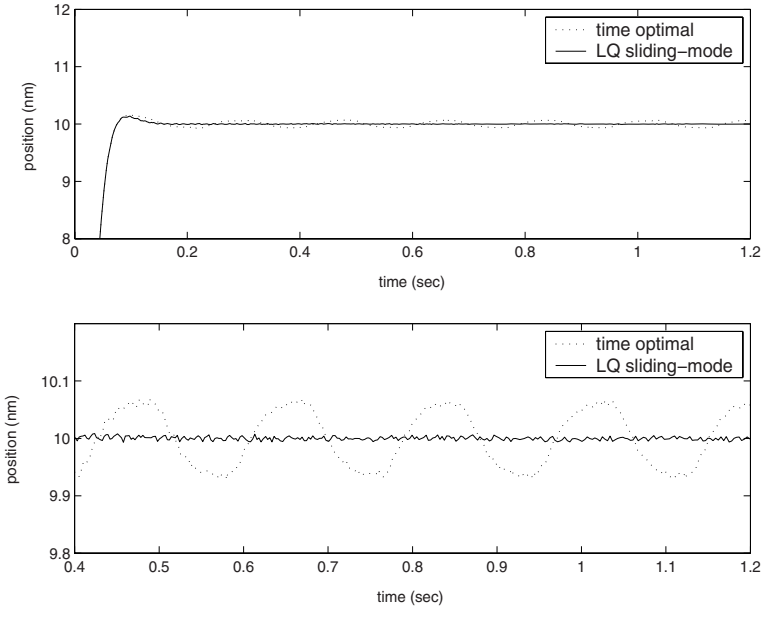


Fig. 2. Comparison of displacement trajectory with a target position of 10 nm

Fig. 2 shows the displacement trajectory comparison for two control algorithms. In this simulation, ΔA, ΔB, ΔC, and ΔD are supposed to be about 10 % of the system parameters, respectively. Under the effect of parameter uncertainty in Eq. (3), the robustness of the positioning system has been tested for two different control algorithms. In Fig. 2, the solid line is a displacement trajectory driven by LQ sliding-mode control, and the dotted line is the one by classical time optimal control. As can be confirmed in this figure, the time optimal control drives the states to approach to a target position (10 nm) rapidly. In Fig. 2, the displacement trajectory by LQ sliding-mode control shows the proximate time optimal features. The rising time is less than 0.07 sec which is similar to that of time optimal one.

Since the focus of the control of ultra-precision positioning system is in nanometric scale around the target position (10 nm), we magnified the region around the target position at the bottom of Fig. 2. In case of the time optimal control,

it is quite sensitive to parameter uncertainty and shows great chattering effect around the target position. Meanwhile, the displacement trajectory driven by LQ sliding-mode control manifests a quite stable maintenance around the target position. We can confirm that the LQ sliding-mode control has better robustness than the classical time optimal control.

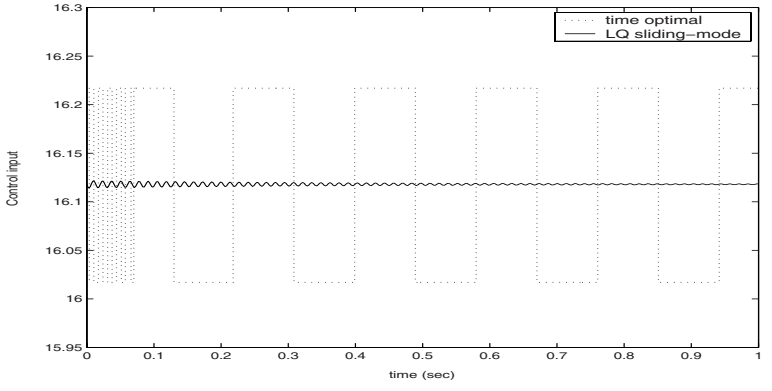


Fig. 3. Comparison of control input values for time optimal control and LQ sliding-mode control

In Fig. 3, the control input values are compared for two control algorithms. In this figure, the solid line is the control input by LQ sliding-mode control, and the dotted line is the one by time optimal control. Again the control input from time optimal control are changing periodically around the target position ($t \geq 0.07$ sec). Hence it makes the states to oscillate around the target position. The control input from LQ sliding-mode control shows quite steady and smooth values.

To prove the robustness to external disturbance, a time maximum disturbance is now applied to the control system. The time maximum disturbance is a test signal designed following the Pontryagin's minimum principle [6]. The time maximum disturbance is known to be the most stressful disturbance for stabilized system. The time maximum disturbance signal is applied to the nano stage model as a secondary input. Hence the LQ sliding-mode control input tries to make the stage to remain at a target position, and the disturbance input makes the stage to oscillate to maximize the severity index, $J = \max \int_{t_0}^{t_f} dt$.

This minimax situation can be considered as a game theoretic approach [7] where the minimization is done with respect to the control selection and the maximization is done with respect to the disturbance selection. The state equation considering the time maximum disturbance is now changed as

$$\begin{aligned} \dot{x}(t) &= (A + \Delta A)x(t) + (B + \Delta B)u(t) + (B + \Delta B)w(t) \\ y(t) &= (C + \Delta C)x(t) + (D + \Delta D)u(t) + (D + \Delta D)w(t) \\ |u(t)| &\leq M_u, \quad |w(t)| \leq M_w \end{aligned} \tag{10}$$

where $w(t)$ is the time maximum disturbance signal to maximize the severity time index. Following the design procedure of time maximum disturbance in [6], the time maximum disturbance is also can be written in bang-bang type as

$$w(t) = M_w \text{sgn}\left[\eta_4(t)\right] \tag{11}$$

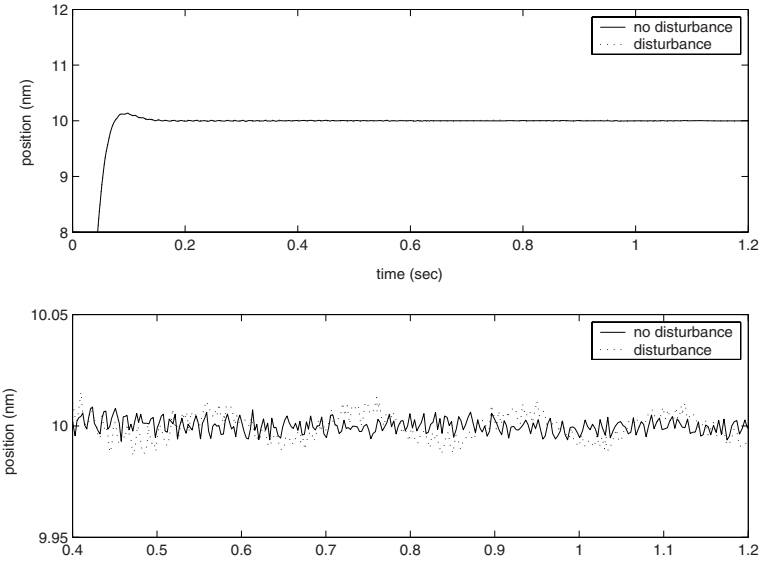


Fig. 4. Displacement trajectory with a time maximum disturbance signal.

Fig. 4 shows the displacement trajectory for LQ sliding-mode control with a time maximum disturbance signal as a secondary input, $w(t)$. The magnitude of the disturbance is chosen as 10% of the LQ sliding-mode control, i.e., $M_w = 0.1M_u$. The solid line is a trajectory result without time maximum disturbance signal, and the dashed line is the one with a time maximum disturbance signal, respectively. From the magnified displacement trajectory shown in the bottom of Fig. 4, the stage can track the target position in spite of the added disturbance signal. The improved robustness can be confirmed using the time maximum disturbance.

5 Conclusion

The LQ sliding-mode control has been presented in this paper to control more accurately in an ultra-precision positioning system. To reduce the chattering effect around the target position, the objective function is determined to minimize the variation of states and input change. The performance of the proposed controller

has been examined through simulations under various parameter uncertainties and disturbances.

Through the simulation results, we can confirm that the proposed control scheme has proximate LQ optimal control performance and reduces chattering effect. The LQ sliding-mode control also prevents the rapid input change around the target position. To test the improved robustness, the time maximum disturbance is designed in bang-bang type and applied to the nano stage model. The time maximum disturbance as a secondary input drives the stage to oscillate. In spite of the applied disturbance signal, the LQ sliding-mode control could maintain the stage within a performance requirement.

Acknowledgement

The authors would like to thank the Korea Science and Engineering Foundation (KOSEF) for financially supporting this research under Contract No. R01-2004-000-10338-0 (2006).

References

1. Topcu, S., Chassagne, L., Haddad, D., Alayli, Y., Juncar, P.: Heterodyne interferometric technique for displacement control at the nanometric scale. Review of Scientific Instruments 74, 4876–4880 (2003)
2. Kwon, Y.K., Park, J.I.: Precise control of a linear pulse motor using neural network. Jour. of Control, Automation, and Systems Engineering 6, 987–994 (2000)
3. Liu, H., Lu, B., Ding, Y., Tang, Y., Li, D.: A motor-piezo actuator for nano-scale positioning based on dual servo loop and nonlinearity compensation. Journal of Micromechanics and microengineering 13, 295–299 (2003)
4. Yi, J.H., Kim, J.A., Kim, S.H., Kwak, Y.K.: Moving fringe detection for ultra precision position measurement. In: IEEE Proc. Instrumentation & Measurement Tech. Conf., vol. 2, pp. 723–728 (2000)
5. Shyu, K.K., Lai, C.K., Tsai, Y.W.: Optimal position control of synchronous reluctance motor via totally invariant variable structure control. IEE Proc. Control Theory and Applications 147, 28–36 (2000)
6. You, K.H., Lee, E.B.: Time maximum disturbance design for stable linear system: a model predictive scheme. IEEE Trans. Automatic Control 46, 1327–1332 (2001)
7. Shaiju, A.J., Dharmatti, S.: Differential games with continuous, switching and impulse controls. Nonlinear Analysis 63, 23–41 (2005)

Integration of Multi-agent Controller and Scheduler for Multi-purpose and Multi-batch Plant

Takashi Hamaguchi[1], Masaki Inoue[1], Takamasa Yoshida[1],
Koji Kawano[2], Hossam A.Gabbar[3], Kazuhiro Takeda[4],
Yukiyasu Shimada[5], Teiji Kitazima[6], and Yoshihiro Hashimoto[1]

[1] Nagoya Institute of Technology, Gokiso-cho, Showa-ku, Nagoya 466-8555, Japan
hamachan@nitech.ac.jp
[2] Mitsubishi Chemical Eng. Corp., 191-1, Shiohama, Yokkaichi, Mie 510-0863, Japan
[3] Okayama University, 3-1-1, Tsushima-Naka, Okayama 700-8530, Japan
[4] Shizuoka University, 3-5-1, Johoku, Hamamatsu, Shizuoka 432-8561, Japan
[5] National Institute of Occupational Safety and Health, 1-4-6, Umezono, Kiyose, Tokyo 204-0024, Japan
[6] Tokyo University of Agriculture and Technology, 2-24-16, Nakamachi, Koganei, Tokyo 204-0024, Japan

Abstract. Recently, there are dynamic changes in customer demands where product lifecycles became shorter and shorter. Although batch process has natural flexibility, however the addition of new products requires many modifications not only in the batch plant but also in the associated batch plant management systems like control, scheduling and monitoring systems. The loads of the change managements of these batch management systems are very heavy and it will be increased extremely according to the increase of number of different produces. Batch plant management system should be improved to catch up these rapid changes and operation challenges easily. Our proposed integration of multi-agent controller and scheduler system can reduce the loads of modification of the scheduling systems.

Keywords: Batch, Control, Scheduler, Simulator, and Multi-agent.

1 Introduction

Batch plants have been employed extensively for the manufacturing of many types of chemicals, particularly those which are produced in small quantities and for which the production processes or the demand patterns are likely to change. Because product lifecycle became shorter, it is essential to shorten the time required to modify batch plants. Therefore, easy modification method becomes increasingly important in chemical industry.

The formal definition of a batch process in batch control standard ANSI/ISA S88.01 is “A process that leads to the production of finite quantities of material by subjecting quantities of input material to an ordered set of process activities over a finite period of time using one or more equipment units”[1]. Batch plant is

B. Apolloni et al. (Eds.): KES 2007/ WIRN 2007, Part II, LNAI 4693, pp. 736–743, 2007.
© Springer-Verlag Berlin Heidelberg 2007

classified by the physical structure. The classification and an example network structure batch plant as shown in Fig.1. Batch plants have various equipment, such as material tanks, reactors, and product tanks, and piping instruments. Batch processing is the management of equipment resources that add value to raw materials for producing finished product. The methods for raw material transformation are identified by a recipe. The recipe identifies many requirements for production such as material, sequence of the control and equipment. For the production of final products, the specified volumes of input materials are processed in each tank separately. After processing, materials must be transferred to next tank for further processing. Batch plants have the ability to be operated in multi-purpose and multi-batch operation modes. The management of material movement from equipment to equipment is complex especially with the high degree of freedom in the network structure of the underlying batch plant. Such complicated operations utilize few equipment and parts as shown in ANSI/ISA-S88.01. To optimize the use of shared resources in complex plant structure, it is widely common to construct integrated control, scheduling and monitoring systems within batch management systems. In current MES systems, plant modification requires cost and time. Intelligent control mechanisms can be implemented in control software systems to support material movement, equipment control, and resources allocation. To achieve such target, the To-Be model for multi-purpose and multi-batch control system for the network structure was proposed [3][4]. In this paper, a method is proposed for integrated batch management system for controller, scheduler, and operational supervision.

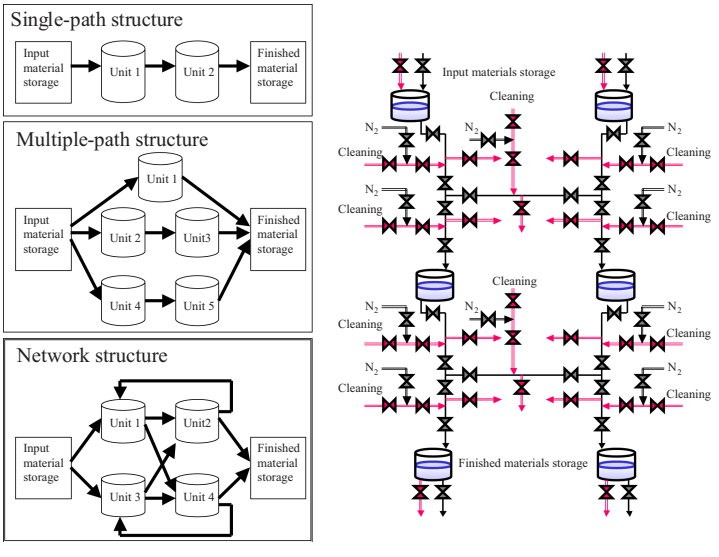


Fig. 1. Classification by Physical Structure

2 Integration of Controller and Scheduler

Information sharing and visualization between plant floor and management layer can support the operation of complex plant. Among the different trials, MES (Manufacture Executing System) has focused on the link system between ERP (Enterprise Resource Planning) and plant floor.

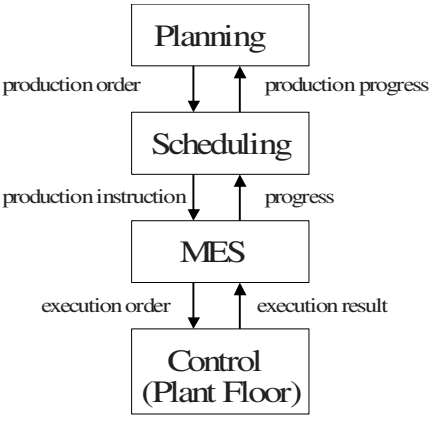


Fig. 2. Integrated Controller and Scheduler through MES

MES might be a realistic approach to provide flexible integration among the developed systems. However, such link has major limitations of flexibility in change management. If the change of the plants physical layout is happened in the batch plant, all system must be modified separately because the systems use different physical models. Modification to current MES might be needed to provide adequate information coordination among all systems. The conventional recipes should be modified by the addition of the new products and their recipe details to be able to get higher profit and market leading position. Such improvement of batch plant operation requires system modification, which lead to higher cost and longer operation time, and consequently, with lower profit.

Based on the above requirements, a new batch operation management system should provide adequate flexibility in change management where it enables the definition of independent recipes using plant physical models. Based on such proposed batch operation management improvement, operation time can be decreased dramatically.

The proposed batch control system is based on multi-agent where it is easy to add and modify physical models, recipes and controllers [3][4]. The proposed system has different types of agents: Recipe agent, Transfer manager agent, Equipment agent and Utility manager agent. In this paper, we have added Cleaning manger agent. Recipe agents move through the batch plant with materials. They show the instruction of the operations in each equipment to Equipment agents. Equipment agents achieve the operation according to the instruction. Equipment agents demand the supply of utility

to Utility agents. Utility agents control the supply and inventory of the utility, such as steam, cooling water, field operator, and so on. Transfer agent controls transfer of materials between equipments. We think that this system should be supervised by human operator. Recipe agent has recipe which is written by XML. The recipe agent is generated by the supervisor who operates batch plant by Man-Machine Interface (MMI). The recipe agent moves with materials in batch plant while communicating with other agents. This agent system is developed using Toshiba's multi-agent Framework named Bee-gent[2]. Using object-oriented programming, agent can generate several instances from various classes. The addition and modification of classes can be programmed separately. The instance information is collected in databases such as recipe database and piping layout database. When access to the specified agent is needed, a mediator agent works to connect them. Consequently, the effect of the modification could be localized.

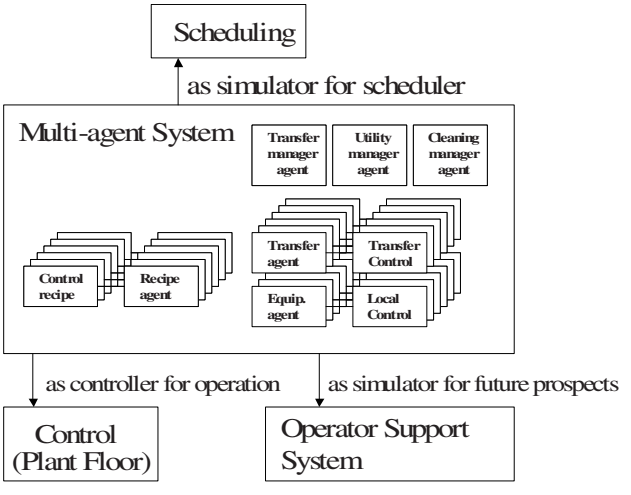


Fig. 3. Integrated Controller, Scheduler and Simulator by Multi-agent

In this paper, we propose the integration of controller, scheduler, and operational supervision systems, as shown in Fig.3. Scheduler and operational supervision system would be able to include the controller model as a simulator.

In the scheduling mode, human operator must assign an adequate start time for each batch. The requirements from each control recipes are set in the queue for agents. Each agent acts to process the input requirements. The agent usually deals with the requests by FIFO. If the requirement can't be processed on the time, it is postponed until next turn. The simulator outputs the results as a feasible detailed schedule. In the operational supervision mode, the initial condition is set by the present situation. The simulator can predict the behavior of the batch plant in the near future. This prediction is important for human worker in the plant floor. This work is based on bottom up integration approach. The activities in the scope of integration are limited between control and dispatching.

3 Integrated Batch Operation Management System

In this section, the purpose and the behavior of each agent in the controller and simulator are explained. The schematic diagram for integrated batch operation management system by multi-agent is shown in Fig.4.

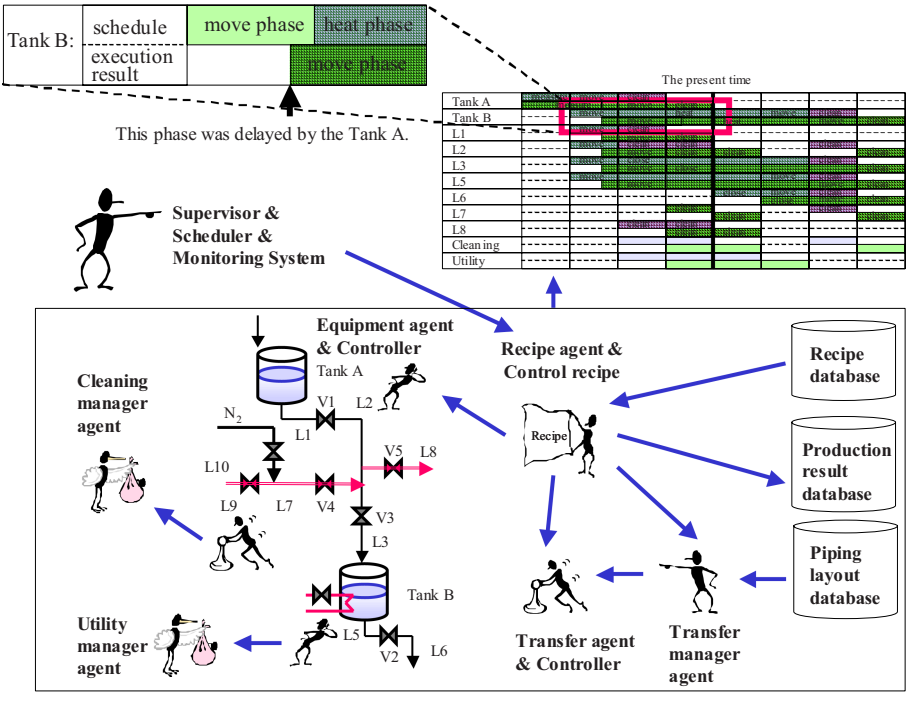


Fig. 4. Integrated Batch Operation Management System

A control recipe is generated from a master recipe in the Recipe database. A new recipe agent is instantiated as an instance from the class definition to mange the control recipe. The production is based on push-type production system by recipe. The recipe agent communicates with other agents by passing data of control recipe in view of requests for batch production. Each request has information about agent's action for batch process, such as start conditions, etc.

In the control mode, the equipment agent behaves as an interpreter for real control system such as PLC. In the schedule mode, the equipment agent behaves as a controller to simulate the batch process and handles controller's events. The transfer manager agent works to manage material transportation.

To transfer raw materials or products, the recipe agent demands the transportation process for the equipment agents for upstream tank, downstream tank, and the transfer manager agent. The transfer manger agent confirms the feasibility of exclusive control for demanded transportation and creates a new instance of the transfer agent to do the necessary control. Related agents prepare for the transportation. When the

preparation is completed, they synchronize and transfer it. In the example shown in Fig. 4, lines L1, L2, L3 and L5 are occupied for transportation from Tank A to Tank B. The ready conditions of all related equipment must be satisfied such as the proper status and cleaning status.

The heating and cooling process need the management of utilities, such as steam and cooling water and cleaning utilities such as water, cleaning liquid, and nitrogen gas. The management of utilities and cleaning is based on pull-type production system by agent calls. A utility manager agent is prepared to manage the use of each utility, properly. When equipment wants to use some utilities, the agent communicates with utility manger agent to get the right use of required volume of utility in the underlying batch process. The utility is not controlled by the utility manager agent but by the equipment agent and controller group because the delay of utility control affects products quality. Feeding lines and purge lines for cleaning utility are connected with the transfer lines for raw materials and products, as shown in Fig.1. Equipment agents and transfer agents know the cleaning procedure of the control parts. When the agent judges that the control part must be cleaned, the agent communicates with the cleaning manger agent to get the right of use about cleaning utility as well as utility control.

The Gantt chart, which is shown in the top of the Fig.4, is used for GUI (Graphical User Interface) for scheduler and operational supervision. Each row corresponds to a sharing resource such as equipment, line, utility and cleaning. These inner rows have two parts in each row.

In operational supervision mode, the schedule generated by scheduler is displayed on the top part of each row. Results and predictions are shown in the bottom part of each row. This information is used to grasp the appropriate batch plant operation in comparison with schedule. Usually, human operator has only to monitor batch processing on GUI because agents judge the situation and act the requested task by the ordered constraints themselves. However, human operator must intervene to modify the agent's behavior whenever needed. The development of the simple and useful GUI for agent's management is important for multi-agent system for rescheduling activity at the plant floor level.

In this paper, the simulator was developed as the discrete event system. The simulator must calculate batch processing time by sequence control. When agent requires some processes and the requirement is admitted, the requested processes are started and start and finish events happen. The accuracy of the event time depends on the model within agents and the repeatability of the recipe. The model can be written by various expressions such as transfer equation or just constant values. Any model is permitted as far as the specification of the interface for input and output is satisfied.

To simulate using the conventional dynamic simulator, human operator must identify, model, and tune the required equations and parameters. This is very hard task for multi-purpose and multi-batch operations. Our proposed system can decrease the time and work required for such task. Moreover, whenever modeling data is not adequate about new products, results of similar products can be used as qualitative information. The analogical reasoning system must be developed to search the similar products and batch process accordingly. When the formulation of required equation is difficult, the accuracy can be modified by learning functions from simulation. However, learning functions should not be used in the control mode. The repeatability

of the control sequence is very important to guarantee the quality of the products. The one example of the application as a scheduler is explained in the flowchart, as shown in Fig. 5.

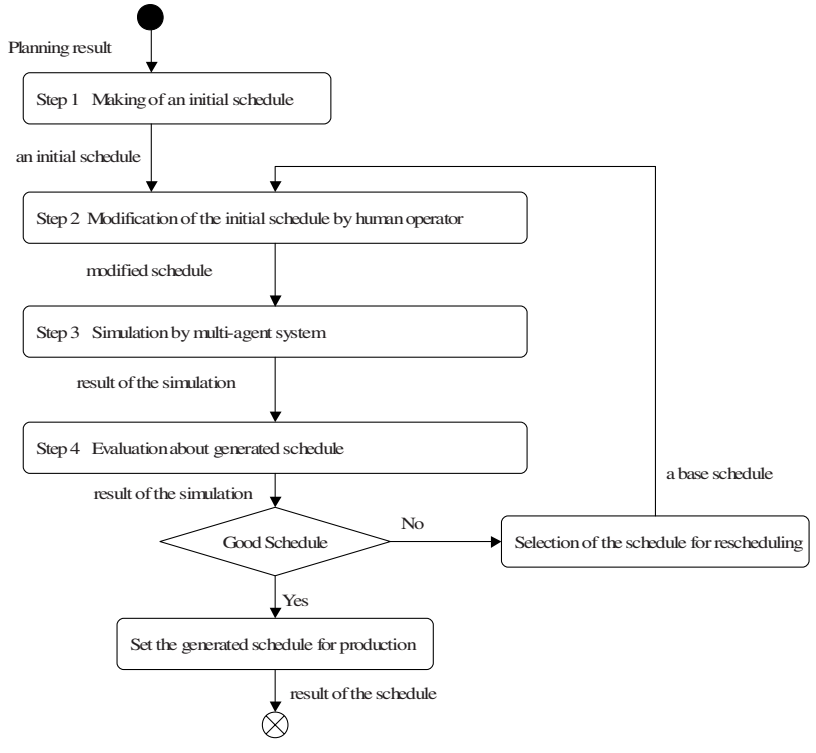


Fig. 5. Flow Chat for Generation of the Schedule

All control recipes for production are assigned in the planning stage. In Step1, detailed information about all control recipes should be added such as start time. As an initial schedule, they are displayed by order sequence on the Gantt chart. The recipes are shown as the combination of task bars. The length of each task bar represents batch process duration. The phases in each recipe are associated with task duration data element, which is used for scheduling. These task durations are not the results of the behavior in the multi-agent simulation.

In Step2, operator picks the overall or parts of the underlying recipe, drags it from side to side and drops it on the desired time. The modification of an initial schedule or a base schedule has constraints such as the conditions of the different phases and resource sharing. GUI highlights unsatisfied conditions using alerts in the user interface. The output data in Step2 is called modified schedule. If some problems exist in the modified schedule, it is possible to simulate it as in Step3.

The purpose of Step3 is to generate an option as feasible detailed schedule for plant floor. It might be generated from the modified schedule, which is the output of Step2. The multi-agent simulation runs using the modified schedule as in Step3. All recipes

add the conditions for all phases for the modified schedule. For example, one of the added conditions is the constraint about start time. This constraint is equal to the location of that task on Gantt chart, as explained in Step2. Agents don't start the task until the constraints are cleared. If the modified schedule in Step2 is too tight, the effect of the delay might propagate over the following batch processes. Therefore, the human operator must evaluate the simulation result and judge whether the result can be used for plant floor execution schedule or not, as in Step4. When results can not be accepted, human operator must decide which schedule to use for rescheduling to find better execution schedule.

4 Conclusion

In this paper, we discussed useful system structure that can be used to shorten the modification time of batch plant management system and proposed the usage of batch controller using multi-agent system, which is used as a simulator for scheduling and operator supervision. The proposed batch controller system can be used as a simulator for real-time scheduler. The proposed scheduling system can be amended automatically by modifying the proposed batch controller system. The realization of the proposed integration between the controller and scheduler of batch plant will provide great flexibility for multi-purpose and multi-batch operation.

Acknowledgements

This research was partially supported by The Nitto Foundation.

References

1. ANSI/ISA-S88.01-1995 Batch Control Part1 - Models and Terminology, ISA (1995)
2. Bee-gent home page, http://www2.toshiba.co.jp/beegent/index.htm
3. Hamaguchi, Y., Hashimoto, Y., Itoh, T., Togari, Y.: Autonomous Decentralized Control System for Batch gent-based Batch Process. In: Proceedings of International Symposium on Design, Operation and Control of Chemical Processes, pp. 207–212 (2002)
4. Sakamoto, M., Eguchi, H., Hamaguchi, T., Ota, Y., Hashimoto, Y., Itoh, T.: Agent-based Batch Process Control Systems. In: Negoita, M.G., Howlett, R.J., Jain, L.C. (eds.) KES 2004. LNCS (LNAI), vol. 3214, pp. 398–404. Springer, Heidelberg (2004)

Optimization Method for Design of Supercritical Water Oxidation Process Using Genetic Algorithm

Hiroya Yamasaki[1], Kazuhiro Takeda[1], Yukiyasu Shimada[2], Takashi Hamaguchi[3], Teiji Kitajima[4], and Tetsuo Fuchino[5]

[1] Shizuoka University, 3-5-1, Johoku, Hamamatsu, Shizuoka 432-8561, Japan
tktaked@ipc.shizuoka.ac.jp
[2] National Institute of Occupational Safety and Health, 1-4-6, Umezono, Kiyose, Tokyo 204-0024, Japan
[3] Nagoya Institute of Technology, Gokiso-cho, Showa-ku, Nagoya 466-8555, Japan
[4] Tokyo University of Agriculture and Technology, 2-24-16, Naka-cho, Koganei, Tokyo, 184-8588, Japan
[5] Tokyo Institute of Technology, 2-12-1, Ookayama, Meguro, Tokyo 152-8552, Japan

Abstract. To decide optimum conditions for a chemical process is difficult task. Design variables of chemical process like the supercritical water oxidation process are discrete variables (ex. structure) and continuous variables (ex. temperature, pressure). This paper develops a method simultaneously optimizes continuous variables and discrete variables in safety region using genetic algorithm. We demonstrate the proposed method using the supercritical water oxidation process in this study.

1 Introduction

In these days, processes using supercritical water oxidation (called SCWO) are regarded to extract energy from organic wastes. The processes need to spend much energy to achieve supercritical state. In contrast, the processes generate much energy through oxidation reaction. Then, it is necessary to optimize conditions of the processes. The conditions consist of discrete variables (ex. structure) and continuous variables (ex. temperature, pressure). To optimize conditions is very difficult task, because discrete variables and continuous variables are simultaneously searched. Some methods have been proposed to optimize that problem, which are mixed integer programming, genetic algorithm (called GA), and so on. GA is tough to optimize a multi-modal problem and a multi-objective problem. Many researchers [1] [2] [3] have proposed the method by use of GA to search discrete variables and continuous variables.

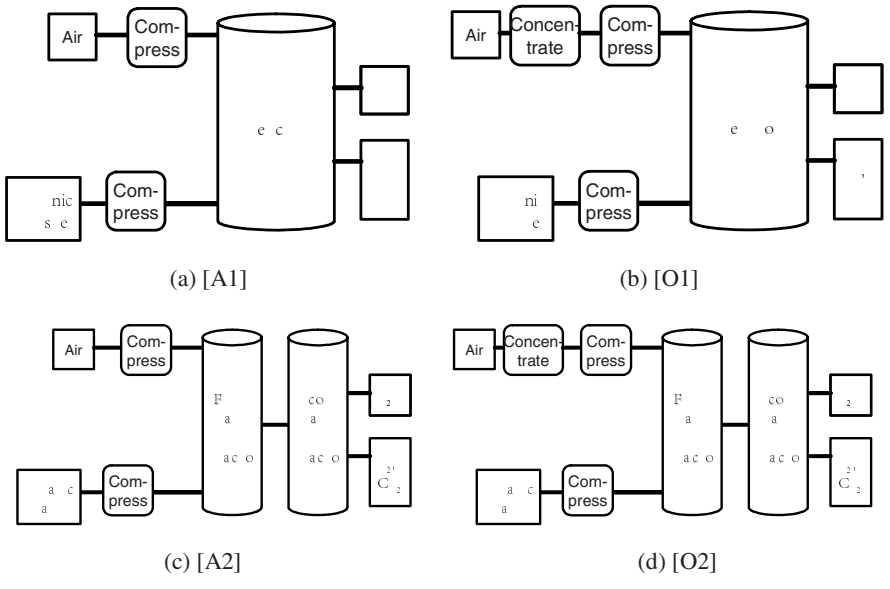
The design problem of SCWO may be multi-modal problem. For each structure, different temperatures may be the best temperature. In this paper, a method is proposed to simultaneously search optimum discrete variable (structure) and continuous variable (temperature) of SCWO process using GA.

B. Apolloni et al. (Eds.): KES 2007/ WIRN 2007, Part II, LNAI 4693, pp. 744–751, 2007.
© Springer-Verlag Berlin Heidelberg 2007

2. Supercritical Water Oxidation Process

2.1 Objective Process

SCWO is an efficient and versatile oxidation process of organic waste to carbon dioxide, nitrogen and water. It is based on a homogeneous reaction carried out at operational conditions above the critical point of pure water. For oxidation, air or concentrated oxygen is compressed and preheated. Organic waste is also compressed and preheated. The process generates much energy through oxidation reaction. The process is constructed by one-pass process or two-pass process. One-pass process uses one reactor in high temperature (about 650C) and high pressure (about 15MPa) to be supercritical state. Two-pass process uses two reactors in middle temperature (about 450C and 300C) and high pressure (about 15Mpa). First-pass is in supercritical state, and second-pass is in subcritical state.



(a) [A1] (b) [O1]

(c) [A2] (d) [O2]

Fig. 1. Schematics of objective processes

Then, following 4 types of processes are considerable.

- [A1] Air is used for oxidation, and one-pass process is used. (Fig. 1 (a)).
- [O1] Concentrated oxygen is used for oxidation, and one-pass process is used. (Fig. 1 (b)).
- [A2] Air is used for oxidation, and two-pass process is used. (Fig. 1 (c)).
- [O2] Concentrated oxygen is used for oxidation, and two-pass process is used. (Fig. 1 (d)).

2.2 Objective Function

In this paper, surplus energy of objective process is evaluated to optimize. Surplus energy is difference between input energy and output energy. More surplus energy process is better process, in this study. To simplify, surplus energy is calculated by use of objective function of temperature.

For each processes, surplus (input and output) energy is calculated as following.

Case O1:

- Input energy: (concentrating oxygen from air) + (compressing oxygen) + (preheating oxygen) + (compressing organic waste) + (preheating organic waste)
- Output energy: (combustion heat of SCWO) + (heat of emission water)

Case A1:

- Input energy: (compressing air) + (preheating air) + (compressing organic waste) + (preheating organic waste)
- Output energy: the same function as output energy of case O1.

Case O2:

- Input energy: the same function as input energy of case O1.
- Output energy: (combustion heat of first-pass of SCWO) + (heat of emission water of first-pass) + (combustion heat of second-pass of SCWO) + (heat of emission water of second-pass)

Case A2:

- Input energy: the same function as input energy of case A1.
- Output energy: the same function as output energy of case O2.

There are the same functions, but the calculated energy may be different, because the different temperatures are used for cases. For example, energy for compressing air is calculated as follows.

$$Qo = -\left[s\, m\, n\, R\, T/(n\text{-}1) \right] \left[\{P2/P1\}^{(n\text{-}1)/s\, n} - 1 \right] / r \tag{1}$$

where Qo [kJ] is energy for compressing air, s is the number of stage of compressing, m [mol] is 1.2 times of required air to break down 1 ton of organic waste, n is cp/cv, R [J/mol K] is gas constant, T [K] is temperature, $P1$ [Mpa] is initial pressure, $P2$ [Mpa] is final pressure, r is power generation efficiency. Qo is not a function of temperature.

Energy for preheating air is calculated as follows.

$$Qh = dH\, m \tag{2}$$

where Qh [kJ] is energy for preheating air, dH is difference enthalpy between initial temperature and final temperature. Qh is a function of temperature.

For two-pass processes, temperature of first-pass is dependent variable. In pilot study, temperature of first-pass is 450C and that of second-pass is 300C. So, it is assumed that (temperature of second-pass) = (temperature of first-pass) – 150.

Furthermore, process structure combination should be considered. As mentioned above, each function corresponding to process structure is needed, respectively. But it is costly tasks to prepare all functions newly. Therefore, each function is constructed by combination of sub-functions corresponding to sub-processes (ex. concentrating, compressing).

3 Optimization Method Using Genetic Algorithm

3.1 Principle of Genetic Algorithm

GA belongs to evolutionary algorithms and is devoted to simulate animals' reproduction. The algorithm is based on that only individuals that are strong may survive and will be able to reproduce. The general schema of GA may be summed up as follows (Fig. 2). First of all, an initial population of potential solutions is generated randomly. One makes also evolve this population by recombination of individuals by the mean of crossover operator and/or mutation operator. A selection procedure based on an objective function enables to choose the individuals candidate for reproduction. The reproduction consists in recombining two individuals by the crossover operator, possibly followed by a mutation of the offspring. Therefore, from the initial population a new generation is obtained. From this new generation, a second new generation is produced by the same process and so on. The stop criterion is based on the number of generations. In the remainder of this section, the details of the settings of the GA we have implemented.

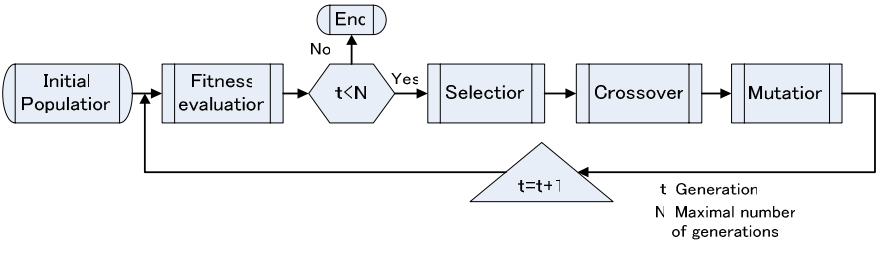


Fig. 2. Flow of genetic algorithm

3.2 Encoding of the Solution

As described above, energy to preheat may be reduced by concentrating oxygen, but energy to concentrate is required. On the other hand, one-pass process needs energy to heat up to about 650C, but two-pass process may need less energy than one-pass process. Therefore, to maximize (optimize) evaluated value, process structure and temperature should be simultaneously considered. Process structure is discrete, and temperature is continuous. In this paper, discrete variables and continuous variables are expressed in one set of genes. Design of genes is displayed in Fig. 3. The genes consist of process structure bits and temperature bits. Process structures are coded using 2 bits. Case O1 is coded as [00], case A1 is codes as [01], case O2 is coded as [10] and case A2 is coded as [11]. Temperature from 400C to 650C is coded using 8

bits. Lower temperature is decided by enough destruction of organic waste, and upper temperature is decided by material and safety limit. In this paper, it was assumed that pressure was 15MPa (constant) and temperature of second-pass of two-pass process was decided by that of first-pass. Therefore, whole the search space (10bits = 1024 patterns) may be searched. But, these assumptions will be resolved in future works. To consider these variables, search space may be large not to be able to search whole the space.

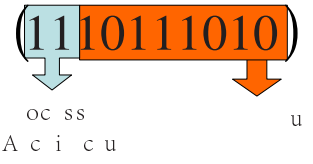


Fig. 3. Design of genes

3.2 Details of the Settings of the Genetic Algorithm

At first, 20 initial solutions are generated randomly. They must be feasible, because of the design of genes. As a selection technique, the elite technique and the roulette technique are employed. Using the elite technique, upper 6 solutions are selected to improve the new successive generations. Using the roulette technique, 14 solutions are randomly selected among 20 solutions to escape from local optimum solutions. Using only the elite technique, the solutions are easy to uniform through generations and search local space. Using the elite technique and the roulette technique, the solutions are able to keep variety to search global space. At crossover, as shown in Fig. 4, select and cross best and worst of solutions, and select and cross second solutions in evaluated value order, and so on. All solutions are crossed. Let [a1, a2, ..., a10] and [b1, b2, ..., b10] be 2 strings encoding 2 solutions. Righter bits than a2 (or b2) mean larger value. To cross large value and small value, the crossover with two crossing point in creating two offspring as follows: [a1, ..., a4, b5, ..., b8, a9, a10] and [b1, ..., b4, a5, ..., a8, b9, b10]. In the crossover, architecture bits and temperature bits are not divided. At every generation, all bits of all solutions may be mutated. The mutation rate is 1/150. Above settings are decided by experiments.

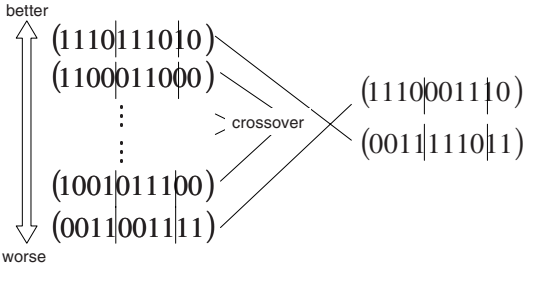


Fig. 4. Crossover operator

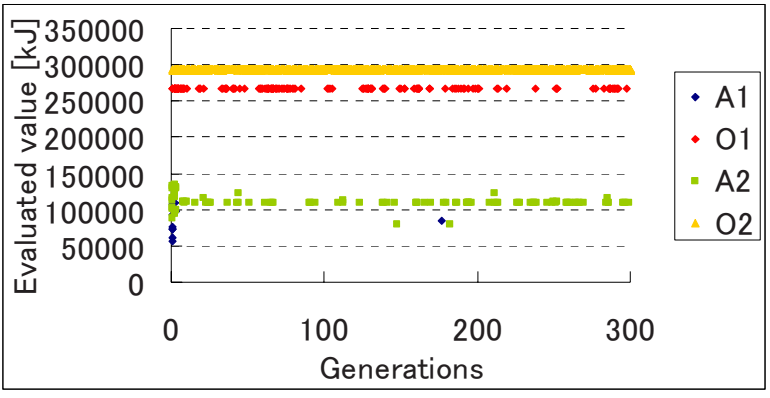


Fig. 5. Evaluated value of every generation

4 Results and Discussion

4.1 Results by Use of GA

Calculation results under above conditions are displayed in Fig. 5. Because of architecture bits of genes, all architectures are considered at the same time. The maximum evaluated value (surplus energy) was 2.93E+5kJ of [O2] process. The optimum process was [O2] process. Best solution of final (300th) generation was coded at 524C. Therefore, the optimum design was that oxygen was concentrated and two-pass process was used for SCWO at 524C. In Fig. 6, solutions of [O2] process successively survive. Various solutions were generated by crossover and mutation. Because of selection, worse solutions were expired. The solutions were converged to the maximum value. But, as shown in Fig. 7, those of the others were randomly generated by mutation and were expired soon.

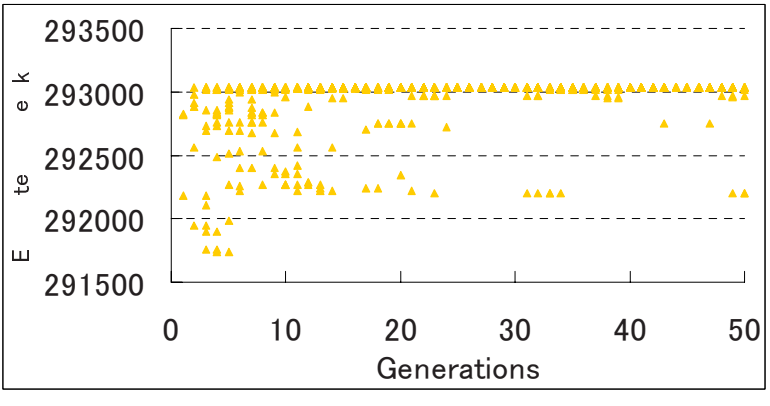


Fig. 6. Evaluated value of every generation of [O2] process

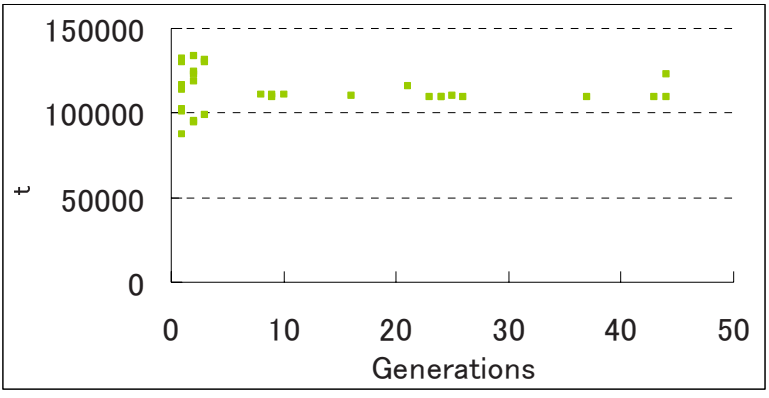


Fig. 7. Evaluated value of every generation of [A2] process

4.2 Result of Reference Experiment

To verify the maximum evaluated value of [O2] process, evaluated value (surplus energy) was plotted against temperature in Fig. 8. The curve was unimodal. The maximum evaluated value was 2.93E+5kJ at 524C. Consequently, the proposed method was able to search the optimum design conditions, which were process structure and temperature.

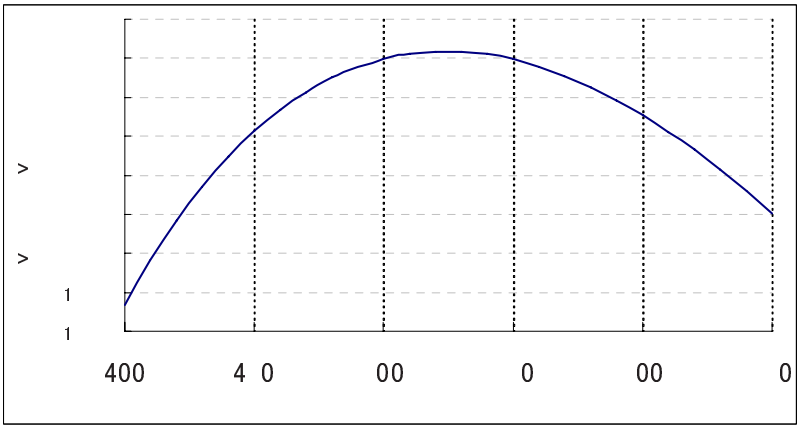


Fig. 8. Temperature vs. evaluated value of [O2] process

5 Conclusion

The search method using GA is proposed to optimize the design of SCWO process. The method is able to optimize discrete variables (structure) and continuous variables (temperature) in safety region, simultaneously. The optimum design was that oxygen was concentrated and two-pass process was used for SCWO at 524C.

In this paper, it was assumed that pressure was 15MPa (constant) and temperature of second-pass of two-pass process was decided by that of first-pass. These assumptions will be resolved in future works. The settings of operations of GA (selection, crossover, mutation) were decided by experiments. For general versatility, it is desirable that these settings are decided by systematic methodology.

Acknowledgments. The authors thank for the support of New Energy and Industrial Technology Development Organization (NEDO) on the project ”Development of elemental technology for bioenergy conversion”.

References

1. Elegbede, C., Adjallah, K.: Availability allocation to repairable systems with genetic algorithms: a multi-objective formulation. Reliability Engineering and System Safety 82, 319–330 (2003)
2. Jansson, N., Wakeman, W.D., Manson, J.-A.E.: Optimization of hybrid thermoplastic composite structures using surrogate models and genetic algorithms. Composite Structures 80, 21–31 (2007)
3. Parmee, I.C.: Genetic Algorithms and Hydropower System Design. Computer-Aided Civil and Infrastructure Engineering 13, 31–41 (1998)

An Implementation Design of a Fine-Grained Database Access Control Policy Consistency Checking Mechanism

Bat-Odon Purevjii[1], Masayoshi Aritsugi[2], Sayaka Imai[1], and Yoshinari Kanamori[1]

[1] Department of Computer Science, Graduate School of Engineering, Gunma University, 1-5-1 Tenjin-cho, Kiryu, Gunma 376-8515, Japan
batodon, sayaka, kanamori@dbms.cs.gunma-u.ac.jp
[2] Computer Science and Electrical Engineering, Graduate School of Science and Technology, Kumamoto University, 2-39-1 Kurokami, Kumamoto 860-8555, Japan
aritsugi@cs.kumamoto-u.ac.jp

Abstract. Policy consistency checker is an essential part of database access control systems. Realizing checking mechanisms for policy inconsistencies is a non-trivial task due to large amounts of rules and various kinds of rule patterns on fine-grained database objects and complicated interrelationships among those rules. We propose a novel rule-based mechanism for checking fine-grained access control policy inconsistencies for relational databases. In particular, we present necessary concepts and implementation design of the inconsistency checking mechanism. We specify policies, with conditions, on fine-grained objects through if-then authorization rules. We compute rule inconsistencies by employing the java-based open-source rule engine Jess.

Keywords: fine-grained database access control, policy consistency.

1 Introduction

Realizing a policy consistency checker within a database access control system is a non-trivial task because of the following reason. Databases normally hold large amounts of data, with various granularities, and large numbers of users interact on the databases [1]. Many rules and various kinds of rule patterns on the fine-grained database objects need to be specified, manipulated, and maintained as well. Complicated interrelationships among those rules or components of rules make the consistency checking process difficult.

To our knowledge, very few efforts have been achieved on deploying fine-grained consistency checker in database access control systems to date. Several research-based fine-grained database access control systems have been proposed [2,3,4,1] and few commercial products have been implemented so far [5,6,7]. None of them discuss how to realize fine-grained policy consistency checker within them. Currently only tool that helps security administrators (SA) analyzing

B. Apolloni et al. (Eds.): KES 2007/ WIRN 2007, Part II, LNAI 4693, pp. 752–760, 2007.
© Springer-Verlag Berlin Heidelberg 2007

policies is Oracle Policy Manager [6,7]. The tool supports SAs a GUI environment to observe policies with corresponding components including tables. Before introducing its limitation, consider the following fact. An end user's access scope on a table has to be calculated depending on all policies specified on it [4]. It seems that we are able to determine which policies correspond to a particular table by employing the grouping facility of the tool. However, the tool currently has no capabilities of describing users' access scopes on a particular table. That means SAs cannot observe increases or decreases of a particular user's access scope during insertion or updates of policies; they seem to be required to observe access scope changes by themselves.

In this paper we present necessary concepts and implementation design of a novel rule-based mechanism for checking fine-grained access control policy inconsistencies for relational databases. We specify policies, with conditions, through if-then authorization rules and check their inconsistencies by employing Jess rule-engine. We introduce the definitions of rule inconsistencies and types of them in the first part. In the second part, we introduce a way of computing the inconsistencies by employing Jess rule-engine [8]. The computation employs its underlying data structure, called the Rete network, and its language constructs. We provide algorithms of computing inconsistencies at fine-grained levels of rules.

2 Preliminaries

We use the following definition of a relational database [4,9]. A relation is a subset of the Cartesian product of a list of domains. The product of domains $D_1, D_2 \dots D_k$, written $D_1 \mathrm{x} D_2 \mathrm{x} \dots \mathrm{x} D_k$, is the set of all k-tuples $(a_1, a_2 \dots a_k)$ such that a_1 is in D_1, a_2 is in D_2, and so on. Each relation that is a subset of some product $D_1 \mathrm{x} D_2 \mathrm{x} \dots \mathrm{x} D_k$ of k domains is said to have *arity* k. A tuple $(a_1, a_2 \dots a_k)$ has k components and is called k-tuple. The set of attribute names for a relation is called *relation schema*. If we name a relation R, and its relation schema has attributes $A_1, A_2 \dots A_k$, we write the relation schema as $R(A_1, A_2 \dots A_k)$. A database schema DS is a set of relation schemas $R_1, R_2 \dots R_k$. A relation instance can be thought of as a *table* in which each tuple is a *row*. A database instance I of a DS is a set of relation instances $R_1(DS), R_2(DS) \dots R_k(DS)$, where each $R_i(DS)$ is a relation instance on R_i.

Authorization rule. Authorization rules on relations are defined through if-$then$ rules in our framework [10,11]. Main constructs of the language is *atomic formulas* (*atoms*), which are predicate symbols with a list of arguments, e.g., $p(A_1, A_2 \dots A_k)$, where p is predicate symbol and $A_1, A_2 \dots A_k$ are arguments. An argument can either be a variable, constant or a function symbol. *Rules* are written in the form $q \leftarrow p_1 \& \dots \& p_n$, where q is the *head* and $p_1 \& \dots \& p_n$ is the *body* of the rule. The symbol $\leftarrow$ denotes "if" condition. A body consists of one or more atoms, called *subgoals*, and the atoms are connected by (logical) AND. The atoms in the body of a rule denote database relations. The atomic formulas can also be constructed with the arithmetic comparison predicates, $>$,

$\geq$, and so on; these predicates are called as *built-in predicates*. The following is the general form of our authorization rule:

$$permitted(s, o(f_1(X_1, X_2, \ldots, X_j), \ldots, f_k(Y_1, Y_2, \ldots, Y_l)), m) \leftarrow N_1, \ldots, N_k.$$

where s and m are expressions, that can be a constant or variable symbol, $o, f_1, \ldots, f_k$ are function symbols and $N_1, ..., N_k$ are subgoals. A subgoal N_i can either be an atom denoting a relation or an atom with a built-in predicate[1]. We assume the domain of s is a set of user IDs, and the domain of m is privilege modes of SQL which has fixed number of members, namely, *select*, *insert*, *delete*, and *update*. When s and m components are variables, they indicate a user group, or just group, and all privileges, respectively. When they are constants, particular users and privileges are indicated by them. The function o with structure $o(f_1(X_1, X_2, \ldots, X_j), \ldots, f_k(Y_1, Y_2, \ldots, Y_l))$ indicates an authorization object. A single component $f_i(Z_1, Z_2, \ldots, Z_m)$ inside o expresses a derived relation. Derived relations are not actually stored in database; instead, they are defined through rules. The difference between functions and predicates is as follows. Functions take domain values as arguments and return domain values. Predicates take domain values as arguments but return truth values. Authorization rules describe which subjects can execute which privilege on what objects. In other words, each rule defines some kind of *access scope* (AS) on relations for users with a particular privilege. There can be horizontal and vertical ASs. *Horizontal scope* (HS) on a relation is defined by the number of arguments in its derived relation in a rule head. *Vertical scope* (VS) of a user on a relation need separately be defined on each attribute since they refer to different domains. Built-in predicates defined on an attribute change the scope or the number of domain elements the attribute refers. Intuitively, an attribute without a built-in predicate has more domain scope than an attribute with a built-in predicate. VS on an attribute expresses domain scope of that attribute users are permitted execute a privilege. Defining different HSs and VSs for different privileges is usual [5,6]. Thus, an AS, either HS or VS, is required to be separately computed for each individual user or group for each individual privilege.

3 Fine-Grained Authorization Rule Inconsistencies

Organizational or *default policies* (DPs) fix the following three constraints:

- Who, that is users or groups, are eligible to interact with database,
- Which privileges the users or groups are eligible to execute,
- How much ASs are assigned to the users or groups with particular privileges.

SAs are responsible for specifying, analyzing and maintaining corresponding authorization rules of the DPs. During the maintenance, SAs must keep the DPs consistent. We assume default rules that are defined through default policies are already stored in rule base (RB). That means the inconsistency types are defined comparatively to default rules. Rules a SA is trying to insert at a

[1] We abbreviate such atoms as built-in predicates in the rest of the paper for simplicity.

moment are called *current* rules. A default and a current rule are denoted as r_d and r_c, respectively, in the definitions. Depending on rule components, the three types of inconsistencies can occur, namely, *subject-inconsistencies*, *privilege-inconsistencies*, and *access-scope-inconsistencies*. If SAs define inconsistent rules they violate the three constraints. When a SA makes mistakes on s component of a rule, s/he will assign a privilege to an improper user or group. This type of inconsistency is called *subject-inconsistencies* (SIs). In particular, when the SA inserts a rule with a new s component indicating a user and a group, *improper-user-assignment-inconsistency* ($IUAI$) and *improper-group-assignment-inconsistency* ($IGAI$) can occur, respectively. When the SA makes mistakes on m component of a rule, s/he will assign an improper privilege to a user or group who has another privilege. This type of inconsistency is called *privilege-inconsistencies* (PIs). When s component is a constant, *improper-privilege-assignment-for-user-inconsistency* ($IPAFUI$) and *improper-all-privilege-assignment-for-user-inconsistency* ($IAPAFUI$) can occur. m is a constant in the former case and a variable in the latter case, respectively. Similarly, when s component is a variable, *improper-privilege-assignment-for-group-inconsistency* ($IPAFGI$) and *improper-all-privilege-assignment-for-group-inconsistency* ($IAPAFGI$) can occur. The final inconsistency type, called *access-scope-inconsistencies* (ASIs), occurs depending on o component and built-in predicates together. ASIs indicate changes of ASs of a particular user or group for a particular privilege. That means rules we consider here have the same s and m components. First, we introduce ASIs between two rules, r_d and r_c, only. General definitions of ASIs are given later based on this. ASs of u_k on relation rel_i, defined through r_d and r_c, are denoted DAS ($default\ AS$) and CAS (*current AS*), respectively. Let us assume DHS and CHS denote default and current HSs of u_k on rel_i. The following three definitions are related to HSs.

Definition 1. r_c has *horizontal-increase-inconsistency* (HII) with r_d on rel_i when DHS and CHS have at least a common and an uncommon attribute. The uncommon attributes of CHS express the increase of DHS.

Definition 2. r_c has *horizontal-movement-inconsistency* (HMI) with r_d on rel_i when DHS and CHS have no common attributes. The uncommon attributes of CHS express the movement of DHS.

Definition 3. r_c has no HS *inconsistencies* (HSIs) with r_d on rel_i when the attributes of CHS are a subset of the attributes of DHS. The uncommon attributes of DHS indicate the decrease of DHS.

The next three definitions are related to VSs. Let us consider a database schema $rel_1(A_1, A_2, A_3)$, $rel_2(B_1, B_2, B_3, B_4)$ and $rel_3(C_1, C_2, C_3, C_4)$. Let us assume DVS and CVS denote $default$ and *current* VSs of u_k on attribute A_j of rel_i.

Definition 4. r_c has *vertical-increase-inconsistency* (VII) with r_d on A_j of rel_i when DVS and CVS refer to at least a common and an uncommon domain element. The uncommon domain elements of CVS indicate the increase of DVS.

Definition 5. r_c has *vertical-movement-inconsistency* (VMI) with r_d on attribute A_j of rel_i when DVS and CVS refer to disjoint domain elements. The uncommon domain elements of CVS indicate the movement of DVS.

Definition 6. r_c has no VS *inconsistencies* (VSIs) with r_d on attribute A_j of rel_i when CVS refer to a subset of DVS's referring domain elements. The uncommon domain elements CVS indicate the decrease of DVS.

General definitions of inconsistencies are as follows.

Definition 7. r_c has no ASIs with r_d when the following conditions are true:

- There are no HSIs between each CHS_i and DHS_i on each rel_i,
- There are no VSIs between each CVS_j and DVS_j on each attribute A_j.

Now let us consider DASs of u_k consists of several default rules and a set of current rules, denoted $r_c^1, r_c^2, \ldots, r_c^n$, are inserted into RB. Then, HS on rel_i and VS on A_j are defined by adding together all rules' corresponding HSs and VSs, respectively. Let us call them *total* HS (THS) and *total* VS (TVS). Then THS_d^i and TVS_d^j denote default rules' THS on rel_i and TVS on A_j, respectively.

Definition 8. Each r_c^i has no ASIs with default rules when each CHS_i and CVS_j has no inconsistencies with each THS_d^i and TVS_d^j, respectively.

4 Computing Rule Inconsistencies

Let us assume $DRS = r_d^1, r_d^2, \ldots, r_d^n$ and $CRS = r_c^1, r_c^2, \ldots, r_c^m$ denote a default rule set and a current rule set, respectively. DRS and CRS indicate rules stored in RB and rules SA tries to insert into RB at a moment, respectively. We assume each rule in CRS hold a distinct user or group. Then, we can compute inconsistencies for a particular user or group at a moment. After doing a computation for a particular user, we can repeat the process for each user. Let us assume r_c^i denotes a current rule we are computing inconsistencies. Let us assume u_i denotes s component of the rule as well. The inconsistency types need to be computed in a fixed order. First, we need to compute SIs. That means we find default rules having the same s component with the r_c^i. The computation checks whether the user u_i is eligible to interact with database. If there is no rule with u_i in DRS, no more computations are needed. The user u_i is not allowed to interact with database by any default rule before. In other words, r_c^i has $IUAI$. The rule must be rejected. The process is similar to s indicating a group. If there are one or more rules with u_i in RB, a set of rules, called DRS_{sc}^i, is produced for further computations. Rules in DRS_{sc}^i now correspond to user u_i only. Second, PIs need to be computed. Let us assume m_i is the particular privilege included in r_c^i. Then, we need to check whether there are rules having privilege m_i in DRS_{sc}^i. If there is no rule with m_i, no more computations are necessary. The user u_i is not allowed to execute the privilege m_i by any default rule before. In other words, r_c^i has $IPAFUI$. The rule must be rejected. The process is similar to m indicating all privilege. If there are one or more rules with m_i in RB, a set of

rules, called DRS^i_{pc}, is produced for further computations. Rules in DRS^i_{pc} now correspond to user u_i and the privilege m_i only. Computations of SIs and PIs are not difficult. We just need to compare s and m components of rules. Thus, we omit the algorithms of the computations from this paper. Finally, ASIs need to be computed. We demonstrate the computations in the following.

Computing Access-Scope-Inconsistencies. The inputs of the computation are DRS^i_{pc} and r^i_c. As we know, sums of HSs and VSs of all rules in DRS^i_{pc} define THS and TVS of the u_i for m_i, respectively. Therefore, first, we need to compute u_i's THS and TVS from rules in DRS^i_{pc}. Then, r^i_c's ASIs can be computed by comparing its HSs with THS and VSs with TVS. The computations are done by using Jess rule-engine [8]; it is an open-source java library. We employ Jess's underlying data structure, called the Rete network [12], to process rules. It is a kind of rooted directed acyclic graph that consists of a network of interconnected nodes. We use $deftemplate$, $defmodule$ and $defrule$ constructs of Jess to insert our rules into a Jess rule engine. Once rules are inserted, the engine builds the corresponding Rete network. We define $deftemplate$s for attributes and $defmodule$s for relations. That allows us to create top level nodes for attributes. Intermediate nodes hold built-in predicates. At the bottom of the network are nodes representing rules. A common node for rules can either be an attribute node or a built-in predicate node. If a built-in predicate on an attribute in one rule is different than another's, a new node is created as branching from the parent. If an attribute or built-in predicates are not defined in a rule, no path goes to the rule.

Computing Vertical-Scope-Inconsistencies (VSIs). The algorithm consists of three modules, namely, $tvsComp$, $vsComp1$ and $vsComp2$. $tvsComp$ computes TVS from the rules in DRS^i_{pc}. For space constraint, we omit the code of $tvsComp$. The module takes a Rete network of rules in DRS^i_{pc} as input and produces $ruleTVS$, a rule indicating TVS of the current user. Then, we can compute VSIs of r^i_c by traversing the network of r^i_c and $ruleTVS$. Input of $vsComp1$[2] is the network of r^i_c and $ruleTVS$ and output is VSIs of r^i_c. $vsComp1$ module traverse from the root and its children; the children nodes correspond to attributes. VSIs are calculated for each attribute $an[j]$ (line 2). If $an[j]$ has a single child and it is leaf, which is a rule (name), no computation is necessary on $an[j]$, because, either r^i_c or $ruleTVS$ does not include $an[j]$ in it (line 5). It is not possible to compare VSs of the two rules in this case. Such differences need to be considered for computations of HSIs. If the child is not leaf, we get children of $an[j]$. The children are built-in predicates. If there is a single child, no computation is necessary (line 7), in contrast, r^i_c has no VSIs on $an[j]$. Because, r^i_c and $ruleTVS$ have the common built-in predicate an_child[1] (line 8); the rules have the same VSs on the attribute. If $an[j]$ has more than one child, more computations need to be done (line 9). If an_child[1] and an_child[2] both are leaves, then r^i_c has no VSIs, because, r^i_c and $ruleTVS$ have the common $an[j]$

[2] r^i_c is denoted as rci in $vsComp1$ and $vsComp2$ modules.

without built-in predicates (line 11). When *an_child*[1] is leaf and *an_child*[2] is not leaf, we need to clarify which one of them is r_c^i. If *an_child*[1] is r_c^i it has VII on $an[j]$ (line 13) because the attribute $an[j]$ is defined without built-in predicate in r_c^i and with built-in predicate in $ruleTVS$. The reason is explained in the preliminaries. In contrast, r_c^i has no VSIs on the attribute (line 14). From line 15, the algorithm computes VSIs when *an_child*[1] is not leaf. If *an_child*[2] is leaf (line 16), we need to check which one of *an_child*[1] and *an_child*[2] is r_c^i. The process between lines 17 and 18 is similar to lines between 13 and 14.

```
module vsComp1(), in: A Rete of rci and ruleTVS, out: VSIs
1.traverse children of the root, store in array an[m]
2.for each attribute node an[j] do
3.  traverse children of an[j], put in an_child[n]
4.    if n = 1 then
5.      if an_child[1] is leaf then print "VS is not considered"
6.      else get children of an_child[1], put in leaf[k]
7.        if k = 1 then print "an[j]'s VS is not considered"
8.        else print "rci has no VSI with ruleTVS on an[j]"
9.   else
10.     if an_child[1] is leaf then
11.       if an_child[2] is leaf then print "rci has no VSI"
12.       else
13.         if an_child[1] is rci then print "rci has VII"
14.         else print "rci has no VSI on an[j]"
15.     else
16.       if an_child[2] is leaf then
17.         if an_child[2] is rci then print "rci has VII"
18.         else print "rci has no VSI on an[j]"
19.       else vsComp2(an_child[1], an_child[2])
```

When *an_child*[1] and *an_child*[2] both are not leaves (line 19), *vsComp*2 module is employed. Different built-in predicates are defined on $an[j]$ of r_c^i and $ruleTVS$ in this case. Depending on built-in operator types and their values, many variations of relationships between them can occur. For simplicity, we currently provide the computation for $=$ and $>$ operators only. The relationships and corresponding VSIs are given in Table 1.

Table 1. Variations of relationships between two built-in predicates on an attribute

r_c^i \ *ruleTVS*	*an[j] = v2*	*an[j] > v2*
an[j] = v1	*v2 = v1*: no *VSIs* *v2 > v1*: *VMI* *v2 < v1*: *VMI*	*v2 = v1*: *VMI* *v2 > v1*: *VMI* *v2 < v1*: no *VSIs*
an[j] > v1	*v2 = v1*: *VMI* *v2 > v1*: *VII* *v2 < v1*: *VMI*	*v2 = v1*: no *VSIs* *v2 > v1*: *VII* *v2 < v1*: no *VSIs*

v1 – value of the built-in predicate of r_c^i on *an[j]* of rel_k

v2 – value of the built-in predicate of *ruleTVS* on *an[j]* of rel_k

```
module vsComp2(), in: an_child[1], an_child[2], out: VSIs on an[j]
1.get operators and values of an_child[1] and an_child[2]
2.assign them to op1, op2, v1 and v2
3.if op1 equal to op2 then
4.  if op2 is "=" then
5.    if v1 = v2 then print "rci has no VSI with ruleTVS on an[j]"
6.    else print "rci has VMI with ruleTVS on an[j]"
7.  else
8.    if v2 > v1 then print "rci has VII on an[j]"
9.    else print "rci has no VSI on an[j]"
10.else
11.  if op2 is "=" then
12.    if v2 > v1 then print "rci has VII on an[j]"
13.    else print "rci has VMI on an[j]"
14.  else
15.    if v2 >= v1 then print "rci has VMI on an[j]"
16.    else print "rci has no VSI on an[j]"
```

Computing Horizontal-Scope-Inconsistencies (HSIs). They can be computed by counting number of shared and non-shared attribute nodes for leaves.

5 Conclusion

We have presented necessary concepts and an implementation design of a novel rule-based fine-grained access control policy consistency checking mechanism. Definition of inconsistencies and types of them are introduced. The way of computing the inconsistencies by Jess rule-engine is presented. The computations are done with the help of its underlying data structure and language constructs.

Checking inconsistencies for rules holding interval-based definitions can be an interesting issue to be investigated.

Acknowledgments

This work was supported in part by the Japan Society for the Promotion of Science, Grant-in-Aid for Scientific Research (C) 18500073.

References

1. Rizvi, S., Mendelzon, A.O., Sudarshan, S., Roy, P.: Extending query rewriting techniques for fine-grained access control. In: SIGMOD Conf., pp. 551–562 (2004)
2. Bertino, E., Jajodia, S., Samarati, P.: Supporting multiple access control policies in database systems. In: IEEE Symposium on Security and Privacy, pp. 94–107 (1996)

3. Jajodia, S., Samarati, P., Subrahmanian, V.S., Bertino, E.: A unified framework for enforcing multiple access control policies. In: SIGMOD Conf., pp. 474–485 (1997)
4. Motro, A.: An access authorization model for relational databases based on algebraic manipulation of view definitions. In: ICDE, pp. 339–347 (1989)
5. Bird, P.: Implementing low level access control with db2 udb. The IDUG Solution Journal 7(3) (2000)
6. Nanda, A., Burleson, D.K.: Oracle Privacy Security Auditing. Rampant (2003)
7. Browder, K., Davidson, M.A.: The virtual private database in Oracle9iR2. Technical report, Oracle Corporation (2002), `http://www.oracle.com/technology/deploy/security/oracle9ir2/pdf/VPD9ir2twp.pdf`
8. Friedman-Hill, E.: (Jess rule engine), `http://herzberg.ca.sandia.gov/jess/`
9. Ullman, J.D.: Principles of Database and Knowledge-Base Systems, Computer Science Press, vol. I (1988)
10. Purevjii, B., Aritsugi, M., Imai, S., Kanamori, Y., Pancake, C.M.: Protecting personal data with various granularities: A logic-based access control approach. In: Hao, Y., Liu, J., Wang, Y.-P., Cheung, Y.-m., Yin, H., Jiao, L., Ma, J., Jiao, Y.-C. (eds.) CIS 2005. LNCS (LNAI), vol. 3802, pp. 548–553. Springer, Heidelberg (2005)
11. Purevjii, B., Aritsugi, M., Kanamori, Y., Pancake, C.M.: A unified framework for evaluating data-dependent access control systems. IPSJ Digital Courier 2, 14–24 (2006)
12. Forgy, C.: Rete: A fast algorithm for the many patterns/many objects match problem. Artif. Intell. 19, 17–37 (1982)

Applying Privacy Preserving Count Aggregate Queries to *k*-Classification

Hidehisa Takamizawa[1] and Masayoshi Aritsugi[2]

[1] Department of Computer Science, Graduate School of Engineering,
Gunma University
1-5-1 Tenjin-cho, Kiryu, Gunma, 376-8515 Japan
[2] Computer Science and Electrical Engineering,
Graduate School of Science and Technology, Kumamoto University
2-39-1 Kurokami, Kumamoto 860-8555, Japan
takamiza@dbms.cs.gunma-u.ac.jp, aritsugi@cs.kumamoto-u.ac.jp

Abstract. It is important to process data effectively while preserving privacy of personal information. In this paper, we propose a technique to reconstruct results of count aggregate queries from a perturbed table for building a decision tree whose target attribute contains more than two classes. Using the conventional technique, we must reconstruct the results of target values from those of each value calculated independently in such the case. In this paper, we borrow and extend the conventional technique to reconstruct the results of target values at once. We also report some experimental results showing that our proposal can reduce reconstruction errors compared to the conventional technique in cases where perturbation ratio is high.

Keywords: Personal Information, Privacy, Data Mining.

1 Introduction

Privacy preserving data mining techniques have been studied [1,2]. Though data mining is used in many areas to extract knowledge from large amounts of data, there are risks that personal information is revealed by data mining results[3]. The *retention replacement perturbation*[2] is known as a perturbation algorithm where elements are retained or replaced randomly for preserving privacy of personal information. The goal of this work is to reconstruct results of count aggregate queries from a perturbed table for building a decision tree whose target attribute contains more than two classes on a model which consists of a server and n clients connected to the server. In this model, each client has a record which consists of $m+1$ attributes $Attr_0, \ldots, Attr_m$. $Attr_0$ is a *target attribute*[4] in classification, and also a categorical attribute. The other attributes $Attr_1, \ldots, Attr_m$ are *conditional attributes*[4]. While no attribute contains any identifier of a person, a combination of attributes may lead to breach of personal information. Each client thus sends to the server a record perturbed with *retention replacement perturbation*, which is described later.

B. Apolloni et al. (Eds.): KES 2007/ WIRN 2007, Part II, LNAI 4693, pp. 761–768, 2007.
© Springer-Verlag Berlin Heidelberg 2007

The server constructs a perturbed table T' which consists of n records received from all clients. Reconstruction techniques for constructing a decision tree whose target attribute contains two classes are proposed in [2]. In this paper, we borrow and extend the reconstruction technique proposed in [2] to reconstruct the results of target values with more than two classes at once.

2 Preparation

2.1 Related Work

Privacy preserving techniques for statistical queries have been studied in the statistical database field [5,6]. However, the statistical database model is not *client-server model*, and thus different from the model in our work.

The privacy preserving model in our work is the same as [1,2]. In [1], each client adds random values to their own records as noise. The server reconstructs a distribution of each attribute in the original table from the perturbed table. However, using the technique proposed in [1], reconstruction processes must be executed on each attribute independently. As mentioned in Section 1, reconstruction techniques for building a decision tree whose target attribute contains two classes are proposed in [2]. We borrow one of the perturbation and the reconstruction techniques proposed in [2].

2.2 Privacy Preserving OLAP

Retention replacement perturbation. *Retention replacement perturbation[2]* replaces values in a record to other values and retains original values at constant probabilities configured for each attribute. The probability, rp_j, is called *retention probability* ($0 \leq rp_j \leq 1$) of $Attr_j$. The *retention replacement perturbation* can be represented as follows:

$$t'_{ij} = \begin{cases} t_{ij} & \text{(with probability } rp_j\text{)}, \\ \text{replace other value} & \text{(with probability } 1 - rp_j\text{)}, \end{cases} \tag{1}$$

where t_{ij} and t'_{ij} denote j^{th} attribute value of ith record of an original table T and a perturbed table T', respectively.

Iterative Bayesian technique. *Iterative Bayesian technique* proposed in [2] reconstructs results of count aggregate queries from a table T' perturbed by the *retention replacement perturbation*. Suppose that we reconstruct a result of count aggregate query $COUNT(P_1 \wedge P_2 \wedge \ldots \wedge P_k)$, where P_j denotes a predicate for j^{th} attribute on table T.

Vectors $\boldsymbol{x} = (x_0, x_1, ..., x_t)$ and $\boldsymbol{y} = (y_0, y_1, ..., y_t)$, which represent results of count aggregate queries on original table T and a perturbed table T', respectively, are represented as follows:

$$\begin{cases} x_i = COUNT(\wedge_{r=1}^{k} Q(r, bit(i, r))) & \text{in } T \text{ for } 0 \leq i \leq t, \\ y_i = COUNT(\wedge_{r=1}^{k} Q(r, bit(i, r))) & \text{in } T' \text{ for } 0 \leq i \leq t, \end{cases} \tag{2}$$

where $t = 2^k - 1$, and $Q(r, i)$ is a predicate for r^{th} attribute represented as $Q(r, i) = \neg P_r$ if $i = 0$, and $Q(r, i) = P_r$ if $i = 1$. Function $bit(i, r)$ is r^{th} bit from the left in the binary representation of the integer i using k bits. If a record satisfies the predicate $\wedge_{r=1}^{k} Q(r, bit(i, r))$, i denotes a status of the record.

A transition probability a_{pq}, which represents a probability that a record perturbed from status p in T to status q in T', can be computed as follows:

$$a_{pq} = \prod_{r=1}^{k} \left\{ (1 - rp_r) \times R_{r,bit(q,r)} + rp_r \times \delta_{(bit(p,r),bit(q,r))} \right\}. \tag{3}$$

In (3), $R_{(r,i)}$ is a function represented as $R_{(r,i)} = 1 - b_r$ if $i = 0$, and $R_{(r,i)} = b_r$ if $i = 1$. And b_r is as follows:

$$b_r = \begin{cases} (high_r - low_r)/(max_r - min_r) & (Attr_r \text{ is numerical}), \\ 1/c_r & (Attr_r \text{ is categorical}), \end{cases} \tag{4}$$

where $high_r$ and low_r denote a maximum value and a minimum value of the predicate P_r for attribute $Attr_r$, respectively, and c_r denotes the number of values in attribute $Attr_r$. $\delta_{(i,j)}$ is represented as $\delta_{(i,j)} = 1$ if $i = j$, and $\delta_{(i,j)} = 0$ if $i \neq j$.

Let $U_1, U_2, ..., U_n$ and $V_1, V_2, ..., V_n$ denote status of each record in table T and T' respectively, both of which consists of n records. Thus, $P(U_i = p) = x_p/n$ and $P(V_i = q) = y_q/n$ are comprised for $0 \leq p, q \leq 2^k - 1$ and $1 \leq i \leq n$. And $P(U_i = p)$ can be also represented as follows:

$$P(U_i = p) = \sum_{q=0}^{t} P(V_i = q) P(U_i = p | V_i = q). \tag{5}$$

Using Bayesian theorem, $P(U_i = p | V_i = q)$ can be represented as follows:

$$\begin{aligned} P(U_i = p | V_i = q) &= \frac{P(V_i = q | U_i = p) P(U_i = p)}{P(V_i = q)} \\ &= \frac{P(V_i = q | U_i = p) P(U_i = p)}{\sum_{r=0}^{t} P(V_i = q | U_i = r) P(U_i = r)} = \frac{a_{pq} x_p}{\sum_{r=0}^{t} a_{rq} x_r}. \end{aligned} \tag{6}$$

We can get the following formula by assigning (6) to (5):

$$x_p^{T+1} = \sum_{q=0}^{t} y_q \frac{a_{pq} x_p^T}{\sum_{r=0}^{t} a_{rq} x_r^T}, \tag{7}$$

where $\boldsymbol{x}^T$ and $\boldsymbol{x}^{T+1}$ denote T^{th} and $T + 1^{th}$ iteration, respectively. We initiate $\boldsymbol{x}^0 = \boldsymbol{y}$ and iterate until the difference of continued $\boldsymbol{x}$ becomes small.

Note that not only *iterative Bayesian technique* but also another reconstruction technique, called *matrix inversion technique*, is proposed in [2]. However, reconstruction errors of *matrix inversion technique* are very large when the number of attributes increase. We thus refer to only *iterative Bayesian technique* in this work.

3 Building a Decision Tree from a Perturbed Table

Generally, objective values such as mutual informationCgini indexCand χ^2 are used for determination of optimal split points for a decision tree [7,8,4]. All objective values can be represented as a function whose argument is a number of records before splitting and after splitting. Thus, the results of count aggregate queries can be used as fixed numbers for determination of optimal splitting.

Here, we explain a procedure to reconstruct the results of count aggregate queries from perturbed table for building a decision tree whose target attribute contains more than two classes using the reconstruction technique explained in Section 2.2. For example, we suppose that a set whose target attribute includes three classes, C_0, C_1, C_2, is split by a predicate P_1. Table 1 shows count aggregate queries for 3-classification for splitting on predicates P_1 and $\neg P_1$ (left), and reconstructable count aggregate queries by the conventional technique at once (right). As shown in Table 1, using conventional technique, we need to reconstruct the results of count aggregate queries for all classes, C_0, C_1, C_2 independently to get the results of count aggregate queries.

Table 1. Count aggregate queries for 3-classification for splitting on P_1(left), and reconstructable count aggregate queries by the conventional technique using predicate P_1 (right)

Target attribute	Count queries
C_0	$COUNT(P_0^0 \wedge \neg P_1)$
	$COUNT(P_0^0 \wedge P_1)$
C_1	$COUNT(P_0^1 \wedge \neg P_1)$
	$COUNT(P_0^1 \wedge P_1)$
C_2	$COUNT(P_0^2 \wedge \neg P_1)$
	$COUNT(P_0^2 \wedge P_1)$

Target attribute	The results of count queries	$\boldsymbol{x}$
C_i	$COUNT(\neg P_0^i \wedge \neg P_1)$	x_{00}
	$COUNT(\neg P_0^i \wedge P_1)$	x_{01}
	$COUNT(P_0^i \wedge \neg P_1)$	x_{10}
	$COUNT(P_0^i \wedge P_1)$	x_{11}

4 Our Proposal

As mentioned above, when reconstructing the results of count aggregate queries from a perturbed table for building a decision tree whose target attribute contains more than two classes, we need to reconstruct all classes included in the target attributes independently with the conventional technique. Therefore, we propose a technique to reconstruct the results of count aggregate queries from a perturbed table for all classes included in the target attribute at once based on the reconstruction technique explained in Section 2.2.

First, we redefine the status of a record defined in Section 2.2. In the conventional technique, a status for the target attribute is represented as a bit; 1 and 0 mean that the value of the target attribute in a record satisfies the predicate P_0 and does not satisfy, respectively. When the target attribute contains two classes, the status can be represented with 1 bit, that is, P_0 means one class,

and $\neg P_0$ means the other one class. However, when the target attribute contains more than two classes, the status of the target attribute cannot be represented with 1 bit. Therefore, we represent the status of the target attribute as *bit-string*.

Suppose that c denotes the number of classes in the target attribute, then the number of bits for representing the status of the target attribute can be calculated as $\lceil \log_2 c \rceil$. Therefore, the status of a record can be calculated by connecting the $\lceil \log_2 c \rceil$ bits representing the status of the target attribute and m bits representing the conditional attributes.

Next, we redefine vectors $\boldsymbol{x} = (x_0, x_1, ..., x_t)$ and $\boldsymbol{y} = (y_0, y_1, ..., y_t)$ as follows:

$$\begin{cases} x_i = COUNT(P_0^{left(i,\lceil \log_2 c \rceil)} \wedge_{r=1}^{k} Q(r, bit(i,r))) & \text{in } T \text{ for } 0 \le i \le t', \\ y_i = COUNT(P_0^{left(i,\lceil \log_2 c \rceil)} \wedge_{r=1}^{k} Q(r, bit(i,r))) & \text{in } T' \text{ for } 0 \le i \le t', \end{cases} \tag{8}$$

where $t' = c \times 2^k - 1$, and $left(i, r)$ denotes r bits from the left in the binary representation of integer i. $left(i, \lceil \log_2 c \rceil)$ means a status of the target attribute, $P_0^{left(i,\lceil \log_2 c \rceil)}$ is thus a predicate to select a target value whose status is $left(i, \lceil \log_2 c \rceil)$. Functions $Q(r, i)$ and $bit(i, j)$ are the same as in Section 2.2.

For example, suppose that a target attribute contains three classes, and we want to reconstruct the results of count aggregate queries for 3-classification splitting by predicates P_1 and $\neg P_1$. The results of count aggregate queries reconstructed using the proposed reconstruction technique are shown in Table 2.

Table 2. Reconstructable count aggregate queries by the proposed technique

Count query	$\boldsymbol{x}$	Target attribute
$COUNT(P_0^0 \wedge \neg P_1)$	x_{000}	C_0
$COUNT(P_0^0 \wedge P_1)$	x_{001}	
$COUNT(P_0^1 \wedge \neg P_1)$	x_{010}	C_1
$COUNT(P_0^1 \wedge P_1)$	x_{011}	
$COUNT(P_0^2 \wedge \neg P_1)$	x_{100}	C_2
$COUNT(P_0^2 \wedge P_1)$	x_{101}	

In this example, the status of the target attribute can be represented by $left(i, \lceil \log_2 3 \rceil) = 2$ bits. Thus, 2 bits from the left in i represents the status of the target attribute and the rest 1 bit represents the status of the conditional attribute.

Then, a transition probability a'_{pq} is represented as follows:

$$\begin{aligned} a'_{pq} = & \left\{ (1 - rp_0) \times b_0 + rp_0 \times \delta_{(left(p,\lceil \log_2 c \rceil), left(q,\lceil \log_2 c \rceil))} \right\} \\ & \times \prod_{r=1}^{k} \left\{ (1 - rp_r) \times R_{r,bit(q,r)} + rp_r \times \delta_{(bit(p,r),bit(q,r))} \right\}. \end{aligned} \tag{9}$$

On the conditional attribute, ranges of predicates P_r and $\neg P_r$ are not exactly the same value. So the probability $1 - b_r$ in which the status of the conditional attribute in a record is changed to other status by perturbation, and the

probability b_r in which the status of the conditional attribute is not changed by perturbation, are not also exactly the same. However, on the target attribute, ranges of P_0^i are all the same value. So the probability in which the status of the target attribute in a record is changed to other status by perturbation, and the probability in which the status of the target attribute in a record is not changed by perturbation, are uniformly $b_0 = 1/c$. Although the change of a status of a conditional attribute can be distinguished by comparison of 1 bit, it needs comparison of $left(i, \lceil \log_2 c \rceil)$ bits to distinguish the change of a status of a target attribute.

Our reconstruction technique is represented as follows:

$$x_p^{T+1} = \sum_{q=0}^{t'} y_q \frac{a'_{pq} x_p^T}{\sum_{r=0}^{t'} a'_{rq} x_r^T}. \tag{10}$$

It is a formula obtained by replacing a_{pq} in the formula (7) to a'_{pq}. Finally we can reconstruct the results of count aggregate queries on table T by calculating $\boldsymbol{x}$ iteratively as well as in Section 2.2.

5 Experimental Results

In this section, we show experimental results to evaluate the effectiveness of our proposal compared to the conventional technique. We define reconstruction errors as values calculated by dividing *l1-norm* of the correct results and reconstructed results by the number of records. We also show reconstruction errors for the results of count aggregate queries on a perturbed table to clarify the effect of reconstruction techniques. We calculate estimated values of $\boldsymbol{x}$ iteratively until the difference between continued values becomes less than 10^{-3}. The reconstruction errors reported in the following are the average values of ten runs.

We configured that all data in each attribute were accordance with Zipf's distribution[9], the number of classes in target attribute was varied with experiments in the range of 2 - 10, and the number of classes in conditional attributes were 1,000, in which predicates of count aggregate queries were ranked in the top 400. Tables consisted of $10,000$ or $100,000$ records, and for all attributes the same retention probability was used. For ease of explanation, let c be the number of values in the target attribute, m be the number of conditional attributes, n be the number of records, and rp be a retention probability.

Figure 1 shows the results for $c = 5$(left) and $c = 10$(right), $m = 1$, $n = 10^4$ with varying rp. In Figure 1, the vertical axis denotes reconstruction errors, l1-norm, for results of count aggregate queries on a perturbed table (randomized), reconstruction results using iterative Bayesian technique (IBT), and reconstruction results using proposed technique (EIBT). The horizontal axis denotes a retention probability (rp). When the retention probability is small, the reconstruction errors of our technique are smaller than those of the conventional technique.

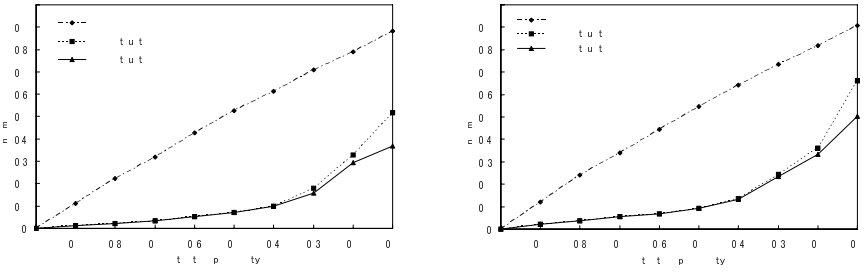


Fig. 1. Errors for $c = 5$(left) and $c = 10$(right), $m = 1$, $n = 10^4$ with varying rp

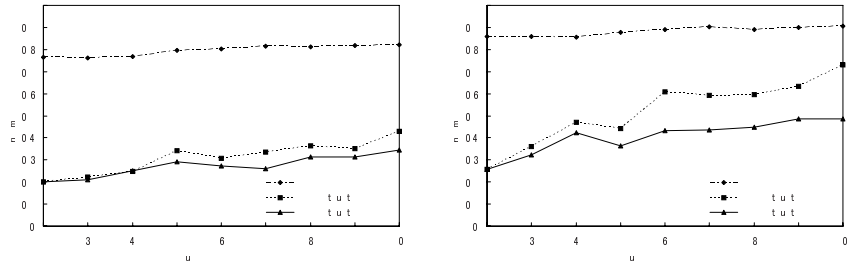


Fig. 2. Errors for $m = 1$, $n = 10^4$, $rp = 0.2$(left) and $rp = 0.1$(right) with varying c

Figure 2 shows the results with varying c in 2 - 10 for $rp = 0.2$(left) and $rp = 0.1$(right). When a target attribute contains two classes, reconstruction algorithms of the conventional and our techniques are the same. Thus reconstruction errors of two techniques are the same value. In other cases ($3 \leq c \leq 10$), reconstruction errors of our proposal are less than those of the conventional technique.

Figure 3 shows the reconstruction errors in the same environment of Figure 1 except for that the number of records is $100,000$. Reconstruction errors of

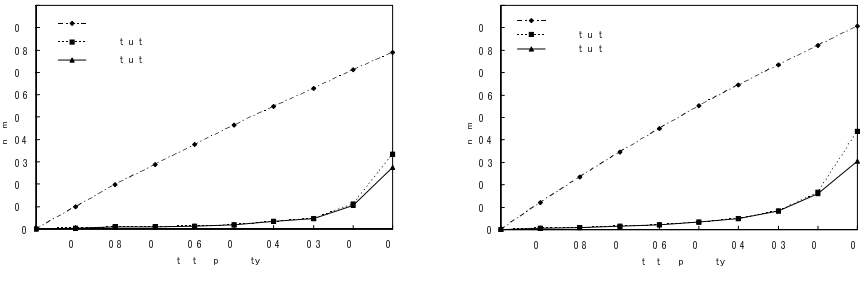


Fig. 3. Errors for $c = 5$(left) and $c = 10$(right), $m = 1$, $n = 10^5$ with varying rp

Figure 3 are smaller than those of Figure 1. The difference between errors of the conventional technique and our technique is also smaller. This is because the influence of independent reconstruction for reconstruction errors is small based on the growth of the number of records.

6 Conclusion

In this paper, we proposed a technique to reconstruct the results of count aggregate queries from a perturbed table for building a decision tree whose target attribute contains more than two classes. We reported some experimental results showing that our proposal can reduce reconstruction errors compared to the conventional technique in cases where perturbation ratio is high.

Acknowledgments

This work was supported in part by the Japan Society for the Promotion of Science, Grant-in-Aid for Scientific Research (C) 18500073.

References

1. Agrawal, R., Srikant, R.: Privacy-Preserving Data Mining. In: SIGMOD Conference, pp. 439–450. ACM, New York (2000)
2. Agrawal, R., Srikant, R., Thomas, D.: Privacy Preserving OLAP. In: SIGMOD Conference, pp. 251–262. ACM, New York (2005)
3. Broder, A.J.: Data Mining, the Internet, and Privacy. In: Masand, B., Spiliopoulou, M. (eds.) Web Usage Analysis and User Profiling. LNCS (LNAI), vol. 1836, pp. 56–73. Springer, Heidelberg (2000)
4. Morimoto, Y., Fukuda, T., Matsuzawa, H., Tokuyama, T., Yoda, K.: Algorithms for Mining Association Rules for Binary Segmentations of Huge Categorical Databases. In: VLDB, pp. 380–391 (1998)
5. Adam, N.R., Wortmann, J.C.: Security-Control Methods for Statistical Databases: A Comparative Study. ACM Comput. Surv. 21(4), 515–556 (1989)
6. Denning, D.E.: Cryptography and Data Security. Addison-Wesley, Reading (1982)
7. Breiman, L., Friedman, J.H., Olshen, R.A., Stone, C.J.: Classification and Regression Trees. Wadsworth (1984)
8. Han, J., Kamber, M.: Data Mining: Concepts and Techniques. Morgan Kaufmann, San Francisco (2000)
9. Zipf, G.K.: Human Behaviour and the Principle of Least Effort: an Introduction to Human Ecology. Addison-Wesley, Reading (1949)

Event-Centralized Management of Geographic Information Collected from Blog

Shoko Yasumura[1], Masakazu Ikezaki[1], Toyohide Watanabe[1], and Taketoshi Ushiama[2]

[1] Department of Systems and Social Informatics,
Graduate School of Information Science, Nagoya University,
Furo-cho, Chikusa-ku, Nagoya, 464-8603, Japan
{yasumura,mikezaki,watanabe}@watanabe.ss.is.nagoya-u.ac.jp
[2] Faculty of Design, Kyushu University,
9-1 Shiobaru 4-chome, Minami-ku, Fukuoka, 815-0032, Japan
ushiama@design.kyushu-u.ac.jp

Abstract. In this paper, we propose a method for automatically extracting events from information networks. Personal experiences can be handled by associating them with time and space information from a viewpoint of managing personal media data. We define an event as an occurrence experienced in the real world, and design a geographic information system that can handle them. We assume that when an event occurs in the real world, blog entries about it are directly registered in the Web. We collect these blog entries and extract appropriate events from them. Useful events are described with time and place. However, problems are caused by two factors. One, identical names are observed in different regions. The other is nonassociated place names are extracted with the event. In order to handle the former problem, a place name dictionary is prepared based on geographic inclusive relationships. In order to handle the latter problem, events are identified by type, time and place.

Keywords: Blog, event.

1 Introduction

Recently, services that associate various media data with maps have appeared. However, such personal media data management is insufficient. An event, which is commonly registered in plural media contents, must be associated with the media data and managed by its contexts [1]. We define an event as an occurrence that can only be experienced for a limited period and design a geographic information system that can handle an event. By handling events on a geographic information system and representing the relation between them, event-oriented media data handling will be realized [2].

For constructing an event-oriented geographic information system, automatically extracting events from Web pages is necessary. When an event occurs in the real world, daily articles (blog entries) are created on the Web by people who experienced that event. In this research, blog entries about events are collected and an event is extracted from them. We define attributes of the event as type, time and a place where it occurred.

B. Apolloni et al. (Eds.): KES 2007/ WIRN 2007, Part II, LNAI 4693, pp. 769–776, 2007.
© Springer-Verlag Berlin Heidelberg 2007

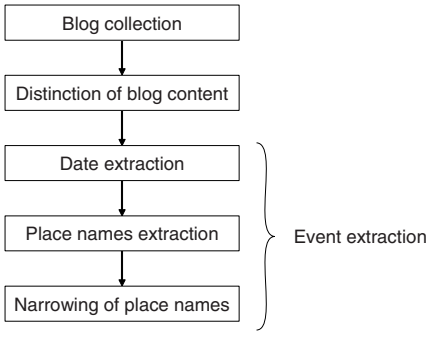


Fig. 1. Procedures

However, extracting the type of event is difficult. Therefore, users name the event by type in the system. Our objective is event extraction from blog entries.

In Section 2, we explain the overview of the entire process of the system. In Section 3, we explain the proposed event extraction method in detail. In Section 4, we discuss the experimental results of event extraction based on the proposed method. In Section 5, we conclude this research and describe future work.

2 Overview

In this research, we collect blog entries that describe events and then extract them form blog entries using the metadata described in the HTML source. The contents of blog entries vary from personal diaries to public messages. In order to collect blog entries about events, the contents of the blog entries must be distinguished.

An event is described with times and place names in a blog entry. The event is identified by the time and the place at which it occurred. We consider event extraction the extraction of a time and place. We assume that the time is almost equal to the last update of the blog entry. The place is a place name character string. However, problems exist when extracting a place of event from a blog entry. First, one place name described in the blog entry might correspond to two or more regions. We handle this problem by place name registration based on geographical inclusive relationships between place names. Second, the place names in the blog entry might not be related to the event. We consider that a place of event is described among blog entries about identical events. Consequently, a place name described in many blog entries is extracted as the place of event. Hence, as more blog entries about the same event are collected, the more easily place names can be narrowed to a place of event. Fig. 1 is a process flowchart.

3 Method

3.1 Blog Collection

Blog entries are collected by Google. A given type name of an event and “blog” are used as search words. Since a type name of the event is the search words, it is possible

to search blog entries by two or more words that declare a type of event. For example, when collecting blog entries about baseball games, a type name of event is "baseball+game".

Then, the body texts of the collected blog entries are extracted, because, in general, information unrelated to the event exists in commercial messages, the blog author's profile, and so on. If metadata exist in the HTML source of the blog entry, the body text can be extracted by using it. If no metadata exist in the HTML source, the HTML tags are deleted from the HTML source. Then, 56 words most commonly used in the blog entry are deleted. They are collected during our research. Since they greatly affect the distinction of the content of the blog entry, they are deleted.

3.2 Distinction of Blog Content

For distinguishing blog entries, event property is created by each type of event. Event property is a tuple of a set of words that exist in the body texts of blog entries, their classes, and their weights. In order to create event property, first, blog entries are collected by Google by type name of the event and "blog" as search words. Then, the morphemes of the body texts of all the collected blog entries are analyzed by ChaSen [3], a morphological analysis tool. Nouns, adjectives, and verbs are collected. Then, words are sorted by their frequencies in use, and forehand 1000 words are weighted by tfidf [4] given by formula (1):

$$tfidf_i = \frac{freq_i}{\sum_{k=1}^{1000} freq_k} \times \ln \frac{blognum_{\max}}{blognum_i} \tag{1}$$

$freq_i$ is the frequency in use of word x_i. $\sum_{k=1}^{1000} freq_k$ is the summation of the use frequencies of words. $blognum_i$ is the number of blog entries in which word x_i is used. $blognum_{\max}$ is the maximum value of the numbers of blog entries.

Blog entries are distinguished by event property. We assume that blog entries collected by type of event as search words usually describe personal experiences of the event. Frequencies in use of words used in these are bigger than frequencies in use of words used in blog entries that do not describe the event. Consequently, their weights are also large. Therefore, when the summation of the weights of words used in the body text of a blog entry is bigger than a threshold, this blog entry can be distinguished as an article about personal experiences of the event. The threshold was determined by a pilot study.

3.3 Date Extraction

The last update of the blog entry is extracted from its metadata. When there are no metadata, the date is extracted from the place of the description that seems to be the date. Here, the day is not extracted. Blog entries are created when an event occurs. However, we cannot predict when blog entries will be created. Consequently, in order

Table 1. Example of target description for date

2000N11	2007.12.30	2007/01/01
2009_1_1	20070101	20071230

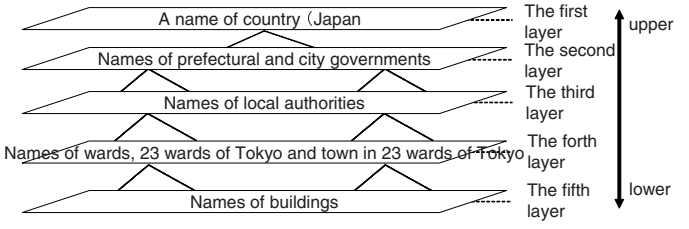


Fig. 2. Layered structure of place names

to disregard the interval of days between the last update of the blog entry and the time of event, the day is not extracted.

We explain the details of the date extraction method when no metadata exist in the HTML source. First, the part of the character strings that represent the years from 2000 to 2009, “200”, is sought from the blog entry. 1900s are not extracted, because blogs were not so common before 2000. When integer X from 0 to 9 follows “200”, the year is “200X”. Second, one of the seven separating charactersi’_’, ’Q’, ’N’, ’:’, ’F’, ’.’jis sought from the next “200X”. If there is a separating character, the following one or two numbers are extracted as the month. If there is no separating character, one from “01”, “02”, c, “12” is sought from the next “200X”. When the description that seems to be the day follows one, it is extracted as the month. Table 1 shows a target date.

3.4 Place Name Extraction

Place names are extracted from the blog entries to associate them with extracted place names. Previously, target place names were manually registered as layered structure based on their geographic inclusive relationships. Fig. 2 is a conceptual diagram of the layered structure of place names.

The geographic inclusive relationship of place names is a relationship in which the region corresponding to the place name in a lower hierarchy is included in the region corresponding to the place name in a higher hierarchy. For example, Nagoya International Exhibition Hall is included in Minato ward, Minato ward is included in Nagoya, Nagoya is included in Aichi prefecture, and Aichi prefecture is included in Japan. Fig. 3 is an example of the geographic inclusive relationship of Nagoya International Exhibition Hall. The layered structure is constructed as follows. The first layer is constructed by a country name. The second layer is constructed by prefecture and city government names. The third layer is constructed by local authority names. The fourth layer is constructed by ward names including Tokyo’s 23 wards and the town names in those 23 wards. The fifth layer is constructed by buildings names.



Fig. 3. Geographic inclusive relation of Nagoya International Exhibition Hall

Place names are sequentially extracted from the body text of the blog entry from place names in the fifth hierarchy to place names in the first hierarchy. By using a layered structure of place names based on their geographic inclusive relationships, the blog entry is also associated with place names in higher hierarchies whose region includes the region corresponding to the extracted place names. Moreover, a problem caused by identical place names that correspond to different regions is solved. When such a place name is extracted, extractions of all the place names in one upper hierarchy are examined. If one of these is extracted, it is extracted as a place name of the lower hierarchy of the extracted one. If some or none of these are extracted, it is ignored.

Furthermore, a problem where one place name might include another place name is also solved. For example, the string "s" (Metropolis of Tokyo) includes the same Chinese characters for the string "s" (Kyoto). Consequently, when an extracted place name is included in another extracted place name, the included place name is deleted from the body text of the blog entry. Then, the included place name is sought from the body text of the blog entry. If not extracted, it is ignored.

3.5 Narrowing of Place Names

Extracted place names are narrowed to one place of event by the date and place names. First, the number of blog entries that have identical extracted date is found by place name. Second, threshold *T* is found by formula (2). *i* is the identifier of the blog entry. *n* is the number of extracted place names. *mappingnum* is the number of blog entries:

$$T_i = \frac{3}{10} \times \sum_{k=1}^{n} mappingnum_k \tag{2}$$

Place names whose number of related blog entries is bigger than *T* are extracted as a place of event. We consider that a place of event is contained in the top three place names ordered by the number of blog entries. Therefore, the coefficient is defined as 3/10. Moreover, when the total number of blog entries related to place names is less than *T*, place names that have the maximum number of blog entries are extracted as a place of event. If extraction of the date is fails, place names are narrowed without considering the date, and the number of blog entries is found by place names. Subsequently, place names are narrowed down in the same way.

Table 2. Experimental results of content distinction of the blog entries

	Success	Failure
concert	65	35
baseball+game	48	52

4 Experiments

4.1 Outline

For event extraction experiments based on the proposed method, we implemented a prototype system by Java. We previously registered 4395 place names. Mostly the names of buildings are names of multipurpose halls, baseball stadiums, and so on. Types of events are "concert" and "baseball+game". We selected randomly 100 of collected and handled about 600 blog entries for each event and checked the following items by visually.

- Distinction of blog content: Is the content of the blog entry created by a person and related to the event?
- Date extraction: Is the date of the last update time of the blog entry correctly extracted?
- Place name extraction: Is a place of event extracted and narrowed?

4.2 Experimental Results of Distinction of Blog Content

Table 2 shows the experimental results of the content distinction of blog entries. The success row shows the number of blog entries created by a person and related to the event. The failure row shows the number of blog entries that were not created by a person or related to the event.

These results show that more than half of the checked blog entries were created by a person and related to the event. However, the success numbers of distinction are different for each type of event. There are also differences in the accuracy of distinction because contents that constitute the majority of the contents of collected blog entries are different for each type of event. For example, for "baseball+game", many blog entries on the draft lottery were collected because this experiment was implemented in off-season. A definition of the threshold for the distinction of blog content in each type of event is necessary. Furthermore, appropriate search words must be used to collect more blog entries about personal experiences of the event.

4.3 Experimental Results of Date Extraction

Table 3 shows the experimental results of date extraction. The success row illustrates the number of blog entries to which the date of the last update was accurately extracted. The failure row shows the number of blog entries to which the date of the last update was not accurately extracted.

From this result, we see that the date when the blog entry was created is extracted with high accuracy. Many blog entries also had metadata. In addition, the proposal date extraction technique is shown to be appropriate.

Table 3. Experimental result of date extraction

	Success	Failure
concert	85	15
baseball+game	90	10

Table 4. Experimental results of place name extraction and narrowing

	Before/After	Success	Unknown	Failure
concert	Before	39	19	42
	After	39	5	56
baseball+game	Before	5	27	68
	After	6	6	88

4.4 Experimental Results of Place Name Extraction and Narrowing

Table 4 shows the experimental results of place name extraction and narrowing. The before/after row shows the before and after of narrowing place names after place name extraction. The success row shows the number of blog entries in which a place of event was extracted. The unknown row shows the number of blog entries in which a place of event and place names unrelated to the event were extracted. The failure row shows the number of blog entries that did not extract a place of event.

From this result of place name extraction, we see that the numbers of success are greatly different for each type of event. A place of event extraction failed in two cases. First, a place of event was not included in the blog entry. Second, place names that were found in the blog entry were not registered. Moreover, place names included in the character strings described in the blog entry were extracted. For successful place name extraction, we must register more place names. In addition, in order to eliminate character strings that include place names, target words should be collected for each type of event. For example, when the type of event is “concert”, the target words should be band names and so on. When the type of event is “baseball+game”, the target words should be baseball terms, player names, and so on.

From the results of narrowing place names, a place of event is approximately deleted due to a lack of blog entries. Furthermore, date extraction has some problems. For example, when the event occurred at the end of the month, the blog entry is often created the next month. In such cases, an extracted date does not equal the date of an event. For applying the proposed method properly, the time of an event must be extracted more strictly.

5 Summary

We focused on blog entries as a media for expressing personal experiences and proposed a method for extracting events. We performed event extraction experiments based on the

proposed method. As a result, some blog entries about events were collected. The year and the month of the last update of the blog entry are extracted with high accuracy as the time of events. However, a place of event is not adequately extracted.

Our future work includes the following. First, more blog entries must be collected. Second, events should be extracted more strictly. By collecting blog entries about identical events, methods for improving the accuracy of event extraction are available. However, program execution time becomes huge. In the experiment, about 600 blog entries were collected that handled each type of event. Calculations required about half a day. Third, more place names should be registered. Fourth, in order to handle events in a geographic information system, place names should be associated with coordinates on a map. Some Web services provide map coordinates as input of place names including building names [5]. We must consider how to display events on maps.

Acknowledgement

Part of this research was executed by a research grant from the Daikou financial group.

References

1. Ushiama, T., Watanabe, T.: An Automatic Indexing Approach for Private Photo Searching Based on E-mail Archive. In: Gabrys, B., Howlett, R.J., Jain, L.C. (eds.) KES 2006. LNCS (LNAI), vol. 4252, pp. 1111–1118. Springer, Heidelberg (2006)
2. Ikezaki, M., Mukai, N.: Toyohide Watanabe, Event Handling Mechanism for Retrieving Spatio-temporal Changes at Various Detailed Level. In: Ali, M., Esposito, F. (eds.) IEA/AIE 2005. LNCS (LNAI), vol. 3533, pp. 353–356. Springer, Heidelberg (2005)
3. ChaSen, http://chasen.naist.jp/hiki/ChaSen/
4. Sparck, J.K.: A statistical interpretation of term specificity and its application in retrieval. Journal of Documentation 28(1), 11–21 (1972)
5. geocoording, http://map.fkoji.com/geo/

Message Flow: Design and Architecture of a Message Flow Management System

Hideaki Ito and Teruo Fukumura

School of Information Science and Technology, Chukyo University
101 Tokodachi, Kaizu-cho, Toyota, Aichi, 470-0393 Japan
{itoh, fukumura}@sist.chukyo-u.ac.jp

Abstract. This paper describes an overview of design and architecture of a message flow management system which manages flows of emails, called Message Flow. Emails are utilized to achieve several types of works. However, to manage email exchanges becomes a heavy load to achieve such works. For reducing such a load it is required to develop the mechanism which manages sending and receiving emails. Then, it is useful that a system provides the mechanisms for specifying a route of email delivery and for monitoring email exchanges. This route is defined as a flow of emails, called a message flow. Message Flow manages them. So, loads with respect to monitoring email exchanges among clients are reduced.

Keywords: emails, message flows, message flow management.

1 Introduction

In current, several types of activities are achieved using emails such as collaborative works, task management, etc.[3,10,11]. Managing received emails and monitoring email delivery are done carefully by users since originally an individual email is exchanged independently. In spite of this feature, many users utilize them for supporting task management and/or collaborative works that require to communicate with some users[10]. To achieve such works, the exchanges of emails are done, and their monitoring and management become a heavy load.

As a system which supports task management, a workflow management system is developed[4,7]. The workflow specifies tasks and routes of the tasks. A person plays a role of a task, if the person is the component of the workflow.

For reducing burdens for monitoring deliveries of emails, it is required that a system supports them. Users specify routes to deliver emails, and the system monitors their deliveries and replies to them. In order to meet these requests, we have been developing a system that manages message exchanges, which is named Message Flow. A user defines clients who receive emails and a route for delivering emails. And the user describes commands to deliver the emails to successive receivers. The sequence of clients and commands becomes a flow, which is called a message flow. The system manages a collection of the flows, and monitors their activities. Emails are used for notifying that a client receives a message flow, messages are delivered together with the message flow.

B. Apolloni et al. (Eds.): KES 2007/ WIRN 2007, Part II, LNAI 4693, pp. 777–784, 2007.
© Springer-Verlag Berlin Heidelberg 2007

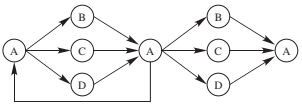


Fig. 1. An example of a message flow

To achieve collaborative works and task management, several types of systems are developed based on emails. Active Mail is developed for a groupware application[5]. But this system does not maintain email exchanges. Some systems try to analyze contents of emails and to find relationships among emails. Taskmaster tries to manage pending tasks and reminding them[1]. TaskView provides mechanisms for showing task date, task overview, and pending tasks[6]. ContactMap[9] provides mechanisms for accessing emails, exchanges with selected contacts, for maintaining reminders, etc. On the other hand, many workflow management systems are developed in business process management[8]. Many types of control flows are specified[2]. Such workflow systems address on automatic execution of tasks. Message Flow executes flows of emails and monitors email exchanges based on defined flows.

This paper is organized as follows. Section 2 describes design of Message Flow. Specification of message flows is presented in Sec. 3. Architectural overview of the system is presented in Sec. 4. Section 5 describes some concluding remarks.

2 Design of Message Flow

Objects to develop Message Flow are that the system reduces a load with respect to operations for email exchange, that it monitors deliveries of emails, and that it manages message flows.

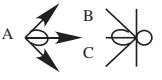
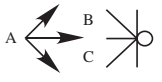
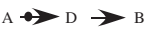
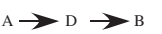
On the other hand, policies for designing this system are as follows;

- the system provides primary mechanisms as a message management system.
- email message exchanges are described in a message flow suitably.
- emails sent from the system are able to be received by an existing mailer.
- the system provides suitable user interface for reading/writing messages, and for defining message flows.
- the system has simple structure, and it works on several types of platforms.

Some emails are exchanged among participants of a collaborative work. One coordinator is established, who starts email exchanges by sending an email to the participants, who puts all replies together, who settles a problem if it happens, and who sends emails which state a result. Such activities impose a burden on the coordinator. Also, the participants have to watch emails carefully.

Figure 1 shows an example of a simple message flow. Let A, B, C and D be members of a collaborative work. Coordinator A needs to get a consensus on a certain topic among other members. A sends emails to B, C and D. They are replied to the emails. The replies are arranged by A. When the participants do

Table 1. Message flow commands for specifying message flows

Commands, symbols and examples	Explanations
(1) next 	After A processes a message, the message flow is delivered to the next receiver B, in sequential.
(2) commit_and 	A message is delivered to both B and C from A. After both B and C process their messages, the message flow is delivered to the next receiver.
(3) commit_or 	A message is delivered to both B and C from A. After one of B and C processes the message, the message flow is delivered to the next receiver.
(4) if 	If the preceding client of command 'if', B, adds a new message to the received message, it return to owner A. After A processes the returned message, the message flow is delivered to the successive client of the 'if' command. Otherwise, the message flow is delivered to the next receiver of B.
(5) if_super 	If the preceding client of 'if_super', B, adds a new message, the message flow is returned to owner A. After A processes the returned message, the message is delivered to the next receiver of the first client A. As this result, a loop is constructed until B does not append a new message. Otherwise, the message is delivered to the next receiver of B.
(6) group_and 	When all members of group D which is defined by owner A process the received message, the message flow is delivered to the next receiver B.
(7) group_or 	When one of the members of group D which is defined by owner A processes the received messages, the message flow is delivered to the next receiver B. The notation of this command is the same with command 'next'.

not agree to an opinion of the topic, email exchanges are repeated until all of the participants agree with each other. Then if one participant does not reply to the received emails, coordinator A sends a reminder.

Message exchanges as shown in the above example are often achieved. The members of the email exchanges are fixed, and the order of the exchanges is able to be expected. But, the number of repetition of exchanges is not supposed. For example, members and orders for receiving mails are able to be assumed in members for making a meeting schedule, and in developers for reaching a consensus on a specification of a certain system, etc.

3 Specification of Message Flows

A message flow is a sequence of commands and clients. A client who defines a message flow is called the owner of the message flow. Table 1 shows the commands and their examples. The system seems that a received message flow is processed by a receiver, when the receiver appends a new message to the

received message or when the receiver confirms to read it. When a message flow is reached to the last receiver, finally the message flow is returned to its owner.

There are seven commands, called message flow commands. 'next' is used to send a message flow to a next receiver sequentially. 'commit_and' and 'commit_or' specifies that a flow is sent to some clients, the flow is split and the split flows are synchronized into one message flow. In 'commit_and' a sender delivers the same message to some clients. When all clients process their received messages, a message is constructed by combining their replies. This constructed message is delivered to successive receivers. While, when one of the receivers replies to the message, the flow is sent to successive receivers, in 'commit_or'. One of the receivers replies to a message, the message flow is progressed. Other receivers are not able to process it, but they are able to read the message. 'if' and 'if_super' specify that a message flow is returned to its owner. If the client who is the predecessor of 'if' writes a new message, the message flow is returned to its owner. After the owner processes it, the message is delivered to the successor of 'if'. If the predecessor of 'if_super' appends a new message, the flow is returned to the owner. After the owner processes it, the message flow is delivered to the next receiver of the owner, that is, the message flow is repeated during the owner and the predecessor of 'if_super'. In both commands, if a receiver does not append a new message, the message flow is delivered to the successive receivers. 'group_and' and 'group_or' are the same as 'commit_and' and 'commit_or', respectively. However, receivers of 'group_and' and 'group_or' are members of a group defined by an owner. The symbols of 'next' and 'group_or' are the same.

In order to achieve smooth deliveries of message flows, it is necessary to meet the cases that a client is away from an email mailer long term, and that a message flow has dead-line. Then, some troubles will occur, e.g., the message flow stops, and the message flow does not finish by the dead-line. To prevent such troubles, it is necessary to skip a certain receiver, and to specify constraints on period of time. The period of time is defined with the message flow. If a client is not able to reply within the definite period of time, the message flow is delivered to successive receivers by skipping a client who interrupts the flow.

4 Architecture of Message Flow

4.1 Modules of Message Flow

Message Flow consists of two modules which are a message flow client module and a message flow server module. The client module is used by an individual client. This module provides the mechanism for processing message flows and for communicating with the server module. The server module monitors activities of entire message flows. An architectural overview of the system is shown in Fig. 2.

When a client uses Message Flow, the message flow client process is activated. This module consists of a message flow processing, a message flow management, and a group management module. The message flow processing module is used for reading a message flow and processing it. If a message flow includes attached

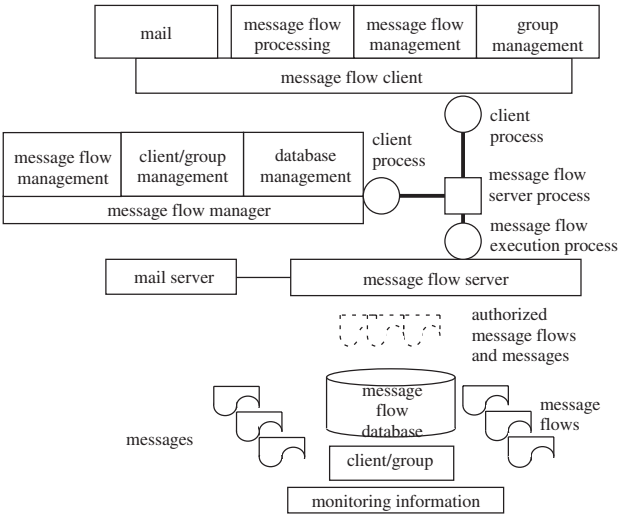


Fig. 2. An architectural overview of Message Flow

files, a client receives them. The message flow management module manages the message flows defined by a client. A client defines new message flows, interrupts message flows, watches reports on message flows still in progress, and deletes message flows, etc. The group management module is used to define a group that is a collection of clients. While, individual clients are managed by a message flow manager, see later. A client is informed that the client receives a message flow by receiving an email issued from Message Flow. The system includes the function to receive emails for supporting operations on message flows, but these emails are able to be received by existing mailer also.

Message Flow is maintained by a superuser called the message flow manager. This manager maintains the components of the system and its database. The manager uses the message flow manager module. This module consists of some functions in addition to a message flow client module. The message flow manager module handles message flows, clients/groups and the database. The message flow management module is used to manage every message flow which is defined by all clients, also. The client/group management module is used to register individual clients and groups, e.g., names of clients, their email addresses, etc. are indicated. The database management module is used for maintaining the database of the system, called the message flow database. This database contains every message flow and message. Moreover, several types of information for monitoring message flows are stored in this database, also.

The message flow server module provides functions to reply requests from a message flow client, to send emails through an email server, and to manipulate the database. When the message flow server is activated, a message flow server process is created. This process receives all requests from the clients. Moreover,

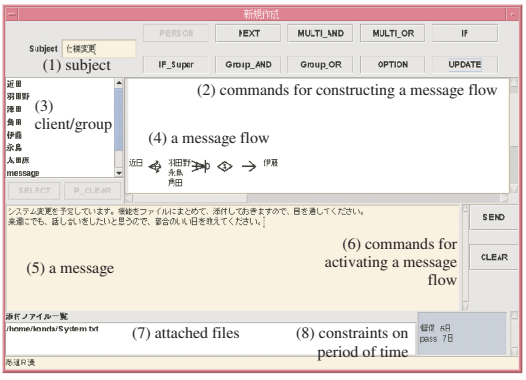


Fig. 3. An example of a window for specifying a new message flow

this server process creates a message flow execution process for one client process, to reply the requests of the client. The execution process handles all message flows related to one client.

On the hand, Message Flow is implemented in Java, its server module works on a Sun workstation. The message flow database is realized using PostgreSQL.

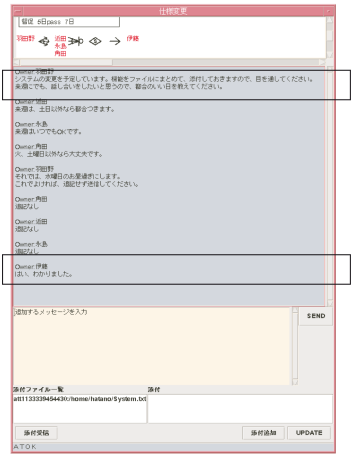
4.2 Specifying and Reading Message Flows

Figure 3 shows an example of the window for defining a new message flow. This window consists of eight parts. Its upper part (1) shows a window for describing the subject of a message flow. (2) shows commands for specifying a message flow. Message flow commands are indicated in this area. (3) shows names of clients and groups. (4) shows a newly defined message flow in graphical. A client describes a message flow by selecting the names of receivers and message flow commands one by one. The message written by an owner is described in area (5). The command for starting the message flow is shown in (6). If attached files are added, their names are shown in (7). (8) shows period of time.

An example of a window for receiving emails is shown in Fig. 4 (a). This window consists of (1) and (2), which indicate message flows and emails, and they are called a message flow box and a mail box, respectively. In the mail box, not only emails that are sent from Message Flow but also emails from other mailers are shown, although only emails from the system are shown in this figure. The emails sent from the system are called message flow emails. Each element in this mail box is one email. One line shows the subject of an email. There are some types of message flow email. For example, they notify; a new message flow arrives, a message flow finishes, an early reply is requested, i.e. a reminder, and a message flow is delayed. On the other hand, the message box shows the received message flows and the message flows defined by a client. One message flow is shown by its owner, message flow ID maintained by the system, and subject.



(a) An example of a window for reading emails which notify a client receives message flows.



(b) An example of a window for reading a received message flow and for replying to it.

Fig. 4. Examples of windows for receiving emails and for processing a message flow

Figure 4 (b) shows an example of a window for reading a message flow and for processing it. This window consists of four parts. In (1), a message flow represented in terms of message flow commands and receivers is shown in graphical. Constraints with respect to period of time defined in the message flow are shown in the same part. (2) shows messages which are written by former receivers. Messages written by the receivers are appended to their received messages in turn. The last message is written by the previous receiver. A message is written using (3) by the receiver. (4) shows attached files which are related to this flow.

5 Concluding Remarks

Message Flow has been developing as a message flow management system. This system provides primary mechanisms for defining message flows, for delivering them based on their definitions, and for monitoring email deliveries, etc. The system is experimentally utilized on the inside of our group, which is used to make meeting schedules, and to make clear specifications of this system.

Now, the system is extended from the viewpoints that the system provides the mechanism to specify flexible and extensible message flows, and that extensibility of a client module is improved. Since a message flow is defined in advance, it is difficult to modify a message flow while it is working. Sometimes to modify the working flows is required for sharing messages with other people. Moreover, there is a case that some message flows are defined in terms of the same receivers and the order of them. If such message flows are stored as a message flow library, such a library is useful. On the other hand, since the client module is implemented in Java, formats of many windows are fixed, moreover, their extensibility are low. For improving their extensibility and meeting the Internet environment, interface provided in the client module is made it work on many browsers in direct.

Acknowledgement

We wish to thank Hatano, Y., Konda, N., Sakakibara, M., Ebi, H., Tunoda, K. and Nagashima, M. for cooperation to implementation of Message Flow.

References

1. Bellotti, V., Ducheneaut, N., Howard, M., Smith, I.: Taking Email to Task: The Design and Evaluation of a Task Management Centered Email Tool. In: Proc. CHI, ACM, New York (2003)
2. Casati, F., Castano, S., Fugini, M., Mirbel, I., Pernici, B.: Using Patterns to Design Rules in Workflows. IEEE Trans. on Software Engineering 26(8), 760–785 (2000)
3. Dabissh, L., Kraut, R., Fussel, S., Kiesler, S.: Understanding Email Use: Predicting Action on a Message. In: Proc. CHI, ACM, New York (2005)
4. Georgakopoulos, D., Hornick, M., Sheth, A.: An Overview of Workflow Management: From Process Modeling to Workflow Automation Infrastructure. Distributed and Parallel Databases 3(2), 119–153 (1995)
5. Goldberg, Y., Safran, M., Shapiro, E.: Active Mail - A Framework for Implementing Groupware. In: Proc. CSCW, ACM, New York (1992)
6. Gwizdka, J.: TaskView - Design and Evaluation of a Task-based Email Interface. In: Proc. of the Conference of the Centre for Advanced Studies on Collaborative Research (2002)
7. van der Aalst, W., van Hee, K.: Workflow Management, Models, Methods, and systems. The MIT Press, Cambridge (2002)
8. van der Aalst, W.M.P.: Business Process Management Demystified: A Tutorial on Models, Systems and Standards for Workflow Management. In: Desel, J., Reisig, W., Rozenberg, G. (eds.) Lectures on Concurrency and Petri Nets. LNCS, vol. 3098, pp. 1–65. Springer, Heidelberg (2004)
9. Whittaker, S., Jones, Q., Nardi, B., Creech, M., Terveen, L., Isaacs, E., Hainsworth, J.: ContactMap: Organizing Communication in a Social Desktop. ACM Transactions on Computer-Human Interaction 11(4), 445–471 (2004)
10. Whittaker, S.: Supporting Collaborative Task Management in Email. Human-Computer Interaction 20, 49–88 (2005)
11. Whittaker, S., Belloti, V., Gwizdka, J.: Email in Personal Information Management. Communications of the ACM, 49(1) (2006)

Scalable 2-Pass Data Mining Technique for Large Scale Spatio-temporal Datasets

Tahar Kechadi and Michela Bertolotto

University College Dublin, Belfield, Dublin 4, Ireland.
{tahar.kechadi, michela.bertolotto}@ucd.ie
Sergio Di Martino, Filomena Ferrucci
Dip. di Matematica e Informatica, Università degli Studi di Salerno, Italy
{sdimartino,fferrucci}@unisa.it

Abstract. In this paper we present a system for mining very large spatio-temporal datasets. The system comprises two main layers: the mining layer and the visualization layer. The mining layer implements a new approach based on a 2-pass strategy to efficiently support the data-mining process, address the spatial and temporal dimensions of the dataset, and visualize and interpret results. In the first pass, the data objects are grouped according to their close similarity. In the second pass these groups are clustered to produce new models or patterns. The main reason for this 2-pass strategy is that the datasets are too large for traditional mining and cannot support the interactivity required by the visualization layer.

1 Introduction

Clustering is one of the fundamental techniques in data mining. It groups data objects based on information about the objects and their relationships. The goal is to optimize the similarity within a group of objects and the dissimilarity between the groups in order to identify interesting structures in the underlying data. However, there are still several open questions in the clustering process. These include the optimized number of clusters, the validity of a given clustering, obtaining different shapes and sizes of clusters (rather than forcing them into spherical shapes according to distance measure functions), finding appropriate clustering structures in a given dataset.

Clustering algorithms can be divided into two main categories, namely partitioning and hierarchical methods. Different clustering concepts have been proposed with different targets and evaluation criteria. Partitioning methods consist of dividing the objects into k groups and iteratively exchange objects between the groups until no improvement can be made. In hierarchical clustering, the clustering is defined in a bottom-up fashion. This assumes that the objects form the initial clusters and iteratively merges the closest pairs until the quality of the clusters does not further improve. Among this class one can find BIRCH [1], CURE [2], and C2P [3]. Density-based approaches tend to cluster objects, which are close to each other and separate them from regions of low density. DBSCAN [4] and OPTICS [5] are the common

B. Apolloni et al. (Eds.): KES 2007/ WIRN 2007, Part II, LNAI 4693, pp. 785–792, 2007.
© Springer-Verlag Berlin Heidelberg 2007

algorithms of this class. These two algorithms differ only in the way they output the results. OPTICS outputs an ordering of objects considered to be the centers of the clusters found in the dataset. Even though these algorithms are very popular they cannot be applied directly to very large spatio-temporal datasets. Indeed, spatial and temporal constraints introduce a very high level of structure in the datasets preventing most of the traditional data mining algorithms from discovering models in such datasets. Much work has been conducted on discovering spatial or temporal models in the data. However, little work has been done in mining spatio-temporal data. Some modifications of the traditional methods have been proposed in [6] to cluster objects of similar trajectories. However, they are very computationally expensive and they cannot deal with distributed and heterogeneous aspects of the data.

In this paper, we propose a new method for mining very large spatio-temporal datasets. Because the raw dataset is too large for any algorithm to process, the idea is to reduce the size of that data by producing a smaller representation of the dataset, as opposed to compressing the data and then uncompressing it later for reuse. Indeed, we want to reduce and transform the data so that it can be managed and mined interactively. Furthermore, we want to exploit the important aspect of spatio-temporal data (i.e., objects that are physically and temporally close tend to be "similar"). This data reduction ("compression") is part of a 2-pass strategy, where firstly the data objects are grouped according to their close similarity and then these groups are clustered by using different clustering techniques based on specific objectives. This strategy is used to design the clustering technique implemented in the first layer of our system.

The paper is organised as follows: Section 2 describes briefly the main layers of our system for mining and managing spatio-temporal datasets. Section 3 describes in details the methodology behind the 2-pass strategy proposed in this paper. Section 4 is dedicated to the visualisation layer and its use. Section 5 presents some experimental results. Finally, in section 6 we conclude and summarise some future work.

2 Spatio-temporal Data Mining Environment

To address the issues of mining and managing spatio-temporal datasets we have proposed a 2-layer system architecture [7,8] including a mining layer and a visualization layer. The mining layer implements a mining process along with the data preparation and interpretation steps. For instance, the data may need some cleaning and transformation according to possible constraints imposed by specific tools, algorithms, or users. The interpretation step consists of using the selected models returned during the mining to effectively study the application behavior. Usually, the model output requires some "post-processing" in order to exploit it. This step can benefit from data visualization, since interactivity and user expertise are very important in the final decision-making and data interpretation. Especially, in spatial data mining, the post-processing (such as automatic and interactive visualization of data) is so important that some researchers considered it as part of modeling, and called it "Interactive Data Mining" (IDM) [9]. It combines both automatic and visual data mining, and received much attention from users as it offers higher degree of satisfaction and confidence [10].

In order to interpret the output of the mining process, in our system we feed the results of the mining process to different visualization tools that provide complementary interacting functionality at the visualization layer. In our first implementation we have developed two powerful visualization tools to support exploratory visual data interpretation. These applications display the output of the mining process in a 3D virtual environment, allowing users to freely change their viewing perspective. More details about the visualization tools can be found in [7,8]. In the following sections we present the novel mining strategy implemented in the mining layer and some details about the visualization layer.

3 2-Pass Strategy

The mining process for spatial data is more complex than for relational data in terms of both the mining efficiency and the complexity of possible patterns that can be extracted from spatial datasets [11]. The reason is that the attributes of neighboring patterns may have significant influence on a pattern and should also be considered. Therefore, new techniques are required to efficiently and effectively mine spatial datasets.

The main problem for analyzing very large datasets is that the hardware resources are not able to deal with the storage (memory) and processing (CPU) within the response time expected by the user to perform interactive queries. For instance, Hurricane Isabel dataset [12] is represented by a space of (500x500x100) with 25x106 data points. So the goal of this strategy is to reduce the amount of memory used during the mining process as well as the processing time of a user query. Ideally, the data reduction or compression should not affect the knowledge contained in the data.

The first task, then, in this strategy is to find the data points that are most similar according to their static (non spatial and temporal) parameters. This part of the strategy is the key to the whole success of the compression, so that we do not lose any important information from the data that could have an adverse effect on the result obtained from mining the data at a later stage.

The second task is to cluster these groups of closely related data points in a meaningful way to produce a new "meta-data" sets so that they are more suitable and acceptable for data mining techniques to analyze and produce results (i.e. models, patterns, rules, etc.).

3.1 First Pass

For the first pass we deal with the raw data. This raw data can be represented by the following parameters; the number of time steps (Nt), the number of data points (N), and the number of static parameters (Ns). For Hurricane Isabel data, Nt = 48 time steps, N = 25x106 data points, and Ns = 13.

The algorithm for the first pass produces clusters of data points that are closely related. The goal here is to produce new data objects, where each object represents one cluster of raw data. Therefore, the main objective is to reduce the size of the initial data without loosing any relevant information. Figure 1 shows a high level view of the steps carried out by the algorithm for this first phase of this approach. It is important

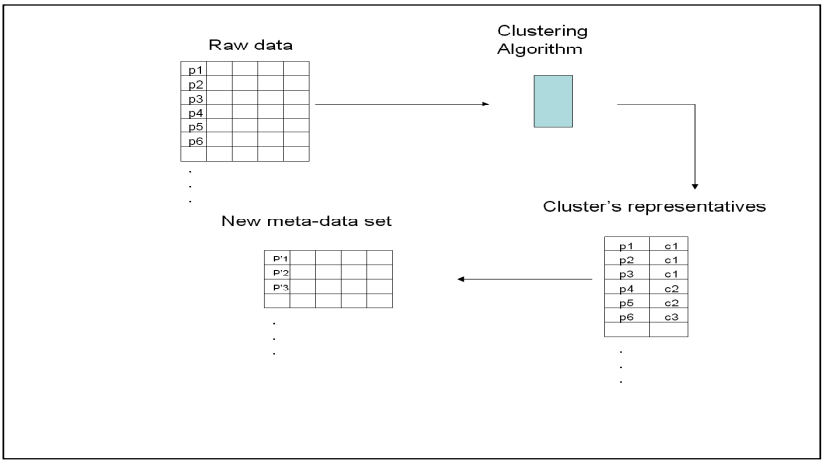


Fig. 1. Step by step view of the first pass of the strategy

to note that only the data points that have a very high similarity between each other will be grouped together. As a result of this pass, the new dataset is much smaller than the original data. It contains more information about individual clusters. This will help the clustering performed during the second pass.

3.2 Second Pass

The second pass involves clustering the tightly grouped data objects to produce new clusters representing new knowledge and ready for evaluation and interpretation. We have implemented two traditional clustering algorithms; DBSCAN and CURE. The first one is more suitable for similarity measures that can be represented by a distance measure. Each cluster is represented by one data object, called *medoid*. The second one can accept any similarity measure and the clusters can be represented by more than one representative. This is very important to represent clusters of different shapes. Figure 2 illustrates all the steps performed in this pass. There are locations in the space that are highly similar; these are represented within each of the small location groups (small circular shapes). It is important that no location group overlaps with another location group so that the integrity of the data is not affected. The next step applies data mining on the meta-data using a clustering algorithm.

4 Visualisation

Once the results are produced by the mining engine, we feed them to the visualization layer. Such a layer can include several complementary tools; for example we have developed a tool that embeds Google-Earth to render in 3D the mining outcomes, extended over a geo-referenced satellite image enhanced by additional informative layers [7]. This approach aims at facilitating domain experts and is particularly useful to highlight spatial relationships among the datasets and their spatial context, in terms of

geographical layers, such as cities, roads, mountain chains, seashores etc., as well as to show how the phenomenon evolves over time.

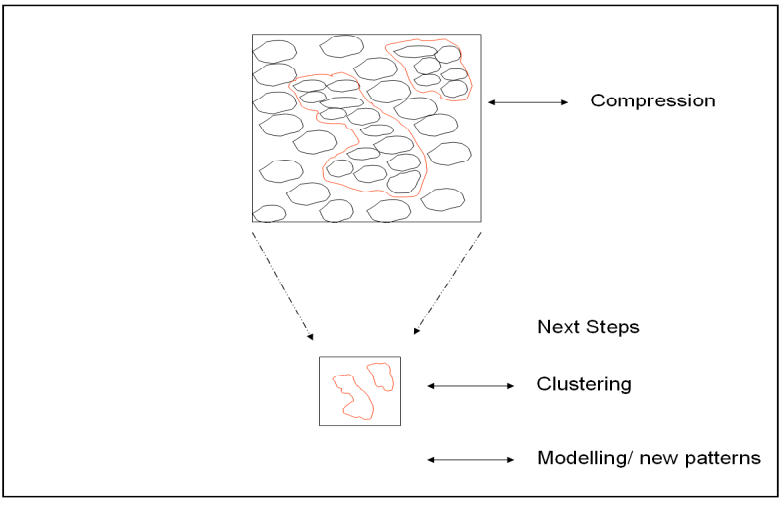


Fig. 2. A synthetic example of clustering meta-data produced in the first pass

We also developed other tools, which are independent interactive graphical user interfaces [7,8]. These provide more advanced user interaction with the mining results in a non-geo-referenced space. Such powerful tools encompass a set of complex features and would be more oriented to data mining experts. The main interesting feature of this visualization environment is the innovative and purpose-built functionality for drawing, upon user request, the "shape" of a rule. By shape of a rule (cluster) we mean the particular region of space in which the rule (cluster) holds – that is, the set of locations in which the cluster (hence all items involved in it) is well supported. This allows discovering specific relationships between the shape of the initial dataset and the shape of specific clusters, i.e. the characteristics of the initial dataset and all uncovered laws controlling the observed event. Standard visualisation tools for geo-spatial data do not provide this functionality.

From a data-mining expert's point of view, visualizing the shape of the cluster not only removes overload of visual information from the screen, but it also helps to highlight the structure of any pattern embedded in the data. Furthermore it focuses the attention only on the subset of the dataset involved in the cluster being studied. This allows for a more efficient and light visualization process, even when displaying millions of points. From a domain expert's point of view not only the content of a cluster is important, but also its shape as it can reveal hidden information about the behaviour of the phenomenon being studied. Thus, being able to narrow the visualization phase only to that shape is extremely valuable in interpreting all the results. To the best of our knowledge, standard visualization tools for geo-spatial data do not provide this functionality.

The two environments we have developed can be used in conjunction, as well as linked to any number of other instances, over an IP-based network, to create a

collaborative and distributed environment, rendering the same datasets and/or rules [7,8]. In the following, we present some results through screen shots illustrating some clusters (rules) that were extracted by our mining strategy.

5 Results

All the experiments presented in this paper are obtained using the Hurricane Isabel dataset [12]. Such a hurricane struck the US east coast in September 2003.

The main strength of our system is the innovative mining strategy coupled with an interactive and user-friendly visualization layer. A cluster or a rule produced at the mining layer is accessed directly and well represented by its shape at the visualization level. This shape is intended to highlight the region of the space where the rule holds. An example of a rule extracted by applying our method to the Hurricane Isabel dataset is illustrated in Figure 3. In this figure the rule extracted is supported everywhere near the sea (and the land), but a narrow air flow (transversal wind component) penetrating the cloud reveals an interesting behaviour: the direction of this stream is exactly perpendicular to that of the hurricane, that at this time-step is steering north to make its landfall and proceed towards Canada. This rule presents a directional infiltration from south-west reaching exactly the eye of the hurricane (the small hole on the left). The antecedent panel provides a top view. This is one of the clusters we extracted, whose shape resembles the shape of the hurricane or one of its features.

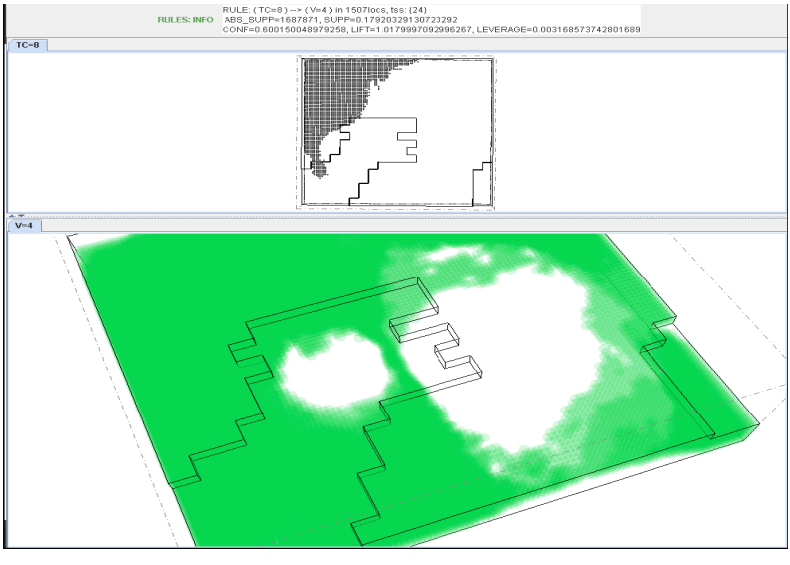


Fig. 3. Clusters representing the eye of the Hurricane and later winds

It is well known that pressure assumes always its lowest value in the eye of the hurricane: however, in this rule we can observe a hidden pattern, showing a 'normal' value for the pressure that implies a very low temperature. The eye of the hurricane

has much importance when predictions have to be made regarding its strength and speed. Therefore, clusters featuring unusual, oriented or hurricane-shaped areas – even if including only a few locations – may be essential to discover new behaviours. This figure is obtained by CURE clustering technique. One can notice that the cluster shapes are not necessarily spherical. This allows clusters to include many details mainly at the border areas of each of them.

Another example is illustrated in Figure 4, where the output of the DBSCAN algorithm is shown. This is a very spherical cluster corresponding to the eye of the Hurricane. Note that it removes all the boundaries of the clusters.

The aim of these clusters is to provide some clues about direction or strength of the hurricane that are can be useful to predict hurricanes' behavior.

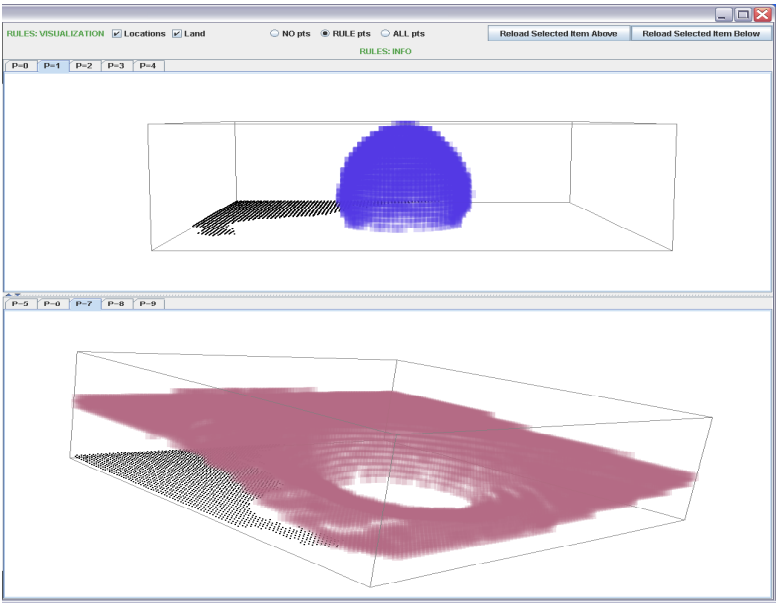


Fig. 4. The eye of the Hurricane in isolation. This is represented by one cluster.

6 Conclusion

In this paper we have described our approach for mining spatio-temporal datasets. We have developed a technique including a 2-pass clustering strategy, which first pre-processes the dataset by clustering it in such a way that it is suitable (in terms of memory and CPU requirements) for any mining algorithm. We have implemented this strategy and so far used it for the application of two known algorithms, CURE and DBSCAN. The technique has been implemented within a data mining system, which includes also a visualisation layer for interactive data interpretation. In this paper we reported some preliminary visual results. A more extensive evaluation is on-going. In the future we intend to apply and compare several different algorithms and improve the interactivity and flexibility of our visualisation tools.

References

1. Zhang, T., Ramakrishnan, R., Livny, M.: BIRCH: An efficient data clustering method for very large databases. In: Proc. of ACM SIGMOD (1996)
2. Guha, S., Rastogi, R., Shim, K.: CURE: An efficient clustering algorithm for large databases. In: Proc. of ACM SIGMOD (1998)
3. Nanopoulos, A., Theodoridis, Y., Manolopoulos, Y.: C2P: Clustering based on closest pairs. In: Proc. of VLDB (2001)
4. Martin, E., Kriegel, H.P., Sander, J., Xu, X.: A density-based algorithm for discovering clusters in large spatial databases with noise. In: Proc. of KDD (1996)
5. Ankerst, M., Breunig, M., Kriegel, H.P., Sander, J.: OPTICS: Ordering points to identify the clustering structure. In: Proc. of ACM SIGMOD, pp. 49–60 (1999)
6. Vlachos, M., Kollios, G., Gunopulos, D.: Discovering similar multidimensional trajectories. In: Proc. of ICDE, pp. 673–684 (2002)
7. Compieta, P., Di Martino, S., Bertolotto, M., Ferrucci, F., Kechadi, T.: Exploratory Spatio-Temporal Data Mining and Visualization. Journal of Elsevier Science, Special Issue on Human-GIS Interaction, 18(3) (June 2007)
8. Bertolotto, M., Di Martino, S., Ferrucci, F., Kechadi, T.: A Visualization System for Collaborative Spatio-Temporal Data Mining. International Journal of Geographical Information Science, 21(7) (July 2007)
9. Spenke, M., Beilken, C.: Visual, Interactive Data Mining with InfoZoom – The Financial Data Set. In: Żytkow, J.M., Rauch, J. (eds.) Principles of Data Mining and Knowledge Discovery. LNCS (LNAI), vol. 1704, pp. 15–18. Springer, Heidelberg (1999)
10. Keim, D.A., Panse, C., Sips, M.: Visual Data Mining in Large Geospatial Poin Sets. IEEE Computer Graphics and Applications 24(5), 36–44 (2004)
11. Roddick, J.F., Lees, B.G.: Paradigms for spatial and spatio-temporal data mining. In: Miller, H.G., Han, J. (eds.) Geographic Data Mining and Knowledge Discovery, Taylor & Francis, London (2001)
12. National Hurricane Center, Tropical Cyclone Report: Hurricane Isabel (2003), http://www.tpc.ncep.noaa.gov/2003isabel.shtml

Synchronizing Music and Video of Query Results in Cross-Media Retrieval System

Teruhisa Hochin[1] and Wen Xue[2]

[1] Kyoto Institute of Technology, Goshokaidocho, Mastugasaki, Sakyo-ku, Kyoto-shi, Kyoto 606-8585, Japan
[2] Osaka Prefecture University, Gakuencho, Sakai-shi, Osaka 599-8531, Japan

Abstract. This paper proposes a method of synchronizing a music clip and a video clip of the query result in the cross-media retrieval system for improving the degree of their fitness. The proposed method changes the speed in playing a video clip. This method decides the synchronization points of a video clip based on the beats in a music clip. The speed in playing a video clip is decided according to the times of the synchronization points. The proposed synchronization method is experimentally evaluated. The evaluation result shows that the proposed method could improve the degree of fitness of a music clip and a video one.

Keywords: Synchronization, Multimedia data retrieval, Impression.

1 Introduction

In recent years, retrieval of multimedia data has extensively been investigated. This research area includes topics on the direct and mutual access to the different kinds of media data[1,4]. This access is called the *mixed-media access* or the *cross-media retrieval*. In the cross-media retrieval, a kind of media data is accessed based on another kind of media data. For example, sound clips are directly accessed by specifying a video clip, and vice versa. Users can freely navigate on the sea of multimedia data. For the purpose of showing the effectiveness of the cross-media retrieval, let us consider the situation that we choose a music clip and a video one for spiritual healing. We have to select a good combination of a music clip and a video one. Two difficulties are encountered in completing this task. The first is occurred in obtaining music clips and video ones. Several retrieval systems have to be consulted because only one type of media data is supported in each retrieval system. In our situation, we have to consult a music retrieval system in order to obtain music clips, and consult a video retrieval system for obtaining video clips. The second difficulty is the adjustment of the fitness of the different kinds of media data. In our case, we have to obtain the most suitable pair. This adjustment process requires skill, and is a time-consuming task. The capability of choosing the good combination of different kinds of media data is required. The cross-media retrieval system could give an answer to this issue. Video clips suitable to a music clip are obtained only by specifying the music clip in the cross-media retrieval system.

B. Apolloni et al. (Eds.): KES 2007/ WIRN 2007, Part II, LNAI 4693, pp. 793–800, 2007.
© Springer-Verlag Berlin Heidelberg 2007

Although this cross-media retrieval system could retrieve video clips suitable to a music clip, how to present them is out of scope. That is, the system only presents a sound clip and a video one. These clips are presented without any relationships, e.g., synchronization. Some research efforts have shown that the synchronization of an audio clip and a video one brings us good impression[5,6]. The cross-media retrieval system has to consider the synchronization of audio clips and video ones in presenting them.

Synchronization of audio and video data has been studied[2]. The synchronization of the audio and the video data treated in these studies is decided a priori. The end of these studies is how to synchronize precisely the time of an audio stream with the corresponding time of a video stream. The synchronization treated in this paper is another type. The synchronization of an audio clip and a video clip is not decided a priori. The system has to decide the synchronization in order to improve the impression in presenting an audio clip and a video clip.

In this paper, a method of synchronizing a music clip and a video one is proposed. The end of this method is the improvement of the degree of fitness of them as the query result in the cross-media retrieval system. The proposed method changes the speed in playing a video clip. The synchronization points of a video clip are decided by using the beats in a music clip. The speed in playing a video clip is decided according to the times of the synchronization points. The proposed synchronization method is experimentally evaluated.

This paper is organized as follows: Section 2 proposes the method of synchronizing a video clip with a music clip. Section 3 presents a prototype system in order to show the feasibility of the proposed method. This section also includes the evaluation of the proposed method. Lastly, Section 4 concludes this paper.

2 Synchronization Method

This section describes a method of synchronization of a music clip and a video one. The synchronization method uses tempo information of a music or sound (audio) clip, and movement information of a video clip. Tempo information is a series of the times of the beats in a music clip. Tempo of a music clip is about thirty beats per minute to two hundred beats per minute. That is, a period between two beats is about 0.3 second to two seconds. Tempo information is derived from a music clip, and is stored into a database with its original music clip. Movement information is a series of the times of the movements in a video clip. Movement information is derived from a video clip, and is stored into a database with its original video clip.

In this paper, the following conditions are assumed.

- The time of the first beat of tempo information is almost the same as that of the first movement of movement information.
- Tempo information of an audio clip may include the lack of beats, and additional beats.
- Movement information of a video clip may also include the lack of movements, and additional movements.

Psychological research efforts[5,6] have revealed that the impression of an audio clip is more dominant than that of a video one. So, a music clip is left as it is, and a video clip is modified in order to be synchronized with a music clip.

Generating a synchronization plan is consisted of three major tasks: the decision of synchronization candidates, the decision of synchronization points, and the calculation of the speed ratios and the start time. The procedure of creating a synchronization plan is as follows.

1. Decide the synchronization candidates from movement information. A synchronization candidate is the movement whose time is closest to the time of a beat in tempo information. A synchronization candidate may also be referred to as a movement corresponding to the beat.
2. Obtain the first beat ta_1 from the tempo information, and obtain the movement tv_1 corresponding to ta_1.
3. Obtain the second beat ta_2, which is the first one whose time is larger than the *standard synchronization time* (SST), from the tempo information. The duration between the beginning of a music clip and ta_2 is called the *real standard synchronization time* (RSST). Obtain the movement tv_2 corresponding to ta_2.
4. Obtain the $(k+1)$th beat ta_{k+1} $(k \geq 2)$ satisfying the following condition: the period between ta_{k+1} and the ta_k is larger than RSST, and the time of ta_{k+1} is the smallest one. Obtain the movement tv_{k+1} corresponding to ta_{k+1}. The segment from tv_k to tv_{k+1} $(k \geq 1)$ is called the kth video segment. The segment from the beginning of a video clip to the first movement tv_1 is called the initial video segment. The segment from the last movement to the end of a video clip is called the terminal video segment.
5. Calculate the speed ratio in playing a video segment by using the formula $(tv_{i+1} - tv_i)/(ta_{i+1} - ta_i)$ where $i \geq 1$. Let the speed ratio of the initial video segment be equal to that of the first video segment, and let the speed ratio of the terminal video segment be equal to that of the last one.
6. Decide the start time st_v of a video clip by using the formula $ta_1 - speed_ratio_1 * tv_1$, where $speed_ratio_1$ is the speed ratio of the first video segment. When st_v is positive (negative, respectively), a video clip begins to play after (before) an audio clip starts.

3 An Example of Implementation

We have constructed a prototype system in order to show the feasibility of the synchronization method described in the previous section. This prototype system is consisted of three modules: the tempo information derivation module, the movement information derivation one, and the synchronization plan generation one. The tempo (movement, respectively) information derivation module derives the tempo (movement) information from a music (video) clip. The synchronization plan generator makes a SMIL (Synchronized Multimedia Integration Language) file[9] from the tempo information and the movement information. A SMIL player could play a pair of a music clip and a video one based on a created

SMIL file in a synchronized manner. From here on, the functions of these three modules are explained.

3.1 Tempo Information Derivation

The WAVE format is used as the file format of an audio clip. Sampled and quantized signals of an audio clip are straightforwardly stored in a WAVE file. As the sampling frequency is very higher than the frequency of the beats in a music clip, signals in 0.01 second are summed. The time where the summed value is locally largest is chosen as the candidate of the time of a beat.

An example of the series of the times of the beats derived is as follows:

$$T_a = \{1.94, 3.59, 4.38, 6.79, 7.55, 10.60, 14.80, 15.08, 16.07, 17.88\}.$$

The program of deriving tempo information is written in Java. The derivation time for an audio clip, which is 1.3MB in size, and is sampled at 8 bit stereo and 22kHz, is about one second under the Windows XP Professional(Version 2002 SP2) on a custom-made PC (Pentium4 3.00GHz, 512MB memory).

3.2 Movement Information Derivation

The MPEG format is used as the file format of a video clip. In the MPEG standard, there are three kinds of frames: I, P, and B frames[2]. I frames are self-contained and coded by using the compression method based on the Discrete Cosine Transformation (DCT). P frames are coded by using forward predictive coding including the motion estimation. B frames are coded by using two reference frames: a past frame and a future one. The coding process of B frames also includes the motion estimation. The movement of a block is represented with a motion vector in a P or B frame.

The motion vectors in B frames are used in order to derive movement information. First, the vertical and the horizontal components of motion vectors in a frame are averaged. The running averages of these averaged components are obtained over every 100 frames in order to delete the dithers of movements. The point where the value of the running average is maximal is the candidate of the movement of a video clip. Here, we could obtain the candidates in four directions: up, down, right, and left. In each direction, the periods between two adjacent points are obtained, and the standard deviation of these periods is calculated. The candidates in the direction where the standard deviation is smallest are adopted as the times of the movements. An example of the series of the times of the movements derived is as follows:

$$T_v = \{1.95, 3.40, 4.21, 5.57, 6.66, 7.75, 8.76, 9.59, 10.66, 12.10, 14.28, 15.31, 16.11, 17.54, 18.28, 19.37, 20.80, 21.84\}.$$

The program of deriving movement information is written in C. The derivation time for a video clip, which is 4MB in size, and 21 seconds in length, is about seventeen seconds under the same environment as that of the measurement of the tempo information derivation.

3.3 Synchronization Plan Generation

Before generating synchronization plans, the SST has to be decided. The pre-examination result showed that the time length in which we could bear the switching of video clips is about 4.5 seconds. So, we decided to use 4.5 seconds as the SST.

Here, T_a (T_v, respectively) described before is used as an example of tempo (movement) information in explaining the synchronization plan generation.

First, the synchronization candidates are decided. The following candidates are obtained by using T_a and T_v:

$$T_v^{cand} = \{1.95, 3.40, 4.21, 5.57, 6.66, 7.75, 10.66, 14.28, 15.31, 16.11, 17.54\}.$$

The times 8.76, 9.59, 12.10, 18.28, 19.37, 20.80, and 21.84 are deleted from T_v because there is no corresponding beat in T_a.

The first beat ta_1 is at 1.94 seconds. The first movement tv_1 is at 1.95 seconds. The second beat ta_2 is decided to be at 6.79 seconds because the SST is 4.5 seconds. The RSST becomes 6.79 seconds. The second movement tv_2 is decided to be at 6.66 seconds because it is the movement corresponding to ta_2. The third beat ta_3 is decided to be at 14.80 seconds because the period between ta_2 and ta_3 (8.01 seconds) is larger than the RSST (6.79 seconds). The third movement tv_3 is decided to be at 14.28 seconds. The music clip ends at 17.88 seconds. As the duration between ta_3 and the end of a music clip is 3.08 seconds, and is smaller than the RSST, there is no more synchronization point. In this example, two video segments except for the initial and the terminal ones are obtained. The one is from tv_1 to tv_2, and the other is from tv_2 to tv_3. The speed ratio of the first video segment is 0.97 which is obtained through the formula $(tv_2 - tv_1)/(ta_2 - ta_1) = (6.66 - 1.95)/(6.79 - 1.94)$. Similarly, the speed ratio of the second segment is 0.95. The start time st_v of a video clip is 0.0485 second which is obtained through the formula $ta_1 - speed_ratio_1 * tv_1 = 1.94 - 0.97 * 1.95$.

The SMIL file created is shown in Fig. 1. An audio stream is played in parallel with two video streams. Two video streams are played sequentially. The first video stream is played at the speed of 0.97 times of the original one. This stream ends at 6.66 seconds from the beginning. The first video stream is created by connecting the initial video segment to the first one because the speed ratio of the initial one is equal to that of the first one. The second video stream is played at the speed of 0.95 times of the original one. The start point of this stream is at 6.66 seconds in the original stream. This video stream is created by connecting the last video segment to the terminal one because their speed ratios are equal to each other.

The proposed method can be said to be a coarse-grained synchronization. This is because a video clip is synchronized with a music clip in every about five seconds. The period of the synchronization points is very long comparing to the processing time of computer system.

```
<smil xmlns="http://www.w3.org/2000/SMIL20/CR/Language">
  <head>
    <meta name="title" content="example"/>
  </head>
  <body>
    <par>
    <audio src="02_24.wav" begin="videoid.begin + 0.0485s" speed="1.0"/>
      <seq>
        <video src="video014.mpg" id="videoid" begin="0s" clip-begin="0s"
               dur="6.66s" speed="0.97" />
        <video src="video014.mpg" clip-begin="6.66s" speed="0.95" />
      </seq>
    </par>
  </body>
</smil>
```

Fig. 1. A SMIL file created for synchronizing an audio clip and a video clip

3.4 Evaluation

Evaluation method. The prototype system is experimentally evaluated. Six pairs of music and video clips shown in Table 1 are prepared. These pairs are obtained through the cross-media retrieval system based on impression[4]. The music clips are royalty free ones. The duration of the music clips are about twenty seconds. Video clips are of the street performance including dancing, and juggling. These were taken by the authors. The duration of the video clips are also about twenty seconds. The *speed* attribute in SMIL is currently not fully supported. Therefore, the speeds of video clips are modified by using Ulead VideoStudio 9[8].

A subject evaluates a pair of a music clip and a video one with the synchronization and the one without it through the Semantic Differential method[3]. The SD method is often used in the psychological testing. In this method, an object is evaluated by the scores of pairs of words. The impression word pairs used in this experiment are shown in Table 2. A score for an impression word pair is -2, -1, 0, 1, or 2. In Table 2, the score put to an impression word is

Table 1. Pairs of audio and video clips used in the experiment

No.	Video ID	Description of video	Music ID	Description of music
1	v052	three girls with white shirts dancing on street	m07	healing music, piano
2	v093	a pierrot juggling with a box	m08	rock'n'roll, rhythmical
3	v072	a boy with red shirts dancing on street	s11	blue grass, banjo
4	v103	a pierrot with red shirts rounding a dish	s11	blue grass, banjo
5	v105	a pierrot giving a man a balloon	s11	blue grass, banjo
6	v112	four girls with sport wear dancing on street	s11	blue grass, banjo

Table 2. Average scores and p values

impression word pair	sync.	no sync.	p value
matched(-2) – unmatched(2)	-0.77	-0.16	0.022
fast(-2) – slow(2)	-0.93	-0.13	0.041
regular(-2) – irregular(2)	-0.53	-0.50	0.81
dynamic(-2) – static(2)	-1.27	-0.80	0.065
beautiful(-2) – ugly(2)	-0.50	-0.20	0.060
smooth(-2) – rough(2)	-0.03	-0.40	0.18
modulated(-2) – monotonous(2)	-0.43	-0.03	0.13
straight(-2) – curved(2)	-0.10	0.03	0.71
bright(-2) – dark(2)	-0.60	-0.50	0.20
pleasant(-2) – unpleasant(2)	-0.87	-0.60	0.025
complex(-2) – simple(2)	-0.13	0.07	0.041
natural(-2) – artificial(2)	0.47	-0.10	0.055
bold(-2) – delicate(2)	-0.20	-0.10	0.41
tense(-2) – relaxed(2)	-0.13	-0.03	0.60
powerful(-2) – powerless(2)	0.10	0.27	0.29
hard(-2) – soft(2)	0.23	0.47	0.37
light(-2) – heavy(2)	-1.23	-0.87	0.006

described with the number in braces following the impression word. For example, a subject gives -2 as a score of the impression word pair *fast–slow* when he/she feels that a pair of a music clip and a video one is very fast. The subjects are five male graduate students.

Evaluation result. The average scores obtained through the evaluation are shown in Table 2. These scores of the synchronized pairs are tested whether these are significantly different from those of the non-synchronized ones. Table 2 also shows p *values* for this test. The average scores whose p value is less than the significant level, e.g., 0.05, are considered to be significantly different from each other.

The average scores of the impression word pairs *matched – unmatched*, *fast–slow*, *pleasant–unpleasant*, *complex–simple*, and *light–heavy* under the synchronized play are significantly different from those under the non-synchronized play. The synchronized pairs give a matched impression to the subjects because the average score of the impression word pair *matched–unmatched* under the synchronized play is significantly different from that under the non-synchronized play. Although the proposed synchronization method is a coarse-grained one, this method could sufficiently synchronize a music clip and a video one.

The audio–video pairs used gave the *fast* and *light* impression to the subjects. This may be caused by the speed in playing a video clip. The average speed ratios of the video clips of the pair No. 1, 2, 3, 4, 5, and 6 are 0.90, 1.16, 1.47, 3.32, 3.18, and 1.41, respectively. Almost all of the video clips are played faster than the original ones. This may result in the *fast* and *light* impression.

4 Conclusion

This paper proposed a method of synchronizing a music clip and a video one of the query result in the cross-media retrieval system. The proposed method is for improving the degree of fitness of a music clip and a video one. This method changes the speed in playing a video clip. This method decides the synchronization points of a video clip based on the beats in a music clip. The speed in playing a video clip is decided according to the times of the synchronization points. The synchronization plan is created in the form of a SMIL file. The proposed synchronization method is experimentally evaluated. The evaluation result showed that the proposed coarse-grained synchronization method is sufficient to improve the degree of fitness of a music clip and a video one.

The synchronization method uses the assumption that the time of the first beat in a music clip is almost the same as that of the first movement in a video clip. The improvement of the method for the purpose of freeing from this assumption is in the future work. In this paper, music clips are used. Handling audio clips including natural sound clips, which may not have clear beats, is also in the future work. The experiment did not evaluate whether users prefer the effects given by the synchronization. The future work includes this evaluation. Adapting the proposed method to the cross-media retrieval system is also included in the future work.

Acknowledgement

This research is partially supported by the Ministry of Education, Science, Sports and Culture, Grant-in-Aid for Scientific Research (C), 18500163, 2006-2008.

References

1. Chen, P., Hearst, M., Kupiec, J., Pedersen, J., Wilcox, L.: Meta-Data for Mixed-Media Access. SIGMOD RECORD 23(4), 64–71 (1994)
2. Furht, B. (ed.): Multimedia Systems and Techniques. Kluwer Academic Publishers, Dordrecht (1999)
3. Gregory, R.J.: Psychological Testing: History, Principles, and Applications, Allyn & Bacon, Inc. (2000)
4. Hochin, T., Tsuji, T.: Mutual Multimedia Access Using Kansei Factors. Kansei Engineering International 2(4), 9–18 (2001)
5. Iwamiya, S.: Interaction between auditory and visual processing when listening to music in an audio visual context: 1. Matching 2. Audio quality, Psychomusicology 13, 133–154 (1994)
6. Lipscomb, S.D.: The perception of audio-visual composites: Accent structure alignment of simple stimuli. Ethnomusicology 12, 37–67 (2005)
7. Mandal, M.K.: Multimedia Signals and Systems. Kluwer Academic Publishers, Dordrecht (2003)
8. Ulead: Ulead VisualStudio 9, `http://www.ulead.co.jp/`
9. W3C: Synchronized Multimedia, `http://www.w3.org/AudioVideo/`

A Fuzzy Approach to Multimodal Biometric Authentication

Antonia Azzini, Stefania Marrara, Roberto Sassi, and Fabio Scotti

Department of Information Technology, University of Milan,
via Bramante 65, 26013 Crema (CR), Italy
{azzini,marrara,sassi,fscotti}@dti.unimi.it

Abstract. In the last few years the security of the user's identity has become of paramount importance. In this paper we investigate the opportunity of using a multimodal biometric system as input of a fuzzy controller designed with the aim of preventing user substitution after the initial authentication process.

Keywords: Biometry, Fuzzy System, Strong Authentication, Security.

1 Introduction

In the last few years, authentication has become an increasingly important issue in modern society. In many consumer applications, as for example financial transactions, it is extremely important to prove that a person is who he/she claims to be. In recent years, traditional methods for authentication such as passwords and PIN numbers have been shown to suffer flaws in security. Such flaws include forgotten or easily guessed passwords, PIN numbers written on the back of cards, etc. Therefore, there have been an intense study of alternative methods to alleviate problems.

One alternative to traditional authentication is biometrics [1]. Biometrics is based on the fact that a person possesses certain characteristics — such as retinal patterns, fingerprint patterns, gait, etc. — that are biologically or behaviorally unique to an individual. It is this characteristic, rather than a forgettable password, which is used to corroborate a user's claimed identity. Because such characteristics are not easily forged, cannot be forgotten, and are not easily guessed, biometrics solve many of the problems that exist in traditional authentication.

Sometimes, in highly sensitive environments, a single, initial authentication may not be sufficient to guarantee security, but it may be necessary to perform strong authentication many times (e.g., at random intervals) to prevent identity substitution after the initial authentication step.

In our research [3,2] we proposed a methodology based on a balanced mix of strong and weak authentication techniques studied to guarantee a high and prolonged-in-time level of security. For this reason, remote access is initially provided by means of biometric devices but then it is granted in time using other authentication methods. In such a scenario, the system must distinguish between

B. Apolloni et al. (Eds.): KES 2007/ WIRN 2007, Part II, LNAI 4693, pp. 801–808, 2007.
© Springer-Verlag Berlin Heidelberg 2007

the initial authentication phase, in which it recognizes the user profile and allows the access, and the subsequent authentication steps, in which the system decides if its trust in user's identity is high enough to allow the user to continue performing the current activity. Focus of this paper is the module that performs the biometric authentication. In particular we study in depth the possibility to perform a multi-modal biometric authentication and we add a new input to the fuzzy controller. The new input represents the reliability of the biometric enrollment since it summarizes the context variables (e.g., illumination, humidity of the hand, etc.) that can affect the quality of the enrollment sample.

The structure of the paper is as follows. Section 2 presents a brief overview of the authentication devices used to ensure the identity of a user and compares advantages and drawbacks of the different techniques, Section 3 describes the general architecture of the fuzzy methodology used to ensure the user identity during time, Section 4 presents the fuzzy rules used by the methodology engines to compute the user identity trust level during time and, finally, Section 5 reviews the conclusions of this work and proposes some future work and open issues.

2 Integration of a Multimodal Biometric System

Multi-modal biometric systems utilize, or are capable of utilizing, more than one physiological or behavioral characteristic for enrollment, verification, or identification. The multimodal biometric systems try to overcome the monomodal systems drawbacks such as non-universality, the dependance from the limits of the adopted biometric technology (noise in sensed data, reliability of the sensor used, limited discriminability in data, upper bound in performance, lack of permanence) and a limited spoofing robustness [4,5]. On the other hand, multimodal systems present costs and verification times higher than the traditional monomodal systems [6]. There exist many possibilities of integration of different biometrics, sensors and techniques, and the choice of the best one strongly depends on the specific application. The main integration schemes, adopted in literature, are:

- *Multiple sensors*; different sensors can be used to acquire the same biometric trait, e.g. optical and capacitive fingerprint sensors used for the same finger, or the acquisition of both 2D and 3D images of the face. In some case, the system reliability is increased since the identification/verification can be performed even if one sensor fails the acquisition.
- *Multiple biometrics*; in these systems different traits are acquired by different sensors. It has the same advantages of the previous case, but in addition, this can enlarge the applicability of the system since not all the biometric traits might be present in the entire population.
- *Multiple snapshots*; the same biometric is acquired by the same sensor but different snapshots are acquired (not independent samples). The system can increment its accuracy with no additional hardware and software costs.
- *Multiple units*; more samples of the same traits are acquired by multiple acquisition units at the same time (e.g. two or more fingerprints of the same

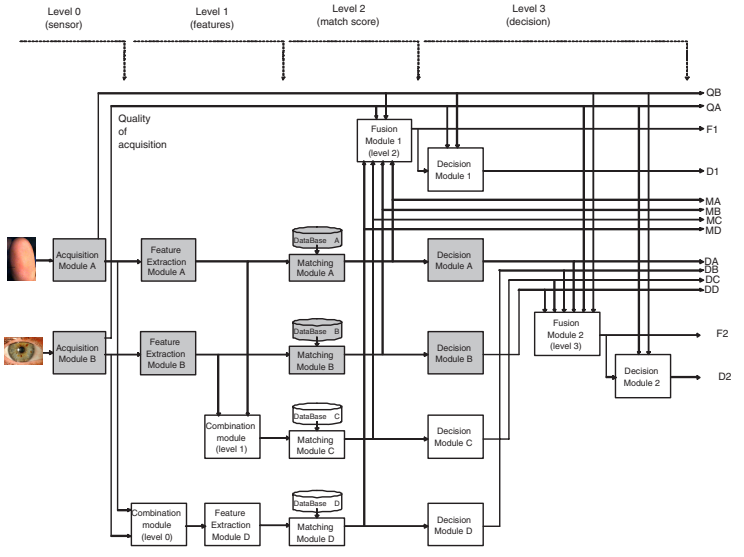


Fig. 1. General architecture of a multi-modal biometric system performing fusion and decision at all the four levels of integration of information

person at the same time); the system can process independent samples of the same person. Again the accuracy of the system and its reliability can be greatly improved.

– *Multiple representations/ matching algorithms*; the same biometric sample is processed using different formats, features and matching algorithms, e.g. the same single fingerprint image can be analyzed using two distinct techniques (such as a minutiae-based system and a technique based on the global orientations of the ridges).

Once the monomodal subsystems composing the multi-modal system have been chosen, for example a fingerprint and a retina recognition device, the information coming from each monomodal module can be integrated at *sensor level*, combining raw data acquired from sensors, *features extraction level*, combining the extracted features from the samples, *matching score level*, merging the matching scores of each different monomodal subsystem, and *decision level*, combining the identity decisions. Figure 1 shows the application of the four levels of integration for a general multi-modal biometric system. Here, the gray modules belong to the monomodal subsystems (for example to commercial monomodal systems). Each subsystem is composed by an *acquisition module*, *a feature extraction module*, *a matching module*, that is interfaced with a database containing the templates, and a *decision module*, that typically produces a boolean output (yes/no) in case of authentication or the individual identifier in case of identification. The white modules in the figure represent the infrastructure of the multi-modal system that is required to combine, fuse and then take decisions about the identity

of the subject acquired from the sensors using all the available levels of integration. Figure 1 shows the general scheme that can be used in identification frameworks; in authentication systems, the DBs are substituted by a single template. The output QA and QB represent the quality measurements of the input samples. Using these values, a sample not correctly acquired can be discarded or its information exploited in the following modules. The output $F1$ (real valued) represents the fusion signal obtained combining all the available matching scores of the monomodal subsystems. It gives a first representation of the identification matching level of the target individual. If a decision module is associated to the $F1$ output, a final multi-modal decision output $D1$ (Boolean) is consequently defined. The multi-modal system can also produce the matching score of all the present monomodal subsystems (MA, ..., MD, real valued). Each decision module produces a boolean signal (DA, ..., DD). These Boolean signals can be then combined (Fusion Module2) in order to produce a multi-modal decision concerning the identification of the user ($D2$) by a final decision module (Decision module2). A general rule for multimodal systems design states that the integration at an early stage of processing (e.g. level 0) leads to systems which might be more accurate than those where the integration is introduced at later stages. Unfortunately, in many cases fusion at sensor level (level 0) is hard to achieve, due to the different natures of the biometric traits, which might be hardly compatible (e.g., fingerprint and face). Moreover, most commercial biometric systems do not provide access to the feature sets vanishing the feasibility of a fusion at feature level (level 1). Consequently, in most applicative contexts, fusions at sensor and feature levels are not performed and the general architecture of figure 1 can be largely simplified. Fusion at matching level (level 2) or at decision level (level 3) does not require the creation of new databases or matching modules (the ones which constitute the monomodal subsystems are employed). Nevertheless, the simplified architecture encompassing only fusion above level 1 can still offer a significant increase of accuracy for the overall biometric system (reducing costs and complexity). In general, fusion at matching level ([7]) is preferable, but robust and efficient normalization techniques are necessary in the Decision Module 1. Fusion at the decision level (level 3) is poor given the limited availability of information which is restricted to the Boolean outputs of the subsystems' decision modules, and, in some cases, to the quality scores of the samples. But it is the only possibility of integration at hand if the match scores of the subsystems are not available. The adoption of a multi-modal biometric system in our model can be useful for different reasons. First of all, in a multi-modal biometric system the error rate in the authentication process consistently decays. Then, we can suppose to have people remotely connecting using a biometric device that can be different on the basis of the location. For example, some locations can be provided with a fingerprint device and other locations can be provided with an iris device. In this case the multi-modal system takes different monomodal inputs and accepts that one or more of these are empty, being however able to produce a decision about the person authentication.

3 A Fuzzy Controller for Multi-modal Authentication

We developed a fuzzy controller computing an output variable expressing trust in the user identification. The novelty of this work is that, by considering the multi-modal biometric system presented in Sect. 2, our controller now bases its decision not only on the matching score, but also on a score that represents the quality (poor/good) of the biometric enrollment, improving the information available about the biometric device considered.

Furthermore, in order to improve the privacy of a user during the overall system, the biometric authentication process is carried out at local level, by using, if needed, on-board biometric authentication devices. In this case the identification data that are sent to the fuzzy model represent the result of the biometric matcher, and correspond to the trustworthiness evaluation parameters that will be processed by the fuzzy controller in order to obtain the trust value.

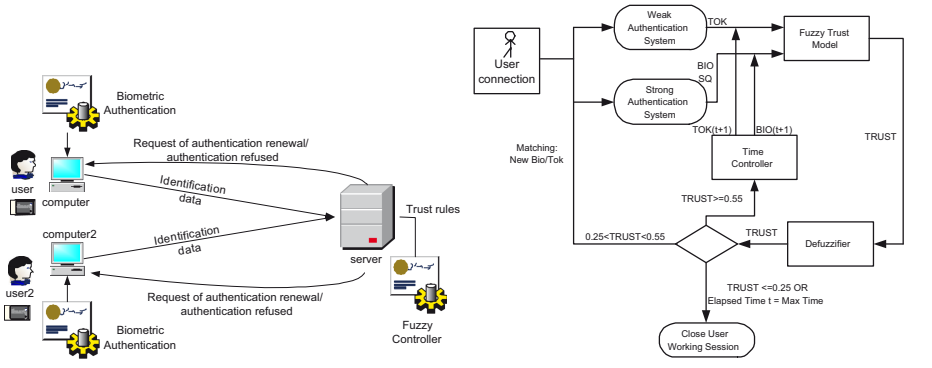
Trust is well-expressed as a linguistic variable. Our controller follows the Mamdani approach with a *center of area* defuzzification. Experimental results confirm the goodness of our choice (see [2]). The controller's operation is described in detail in Sect. 3.1. Briefly, we suppose that, after the initial authentication, in this paper trust in user identity is computed on the basis of biometric matching score, biometric quality score and password check (accept/reject). If the password is correct and the biometric matching and quality scores are higher than a pre-set threshold, the session starts with a certain trust value. Then, biometric and password trustworthiness start to decay and, after a certain time, a new trust value is computed. On the basis of this value the system may decide to close the session altogether, continue the session without asking for a new authentication or asking the user to authenticate again. If a new authentication is asked, a new trust value is computed and the session goes on. This cycle ends when the session is closed by the system, or ended by the user.

3.1 Architecture of the Fuzzy System

Our approach includes a trust evaluation process which continuously checks the identity of the user who is performing a certain activity. Figure 2(a) shows the basic steps of our process: after an initial authentication, the server can require further authentication steps based on the level of trust previously computed and the time passed from the last authentication. We suppose the first authentication to have been performed using both techniques. Indeed the userID is used to choose in the database the template to be used in the matching of the biometric acquisitions, because in a matching one-to-one the error rates are significantly reduced. The following steps, instead, can be acquired by strong or weak techniques on the basis of the trust level.

3.2 Trustworthiness Evaluation Parameters

After receiving an initial strong authentication, the server accepts or refuses the user on the basis of the biometric matching value (BIO) computed by the



(a) (b)

Fig. 2. Context for trust evaluation model (a), and architecture of the Multi-modal Fuzzy Trust Model (b)

multimodal biometric system in input. *BIO* has to be higher than a certain threshold (*th*), which is fixed for the application. Also the biometric quality score (*QS*) that indicates the reliability of the biometric acquisition on the basis of the quality of the enrollment is considered. In case the user is authenticated, the system receives two fuzzy values (e.g., 0.85 and 0.79) which represent how the biometric enrollment matches the user's template and how reliable this measure is. Then we use another authentication parameter, *TOK*, which corresponds to the boolean output (low/high) of the weak authentication system which supports the evaluation of the trust in the user's identity during the activity. In our prototype we used a UserId/password system. At the initial authentication step, the weak technique is involved to enforce the biometric acquisition and the parameter *TOK*, if the authentication successes, is set to *high*, i.e. equal to 1. Two *aging* parameters, respectively for the biometric and the token parameter values, are defined in order to measure how the system's trust in the identity of the user decays in time, by considering separately two straight lines for *BIO* and for *TOK*. All parameters values are empirically decided and are application dependant, the curves have been detailed in [3].

3.3 Fuzzy Controller Operation

The entire process implemented in our approach is shown in Figure 2(b).

At the first step, the information obtained by the biometric engine at local level, i.e. the value *BIO*, the quality score *QS* and the parameter *TOK* are fed into a fuzzy inference engine *Fuzzy Trust Model* in order to calculate a trust value *TRUST* that expresses the level of trust of the system in the user's identity after the initial authentication at time t_0. Of course, prior to process the inputs, it is necessary to define fuzzy membership functions which characterize the degree of membership of each input parameter in the context of the proposed model. Also, it is necessary to define the controller's fuzzy rules. We shall deal

Table 1. The Fuzzy Rules defined for the Fuzzy Trust Model

Fuzzy Rules of Trust Model
IF BIO is high **AND** QS is good **AND** TOK is high **THEN** TRUST is high
IF BIO is high **AND** QS is med. **AND** TOK is high **THEN** TRUST is med.
IF BIO is high **AND** QS is poor **AND** TOK is high **THEN** TRUST is low
IF BIO is high **AND** TOK is low **THEN** TRUST is low
IF BIO is med. **AND** QS is good **AND** TOK is high **THEN** TRUST is med.
IF BIO is med. **AND** QS is med. **AND** TOK is high **THEN** TRUST is med.
IF BIO is med. **AND** QS is poor **AND** TOK is high **THEN** TRUST is low
IF BIO is med. **AND** TOK is low **THEN** TRUST is low
IF BIO is low **AND** QS is good **AND** TOK is high **THEN** TRUST is low
IF BIO is low **AND** QS is med. **AND** TOK is high **THEN** TRUST is low
IF BIO is low **AND** QS is poor **AND** TOK is high **THEN** TRUST is very low
IF BIO is low **AND** TOK is low **THEN** TRUST is very low

with these aspects in the next section. At time t_0, *BIO* and *TOK* are initialized; at each time t_i ($i > 0$), the decay rate of these values will depend on the corresponding aging parameters. The *TRUST* value is then defuzzified through a *Defuzzifier* engine, using the standard centroid-of-area technique to compute the final level of trust at each time t_i. If the output trust is *low* the system asks for trust enforcement by going through the *Matching* phase. In this case the system asks for a user re-authentication, that can be biometric or knowledge-based, depending on the *BIO* and *TOK* values at that time t_i. In particular, the system re-acquires the parameter whose value at time t_i is less than a corresponding minimum threshold, previously defined, while maintains the same value at time t_i if it is more than the corresponding threshold. If the trust output is considered *medium* or *high* the system checks, through the *Time Controller* module, how the trust acquired at time t_i has been affected by the decay rate of the *BIO* and *TOK* authentication timeless functions, giving the new, decreased, parameter values for *BIO* and *TOK* at time t_{i+1}. These values are fed into the *Fuzzy Trust Model* in order to obtain the new trust value at time t_{i+1}. When the trust level decays to the value of *very low*, the user inserts two wrong passwords in the same weak authentication step, or when the maximum value of the examination time is reached, the execution step goes to the *Close User Working Session* and the process stops.

4 Trust Model Rules

The model presented in the previous section has been implemented with a set of fuzzy rules, in order to control the trust value at each time step t_i with respect to different evolved parameters of *BIO* and *TOK*. Moreover, the fuzzy controller considers the biometric quality score *SQ* when evaluating the reliability of the parameter *BIO* obtained from the biometric authentication step carried out at local level. The fuzzy rules are reported in Table 1.

The Fuzzy Trust Model is carried out in order to give a trustworthy value depending on biometric and token based acceptance rates, acquired from the *Strong* and *Weak* authentication systems. The *TRUST* value defuzzified is then used to decide the next steps during the user permission checks.

5 Conclusions and Future Work

In this paper we discuss a biometric approach for in-time user authentication by using a fuzzy controller. We investigate the possibility of using a biometric multimodal system to perform the strong authentication phase and we present different possibilities of integration of biometrics, sensors and techniques, showing how the available information coming from each monomodal module that compose the multimodal system can be integrated in order to improve the accuracy of the authentication process. In this paper our controller bases its decision also on a quality score, representing the goodness of the biometric enrollment. Moreover we improve the privacy protection of the biometric features, moving their evaluation on the client side instead of the server side. The biometric template can be stored on a smartcard, that is owned by the user, and then no biometric data is stored elsewhere. Actually we are working on techniques to prevent biometric attacks during the overall section time.

Acknowledgments

This work was partly funded by the Italian Ministry of Research under FIRB contract n. `RBNE05FKZ2_004`, `TEKNE`.

References

1. O'Gorman, L.: Comparing Passwords, Tokens, and Biometrics for User Authentication. Proceedings of the IEEE 91(12), 2021–2032 (2003)
2. Azzini, A., Damiani, E., Marrara, S.: Ensuring the identity of a user in time: a multimodal fuzzy approach IEEE Computational Intelligence for Measurement Systems and Applications, June 27-29, Ostuni, Italy (to appear, 2007)
3. Azzini, A., Marrara, S.: A Fuzzy Trust Model Proposal to Ensure the Identity of a User in Time. In: Proceedings of the 9th Fuzzy Days, September 18-20, Dortmund, Germany (2006)
4. Maltoni, D., Maio, D., Jain, A.K., Prabhakar, S.: Handbook of Fingerprint Recognition. Springer, New York (2003)
5. Bolle, R.M., Connell, J.H., Pankanti, S., Ratha, N.K., Senior, A.W.: Guide to Biometrics. Springer, New York (2004)
6. Hong, L., Jain, A.: Multimodal Biometrics ch. 16. In: Jain, A., Bolle, R., Pankanti, S. (eds.) Biometrics: Personal Identification in Networked Society, Kluwer Academic Publishers, Norwell (1999)
7. Tax, D.M.J., Breukelen, M.V., Duin, R.P.W., Kittler, J.: Combining multiple classifiers by averaging or by multiplying? Pattern Recognition 33, 1475–1485 (2000)

An Open Source Java Framework for Biometric Web Authentication Based on BioAPI

Elisardo González Agulla, Enrique Otero Muras, José Luis Alba Castro, and Carmen García Mateo

Department of Signal Theory and Communications, University of Vigo
Lagoas Marcosende, 36310, Vigo, Spain
eli@gts.tsc.uvigo.es

Abstract. One of the major early problems that biometrics faced was the lack of interoperability between different software applications and devices developed by different vendors. The BioAPI Specification has defined an open system standard application program interface (API) which allows software applications to communicate with a broad range of biometric technologies. This paper describes the design and implementation of an open source Java framework intended to provide single sign-on web authentication based on BioAPI-compliant biometric software or devices. The open-source principle makes this system a novel and practical development tool for testing and integrating biometric modules and devices from third parties.

Keywords: Biometrics, web authentication, BioAPI, open source.

1 Introduction

The popularity of broadband internet and cell phone connections has introduced many high added-value remote services where it is important to verify the identity of the authorized user. Classical techniques for electronic person authentication have several drawbacks in terms of performing reliable and user-friendly identity recognition; this occurs particularly with remote operations, where hacker attacks add to forgotten, shared, lost or stolen passwords or cards. Automatic identity verification, based on distinctive anatomical features (e.g., face, voice, fingerprint, iris, etc.) and behavioral characteristics (e.g., online/offline signature, keystroke dynamics, etc), is becoming an increasingly reliable standalone solution and attracting a great deal of attention as far as remotely-based applications are concerned. Nevertheless, this kind of biometric authentication implies new technological challenges that must be addressed to ensure full acceptance in many remote services [1].

One of the major early problems that biometrics faced was the lack of interoperability between different software applications and devices developed by different vendors. For this reason, the BioAPI Consortium was founded to develop a biometric Application Programming Interface (API) that would bring platform and device independence to application programmers and Biometric

B. Apolloni et al. (Eds.): KES 2007/ WIRN 2007, Part II, LNAI 4693, pp. 809–815, 2007.
© Springer-Verlag Berlin Heidelberg 2007

Service Providers (BSPs). The BioAPI Consortium developed a specification and reference implementation for a standardized API (BioAPI 1.1) [2] that is compatible with a wide range of biometric application programs and a broad spectrum of biometric technologies [3]. The BioAPI Specification defines an open system standard application program interface (API) which allows software applications to communicate with a broad range of biometric technologies.

This paper describes an open-source Java system intended to provide single sign-on web authentication based on BioAPI-compliant biometric software or devices. Integration of multiple biometric devices and algorithms (interoperability) is quite straightforward, as we will demonstrate in the sections below. We have paid a great deal of attention to the design of a client-server architecture capable of controlling any kind of biometric acquisition device over a secure TCP/IP connection.

Otherwise, we have focused on driving user interaction with the system through an easily configurable human-machine dialogue. Thus, verification tasks are modeled and specified by an XML document which describes the sample acquisition process and the biometric verification mode.

The remainder of this paper is organized as follows. Section 2 is devoted to the description of the main design guidelines followed in the development of our system for single sign-on web authentication based on server-side biometric verification. Section 3 focuses on the integration of the BioAPI framework within our system for allowing server-side biometric authentication in web environment. Section 4 presents the overall system, described both from a structural and functional point of view. Finally, Section 5 describes our conclusions and future research lines.

2 Design Guidelines

This section describes the main design decisions taken in creating a web-based system for biometric authentication that complies with the principles of security, interoperability and usability.

Any biometric recognition procedure can be divided into a number of stages:

1. Acquisition of a biometric sample (device-dependent processing).
2. Extraction of a biometric template (signal processing: pre-processing, feature extraction and user-template creation).
3. Biometric template matching (pattern recognition processing).

The third of these stages usually requires prior training of user models and thresholds set up in an enrolment process whenever a new user is added to the system.

When dealing with biometric authentication in client-server architectures, the three processes described above will provide different configurations depending on where the processes are executed. Although it is obvious that the acquisition process must be executed on the client side, there are three options remaining for extracting biometric templates and matching user reference templates: both

these processes are performed in the client machine (pull model); the biometric template is extracted on the client side, sent to the server and matched there against the user reference templates (push model); or both processes are performed in the server (a variant of the push model).

The system we have developed is based on the third of the above configurations. The client-side biometric application is in charge of multi-biometric sample acquisition, encryption and secure transaction. On the server side, there is a centralized authentication server with a biometric authentication module that is in charge of extracting the biometric template, matching, and checking access privileges.

The main guidelines for the construction of our system for biometric authentication have been the following:

- Maximized portability
  - of the client application, by using Java as programming language. Concretely, the biometric client application uses the Java Web Start technology, a Java-based environment for running desktop applications.
  - of the verification server, by following the standardized Java Enterprise Edition (EE) specification in web application development, permitting deployment in any Java EE compliant server.
- Maximized interoperability
  - with different biometric software or devices, by using a Java Native Interface wrapper for the BioAPI framework that is compatible with the BioAPI 1.1 reference implementations on Linux and Win32 operating systems.
  - with different software platforms, by using web services as middleware, permitting the development of alternative client applications that comply with our system.
- Use of free open source software: e.g. JBoss Application Server, MySQL database...
- Maximized security in managing biometric data, trying to comply with the Core Security Requirements of the ANSI X9.84 standard for Biometric Information Management and Security [4]. For this purpose SSL connections are used, and local disk writing of user samples is avoided, for instance.
- Maximized code reusability and maintainability, by using object-oriented software and design patterns.
- Multilingual support both in the web and in the client desktop application.

3 Integrating the BioAPI Framework into Our Architecture for Server-Side Biometric Verification

In order to allow server-side biometric verification based on BioAPI-compliant modules, the server contains a BioAPI instance running in a Linux platform and allowing third party biometric modules compliant with BioAPI 1.1 and compiled in C/C++ to be integrated in our system as Linux shared objects.

All these biometric modules can be controlled through this server-side BioAPI. BioAPI-compliant modules are also called Biometric Service Providers or BSPs following BioAPI nomenclature.

Moreover, if a user has a BioAPI installed locally in the client machine (Linux or Win32 platforms), the corresponding local BSPs will be also driven by our client application. Nevertheless, for a complete integration of any BioAPI 1.1 compliant BSP within our authentication system by using server-side verification for more security, compiled versions of the BSP libraries for both Linux and Win32 platforms are required.

Each BSP involved in the verification or enrolment process is properly indicated in an XML document that contains a high-level description of the protocol for the current session of verification or enrolment. This document is securely transferred from server to client when starting the session.

The low-level processing is composed of the primitives `BioAPI_Capture`, `BioAPI_CreateTemplate`, `BioAPI_Process` and `BioAPI_VerifyMatch`. In the more secure configuration, all these low-level primitives except `BioAPI_Capture` are executed as calls to a BSP in the server through the remote BioAPI. The `BioAPI_Capture` primitive is executed as a call to the BioAPI and BSP installed in the client machine and the resulting Biometric Identification Record (BIR) is sent back to the server (see Figure 1).

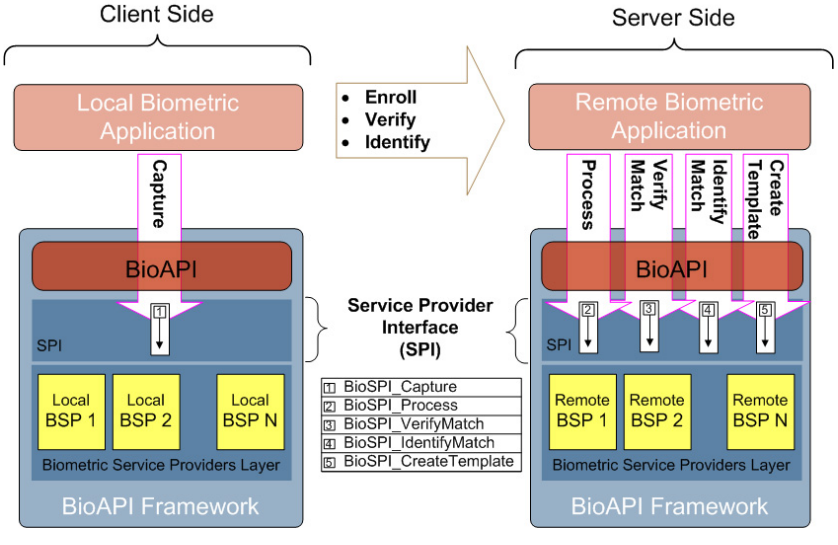


Fig. 1. Building-blocks of the BioAPI model in client/server environment

Java management of the different BSPs through local and remote BioAPI instances is performed using a Java wrapper that manages the calls to native code based on the Java Native Interface technology. Concretely, an open source Java Native Interface wrapper for the BioAPI framework on Linux/Unix has

been used [5]. To integrate into our system, this Java wrapper has been extended to include Windows support and access to low-level BioAPI primitives.

With the advent of the standard BioAPI 2.0, the definition of an additional layer of components below the BSPs (called Biometric Function Providers or BFPs) will allow our system to relax some of the interoperability conditions for including new biometric software for server-side verification. Only the corresponding implementation (whether Windows or Linux) of the BFP in charge of the capture process will be necessary on the client side (called Sensor BFPs), and only the Linux implementation of the BFPs in charge of the remaining processes will be necessary on the server side (archive, processing-algorithm and matching-algorithm BFPs). Thus, the server can perform verification without forcing the client to have a duplicate of the complete BSP.

Otherwise, we use a web service as middleware for the communication between the clients and the server. Web services with SOAP (Simple Object Access Protocol) allow interoperability between heterogeneous clients, permitting the development of new client applications that are fully compatible with our system and that do not need to be Java-based, merely compliant with the agreed interface. This functionality is particularly suitable for client terminals without a Java Virtual Machine.

4 The Overall Framework

As mentioned above, an XML document specifies the biometric sample acquisition protocol in an enrolment or verification session. This document represents the man-machine dialogue. Due to the inexistence of a universally agreed markup language for biometric acquisition we have developed an extension to the VoiceXML standard suited to our necessities. Our system can be easily translated into any other future markup language.

Figure 2 depicts a block diagram and functionality description for the system. Starting from a client verification or enrolment request, the successive actions and functionalities are explained as follows (see diagram numbering):

1. The biometric client application obtains, from the server, the XML dialogue that contains the enrolment or verification process description.
2. The client application interprets the protocol contained in the XML dialogue, prompts the corresponding information to the user, acquires the biometric sample, and performs an enrolment or verification.
3. Each time a biometric sample is required, the sample is captured from the corresponding BioAPI-compliant module. For this purpose, the client application calls the `BioAPI_Capture` primitive using a Java Native Interface wrapper for the BioAPI framework.
4. The result of the acquisition process is sent to the server bound to an enrolment or verification request.
5. The enrolment or verification process is executed in the server as a sequence of BioAPI calls.

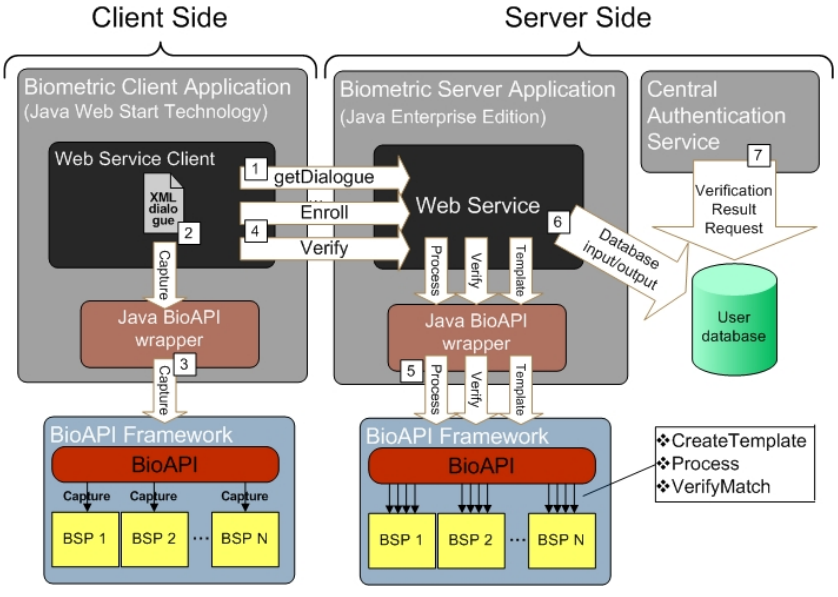


Fig. 2. Building-blocks and functionality description of the overall architecture

6. The verification result or the enrolment template is stored in the server database.
7. The database with the biometric verification results will be available to finally authenticate users for the web through the Centralized Authentication Service (CAS).

The main objective of the system was to offer biometric authentication capabilities to other web-based applications. For this purpose, we have integrated our system within a widely accepted Java-based open-source authentication solution known as Central Authentication Service (CAS) [6] [7]. CAS is a single sign-on solution originally developed at Yale University and later placed under the auspices of the Java Architectures Special Interest Group. CAS has quickly become the most popular single sign-on solution for universities. The main idea behind the integration of biometric verification functionality within the Central Authentication Service was to take advantage of the infrastructure provided by CAS to offer single sign-on web authentication, while improving security beyond basic mechanisms based on login and password, by adding biometrics. Thus, we make possible that any application prepared to use CAS for authenticating its users can also use our biometric system for this purpose, supporting any BioAPI-compliant biometric software or device in order to perform the authentication. The open-source e-learning platforms Moodle, ILIAS, or Claroline are well known examples of web applications that are yet capable of relying the authentication task on CAS. We had used Moodle and ILIAS to demonstrate the usability of our biometric extension of CAS within a common web application.

5 Conclusion and Future Work

In this paper we have presented and explained the design and implementation of an open source Java framework intended to provide single sign-on web authentication based on BioAPI-compliant biometric software or devices.

In order to control any BioAPI-compliant device from our Java architecture we have used an open-source Java Native Interface wrapper for the BioAPI framework that is compatible with both BioAPI 1.0 reference implementations of Linux and Windows platforms.

Moreover, diverse open-source solutions have been carefully integrated into our system attending to flexibility, portability, and interoperability issues.

Current version of our open-source biometric authentication system is available under LGPL license on [8]. Regarding future lines of work, we are analyzing the viability of porting our biometric system for web authentication to mobile devices.

Acknowledgments. This project has been partially supported by Spanish MEC under the project PRESA TEC2005-07212, Xunta de Galicia with the projects PGIDIT05TIC32202PR and PGIDIT05PXIC32202PM, and the European NoE BioSecure.

References

1. Jain, A., Bolle, R., Pankanti, S.: Introduction to Biometrics. In: Biometrics. Personal Identification in Networked Society, Kluwer Academic Publishers, Dordrecht (2002)
2. BioAPI Consortium (ANSI/INCITS 358-2002), `http://www.bioapi.org`
3. Yuan, X., Hui, S.C., Leung, M.H.K., Gao, Y.: Towards a BioAPI compliant face verification system. Computer Standards & Interfaces 26, 289–299 (2004)
4. ANSI X9.84-2003, Biometric information management and security for the financial services industry. American National Standards Institute, New-York (USA) (2003)
5. JBioAPI, a library of tools for accessing BioAPI-compliant biometric service providers in Java: `http://code.google.com/p/jbioapi/`
6. JA-SIG (Java Architectures Special Interest Group) Central Authentication Service (CAS): `http://www.ja-sig.org/products/cas/`
7. Aubry, P., Mathieu, V., Marchal, J.: ESUP-Portail: open source Single Sign-On with CAS (Central Authentication Service). In: Proceedings of EUNIS04 - IT Innovation in a Changing World, Bled (Slovenia) (July 2004)
8. Biometrics for Web Authentication: `http://sourceforge.net/projects/biowebauth/`

Feature Distribution of the Fingerprint Template Generated by the Geometric Hashing-Based Fuzzy Vault

Sungju Lee[1], Daesung Moon[2], and Yongwha Chung[1,*]

[1] Department of Computer and Information Science, Korea University, Korea
{peacfeel,ychungy}@korea.ac.kr
[2] Biometrics Technology Research Team, ETRI, Daejeon, Korea
daesung@etri.re.kr

Abstract. Recently, a cryptographic construct, called *fuzzy vault*, has been proposed for crypto-biometric systems, and some implementations for fingerprint have been reported to protect the stored fingerprint template by hiding the fingerprint features, *i.e.*, adding "chaff" minutiae into the original minutiae set. Also, solutions to the most challenging issue in applying the fuzzy vault to fingerprint(*i.e.*, "align" the input fingerprint with the stored fingerprint template in the domain of the fuzzy vault) have been proposed. One of the solutions exploits the idea of the *geometric hashing* to solve the auto-alignment problem. In this paper, we analyze the feature distribution of the generated hash table as a stored fingerprint template and improve the security of the geometric hashing-based fuzzy vault by uniformly distributing the features in the stored fingerprint template. The goals of this addition make the feature distribution of the hash table for the chaff minutiae similar to that of the real minutiae as well as the feature distribution of the fingerprint image for both types of minutiae uniform. Based on the experimental results, we confirm that the proposed approach can perform the fingerprint verification more securely without a significant degradation of the verification accuracy.

1 Introduction

Recently, Juels and Sudan[1] proposed a scheme called **fuzzy vault**. Also, solutions to the most challenging issue in applying the fuzzy vault to fingerprint(*i.e.*, "align" the input fingerprint with the stored fingerprint template in the domain of the fuzzy vault) have been proposed[4-6]. One of the solutions[6] exploits the idea of the **geometric hashing**[7] to solve the **auto-alignment problem**.

In the domain of the fuzzy fingerprint vault, the **security level** of the fuzzy vault scheme depends on how difficult it is to separate the chaff minutiae from the legitimate features in the vault[8]. In this paper, we propose an approach to enhance the security level of the geometric hashing-based fuzzy fingerprint vault which can solve the auto-alignment problem. To enhance the security level of the straightforward approach, we modify the way to add the chaff minutiae in the "fingerprint image

* Corresponding author.

B. Apolloni et al. (Eds.): KES 2007/ WIRN 2007, Part II, LNAI 4693, pp. 816–823, 2007.
© Springer-Verlag Berlin Heidelberg 2007

domain" such that the feature distribution of the "generated hash table domain" for the chaff minutiae is similar to that of the real minutiae. Simultaneously, we also consider the feature distribution of the "fingerprint image domain" such that the features of the fuzzy fingerprint vault are distributed uniformly and the possible degradation of the verification accuracy is minimized.

The rest of the paper is structured as follows. Section 2 explains the overview of the fuzzy fingerprint vault and the geometric hashing-based fuzzy fingerprint vault which can solve the auto-alignment problem. Section 3 describes the proposed approach to guarantee the similar distribution of the real and chaff features as well as the uniform distribution of the both features in the stored fingerprint template in order to perform the fingerprint verification more securely. The experimental results are given in Section 4, and conclusions are made in Section 5.

2 Background

Based on the *fuzzy vault* proposed by Juels and Sudan[1], some implementation results for fingerprint have been reported[3]. In the fuzzy vault scheme, the secret k is locked by a user's biometric(set A) using a probabilistic LOCK function, resulting in a vault V_A. The corresponding decryption algorithm UNLOCK takes as input a vault V_A, and a decryption biometric(set B) and outputs k if B is close enough to A, or null, otherwise. The authors argued that in a minutiae-based fingerprint matching system, if a real minutiae template is augmented with a larger number of "chaff" minutiae that constitute random noise, the secrecy of the fingerprint features as well as the secret k is strengthened.

Among the implementation results of the fuzzy fingerprint vault, some results solved the most challenging issue in applying the fuzzy vault to fingerprint (*i.e.*, "align" the input fingerprint with the stored fingerprint template in the domain of the fuzzy vault)[4-6]. For example, Yang and Verbauwhede[4] defined *reference minutiae* that were extracted during both the vault encoding and decoding. If these two reference minutiae are the same, the origins of the coordinate frames used during the locking and unlocking of the vault would be the same, and hence, the alignment could be established. Another solution proposed by Uludag and Jain[5] considered the small amount of data used for alignment without any information about the feature-based fingerprint template. However, both solutions have some problems in the verification accuracy.

Chung, *et al.*[6] proposed the method to solve the auto-alignment problem using the idea of the *geometric hashing*[7]. In typical fingerprint verification systems, minutiae are stored in the template file, and input minutiae are compared with the template minutiae after aligning them. Similarly, the geometric hashing consists of two procedures - preprocessing(or enrollment) and recognition(or identification).

In comparing the enrollment and verification minutiae tables, the transformed minutiae pairs with the same coordinates, the same angle, and the same type are determined. The minutiae pairs having the maximum number and the same basis are selected as the subset of real minutiae(G). Also, any additional alignment process is not needed

because the pre-alignment with each minutia is already computed in the enrollment and verification minutiae table generation. Additional details can be found in [6].

3 Feature Distribution of Geometric Hashing-Based Fuzzy Vault

As we explained, the goal of an attacker is to separate the chaff minutiae from the real minutiae in the vault, and the security level of the fuzzy fingerprint vault depends on the difficulty of this separation[8].

3.1 A Non-uniform Distribution of the Geometric Hashing-Based Fuzzy Vault

Applying the geometric hashing technique straightforwardly has some limitations in solving the auto-alignment problem. First, when the outside minutia is selected as a basis, the minutiae of the opposite side are located out of the hash table area. That is, the geometric hashing technique losses some outside information.

Second, most of the transformed minutiae are located at the center of the generated hash table because the minutiae transformed by the basis are superposed on the enrollment hash table. Fig. 1 shows the distribution of the generated hash table. As shown in Fig. 1, we can know that the most minutiae are concentrated in the center of the hash table. Therefore, the hash table results in a non-uniform, Gaussian distribution. Fig. 2 shows that is the geometric hash table generated by real minutiae only and chaff minutiae only, respectively.

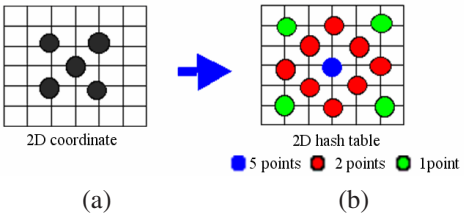


Fig. 1. An Illustration of the Center-Concentrated Geometric Hash Table. (a) An Example of Input Minutiae (b) A Geometric Hash Table generated with Fig. 1(a).

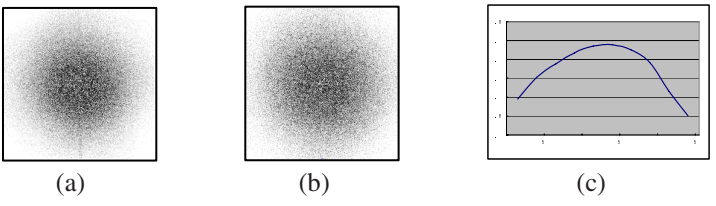


Fig. 2. Characteristics of the Geometric Hash Table; (a) Feature Distribution of the Geometric Hash Table generated by Real Minutiae. (b) Feature Distribution of the Geometric Hash Table generated by Chaff Minutiae. (c) Probability that the Selected Minutiae will be a Real (*i.e.*, $\text{probability}_{\text{real}}$ = # of real / # of (real + chaff)) in the Hash Table Domain.

Especially, Fig. 2(c) shows that the probability of a real(*i.e.*, probability that a selected minutia is a real), denoted as **probability$_{real}$**, varies with the location of the selected minutia. Because of this characteristic of the generated hash table, attackers may try to find the real minutiae starting from the center of the hash table. For the purpose of explanation, we denote this type of the attack considering the probability$_{real}$ as **geometric attack**. For example, the probability$_{real}$ at the center is 0.17, whereas that of the boundary is 0.09. Furthermore, in case of the *n*-degree fuzzy vault where the attacker tries to find *n*+1 real minutiae at once, the geometric attack becomes more effective. For example, in Table 2, the probability of finding 10 real minutiae in the degree-9 polynomial is $5.766 \times 10^{-9} (=0.15^{10})$ with the brute-force attack. However, the probability can be reduced to 2.015×10^{-8} $(=0.17^{10})$ with the geometric attack. Therefore, we need to generate uniformly distributed probabilities$_{real}$ in the hash table domain to protect the geometric attack.

Table 1. Effectiveness of the Geometric Attack (200 Chaff)

Type of Attack \ Degree-*N* Polynomial	Degree- 9 Polynomial	Degree-10 Polynomial
brute-force attack	5.766×10^{-9}	8.649×10^{-10}
geometric attack	2.015×10^{-8}	3.427×10^{-9}

3.2 A Proposed Algorithm for Adding Chaff Minutiae

Generally, the chaff minutiae are added into the fingerprint image after being randomly generated without any consideration of the locations. For example, Clancy, *et al.*[2] generated a predetermined number of the chaff minutiae and added them into the fingerprint image randomly such that the authentic user can unlock the vault efficiently. Recently, Chang, *et al.*[8] proposed a different approach to add the chaff minutiae for more secure feature distribution. That is, they added a maximum number of the chaff minutiae and then removed some near-by chaff minutiae such that the "free area" was minimized(*i.e.*, the chaff minutiae were distributed more uniformly). Although they ignored the complexity of the unlocking, their observation based on the notion of "free area" needs to be considered in adding the chaff minutiae for the secure, geometric hashing-based fuzzy fingerprint vault. Furthermore, the security level of the geometric hashing-based fuzzy fingerprint vault needs to be improved by making the feature distributions of the hash table generated by both the real and chaff minutiae similar, as we mentioned in Section 3.1.

In this paper, we propose an approach to add the chaff minutiae for enhancing the security level of the geometric hashing-based fuzzy fingerprint vault without degradation of the verification accuracy. For the purpose of explanation, we denote the possible attack which exploits the notion of "free area" as **Chang's attack**. Note that, the geometric hashing-based fuzzy fingerprint vault also needs to consider the Chang's

attack because the input fingerprint image can be derived back from the generated hash table.

We first describe our approach to resist to the geometric attack by making the feature distributions of the hash table generated by both the real and chaff minutiae similar, *i.e.*, probabilities$_{real}$(probability that a selected minutia is a real) are almost the same across the hash table domain. As we explained in Section 3.1, most real minutiae in the geometric hash table are concentrated at the center. Thus, most chaff minutiae in the geometric hash table need to be concentrated at the center, too. For the purpose of explanation, we define the notion of *sensor area*, *minutiae area* and *feature area*. The sensor area means the region of the fingerprint image captured by a fingerprint sensor, whereas the minutiae area is the region whose boundaries are determined by the most outside real minutiae in the fingerprint image. Finally, the feature area is the region where we can add the chaff minutiae.

Note that, we have two requirements in defining the feature area. The first requirement is that the feature distribution of the generated hash table domain for the chaff minutiae is similar to that of the real minutiae. One possible solution to this requirement is to set the boundary of the feature are to be equal to that of the minutiae area. However, in this solution, an attacker can easily find some real minutiae located at the boundary of the minutiae area. To hide these real minutiae, we need to define the feature area differently from the minutiae area. The second requirement for the feature area is that it should be enough for large numbers of chaff minutiae. Note that, the typical number of chaff minutiae(*i.e.*, > 200) is much larger than the typical real minutiae(*i.e.*, 30). Consequently, we need to set the feature area larger than the minutiae area. In this paper, we set the feature area larger than the minutiae area by the amount of ($2\times M$). Note that, M is defined as the **acceptable margin** and explained in the following.

Now, we consider the Chang's attack simultaneously. To determine the acceptable margin, we define some notations. S and R are defined as the total number of pixels in the feature area and the total number of minutiae, respectively. B is defined as the area of the **matching bound** which is computed by $(\Delta M+1)^2$. Note that, the newly added chaff minutiae should be located outside of the matching bound of any chaff or real minutiae in order to avoid both the Chang's attack and the possible degradation of the verification accuracy. Both S and R are determined by the given fingerprint, and S can be represented by $(\Delta x+4\Delta M)\times(\Delta y+4\Delta M)$ where Δx is the x-axis length of the minutiae area and Δy the y-axis length. Also, $S\approx B\times R$. Thus, $(\Delta x+4\Delta M)\times(\Delta y+4\Delta M)\approx(\Delta M+1)^2\times R$. By solving this equation, we can get the acceptable margin represented by Eq. (1). Then, the last step of our approach is to add the randomly generated chaff minutiae into the features area defined by the acceptable margin.

$$\Delta M = \frac{\sqrt{(2(\Delta x+\Delta y)-R)^2-(16-R)(\Delta x\Delta y-R)}-(R-2(\Delta x+\Delta y))}{R-16} \qquad \text{Eq. (1)}$$

Fig. 3 illustrates the feature area defined by the acceptable margin and the result of the uniform distribution of the chaff minutiae by applying the proposed approach.

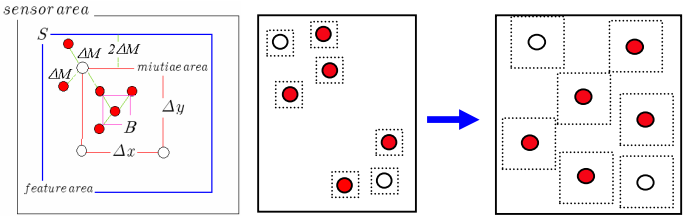


Fig. 3. Illustrations of the Feature Area and the Distribution of the Chaff and Real Minutiae. (a) Illustration of the Feature Area. (b) Non-uniform Distribution of the Chaff Minutiae. (c) Uniform Distribution of the Chaff Minutiae.

4 Experimental Results and Security Analysis

For the purpose of evaluating of the proposed approach, a data set of 400 fingerprint images composed of four fingerprint images per one finger was collected from 100 individuals by using the optical fingerprint sensor[9]. The resolution of the sensor was 500dpi, and the size of captured fingerprint images was 248×292. The average number of minutiae in our data set is 36, and the size of the enrollment minutiae hash table is 256×256.

Since the geometric attack is possible in the straightforward, geometric hashing-based fuzzy fingerprint vault, the security level against the brute-force attack may be degraded with the geometric attack. Fig. 4 shows the distributions of the probability$_{real}$, where the x-axis represents the geometric hash table and y-axis means the probability$_{real}$. Fig. 4(a) shows that the probability$_{real}$ was 0.17 at the center and 0.09 at the boundary of the hash table, and the geometric attack is effective in the straightforward approach. On the contrary, the probabilities$_{real}$ shown in Fig, 4(b) are distributed more uniformly so that the geometric attack is not effective in the proposed approach.

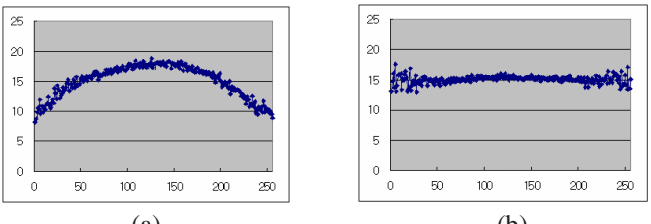


Fig. 4. Effectiveness of the Geometric Attack(200 Chaff). (a) Non-Uniform Distribution of the Probability$_{real}$ with the Straightforward Approach. (b) Uniform Distribution of the Probability$_{real}$ with the Proposed Approach.

Next, we show the resistance of the proposed approach to the Chang's attack. Fig. 5 shows the feature distribution in the input fingerprint image. Because the features in the fingerprint image with the proposed approach is more uniformly distributed than the straightforward approach, the Chang's attack is not effective in the proposed approach.

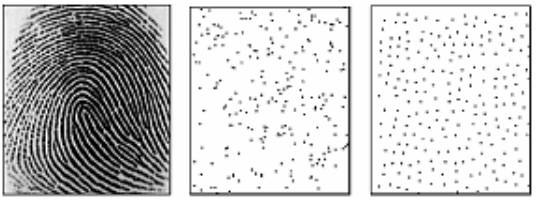


Fig. 5. Feature Distributions in the Input Fingerprint Image(200 Chaff). (a) Input Fingerprint Image. (b) Non-Uniform Distribution of the Features with the Straightforward Approach. (c) Uniform Distribution of the Features with the Proposed Approach.

5 Conclusions

In this paper, we focused on the fuzzy fingerprint vault based on the geometric hashing technique and analyzed the feature distribution and its security. In the previous implementation of the geometric hashing-based fuzzy vault, the probabilities of finding real minutiae from the hash table depend on the locations of the checked minutiae, and this type of geometric characteristics of the generated hash table can be utilized by attackers. We defined this type of attack as geometric attack.

To protect this geometric attack, we first analyzed the geometric transform from the real and chaff minutiae in "the fingerprint image domain" to "the generated hash table domain" and found that the feature distribution of the real minutiae in the hash table was different from that of the chaff minutiae. To make the two feature distributions in "the hash table domain" similar, we modified the way to add the chaff minutiae in "the fingerprint image domain" by considering both the relations among the real minutiae and the possible degradation of the fingerprint verification accuracy. Our approach also considered the feature distribution of both real and chaff minutiae in "the fingerprint image domain" such that the distribution was uniform.

Based on the experimental results, we confirmed that the proposed approach can be resistant to both the geometric attack and the Chang's attack and perform the fingerprint verification more securely without a significant degradation of the verification accuracy.

References

1. Juels, A., Sudan, M.: A Fuzzy Vault Scheme. In: Proc. of Symp. on Information Theory, p. 408 (2002)
2. Clancy, T., Kiyavash, N., Lin, D.: Secure Smartcard-based Fingerprint Authentication. In: Proc. of ACM SIGMM Multim., Biom. Met. & App., pp. 45–52 (2003)
3. Uludag, U., Pankanti, S., Jain, A.: Fuzzy Fingerprint Vault. In: Proc. of Workshop Biometrics: Challenges arising from Theory to Practice, pp. 13–16 (2004)
4. Yang, S., Verbauwhede, I.: Automatic Secure Fingerprint Verification System based on Fuzzy Vault Scheme. Proc. of IEEE 5, 609–612 (2005)
5. Uludag, U., Jain, A.: Securing Fingerprint Template: Fuzzy vault with Helper Data. In: Proc. of Conf. on Computer Vision and Pattern Recognition Workshop, pp. 163–172 (2006)

6. Chung, Y., et al.: Automatic Alignment of Fingerprint Features for Fuzzy Fingerprint Vault. In: Feng, D., Lin, D., Yung, M. (eds.) CISC 2005. LNCS, vol. 3822, pp. 358–369. Springer, Heidelberg (2005)
7. Wolfson, H., Rigoutsos, I.: Geometric Hashing: an Overview. IEEE Computational Science and Engineering 4, 10–21 (1997)
8. Chang, E., Shen, R., Teo, F.: Finding the original Feature Set Hidden among Chaff. In: Proc. of ASIACCS on Computer and Comm. Security, pp. 182–188 (2006)
9. Ahn, D., et al.: Specification of ETRI Fingerprint Database (in Korean). Technical Report – ETRI (2002)

Hand Geometry Verification Using Time Series Representation

Vit Niennattrakul, Dachawut Wanichsan, and Chotirat Ann Ratanamahatana

Department of Computer Engineering, Chulalongkorn University
Phayathai Rd., Pathumwan, Bangkok 10330 Thailand
{g49vnn,g49dwn,ann}@cp.eng.chula.ac.th

Abstract. Biometric authentication based on human physical traits has recently been heavily studied; these biometric sources include face, hand geometry, voice, fingerprint, iris, retina, etc. The hand geometry is one of the most conventional biometric since it is fairly easy to implement and acquire the data, comparing to other biometrics such as retina, iris, or DNA sequences. In this work, we propose a novel time series representation for hand geometry system by converting raw images into time series data, where this representation can gracefully handle variability of hand's position, translation, and rotation, especially in a peg-free system with the help of a Dynamic Time Warping similarity measure. We demonstrate the utility of our approach by implementting the real hand geometry verification/identification system, and it has proven to work effectively and competitively with low false acceptance and false rejection rates.

Keywords: Biometric, Hand geometry, Verification System, Time Series.

1 Introduction

At present, biometric authentication has been widely accepted and used in place of well-known traditional methods, such as password authentication and a use of key because passwords can be forgotten and keys can be misplaced or stolen. Therefore, many biometric authentication systems have been implemented [1], e.g., face, retina, DNA, voice, iris, fingerprint, hand geometry, etc. Each biometric has different strengths and weaknesses with characteristics categorized into seven categories [3], i.e., universality, uniqueness, permanence, collectability, performance, acceptability, and circumvention. Uniqueness is perhaps one of the most important factors in a good verification and identification system; however, it also comes with a cost. Hand geometry, on the other hand, may not achieve the highest score on uniqueness, but its strong advantages include the ease to acquire data, low-cost hardware, and convenience.

In general, most hand geometry verification systems have been designed and proposed to work in either one of the two environments:- a peg-fixed system [9][13] and a peg-free system [5][6][7][11][12]. The main difference between the peg-fixed system and the peg-free system is that in the peg-fixed system, users have to align

B. Apolloni et al. (Eds.): KES 2007/ WIRN 2007, Part II, LNAI 4693, pp. 824–831, 2007.
© Springer-Verlag Berlin Heidelberg 2007

their hand to abut against pins that are fixed on a plate, but in the peg-free system, users can place their hand freely on a surface. These systems usually utilize extracted geometric features from the hand, e.g., hand width, finger width, finger length, and fingertip regions [6][7][9][12] (shown in Figure 1).

In this paper, we propose a new representation of hand geometry; instead of several extracted features. We consider two techniques in converting shapes to time series, an angle-based technique [2] and a centroid-based technique [4]. To compare the similarity of two time series, we use Dynamic Time Warping distance measure with global constraint to control the warping path.

The rest of the paper is organized as follows. Section 2 describes our approach of converting raw images into time series data, and describes our similarity measurement and verification method. Section 3 discusses our evaluation method and shows the experimental results. Finally, in Section 4, we conclude our work and provide some suggestions.

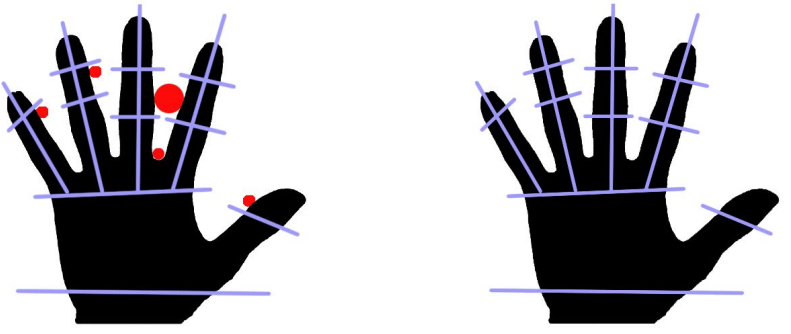


Fig. 1. (a) An image acquired from the peg-fixed system, and (b) an image acquired from the peg-free system

2 Methodology

This section will first describe techniques we use to transform raw hand images into time series representation, and then discuss our authentication process.

2.1 Time Series Conversion

To convert an image to time series data, we need three main pre-processing steps, i.e., brightness and contrast adjustment, binarization, and edge extraction, before time series data can be achieved. Figure 2 shows the block diagram describing the processes and the output in each stage.

Brightness and Contrast Adjustment. After an original color raw image is obtained, it will be transformed into a grayscale image. Then, the system will adjust its brightness and contrast to the quality that is suitable for the binarization step.

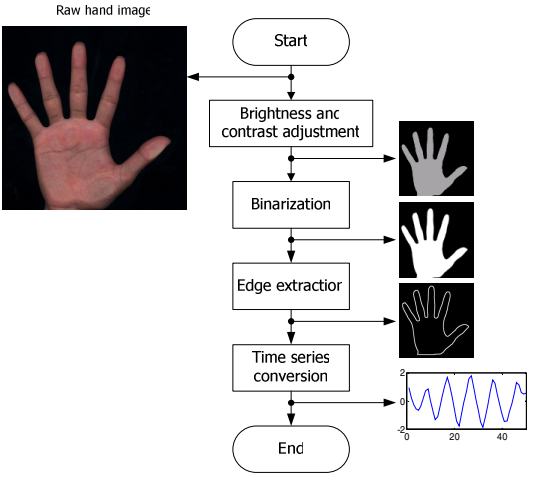


Fig. 2. A diagram describes process steps which converting raw images to time series

Binarization. After brightness and contrast adjustment, some image pixels may not be pure black or white. So, we binarize each pixel into solid black or white, represented by '0' and '1', respectively. The binarization function is defined as follows

$$B(x,y) = \begin{cases} 1 & if\ I_{xy} \geq t \\ 0 & \end{cases} \tag{1}$$

where I_{xy} is the intensity, which ranges from 0 to 1, at pixel (x, y) and t is the specific threshold for binarization. The default threshold value is set to 0.5.

Edge Extraction. The goal in this step is to find an edge sequence from a binarized hand image by using boundary extraction algorithm [2]. We consider two different approaches in this work. The traditional algorithm precisely starts the scanning of each pixel from the top left of the image to the bottom right of the image. Once it finds a first black pixel, it stops scanning, and then traces along the edge in a clockwise direction until it returns to the starting pixel. An example of an extracted edge is shown in Figure 3(a). Unlike the traditional algorithm, our proposed algorithm will not include the edge of the wrist at the bottom border of the image; our scanning starts from the bottom left to the right. After it has found the first pixel, it travels along the edge in a clockwise direction, and the algorithm stops when it touches the image boundaries again. Figure 3(b) shows an example of our extracted edge.

Time Series Conversion. In this step, we calculate the edge sequence, and transform hand's shape into time series data by using two techniques, i.e., an angle-based technique [2] and a centroid-based technique [4]. In the angle-based technique, for each pixel index i, we create two tangent lines – forward and backward tangents. The forward line is created by drawing the straight line from the pixel index i to the pixel



Fig. 3. Examples of extracted edge using (a) the original algorithm and (b) our proposal algorithm

index $i+\delta$, and the backward line is create by drawing the straight line from the pixel index i to the pixel index $i-\delta$. Note that the δ value depends on the size of the image; the larger the δ value the smoother the time series, and vice versa. In our experiment, the default value of δ is set to 10. After that we record the angle formed by these two lines as time series amplitude, as shown in Figure 4 (a). In the centroid-based technique, we first locate the hand's centroid. Once the centroid is obtained, we slightly plot the Euclidean distance from each pixel position to the centroid position. The general idea of centroid-based conversion is shown in Figure 4 (b).

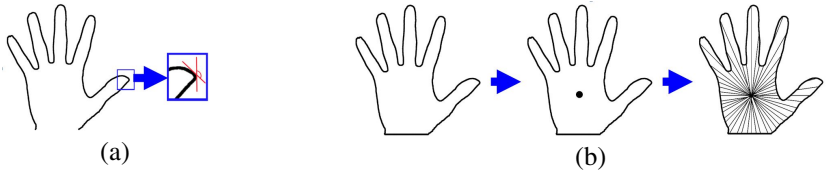


Fig. 4. The images show the general idea of (a) angle-based conversion technique and (b) controid-based conversion technique

2.2 Finding User's Threshold and Making Decision for Authentication

Finding User's Threshold (Training Phase). To set the threshold parameter to accept or reject a new hand image inquiry, we start collecting several hand images from a user. The Dynamic Time Warping distance measure is used to calculate the distance between the new hand image inquiry and every possible pairing of the user's templates. Once all the distance calculation is completed, the maximum value will be stored as the user's threshold.

Making Decision (Testing Phase). In verification step, once the system obtains a new hand image, it will transform the input image to time series, and will retrieve a set of that particular user's templates that he/she claims to be. After that, it will calculate the best distance between the converted time series and the set of time series templates. If this best distance is smaller than the weighted user's threshold, the access will be granted, or rejected otherwise.

Dynamic Time Warping. The Dynamic Time Warping (DTW) distance measure is used to determine the distance between two time series. Unlike the Minkowski distance function, DTW breaks the limitation of one-to-one alignment. It first finds all possible paths, and then selects the one that yields a minimum distance between the two time series using a distance matrix, where each element in the matrix is a cumulative distance of the minimum of the three surrounding neighbors. Suppose we have two time series, a sequence $Q = q_1, q_2, \ldots, q_i, \ldots, q_n$ and a sequence $C = c_1, c_2, \ldots, c_j, \ldots, c_m$. We first create an n-by-m matrix where every (i, j) element of the matrix is the cumulative distance of the distance at (i, j) and the minimum of the three elements neighboring the (i, j) element, where $0<i\leq n$ and $0<j\leq m$. We can define the (i, j) element as:

$$e_{ij} = d_{ij} + \min\{e_{(i-1)(j-1)}, e_{(i-1)j}, e_{i(j-1)}\} \tag{2}$$

where $d_{ij} = (c_i - q_j)^2$ and e_{ij} is (i, j) element of the matrix which is the summation between the squared distance of q_i and c_j, and the minimum cumulative distance of three elements surrounding the (i, j) element. Then, to find an optimal path, we have to choose the path that gives minimum cumulative distance at (n, m). The distance is defined as:

$$D_{DTW}(Q,C) = \min_{\forall w \in P}\left\{\sqrt{\sum_{k=1}^{K} d_{w_k}}\right. \tag{3}$$

where P is a set of all possible warping paths, and w_k is (i, j) at k^{th} element of a warping path and K is the length of the warping path. Due to space limitations, we refer interest readers to consult [8] for more details on DTW.

In reality, DTW may not give the best alignment that fits our purpose because it will try its best to find the minimum distance, and may generate unwanted paths. We can resolve this problem by limiting the warping path using a global constraint. A large number of researches have been adopting the Sakoe-Chiba Band [10] for the global constraint with the 10% warping window size (percentage of time series length) as a typical value. However, recent work has demonstrated that it is not always the case that using 10% band will result in best accuracies [8]. In this work, we allow the warping window size to vary from 0% to 100%.

3 Experimental Evaluation

To evaluate our proposed system, we collect hand images by using a color scanner from 22 people with 6 to 7 images for each person. The image resolution is set to 120 dpi, and its size is 1,200×1,200 pixels, as shown in Figure 5. After that, we transform all images to four time series types, i.e., two extraction methods and two conversion techniques. The length of the transformed data is between 3,000-5,000 data points, which appears to be quite oversampled. To reduce the processing time while

maintaining the crucial verification features, we could downsample our time series to as short as 50 data points. Figure 6 shows examples of different combinations of transformation methods.

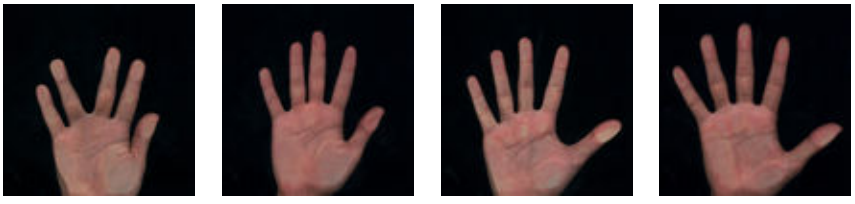


Fig. 5. Examples of four hand images from four different persons

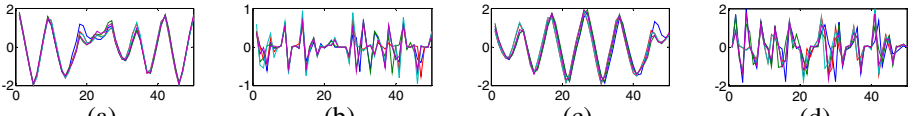


Fig. 6. Examples of four different combinations of time series transformation methods (downsampled to 50 data points) (a) traditional edge extraction with centroid-based conversion technique (b) traditional edge extraction with angle-based conversion technique (c) our proposed edge extraction with centroid-based conversion technique (d) our proposed edge extraction with angle-based conversion technique.

To evaluate the performance, we use three following error measures criteria:- False Acceptance Rate (FAR), False Rejection Rate (FRR), and Total Success Rate (TSR).

$$FRR = \frac{\#\text{RejectGenuineClaims}}{\text{Total}\#\text{GenuineAccess}} \times 100\% \tag{4}$$

$$FAR = \frac{\#\text{AcceptImposterClaims}}{\text{Total}\#\text{ImposterAccess}} \times 100\% \tag{5}$$

$$TSR = \left(1 - \frac{FAR + FRR}{\text{Total}\#\text{Access}}\right) \times 100\% \tag{6}$$

To evaluate this system, we vary three parameters, i.e., the time series transformation method, the width of the global constraint, and the weighted user's threshold. Table 1 shows the performance among four transforming combinations at EER point (Equal Error Rate - FAR≈FRR) and at the point where TSR is maximum. To further illustrate our overall system performance, we also show their ROC (Receiver Operating Characteristics) curves (shown in Figure 7.)

Table 1. The comparison of FAR, FRR, and TSR among several approaches

	Window size (%)	Weighted user's threshold	FAR (%)	FRR (%)	TSR (%)
Traditional Extraction + Centroid-based Technique (EER)	2	0.6	17.03	16.41	98.76
Traditional Extraction + Centroid-based Technique (max TSR)	14	0.55	5.39	21.09	99.01
Proposed Extraction + Centroid-based Technique (EER)	0	0.35	24.65	28.91	98.25
Proposed Extraction + Centroid-based Technique (max TSR)	8	0.5	5.39	41.41	98.26
Traditional Extraction + Angle-based Technique (EER)	2	0.65	26.45	31.25	98.25
Traditional Extraction + Angle-based Technique (max TSR)	0	0.6	0.78	99.22	97.85
Proposed Extraction + Angle-based Technique (EER)	8	0.7	29.14	30.47	97.78
Proposed Extraction + Angle-based Technique (max TSR)	0	0.6	25.23	28.91	97.99

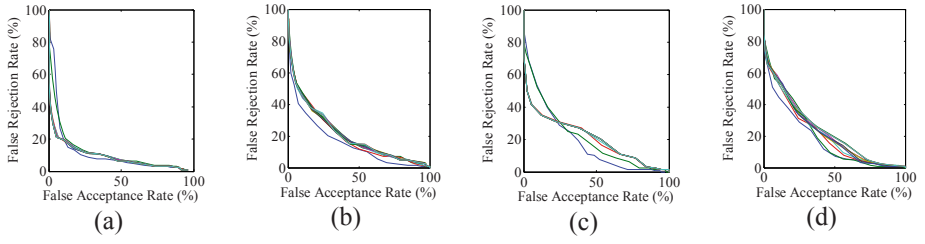


Fig. 7. ROC of (a) traditional edge extraction with centroid-based conversion technique (b) traditional edged extraction with angle-based conversion technique (c) our proposed edge extraction with centroid-based conversion technique (d) our proposed extraction with angle-based conversion technique.

For the best TSR, our system achieves 99.01% by using the traditional edge extraction and the centroid-based technique with 2% warping window and a weighted user's threshold of 0.55. An EER, the best FAR and FRR is approximately 16.50% by using the traditional edge extraction and the centroid-based technique with 2% warping window and a weighted user's threshold of 0.6. Note that the weighted user's threshold and the warping window size depends on the system environments such as image resolution, image size, and hand shapes in the database.

4 Conclusion

In this paper, we have demonstrated the utility of our novel time series representation for hand geometry system by converting raw images into time series data, where this representation can gracefully handle variability of hand's position, translation, and rotation, especially in a peg-free system with the help of a Dynamic Time Warping

similarity measure. We implement the real hand geometry verification system, and it has proven to work effectively and competitively with low false acceptance and false rejection rate.

We do hope that this work will be a good motivation for exploiting time series representation in a much wider range in biometric authentication such as face recognition, fingerprinting, and especially hand geometry which allows a large reduction in computational effort and storage space.

References

1. Jain, A., Bolle, R., Pankanti, S.: Biometrics Personal Identification in Networked Society: Personal Identification in Networked Society (1998)
2. Gandhi, A.: Content-based Image Retrieval: Plant species Identification. MS Thesis, Oregon State U. (2002)
3. Jain, A., Hong, L., Pankanti, S.: Biometric identification. Communication of the ACM 43, 90–98 (2000)
4. Keogh, E., Wei, L., Xi, X., Lee, S.-H., Vlachos, M.: LB_Keogh Supports Exact Indexing of Shapes under Rotation Invariance with Arbitrary Representations and Distance Measures. In: Proceedings of the 32nd VLDB, pp. 882–893 (2006)
5. Kumar, A., Wong, D.C., Shen, H.C., Jain, A.K.: Personal Verification Using Palm print and Hand Geometry Biometric. In: Kittler, J., Nixon, M.S. (eds.) AVBPA 2003. LNCS, vol. 2688, pp. 668–678. Springer, Heidelberg (2003)
6. Kumar, A., Wong, D.C., Shen, H.C., Jain, A.K.: Personal authentication using hand images. Pattern Recognition Letters 27(13), 1478–1486 (2006)
7. Ong, M.G.-K., Connie, T., Jin, A.T.-B., Ling, D.N.-C.: A Single-Sensor Hand Geometry and Palmprint Verification System. In: Proceedings of the 2003 ACM SIGMM Workshop on Biometrics Methods and Applications, pp. 100–106. ACM Press, New York (2003)
8. Ratanamahatana, C.A., Keogh, E.: Everything you know about Dynamic Time Warping is Wrong. In: Proceedings of 3rd SIGKDD Workshop on Mining Temporal and Sequential Data, at 10th ACM SIGKDD, ACM Press, New York (2004)
9. Sanchez-Reillo, R., Sanchez-Avila, C., Gonzalez-Marcos, A.: Biometric identification through hand geometry measurements. IEEE Transactions on Pattern Analysis and Machine Intelligence 22(10), 1168–1171 (2000)
10. Sakoe, H., Chiba, S.: Dynamic programming algorithm optimization for spoken word recognition. IEEE Transactions on Acoustics, Speech, and Signal Processing, 43–49 (1978)
11. Toh, K.A., Xiong, W., Yau, W.-Y., Jiang, X.: Combining Fingerprint and Hand-Geometry Verification Decisions. In: Kittler, J., Nixon, M.S. (eds.) AVBPA 2003. LNCS, vol. 2688, pp. 688–696. Springer, Heidelberg (2003)
12. Wong, A.L.-N., Shi, P.: Peg-Free Hand Geometry Recognition Using Hierarchical Geometry and Shape Matching. In: IAPR Workshop on Machine Vision Applications (2002)
13. Zunkel, L.R.: Hand Geometry Based Verification. In: Biometrics: Personal identification in networked society, pp. 87–101. Kluwer Academic Publishers, Dordrecht (1999)

Navigation Dynamics as a Biometrics for Authentication

Ernesto Damiani and Gabriele Gianini

Department of Information Technology - University of Milan
via Bramante 65, I-26013 Crema (CR) - Italy
{damiani,gianini}@dti.unimi.it

Abstract. In this work we propose a non-invasive behavioural biometrics based on human-computer interaction: we show that the time-texture of the user navigation within his/her own file-system or on the web, and of the develper's navigation within his/her IDE, can be used as an identity indicator, suitable to complement other traditional authentication biometrics.

Keywords: Biometrics, authentication, human-computer interaction, Markov chains, Bayesian methods.

1 Introduction

When a person talks, walks, drives a car or navigates the web, in her actions there are features which that person shares with other people, and features which, on the contrary, express the person's individuality, the latter, when suitably identified, could be considered as candidates metrics for identification or for authentication. Authentication (or identity verification) is the process through which a user claims who she is before the system accepts or declines her claim, whereas identification occurs when the system establishes the person identity, or fails to do it, without any prior claim.

In this paper we show that there are elements of individuality in the way a person navigates the web or her file system, and that some of them could be extracted from the corresponding time texture, by capturing the short range dynamics of the process through a Markov model; we propose to use this biometrics to strengthen user authentication.

The paper is organized as follows: in the next section we recall the definition and desirable features of a biometrics (Section 2); in the following (Section 3) we state the definition of Markov Chain (a.k.a. Markov model or Markov machine), and show first how a Markov model can be learned from sample sequences according to the Bayesian approach (this technique will be used the training phase of biometrics authentication) and, subsequently, how two Markov models learned from sample data can be compared (this technique will be used in the identity verification phase of biometrics authentication). Then we show the results of the application of these techniques to real data collected from different users (Section 4). Finally we discuss the characteristics of this form of personal time

B. Apolloni et al. (Eds.): KES 2007/ WIRN 2007, Part II, LNAI 4693, pp. 832–839, 2007.
© Springer-Verlag Berlin Heidelberg 2007

pattern characterization from the point of view of the desirable characteristics for a biometrics and propose to use this biometrics to strengthen user authentication in conjunction with other authentication techniques.

2 Biometrics for Authentication

A biometrics is a measurable biological (anatomical and physiological) or behavioural characteristic that can be used for automated recognition (for a survey see [1]). Biometrics are excellent candidates for identity verification mainly because, unlike keys or passwords, cannot be lost, stolen, or overheard.

Physiological static characteristics, such as fingerprints, palm print, hand geometry, fingernail bed, hand vein, iris or retinal patterns, ear shape and DNA, represent good biometrics because they are practically unique to individuals; furthermore they are relatively stable in time. Behavioural traits (e.g. signature, voiceprint, keystroke dynamics) on the other hand, have both some physiological and some psychological basis and are in general characterized by some remarkable variability; they form the basis of non-static biometrics systems.

A Biometrics system can be based on any physiological or behavioural characteristics as long as the following properties are fulfilled:

(i) *Universality* - Every person should have the characteristics.
(ii) *Uniqueness* - No two persons should be the same in terms of the biometrics.
(iii) *Permanence* - The characteristics should be relatively invariant with time.
(iv) *Collectability* - The characteristics must be quantitative and easy to acquire.
(v) *Performance* - The biometrics technique should have a good accuracy level.
(vi) *Acceptability* - The level of user acceptance of the system should be good.
(vii) *Circumvention* - The level of difficulty to forge an identification / authentication should be high.

Biometric technologies have been applied to a broad range of application fields such as: forensic science, financial and trade security, physical access control, information systems security. They are gaining popularity because they can be used together with traditional authentication methods to obtain an extra level of security.

The performance of a biometrics in terms of classification accuracy can be quantified in terms of ratios of correctly or incorrectly accepted or rejected candidates. The most commonly used accuracy metrics are the following rate error metrics:

(i) False Rejection Rate (FRR) - rate of accesses where a legitimate user is rejected by the system; it is given by the number of incorrectly rejected users divided by the total number of legitimate users who attempted to access the system.
(ii) False Acceptance Rate - rate of accesses where an impostor is accepted by the system; it is given by the number of incorrectly accepted users divided by the total number of impostors who attempted to access the system.

(iii) Equal error rate (EER) - the value at which FAR and FRR are equal; represents a standard measure of the trade-off between FRR and FAR.
(iv) Receiving operating curve (ROC) (the graph of FAR as a function of FRR); the whole set of possible trade-offs.

In this paper we propose a new behavioural biometrics technique based on the time texture of user interaction during common activities such as web navigation or IDE navigation. This biometrics is universal and relatively invariant with time; it can be acquired by installing a suitable plug-in in the working environment: after this operation the user can work undisturbed. We have also explored in practice the degree of uniqueness and the accuracy of the metrics and have found that its performance is pretty good for long training sets and in any case suitable to be used in conjunction with other authentication techniques.

3 Learning and Comparing Markov Models

In order to characterize the activity of a user based on the time texture of its activity – as it is visible from process logs – we focus on extracting the local, short-range, stochastic dynamics of the event lengths, which in our context are equivalent to inter-event waiting times. We model the process as a Markov Chain – that is, a process model where the probability of a transition of the system to a given state depends only on the states the system comes from, i.e. on the value of the previous states in the sequence (in this way the discovery of the short or medium range dynamics of the process is cast in terms of the learning of Markov Models of different orders, i.e. in terms of learning the corresponding transition matrices). As we will see the different states of our Markov model will be represented by different waiting times.

A given Markov model is characterized by its matrix of state transition probabilities, whose matrix elements p_{ij} represent the probability of the occurrence of a transition from state i to state j: in a first-order Markov models with r states this is a $(r \times r)$ square matrix, in an m-th order Markov model by an $r^m \times r$ matrix. To learn the matrix elements from the data we will adopt the Bayesian approach, which is more robust with respect to low statistics.

3.1 Bayesian Learning

In the Bayesian approach, learning a Markov model from a training sequence consists in using the sequence data for updating the probability of the full class of possible models from an *a priori* probability distribution to the *a posteriori* distribution. If the experimental data consist in a sample state sequence, the calculation of the *a-posteriori* probability density of the parameters can be carried on by counting the number of occurrences k_{ij} of that kind of transition: according to Bayes' theorem

$$P(p_{i1}, \ldots, p_{ir} | k_{i1}, \ldots, k_{ir}) = \frac{P(k_{i1}, \ldots, k_{ir} | p_{i1}, \ldots, p_{ir}) \cdot P(p_{i1}, \ldots, p_{ir})}{P(k_{i1}, \ldots, k_{ir})}$$

for $i = 1, \ldots, r^m$.

Assuming a multinomial forward probability setting and a uniform prior distribution, we have thus r^m Dirichlet posterior density distributions for the parameters

$$P(p_{i1},\ldots,p_{ir}|k_{i1},\ldots,k_{ir}) = \frac{\prod_{j=1}^{r} p_{ij}^{k_{ij}}}{\int_{\Omega}\prod_{j=1}^{r} p_{ij}^{k_{ij}}\,dp_{ij}} = \frac{\prod_{j=1}^{r}\Gamma(k_{ij}+1)}{\Gamma(n_i+r)}\cdot\prod_{j=1}^{r} p_{ij}^{k_{ij}}$$

for $i=1,\ldots,r^m$, where $n_i=\sum_j k_{ij}$, while Ω is the sub-region fulfilling the constraint $\sum_{j=1}^{r} p_{ij}=1$, and where $\Gamma(n)=(n-1)!$ is Euler's Gamma function.

3.2 Model Comparison

The problem of performing validation of a test sequence based on a given training sequence can be reduced to the problem of computing *the probability that training and test sequences come from the same model*. This can be done by quantifying the degree of superposition of the posterior densities of the model parameters for the $i-$th pair of states, by the scalar product

$$s_i = \int P(p_{i1},\ldots,p_{ir}|k'_{i1},\ldots,k'_{ir})P(p_{i1},\ldots,p_{ir}|k''_{i1},\ldots,k''_{ir})dp_{i1}\ldots dp_{ir}$$

where k'_{ij} and k''_{ij} represent the transition counts of the first and second sequence respectively. The above scalar product can be indicated by $s=<P',P''>$, with clear symbol interpretation. By substituting the Dirichelet densities and integrating we obtain

$$s_i = \frac{\prod_{j=1}^{r}\Gamma(k'_{ij}+1)\Gamma(k''_{ij}+1)}{\Gamma(n'_i+r)\Gamma(n''_i+r)}\,\frac{\prod_{j=1}^{r}\Gamma(k'_{ij}+k''_{ij}+1)}{\Gamma(n'_i+n''_i+r)}$$

where $n'_i=\sum_{j=1}^{r} k'_{ij}$ and $n''_i=\sum_{j=1}^{r} k''_{ij}$.

In [2] a superposition measure renormalized by the densities' Euclidean norms – which can synthetically be defined as follows $s^*=<P',P''>/(|P'||P''|)$ – has been used to obtain a value of the degree of superposition ranging from 0 (when there is no superposition) to 1 (when the two sequences, of the same length in that paper, are compatible with having been produced by the same model). The explicit expression for this superposition measure is

$$s_i^* = \left\{\frac{\Gamma(2n'_i+r)}{\prod_{j=1}^{r}\Gamma(2k'_{ij}+1)}\right\}^{\frac{1}{2}} \times \left\{\frac{\Gamma(2n''_i+r)}{\prod_{j=1}^{r}\Gamma(2k''_{ij}+1)}\right\}^{\frac{1}{2}} \times \left\{\frac{\prod_{i=1}^{r}\Gamma(k'_i+k''_i+1)}{\Gamma(n'_i+n''_i+r)}\right\}$$

The complete superposition of two Markov Models, is obtained by computing s_i^* for each of the state pairs, and then by multiplying them together $s^*=\prod_{i=1}^{r} s_i^*$.

The quantity s^* can be used to quantify the amount of correspondence the test instance and the training instance during the authentication process.

4 Results

We specialized the above general procedure, for model learning and comparison, to the case of the time texture, i.e. event-distance – in our case, equivalently event-length. Focussing on the event length we adopted the correspondence *event-length* $\leftrightarrow$ *Markov state*: given two subsequent events in a sequence, we identified the transition from an event length to another event length with a transition from a state to another state of the corresponding Markov machine.

Faced with the problem of characterizing the time dynamics in a compact way we adopted the choice of dividing the time lengths of the events into three categories, labelled *Short* (S), *Medium* (M) or *Long* (L): the time thresholds representing the boundaries between categories were decided on the basis of the user inter-event time distribution – the S label would be assigned to the first third of the events, the label M to the second and the label L to the third.

In this way each person's event-sequence time-texture could be captured by a long "'text"', consisting in the alternation of the three symbols S,M and L. From these sequences we intended to learn matrix elements of the corresponding three state Markov machine.

The learning and comparison of the Markov models described above were performed using data collected by means of non-intrusive software probes integrated in the working environment of a few users [3,4,5,6,7]. The probes collected the timing of every event (be that of a click on al link in the web browser or a location change in a Java development project) and all the relevant information associated with it, and available to the OS and the IDE. Every event was logged in a distinct XML file. We retained only the time-stamp information, in order to focus only on the time structure of the event-sequence, as described above.

We have been considering the log files produced by five users (three Java programmers and two generic users) working on their own machine and empirically counting the occurrences of each transition.

In order to validate our learning and comparison method, we randomly chose subsequences from the main sequence of each user; then calculated the measure of superposition s^* for pairs of subsequences coming from the same user and for pairs of subsequences coming from different users. For sake of simplicity we used training and test sequences of equal lengths. The lengths considered were of 10, 100, 1000 and 10000 symbols. Then for each length we histogrammed the values of s^* both for the case of same-user subsequence pairs and different user subsequence pairs. As expected with 10-symbol subsequences the two distributions would appear virtually undistinguishable.

However the two distributions appeared remarkably separated already with 100-symbol sequences: the measure s^* would yield values comparatively close to 1 when the two sub-sequences were coming from the same programmer, whereas it would yield values comparatively closer to 0 when the two sub-sequences were coming from two different programmers. The separation increased with the length of the sequences used for comparison.

This separation can be better appreciated with a logarithmic abscissa: in the figures we use the variable $u = -log_{10}(s^*)$ and plot the False Rejection Rate and

the False Acceptance Rate as a function of u: the FRR curve corresponds to the cumulative distribution function of the same-user distribution, whereas the FAR curve corresponds to the complementary cumulative distribution function of the different-user distribution. The figures show the FRR and FAR curves obtained by the comparison of pairs of first-order Markov models learned from of 100-symbols length subsequences (Fig.1(a)) and 1000-symbols length subsequences (Fig.1(b)) respectively.

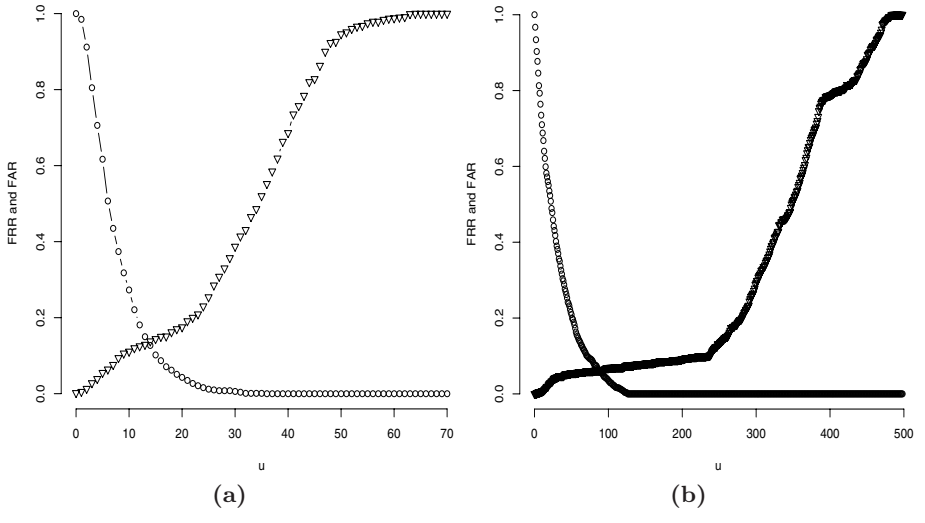


Fig. 1. FRR (circles) and FAR (triangles) as a function of u obtained by the comparison of pairs of first-order Markov models learned (a) from of 100-symbols length subsequences and (b) from of 1000-symbols length subsequences

We also made sequence comparisons using higher order Markov models. It is interesting to note that, even for sequences of length of 100 symbols, the separation of the two distributions slightly increases initially with the order of the Markov machine, but soon it deteriorates (approximately at order five the number of cells of the transition matrices becomes high enough to render the distinction of the different dynamics practically impossible). A synthesis of the results is given in Table 1 in terms of Equal Error Rate (EER) – i.e. in terms of the value where the FAR equals the FRR – for Markov models of order 1,2

Table 1. Equal Error Rate (ERR) for some representative sequence lengths and Markov model orders

Order of Markov model learned	1	2	3
EER for Subsequences of length 100	.128	.233	.750
EER for Subsequences of length 1000	.062	.069	.125

and 3. In synthesis, the comparison method proposed in the previous section looks effective in allowing the distinction of sub-sequences produced by the same or by a different developer already for sequences in the typical length of 100 symbols with the lower orders of Markov machines.

5 Summary and Discussion

In this paper we study the possibility of using user navigation dynamics as a biometrics for authentication. The dynamics – in our case the rhythmic patterns of the users interacting with her own working environment – can be captured by learning an event-length based Markov model from sequences of events (training phase); a model superposition measure can then be used for quantifying the degree of similarity of sequences (authentication phase).

One of the advantages of this superposition measure comes from its Bayesian character and lies in the fact that it can be used even when the statistics is very low, this allows the study of higher order Markov models - despite the fact that the number of parameters to be inferred grows exponentially with the order of the model - and therefore helps in taking into account more accurate details of the user navigation dynamics.

The approach displays, on the studied samples, a good power of separation, which however is not comparable, alone, with other high accuracy biometrics authentication techniques. Nonetheless, its inexpensive character and its irrelevant invasivity make it a good candidate to be used in conjunction with other authentication methods. Most of the methods currently considered for biometrics authentication are in fact independent from this one and could benefit from a joint utilization. Due to its progressive character, the method, as several other behaviour based methods, is suitable for a "'continuous authentication"', whereby a user continuously works on his/her environment: this method can detect the substitution of the authenticated user from the progressive degradation of the superposition measure.

This biometrics can be given a comprehensive evaluation using the standard criteria, mentioned above:

(i) *Universality* - Every person has the characteristics: indeed every user working inside an environment can produce sequences of some kind of events characterized by some time-stamp.
(ii) *Uniqueness* - No two persons should be the same in terms of the biometrics characteristics: we have shown that there is a good degree of separation between users in our restricted sample of users. We look forward to check that this desirable feature holds also for larger user samples.
(iii) *Permanence* - The characteristics should be essentially invariant with time: for sequences long enough the elements of the transition matrices look rather stable, as testified by the distribution of same-user subsequences, chosen randomly from the main sequence. Some degree of variability is intrinsically tolerated by the method.

(iv) *Collectability* - The characteristics should measurable quantitatively and easy to acquire: this is true for our method, as discussed above, because the method is quantitative and the data collection can be performed with a simple plug-in.
(v) *Performance* - The biometrics technique should have a good accuracy level: the accuracy level is good and improves with the length of the event-sequence considered, for sequences of few tens of events it is not completely satisfactory when used alone, however it can be used in conjunction with other authentication metrics.
(vi) *Acceptability* - The level of user acceptance of the biometrics system is good, indeed, as already mentioned, it is non invasive, and the privacy can be reduced at will.
(vii) *Circumvention* - The level of difficulty in order to forge an identification / authentication is high: indeed this is probably the strongest point about this metrics, it is virtually impossible to fake the personal rhythmic patterns of another person for a long time.

We plan to extend this work in the future by using a larger and more representative sample of users, and to investigate not only short-range correlations but also long-range correlations.

Acknowledgments. This work was partially supported by contract/grant sponsor FIRB research fund of MIUR, research project TEKNE grant n.RBNE05FKZ2).

References

1. Biometrics Glossary, `www.biometrics.org`
2. Colombo, A., Damiani, E., Gianini, G.: Discovering the software process by means of stochastic workflow analysis. Journal of Systems Architecture 52(11), 684–692 (2006)
3. Angioni, M., Sanna, R., Soro, A.: Defining a Distributed Agile Methodology for an Open Source Scenario. In: Scotto, M., Succi, G. (eds.) Proceedings of the First International Conference on Open Source Systems Genova, 2005, pp. 209–214 (2005)
4. Sillitti, A., Janes, A., Succi, G., Vernazza, T.: Collecting, Integrating and Analyzing Software Metrics and Personal Software Process Data. EUROMICRO, 336 (2003)
5. Sillitti, A., Janes, A., Succi, G., Vernazza, T.: Monitoring the Development Process with Eclipse. In: ITCC'04. International Conference on Information Technology: Coding and Computing, vol. 2, pp. 133–134 (2004)
6. Scotto, M., Sillitti, A., Succi, G., Vernazza, T.: A non-invasive approach to product metrics collection. Journal of Systems Architecture 52(11), 668–675 (2006)
7. Sillitti, A., Succi, G., De Panfilis, S.: Managing non-invasive measurement tools. Journal of Systems Architecture 52(11), 676–683 (2006)

A Cooperative Grid Computing Approach to Automated Theorem Finding and Automated Problem Proposing

Jingde Cheng, Yuichi Goto, Shinsuke Nara, and Takahiro Koh

Department of Information and Computer Sciences, Saitama University
338-8570 Saitama, Japan
{cheng,gotoh,nara,t_koh}@aise.ics.saitama-u.ac.jp

"In mathematics the art of proposing a question must be held of higher value than solving it."
– G. F. L. P. Cantor, 1867.

Abstract. This paper proposes a novel research direction: to build the Theory Grid as cooperatively shared formal theories within a virtual organization, and then to implement a crowd of automated reasoning programs working based on the Theory Grid such that as the wisdom of crowds a grid theorist has the ability to find new theorems and propose new questions automatically or semi-automatically. The paper also presents an architecture of the Theory Grid and grid theorists, and major scientific challenges and technical issues in the implementation of the Theory Grid and grid theorists.

Keywords: Theory grid, Grid theorists, Knowledge discovery, Knowledge creation, Knowledge sharing, Knowledge web.

1 Introduction

Wos proposed 33 basic research problems in automated reasoning [20, 21]. The 31st one is the problem of Automated Theorem Finding (**ATF** for short): "What properties can be identified to permit an automated reasoning program to find new and interesting theorems, as opposed to proving conjectured theorems?" The **ATF** problem is still completely open until now. The most important and difficult requirement of the problem is that, in contrast to proving conjectured theorems supplied by the user, it asks properties and/or criteria such that an automated reasoning program can use them to find some theorems in a field that must be evaluated by theorists of the field as new and interesting theorems. The significance of solving the problem is obvious because an automated reasoning program satisfying the requirement can provide great assistance for scientists in various fields. Note that the **ATF** problem requires general properties and/or criteria that an automated reasoning program should be obedient to, but not some concrete theorems in a special field. Therefore, any of those automated reasoning programs [11, 17, 18, 22] to discover or prove some concrete theorems in a special field by heuristic search has nothing to do with the **ATF** problem. The **ATF** problem also requires that an automated reasoning program can use the properties and/or criteria to find "new and

B. Apolloni et al. (Eds.): KES 2007/ WIRN 2007, Part II, LNAI 4693, pp. 840–851, 2007.
© Springer-Verlag Berlin Heidelberg 2007

interesting" theorems, but not those known or trivial ones. Finally, from the description of the **ATF** problem, we can see that it explicitly requires an individual automated reasoning program to find new and interesting theorems.

Any scientific discovery was not completely based on only one scientist's wisdom and experiences but must have achieved as a result of endeavors by many scientists. This observation should be really true in the world of today. On the other hand, the concept of grid in Grid Computing discipline is defined as the controlled and coordinated resource sharing and problem solving in dynamic, multi-institutional virtual organizations [12-15]. The Grid Computing technologies and/or activities involve networking services, connections, and communication of an unlimited number of resources within a virtual organization, and therefore, the Grid Computing has been refereed as the world's largest computer [12-15].

This paper considers a revised edition of the **ATF** problem: "What properties can be identified to permit a crowd of automated reasoning programs to find new and interesting theorems, as opposed to proving conjectured theorems?" We also consider the problem of Automated Problem Proposing (**APP** for short): "What properties can be identified to permit a crowd of automated reasoning programs to propose new and interesting questions?" We propose a novel research direction: to build the Theory Grid as cooperatively shared formal theories within a virtual organization, and then to implement a crowd of automated reasoning programs working based on the Theory Grid such that as the wisdom of crowds a grid theorist has the ability to find new theorems and propose new questions automatically or semi-automatically. We also present an architecture of the Theory Grid and grid theorists, and major scientific challenges and technical issues in the implementation of the Theory Grid and grid theorists.

2 Basic Notions and Notations

A ***formal logic system*** **L** consists of a formal language, called the ***object language*** and denoted by F(**L**), which is the set of all ***well-formed formulas*** of **L**, and a ***logical consequence relation***, denoted by meta-linguistic symbol $\vdash_{\mathbf{L}}$, such that for $P \subseteq$ F(**L**) and $c \in$ F(**L**), $P \vdash_{\mathbf{L}} c$ means that within the framework of **L**, c is a valid conclusion of premises P, or given P as premises, c as a valid conclusion follows from P. For a formal logic system (F(**L**), $\vdash_{\mathbf{L}}$), a ***logical theorem*** t is a formula of **L** such that $\phi \vdash_{\mathbf{L}} t$ where ϕ is the empty set. We use Th(**L**) to denote the set of all logical theorems of L. Th(**L**) is completely determined by the logical consequence relation $\vdash_{\mathbf{L}}$. According to the representation of the logical consequence relation of a logic, the logic can be represented as a Hilbert style axiomatic system, a Gentzen natural deduction system, a Gentzen sequent calculus system, or other type of formal system. A formal logic system **L** is said to be ***explosive*** if and only if $\{A, \neg A\} \vdash_{\mathbf{L}} B$ for any two different formulas A and B; **L** is said to be ***paraconsistent*** if and only if it is not explosive.

Let (F(**L**), $\vdash_{\mathbf{L}}$) be a formal logic system and $P \subseteq$ F(**L**) be a non-empty set of ***sentences*** (i.e. ***closed well-formed formulas***). A ***formal theory*** with premises P based on **L**, called a ***L-theory with premises P*** and denoted by $T_{\mathbf{L}}(P)$, is defined as $T_{\mathbf{L}}(P) =_{df}$ Th(**L**) $\cup$ $Th_{\mathbf{L}}^{e}(P)$, and $Th_{\mathbf{L}}^{e}(P) =_{df} \{et \mid P \vdash_{\mathbf{L}} et$ and $et \notin$ Th(**L**)$\}$ where Th(**L**) and $Th_{\mathbf{L}}^{e}(P)$ are called the ***logical part*** and the ***empirical part*** of the formal

theory, respectively, and any element of $\mathrm{Th}_{\mathbf{L}}^{e}(P)$ is called an ***empirical theorem*** of the formal theory. A formal theory $\mathrm{T}_{\mathbf{L}}(P)$ is said to be ***directly inconsistent*** if and only if there exists a formula A of **L** such that both $A \in P$ and $\neg A \in P$ hold. A formal theory $\mathrm{T}_{\mathbf{L}}(P)$ is said to be ***indirectly inconsistent*** if and only if it is not directly inconsistent but there exists a formula A of **L** such that both $A \in \mathrm{T}_{\mathbf{L}}(P)$ and $\neg A \in \mathrm{T}_{\mathbf{L}}(P)$. A formal theory $\mathrm{T}_{\mathbf{L}}(P)$ is said to be ***consistent*** if and only if it is neither directly inconsistent nor indirectly inconsistent. A formal theory $\mathrm{T}_{\mathbf{L}}(P)$ is said to be ***explosive*** if and only if $A \in \mathrm{T}_{\mathbf{L}}(P)$ for any $A \in \mathrm{F}(\mathbf{L})$; $\mathrm{T}_{\mathbf{L}}(P)$ is said to be ***paraconsistent*** if and only if it is not explosive. An explosive formal theory is not useful at all. Therefore, any meaningful formal theory should be paraconsistent. Note that if a formal logic system **L** is explosive, then any directly or indirectly inconsistent **L**-theory $\mathrm{T}_{\mathbf{L}}(P)$ must be explosive.

In logic, a sentence in the form of 'if ... then ...' is usually called a ***conditional proposition*** or simply ***conditional*** which states that there exists a relation of sufficient condition between the 'if' part and the 'then' part of the sentence. In general, a conditional must concern two parts which are connected by the connective 'if ... then ...' and called the ***antecedent*** and the ***consequent*** of that conditional, respectively. The truth of a conditional depends not only on the truth of its antecedent and consequent but also, and more essentially, on a necessarily relevant and conditional relation between them. The notion of conditional plays the most essential role in reasoning because any reasoning form must invoke it, and therefore, it is historically always the most important subject studied in logic and is regarded as the heart of logic. In a formal logic system with a primitive connective to represent the notion of conditional, a formula may have a structure of deeply nested conditional. With the depth of nested conditional, a formal logic system can be divided into different fragments about different degree of nested conditional.

For a formal logic system where the notion of conditional is represented by a primitive connective, say '$\Rightarrow$', a formula is called a ***zero degree formula*** if and only if there is no occurrence of $\Rightarrow$ in it; a formula of the form $A \Rightarrow B$ or $Q(A \Rightarrow B)$ where Q is the quantifier prefix of $A \Rightarrow B$ is called a ***first degree conditional*** if and only if both A and B are zero degree formulas; a formula A is called a ***first degree formula*** if and only if it satisfies the one of the following conditions: (1) A is a first degree conditional, (2) A is in the form $+B$ or $Q(+B)$ where + is a one-place connective such as negation and so on and Q is the quantifier prefix of $+B$ such that B is a first degree formula, and (3) A is in the form $B*C$ or $Q(B*C)$ where * is a non-conditional two-place connective such as conjunction or disjunction and so on and Q is the quantifier prefix of $B*C$ such that both of B and C is first degree formulas, or one of B and C is a first degree formula and another is a zero degree formula.

The notion of degree of nested conditional can also be generally defined as follows. Let A, B, C are formulas. The degree of conditional in A, denoted by $\mathrm{D}_{\Rightarrow}(A)$, is inductively defined as follows: (1) $\mathrm{D}_{\Rightarrow}(A) = 0$ if and only if there is no occurrence of $\Rightarrow$ in A, (2) if A is a conditional of the form $B \Rightarrow C$, then $\mathrm{D}_{\Rightarrow}(A) = \max\{ \mathrm{D}_{\Rightarrow}(B), \mathrm{D}_{\Rightarrow}(C) \} + 1$, (3) if A is in the form $+B$ where + is a one-place connective such as negation and so on, then $\mathrm{D}_{\Rightarrow}(A) = \mathrm{D}_{\Rightarrow}(B)$, (4) if A is in the form $B*C$ where * is a non-conditional two-place connective such as conjunction or disjunction and so on, then $\mathrm{D}_{\Rightarrow}(A) = \max\{ \mathrm{D}_{\Rightarrow}(B), \mathrm{D}_{\Rightarrow}(C) \}$, and (5) if A is in the form QB where Q is the

quantifier prefix of B, then $D_{\Rightarrow}(A) = D_{\Rightarrow}(B)$. Let k be a natural number. A formula A is called a ***k^{th} degree formula*** if and only if $D_{\Rightarrow}(A) = k$, in particular, A is called a ***k^{th} degree conditional*** if it is a conditional.

Let $(F(\mathbf{L}), \vdash_{\mathbf{L}})$ be a formal logic system and k be a natural number. The ***k^{th} degree fragment*** of $\mathbf{L}$, denoted by $Th^k(\mathbf{L})$, is a set of logical theorems of $\mathbf{L}$ which is inductively defined as follows (in the terms of Hilbert style axiomatic system): (1) if A is an axiom of $\mathbf{L}$ and $D_{\Rightarrow}(A) \leq k$, then $A \in Th^k(\mathbf{L})$, (2) if A is the result of applying an inference rule of $\mathbf{L}$ to some members of $Th^k(\mathbf{L})$ and $D_{\Rightarrow}(A) \leq k$, then $A \in Th^k(\mathbf{L})$, and (3) Nothing else are members of $Th^k(\mathbf{L})$, i.e., only those obtained from repeated applications of (1) and (2) are members of $Th^k(\mathbf{L})$. Note that the kth degree fragment of logic $\mathbf{L}$ does not necessarily include all k^{th} degree logical theorems of $\mathbf{L}$ because it is possible for $\mathbf{L}$ that deductions of some k^{th} degree logical theorems of $\mathbf{L}$ must invoke those logical theorems whose degrees are higher than k. On the other hand, the following holds obviously: $Th^0(\mathbf{L}) \subset Th^1(\mathbf{L}) \subset ... \; Th^{k-1}(\mathbf{L}) \subset Th^k(\mathbf{L}) \subset Th^{k+1}(\mathbf{L}) \subset ...$.

Based on the notion of degree of nested conditional and the notion of degree of logic fragment about conditional, we can introduce the notion of degree of formal theory about conditional.

Let $(F(\mathbf{L}), \vdash_{\mathbf{L}})$ be a formal logic system, $P \subset F(\mathbf{L})$ $(P \neq \phi)$, and k and j be two natural numbers. A formula A is said to be ***j^{th}-degree-deducible from P based on $Th^k(\boldsymbol{L})$*** if and only if there is a finite sequence of formulas $f_1, ..., f_n$ such that $f_n = A$ and for all i $(i \leq n)$, (1) $f_i \in Th^k(\boldsymbol{L})$, or (2) $f_i \in P$, or (3) f_i is the result of applying an inference rule to some members $f_{j1}, ..., f_{jm}$ $(j1, ..., jm < i)$ of the sequence and $D_{\Rightarrow}(f_i) \leq j$. The set of all formulas which are j^{th}-degree-deducible from P based on $Th^k(\mathbf{L})$ is called the ***j^{th} degree fragment of the formal theory with premises P based on $Th^k(\boldsymbol{L})$***, denoted by $T^j_{Th^k(\mathbf{L})}(P)$. A formula is said to be ***j^{th}-degree-deductive from P based on $Th^k(\boldsymbol{L})$*** if and only if it is j^{th}-degree-deducible from P based on $Th^k(\mathbf{L})$ but not $(j-1)^{th}$-degree-deducible from P based on $Th^k(\mathbf{L})$. Note that in the above definitions, we do not require $j \leq k$. The notion of jth-degree-deductive can be regarded as a metric to measure the difficulty of deducing an empirical theorem from given premises P based on logic $\mathbf{L}$. The difficulty is relative to the complexity of problem being investigated as well as the strength of underlying logic $\mathbf{L}$.

The notion of degree of nested conditional can be generally extended to other logic connectives and modal operators as follows: Let θ be an arbitrary n-ary $(1 \leq n)$ connective or modal operator of logic $\mathbf{L}$ and A be a formula of $\mathbf{L}$, the degree of θ in A, denoted by $D_\theta(A)$, is defined as follows: (1) $D_\theta(A) = 0$ if and only if there is no occurrence of θ in A, (2) if A is in the form $\theta(a_1, a_2, \ldots, a_n)$ where $a_1, a_2, \ldots, a_n$ are formulas, then $D_\theta(A) = \max\{ D_\theta(a_1), D_\theta(a_2), \ldots, D_\theta(a_n) \} + 1$, (3) if A is in the form $\sigma(a_1, a_2, \ldots, a_n)$ where σ is a connective or modal operator different from θ and $a_1, a_2, \ldots, a_n$ are formulas, then $D_\theta(A) = \max\{ D_\theta(a_1), D_\theta(a_2), \ldots, D_\theta(a_n) \}$, and (4) if A is in the form QB where B is a formula and Q is the quantifier prefix of B, then $D_\theta(A) = D_\theta(B)$.

The notion of degree of logic fragment about conditional can also be generally extended to other logic connectives and modal operators as follows: Let $\theta_1, \theta_2, \ldots, \theta_n$ be connectives or modal operators of logic $\mathbf{L}$ and $k_1, k_2, \ldots, k_n$ be natural numbers, the fragment of $\mathbf{L}$ about $\theta_1, \theta_2, \ldots, \theta_n$ and their degrees $k_1, k_2, \ldots, k_n$, denoted by $Th^{(\theta_1, k_1, \theta_2, k_2, \ldots, \theta_n, k_n)}(\mathbf{L})$, is a set of logical theorems of $\mathbf{L}$ which is inductively defined as follows

(in the terms of Hilbert style axiomatic system): (1) if A is an axiom of **L** and $D_{\theta 1}(A) \leq k_1$, $D_{\theta 2}(A) \leq k_2$, …, $D_{\theta n}(A) \leq k_n$, then $A \in Th^{(\theta_1, k_1, \theta_2, k_2, \ldots, \theta_n, k_n)}(\mathbf{L})$, (2) if A is the result of applying an inference rule of **L** to some members of $Th^{(\theta_1, k_1, \theta_2, k_2, \ldots, \theta_n, k_n)}(\mathbf{L})$ and $D_{\theta 1}(A) \leq k_1$, $D_{\theta 2}(A) \leq k_2$, …, $D_{\theta n}(A) \leq k_n$, then $A \in Th^{(\theta_1, k_1, \theta_2, k_2, \ldots, \theta_n, k_n)}(\mathbf{L})$, and (3) Nothing else are members of $Th^{(\theta_1, k_1, \theta_2, k_2, \ldots, \theta_n, k_n)}(\mathbf{L})$. Similarly, the notion of degree of formal theory about conditional can also be generally extended to other logic connectives and modal operators.

3 The Theory Grid and Grid Theorists

Conceptually, the ***Theory Grid*** is a formal theory infrastructure that coordinates various formal theories represented by and based on various formal logic systems in a distributed way using standard, open, general-purpose protocols and interfaces to meet demands of its application programs for theorem discovery and/or question proposition.

The concept of Theory Grid differs from the concepts of Computational Grid and Data Grid in that the Theory Grid concerns coordinated formal theory sharing for theorem discovering and/or question proposing. Unlike Computational Grid and Data Grid where computational power and data can be shared in a simple way of "put together", the Theory Grid must exactly manage formally and precisely defined relationships among various formal theories in order to show the foundation, grounds, and evidence of theorems and questions to its users. On the other hand, from the viewpoint of architecture, the Theory Grid should be built on the top of a computational and/or data grid computing environment. Therefore, if the intrinsic function of a computational and/or data grid computing environment is to provide its users with standard, open, general-purpose protocols and interfaces for computational power and/or data sharing, then the intrinsic function of the Theory Grid is to provide its users with foundation, grounds, and evidence of theorems and questions for formal theory sharing by using standard, open, general-purpose protocols and interfaces provided by a computational and/or data grid computing environment.

Conceptually, a ***grid theorist*** is a crowd of (an unlimited number of) automated reasoning programs working together on the Theory Grid in order to discover new and interesting theorems and/or propose new and interesting questions in a special field. A ***successful grid theorist*** should have the ability to find new and interesting theorems and/or propose new and interesting questions by a systematic method that can be learnt by other grid theorists.

Thus, the fundamental questions we have to answer are: How can we build the Theory Grid on a grid computing environment? Can we implement a successful grid theorist? How can we implement a successful grid theorist?

To implement the Theory Grid and grid theorists, we have designed and are implementing an architecture of the Theory Grid and grid theorists. Fig. 1 shows our architecture. The Theory Grid consists of two groups of components: the server group that includes a number of databases (knowledge-bases) and functional components to implement the Theory Grid, and the client group that includes an unlimited number of clients to behave grid theorists working on the Theory Grid. All databases (knowledge-bases), functional components, and clients are connected by soft system buses [6] and distributed over the Internet.

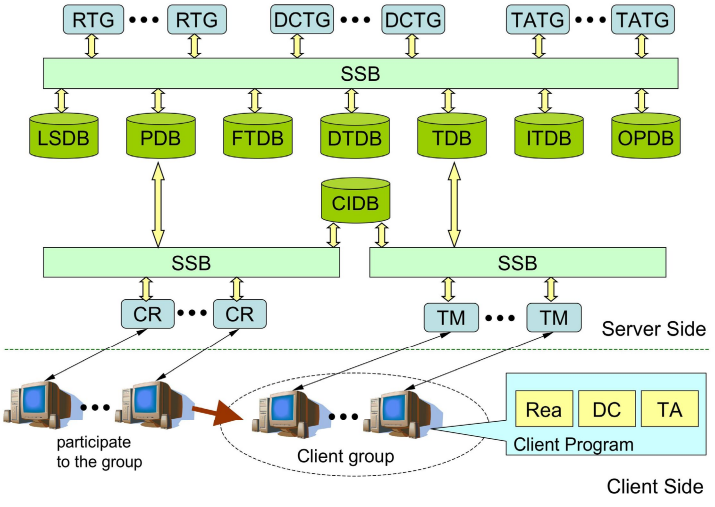


Fig. 1. An Architecture of the Theory Grid and Grid Theorists

The databases (knowledge-bases) and functional components of the server group are as follows:

Logic System Database (LSDB): The database managing various fragments of underlying logic systems and various inference rules.

Formal Theory Database (FTDB): The database managing previously known empirical premises, empirical theorems, and empirical inference rules of the target formal theory.

Open Problem Database (OPDB): The database managing previously known open problems.

Deduced Theorem Database (DTDB): The database managing new empirical theorems deduced by clients.

Intermediate Theorem Database (ITDB): The database managing those empirical theorems which are intermediate results need to be checked.

Client Information Database (CIDB): The database managing information about all clients.

Task Database (TDB): The database managing information about all reasoning tasks.

Project Database (PDB): The database managing information about the current project of **ATF**, e.g., a fragment of the underlying logic system, inference rules, degree of the nest of logic connectives, and so on, which is specified by an organization of **ATF**.

Client Register (CR): A client register CR registers clients such that it gets setting information, a fragment of the underlying logic system and inference rules from the Project Database PDB, sends a client reasoning program, its setting information, the fragment of the underlying logic system and the inference rules to a client, and sends information about registered clients to the Client Information Database CIDB. There may be a number of client registers implemented in a distributed environment.

Reasoning Task Generator (RTG): A reasoning task generator RTG generates reasoning tasks continuously and passes the tasks into Task Database TDB. A reasoning task is defined by premises, inference rules, and degree of the nest of logic connectives. There may be a number of reasoning task generators implemented in a distributed environment.

Duplication Check Task Generator (DCTG): A duplication check task generator DCTG generates duplication check tasks continuously and passes the tasks into Task Database TDB. A duplication check task is defined by an empirical theorem (or schema) and a set of intermediate theorems. There may be a number of duplication check task generators implemented in a distributed environment.

Theorem Analysis Task Generator (TATG): A theorem analysis task generator TATG generates theorem analysis tasks continuously and passes the tasks into Task Database TDB. A theorem analysis task is defined by a previously known theorem or a previously known open problem and a set of deduced theorems. There may be a number of theorem analysis task generators implemented in a distributed environment.

Task Manager (TM): A task manager TM manages reasoning tasks, duplication check tasks, and theorem analysis tasks such that it takes tasks from TDB, certifies clients, assigns tasks to clients, receives results of tasks from clients, and sends the results to TDB. There may be a number of task managers implemented in a distributed environment.

A client is basically a "worker" with a general-purpose reasoning engine and a general-purpose pattern matching checker [8]. An authenticated client may perform any of three types of tasks, i.e., reasoning tasks, duplication check tasks, and theorem analysis tasks. For a reasoning task, it receives the reasoning task from a task manager, reasons about empirical theorems of the target formal theory with specified the fragment of the underlying logic system, inference rules, and premises, and sends the deduced empirical theorems to the task manager. For a duplication check tasks, it receives the duplication check task from a task manager, checks whether the intermediate theorems are duplicated against previously deduced empirical theorems or not, and sends the check results to the task manager. For a theorem analysis task, it receives the theorem analysis task from a task manager, checks whether one of the deduced theorems is a previously known theorem (in this case, the theorem is rediscovered) or the solution of a previously known open problem (in this case, the problem is resolved) or not, and sends the analysis results to the task manager. The number of clients is not limited.

Thus, the cooperative grid computing approach to **ATF** and **APP** may consist of the following steps:

(1) The principal investigator (or the principal group of investigators) decides the underlying logic system at first, then prepares the various fragments of the underlying logic system, prepare various fragments of the target formal theory based on the underlying logic system, prepares databases (knowledge-bases) LSDB, FTDB, and OPDB, defines the current project, prepares databases PDB, and finally sends call for participation to the world.

(2) Any individual or organization wanting to contribute to the project submits registration data of computer(s) to a client register anytime anywhere and then gets all necessary setting information, the programs of a general-purpose reasoning engine

and a general-purpose pattern matching checker, the fragment of the underlying logic system, and the inference rules to work as a client or clients.

(3) Task generators, task managers, and clients work cooperatively to deduce empirical theorems by forward deduction or generate questions by induction and abduction, and present candidates of new empirical theorems and new questions to the principal investigator (or the principal group of investigators).

(4) For those candidates of new empirical theorems and new questions, the principal investigator (or the principal group of investigators) evaluates their significance and/or interestingness.

4 Scientific Challenges and Technical Issues

To implement the Theory Grid and grid theorists, and ultimately find new theorems and propose new questions automatically or semi-automatically, there are many scientific challenges and technical issues we have to solve.

The first scientific and theoretical fundamental problem that is intrinsically important to **ATF** and **APP** is to adopt what logic system as the fundamental logic system to underlie representing and reasoning about various formal theories. Although the concept of the Theory Grid is independent of underlying logic systems of formal theories, for the fundamental logic system to underlie representing and reasoning about various formal theories, at present, we consider that the best candidate is relevant (relevance) logics, in particular, strong relevant (relevance) logics which require the strong relevance between the premises and conclusion of a valid argument and reasoning, because they are really suitable and satisfactory to underlying reasoning and discovery with incomplete and sometime inconsistent knowledge [2, 3, 5]. While it is not suitable and satisfactory to adopt classical mathematical logic or any of its various classical conservatives as the fundamental logic system to underlie representing and reasoning about various formal theories because of the following facts: (1) any meaningful formal theory should be paraconsistent but not explosive, (2) there is no guarantee that any empirical theory must be complete and consistent (in fact, an empirical theory, in particular, a group of empirical theories, may be inconsistent in many ways), (3) if a formal logic system **L** is explosive, then any directly or indirectly inconsistent **L**-theory $T_{\mathbf{L}}(P)$ must be explosive, (4) classical mathematical logic and its various classical conservatives extensions are explosive but not paraconsistent, and (5) a reasoning based on any of classical mathematical logic and its various classical conservatives extensions is not necessarily relevant, i.e., the conclusion of the reasoning may be not relevant to its premises at all. However, the strong relevant logics themselves have some open problems and there are many challenges for us to apply the strong relevant logics and their various extensions to representing and reasoning about various formal theories.

The second scientific and theoretical fundamental problem is how to find hopeful candidates for new theorems and questions from infinite formulas. Since a formal theory $T_{\mathbf{L}}(P)$ is generally an infinite set of formulas, even though premises P are finite, we have to find some method to limit the range of candidates for new theorems and questions to a treatable finite set of formulas. The strategy we adopted is to sacrifice the completeness to get the treatable finite set of candidates. This is based on

Cheng's conjecture that almost all new theorems and questions of a formal theory can be deduced from the premises of that theory by finite inference steps concerned with finite number of low degree entailments [2, 3]. More formally and precisely speaking, we take $T^{j}_{Th^{k}(\mathbf{L})}(P)$, the jth degree fragment of the formal theory with premises P based on $Th^{k}(\mathbf{L})$, but not the whole of formal theory $T_{\mathbf{L}}(P)$ as the treatable finite set of candidates for new theorems and questions. Similarly, in the case concerning not only the notion of conditional but also other logic connectives and modal operators, we take the jth degree fragment of the formal theory with premises P based on $Th^{(\theta_1, k_1, \theta_2, k_2, \ldots, \theta_n, k_n)}(\mathbf{L})$ but not the whole of formal theory $T_{\mathbf{L}}(P)$ as the treatable finite set of candidates for new theorems and questions.

As the third scientific and theoretical fundamental problem, we have to consider how to provide foundation, grounds, and evidence of new theorems and questions. To build the Theory Grid on a grid computing environment, it is obviously necessary to generate and distribute various formal theories over the grid computing environment. However, from the viewpoint of **ATF** and **APP**, it is rather more important to formally define and precisely manage and update relationships among various formal theories in order to provide foundation, grounds, and evidence of new theorems and questions. We consider that the Theory Grid can be modeled as a finite semilattice such that its different elements represent different fragments of formal theories respectively, its partial order among elements represents the inclusion relation between two fragments of formal theories, and its minimum element represents a formal theory of the axiomatic set theory. There may be various types of the inclusion relation between two fragments of formal theories. For examples, a formal theory $T_{\mathbf{L}}(P)$ may be a subset of another formal theory $T_{\mathbf{L}}(P')$ if $P \subset P'$, a formal theory $T_{\mathbf{L}}(P)$ may be a subset of another formal theory $T_{\mathbf{L}'}(P)$ if $Th(\mathbf{L}) \subset Th(\mathbf{L}')$, a formal theory $T_{\mathbf{L}}{}^{k}(P)$ may be a subset of another formal theory $T_{\mathbf{L}}{}^{k+1}(P)$ if $Th^{k}(\mathbf{L}) \subset Th^{k+1}(\mathbf{L})$, the $(j-1)^{\text{th}}$ degree fragment of the formal theory with premises P based on $Th^{(\theta_1, k_1, \theta_2, k_2, \ldots, \theta_n, k_n)}(\mathbf{L})$ may be a subset of jth degree fragment of the formal theory with premises P based on $Th^{(\theta_1, k_1, \theta_2, k_2, \ldots, \theta_n, k_n)}(\mathbf{L})$, and so on. An intrinsic characteristic of **ATF** is that once a theorem is found by a process of **ATF**, there has been a proof of that theorem, i.e., the trace of finding the theorem. However, this is true only if we can certainly get a correct trace for any process of **ATF**. For a process of **ATF** in a single formal theory, it is not so difficult to record and backtrack a theorem finding trace, while for a process of **ATF** concerning a group of formal theories, it is very important to correctly define, manage, and update the partial order among the formal theories.

The last scientific and theoretical fundamental problem, probably the most difficult one in **ATF** and **APP**, is how to evaluate the interestingness of new theorems and questions. It is very difficult, if not impossible, to establish some general criteria to evaluate the interestingness of theorems and questions. The approach we adopted at present is to ask users themselves to evaluate the interestingness of new theorems and questions.

Besides the above scientific challenges, to implement the Theory Grid and grid theorists, and ultimately find new theorems and propose new questions automatically or semi-automatically, there are many technical issues we have to deal with.

In general, **ATF** by forward deduction has to deal with the following three types of process: reasoning process, duplication check process, and theorem analysis process.

A complete process of **ATF** by forward deduction is a repetitive sequence of the processes such that a reasoning process deduces, based on the underlying logic system, some theorems from given premises at first, then a duplication check process checks whether or not the theorems deduced by the reasoning process are really new to current premises, a theorem analysis process analyzes whether or not those new theorems are previously known theorems or solutions of open problems, and finally those new theorems are added to premises of the next reasoning process. Therefore, there are some dependence relationships among the processes. How to reduce the dependence relationships to a minimum is an important technical issue in designing and implementing a distributed model of **ATF** by forward deduction on a large scale grid computing environment.

An important feature of **ATF** by forward deduction is that in general a forward deduction has not an explicitly defined solution as its goal, and the number of intermediate theorems deduced in the deduction is unpredictable. Therefore, we cannot distribute all data and assign all tasks to computing nodes before the deduction but only can do that at run-time dynamically. This leads another important technical issue in designing and implementing a distributed model of **ATF** by forward deduction on a large scale grid computing environment: how to dynamically distribute data and assign tasks to computing nodes such that the efficiency of the whole process of **ATF** by forward deduction can be elevated to a maximum?

In fact, our architecture of the Theory Grid and grid theorists presented in Section 3 is a result of our considerations on the above important technical issues. Other technical issues we have to consider are as follows.

How can we build and update all databases (knowledge-bases) and maintain the consistency among them? The databases are essential elements of the Theory Grid and the consistency among them is the fundamental guarantee of correctness of new theorems found by grid theorists. Therefore, we have to collect knowledge as sufficient as possible at first, then formalize the knowledge as formal theories, and finally use the formal theories to construct all databases. We also have to find a way to check and maintain the consistency among the databases.

How can we dynamically create and maintain RTGs, DCTGs, and TATGs such that they have enough capacities for task generations? In the process of **ATF** based on our architecture of the Theory Grid and grid theorists, all tasks are dynamically created by RTGs, DCTGs, and TATGs. On the other hand, RTGs, DCTGs, and TATGs themselves also need to be created and maintained dynamically. Therefore, we have to find an effective way to dynamically create and maintain RTGs, DCTGs, and TATGs such that they work cooperatively and efficiently.

How can we dynamically create and maintain TMs such that they work cooperatively and efficiently? In the process of **ATF** based on our architecture of the Theory Grid and grid theorists, the TMs will play a crucial role because they are coordinators between servers and clients. The number of TMs, their processing powers, and their cooperation are one of the key issues to achieve the efficiency of the whole process where tens of thousands, or even millions of clients are working cooperatively. Therefore, we have to find an effective way to dynamically create and maintain TMs such that they work cooperatively and efficiently.

How can we define a priority among tasks such that a TM can use it to decide that which task should be assigned to a client? Because of dependence relationships

among tasks and differences between computing powers of clients, different performance orders of tasks may lead to greatly different effects on the efficiency of whole process of **ATF**. Therefore, we have to define a priority among tasks for TMs to decide performance orders of tasks dynamically.

How can we reduce the load of network to a minimum? When millions of clients are working together for a project of **ATF** based on the Theory Grid, the load of network may be huge. Therefore, we have to consider some way, e.g., data compression technology, to reduce the load of network to a minimum.

5 Concluding Remarks

Historically, scientific discovery is the most important subject studied in history of science and philosophy of science. From 1980s, it became a research subject of some cognitive scientists and computer scientists who believe that the process of a scientific discovery can be described and modeled in a normal way and therefore it can be simulated by computer programs automatically [16, 19].

Wos's **ATF** problem can be regarded as an attempt to find a systematic method in automated reasoning area. From the 1970s, there have been some works on mathematical theorem discovery and proof [11, 17, 18, 22]. But, as we have mentioned in Section 1, these works have nothing to do with the **ATF** problem.

A computational grid with distributed storage systems and execution platforms intends to provide users a powerful computing environment for large-scale computational tasks [12-15]. A data grid, usually constructed based on a computational grid, intends to provide users a global infrastructure for cooperative sharing of information in distributed data resources [12-15]. A knowledge-based grid intends to provide users a high-level, abstract reference architecture for knowledge discovery and data mining on data grids [1, 9, 10, 23].

The Theory Grid presented in this paper, which is a complete version of our early position papers [4, 7], intends to provide scientists a global, computational infrastructure for scientific discovery cooperatively performed on formal theories. The position of this paper in Computer Science as well as Artificial Intelligence is to show a novel research direction to find a systematic method for **ATF** and **APP** using the wisdom of crowds coordinately shared by grid computing technologies.

References

1. Berman, F.: From TeraGrid to Knowledge Grid. Communications of the ACM 44(11), 27–28 (2001)
2. Cheng, J.: A Strong Relevant Logic Model of Epistemic Processes in Scientific Discovery. In: Kawaguchi, E., Kangassalo, H., Jaakkola, H., Hamid, I.A. (eds.) Information Modelling and Knowledge Bases XI. Frontiers in Artificial Intelligence and Applications, vol. 61, pp. 136–159. IOS Press, Amsterdam (2000)
3. Cheng, J.: Automated Knowledge Acquisition by Relevant Reasoning Based on Strong Relevant Logic. In: Palade, V., Howlett, R.J., Jain, L. (eds.) KES 2003. LNCS, vol. 2773, pp. 68–80. Springer, Heidelberg (2003)

4. Cheng, J.: Theory Grid and Grid Theorists (Position Paper). In: Proc. IPSJ SIGGN 2005 Workshop on Groupware and Network Services. IPSJ Symposium Series, vol. 14, pp. 57–58 (2005)
5. Cheng, J.: Strong Relevant Logic as the Universal Basis of Various Applied Logics for Knowledge Representation and Reasoning. In: Kiyoki, Y., Henno, J., Jaakkola, H., Kangassalo, H. (eds.) Information Modelling and Knowledge Bases XVII. Frontiers in Artificial Intelligence and Applications, vol. 136, pp. 310–320. IOS Press, Amsterdam (2006)
6. Cheng, J.: Persistent Computing Systems Based on Soft System Buses as an Infrastructure of Ubiquitous Computing and Intelligence (invited paper). Journal of Ubiquitous Computing and Intelligence 1(1), 35–41 (2007)
7. Cheng, J., Nara, S., Koh, T., Goto, Y.: The Theory Grid and Grid Theorists. In: Proc. 2nd International Conference on Semantics, Knowledge and Grid, IEEE Computer Society Press, Los Alamitos (2006)
8. Cheng, J., Nara, S., Goto, Y.: FreeEnCal: A Forward Reasoning Engine with General-Purpose. In: Knowledge-Based Intelligent Information and Engineering Systems. 11th International Conference, KES 2007, Vietri sul Mare, Italy, 12-14 September, 2007. LNCS (LNAI), Springer, Heidelberg (2007)
9. Cannataro, M., Talia, D.: The Knowledge Grid. Communications of the ACM 46(1), 89–93 (2003)
10. Cannataro, M., Talia, D.: Semantics and Knowledge Grids: Building the Next-Generation Grid. IEEE-CS Intelligent Systems 19(1), 56–63 (2004)
11. Colton, S.: Automated Theory Formation in Pure Mathematics. Springer, Heidelberg (2002)
12. Foster, I.: What is the Grid? A Three Point Checklist. GRID Today 1(6) (2002)
13. Foster, I., Kesselman, C.: Globus: A Metacomputing Infrastructure Toolkit. International Journal of Supercomputer Applications and High Performance Computing 11(2), 115–128 (1997)
14. Foster, I., Kesselman, C. (eds.): The Grid: Blueprint for a New Computing Infrastructure. Morgan Kaufmann, San Francisco (1999)
15. Foster, I., Kesselman, C. (eds.): The Grid 2: Blueprint for a New Computing Infrastructure, 2nd edn. Morgan Kaufmann, San Francisco (2003)
16. Langley, P., Simon, H.A., Bradshaw, G.L., Zytkow, J.M.: Scientific Discovery Computational Explorations of the Creative Processes. MIT Press, Cambridge (1987)
17. Lenat, D.B.: AM: An Artificial Intelligence Approach to Discovery in Mathematics as Heuristic Search, Ph. D. Thesis. Stanford University (1976)
18. Quaife, A.: Automated Development of Fundamental Mathematical Theorems. Kluwer Academic Publishers, Dordrecht (1992)
19. Shrager, J., Langley, P.: Computational Approaches to Scientific Discovery. In: Shrager, J., Langley, P. (eds.) Computational Models of Scientific Discovery and Theory Formation, pp. 1–25. Morgan Kaufmann, San Francisco (1990)
20. Wos, L.: Automated Reasoning: 33 Basic Research Problems. Prentice-Hall, Englewood Cliffs (1988)
21. Wos, L.: The Problem of Automated Theorem Finding. Journal of Automated Reasoning 10(1), 137–138 (1993)
22. Wu, W.: Mathematics Mechanization: Mechanical Geometry Theorem-Proving, Mechanical Geometry Problem-Solving, and Polynomial Equations-Solving. Science Press / Kluwer Academic Publishers, Dordrecht (2000)
23. Zhuge, H.: The Knowledge Grid. World Scientific, Singapore (2004)

A Logic Grammar for Circuit Analysis
Problems of Recursive Definition

Takushi Tanaka

Department of Computer Science and Engineering
Fukuoka Institute of Technology
3-30-1 Wajiro-Higashi Higashi-ku, Fukuoka 811-0295, Japan
tanaka@fit.ac.jp
http://www.fit.ac.jp/~tanaka

Abstract. We have developed a logic grammar for circuit analysis. Circuits are viewed as sentences, and their elements as words. Circuit structures are defined by logic grammar called DCSG. A set of grammar rules, when converted into Prolog clauses, forms a logic program which execute top-down parsing. When we apply our methods to fundamental circuit analysis, problems of recursive definition occur. Recursive definitions often cause looping problems in logic programming. We will discusses several problems of recursive definition in circuit analysis.

Keywords: Logic Grammar, Circuit Analysis, Parsing Circuits, DCG, DCSG.

1 Introduction

Electronic circuits are designed as hierarchical structures of functional blocks. As the hierarchical structures were analogous to syntactic structures of language, we developed a logic grammar for parsing electronic circuits[4]. In the study, each circuit was viewed as a sentence and its elements as words. Structures of functional blocks were defined by logic grammar called DCSG (Definite Clause Set Grammar)[3]. Here, the DCSG is a DCG[2]-like logic grammar developed for analyzing word-order free languages.

Many grammar rules were defined for functional blocks of analog bipolar-IC, and given circuits were parsed as grammatical sentences[4]. But recursively defined grammar rules were not important in the study, because the bipolar-IC did not use recursive structures. Problems of recursive definition occurred when we apply our methods to fundamental circuit analysis. Recursive definitions often cause looping problems in logic programming. This paper first introduces DCSG, then discusses several problems of recursive definition in circuit analysis.

2 Definite Clause Set Grammar

2.1 How DCSG Differs from DCG

DCG (Definite Clause Grammar)[2] is a method for expressing context-free grammars in logic programming. A set of grammar rules, when converted into

B. Apolloni et al. (Eds.): KES 2007/ WIRN 2007, Part II, LNAI 4693, pp. 852–860, 2007.
© Springer-Verlag Berlin Heidelberg 2007

Prolog clauses, forms a logic program which performs top-down parsing. Inspired by the method of DCG, we developed DCSG, which is a definite clause grammar for word-order free languages. DCG uses difference lists of words to represent non-terminal symbols, because each non-terminal symbol generates a sub-string of a sentence, whereas DCSG uses difference sets (complementary sets) of words to represent non-terminal symbols. This is because each non-terminal symbol generates an order-free set of words in a word-order free sentence.

DCSG uses the predicates *subset* and *member* to convert grammar rules into Prolog clauses, but the differences between DCG and DCSG are very small. If we replace the predicate *subset* with *substring* and remove the clause (3) from the definition of *member* in Section 2.3, the conversion is equal to DCG conversion, although ordinary DCG does not use the predicate of *substring* for simplification.

2.2 Word-Order Free Language

A word-order free language **L(G')** is defined by modifying the definition of a formal grammar[3]. We define a context-free word-order free grammar **G'** to be a quadruple $< V_N, V_T, P, S >$ where: V_N is a finite set of non-terminal symbols, V_T is a finite set of terminal symbols, P is a finite set of grammar rules of the form:

$$A \longrightarrow B_1, B_2, ..., B_n. \quad (n \geq 1)$$
$$A \in V_N, \quad B_i \in V_N \cup V_T \quad (i = 1, ..., n)$$

and $S(\in V_N)$ is the starting symbol. The above grammar rule means rewriting a symbol A not with the string of symbols "$B_1, B_2, ..., B_n$", but with the set of symbols $\{B_1, B_2, ..., B_n\}$. A sentence in the language **L(G')** is a set of terminal symbols which is derived from S by successive application of grammar rules. Here the sentence is a multi-set which admits multiple occurrences of elements taken from V_T. Each non-terminal symbol used to derive a sentence can be viewed as a name given to a subset of the multi-set.

2.3 DCSG Conversion and Parsing

The general form of the conversion procedure from a grammar rule

$$A \longrightarrow B_1, B_2, ..., B_n. \quad (1)$$

to a Prolog clause is:

$$\begin{aligned} subset(A, S_0, S_n) \ :- \ & subset(B_1, S_0, S_1), \\ & subset(B_2, S_1, S_2), \\ & ... \\ & subset(B_n, S_{n-1}, S_n). \end{aligned} \quad (1)'$$

Here, all symbols in the grammar rule are assumed to be a non-terminal symbol. If "$[B_i]$"$(1 \leq i \leq n)$ is found in the right hand side of grammar rules, where

"B_i" is assumed to be a terminal symbol, then "$member(B_i, S_{i-1}, S_i)$" is used instead of "$subset(B_i, S_{i-1}, S_i)$" in the conversion.

The arguments $S_0, S_1, ..., S_n$ in $(1)'$ are multisets of V_T, represented as lists of elements. The predicate "*subset*" is used to refer to a subset of an object set which is given as the second argument, while the first argument is a name of its subset. The third argument is a complementary set which is the remainder of the second argument less the first; e.g. "$subset(A, S_0, S_n)$" states that "A" is a subset of S_0 and that S_n is the remainder.

The predicate "*member*" is defined by the following Prolog clauses (2) and (3). It has three arguments. The first is an element of a set. The second is the whole set. The third is the complementary set of the first.

$$member(M, [M|X], X). \tag{2}$$
$$member(M, [A|X], [A|Y]) :- \ member(M, X, Y). \tag{3}$$

When the clause (1)' is used in parsing, an object sentence (multiset of terminal symbols) is given to the argument S_0. In order to find the subset A in S_0, the first subgoal finds the subset B_1 in S_0 then put the remainder into S_1, the next subgoal finds B_2 in S_1 then put the remainder into S_2, ..., and the last subgoal finds B_n in S_{n-1} then put the remainder into S_n. That is, when a grammar rule is used in parsing, each non-terminal symbol in the grammar rule makes a new set from the given set by removing itself as its subset. While, each terminal symbol used in the grammar rule also makes a new set from the given set by removing itself as its member.

2.4 Backward Chaining and Topdown Parsing

Usually, a Prolog program consists of two kinds of definite clauses, called facts $\{F_1, F_2, ..., F_n\}$ and rules $\{R_1, R_2, ..., R_m\}$ both viewed as axioms. The execution of a Prolog program can be viewed as a process of deriving a theorem by backward chaining from the axioms. The top-down parsing of a word-order free sentence somewhat resembles the process of backward chaining. The object sentence is given as a set of terminal symbols $\{W_1, W_2, ..., W_n\}$. The starting symbol "S" is decomposed into terminal symbols using grammar rules $\{G_1, G_2, ..., G_m\}$ until they coincide with the given sentence. That is, the set of facts corresponds to the sentence, and the set of backward chaining rules corresponds to the set of grammar rules. Deriving theorems in backward chaining corresponds to identifying non-terminal symbols in the top-down method.

There is an important difference between backward chaining and top-down parsing. The backward chaining allows multiple use of the same fact to derive a theorem, while in a context-free language, each terminal symbol in a sentence contributes only once to the reduction of non-terminal symbols. This characteristic is very useful for avoiding a common looping problem in backward chaining, the problem which is caused by multiple use of the same fact.

3 Finding Structures on Circuits

3.1 Circuit Representation

The circuit $ca49$ in Figure 1 is represented by the word-order free sentence (4). The compound term $resistor(r1, 1, 3)$ as a word represents the resistor $r1$ connected to the nodes 1 and 3. The compound term $battery(b1, 1, 2)$ represents the battery $b1$ connected its positive terminal to the node 1 and its negative terminal to the node 2.

$$ca49([resistor(r1, 1, 3), resistor(r2, 2, 3), resistor(r3, 1, 4), \\ resistor(r4, 2, 4), resistor(r5, 3, 4), battery(b1, 1, 2)]). \quad (4)$$

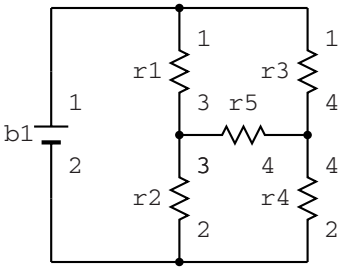


Fig. 1. Circuit $ca49$

3.2 Grammar Rules Without Recursion

As a resistor is a non-polar element, we want to refer the resistor regardless of its node order. The non-terminal symbol $res(R, A, B)$ can refer either $resistor(R, A, B)$ or $resistor(R, B, A)$. Here, DCSG allows the symbol ";" as abbrevation of two grammar rules with the same lefthand side. The non-terminal symbol $batt(...)$ enables to refer the same battery with different names. If a battery is referred with reverse node-order, the name of battery is prefixed by a minus symbol. The non-terminal symbol $anyElm(X, A, _)$ can refer any element X connected to the node A.

$$res(R, A, B) \longrightarrow [resistor(R, A, B)]; \\ [resistor(R, B, A)].$$

$$batt(E, A, B) \longrightarrow [battery(E, A, B)]. \\ batt(-E, A, B) \longrightarrow [battery(E, B, A)].$$

$$anyElm(X, A, B) \longrightarrow res(X, A, B); \\ batt(X, A, B).$$

3.3 All Elements Connected to a Node

In circuit analysis, Kirchhoff's current law (KCL) requires the sum of all branch currents into a node must be zero. In order to apply KCL on a specific node, we must find all elements connected to the node. The non-terminal symbol $allElm(X, A)$ successfully finds these elements connected to the node A. Here, X is substituted by a list of the all elements. If the same rules were written in bacward chaining by replacing the symbol "$\longrightarrow$" with "`:-`", a looping problem due to using the same data repeatedly would occur.

$$\begin{aligned} allElm([\,], A) &\longrightarrow not\ anyElm(_, A, _). \\ allElm([X|Y], A) &\longrightarrow anyElm(X, A, _), \\ &\quad\ allElm(Y, A). \end{aligned}$$

The following goal finds all elements connected to the node 1 in the circuit $ca49$ as:

$$?-\ ca49(CT), subset(allElm(X, 1), CT, _).$$

$$X = [r1, r3, b1]$$

3.4 Paths and Loops

According to Kirchhoff's voltage law (KVL), the sum of branch voltages along to a loop must be zero. In order to find loops in a circuit, we first define the non-terminal symbol $path(X, A, B)$ which finds routes between two nodes A and B.

$$\begin{aligned} path([X], A, B) &\longrightarrow anyElm(X, A, B). \\ path([X, B|Y], A, C) &\longrightarrow anyElm(X, A, B), \\ &\quad\ path(Y, B, C). \end{aligned}$$

The variable X in $path(X, A, B)$ is substituted by elements and nodes from the starting node A to the ending node B. The following goal attempts to find all routes from the starting node 1 to the ending node 2 in the circuit $ca49$.

$$?-\ ca49(CT), subset(path(X, 1, 2), CT, _).$$

$$\begin{aligned} &X = [b1]; \\ &X = [r1, 3, r2]; \\ &X = [r1, 3, r5, 4, r4]; \\ &X = [r1, 3, r5, 4, r3, 1, b1] \end{aligned}$$

The first answer $[b1]$ shows the path from the node 1 via element $b1$ to the node 2. The second answer shows the path from the node 1 via $r1$, node 3, and $r2$ to the node 2. The first three answers show paths from the node 1 to the node 2, but the fourth answer is not intended. It goes back to the starting node 1 then go to the ending node 2 via $b1$. As parsing sentence does not use the same word twice,

no element in the circuit appeared twice in the answer. But, the definition did not inhibit to use the same node twice.

As we want to get right answers which do not include loops in a path, we will modify the grammar rules as:

$$\begin{aligned} &path([X], A, B, _) \longrightarrow anyElm(X, A, B). \\ &path([X, B|Y], A, C, T) \longrightarrow anyElm(X, A, B), \\ &\qquad\qquad\qquad\qquad\quad quote\ \ not\ member(B, T, _), \\ &\qquad\qquad\qquad\qquad\quad path(Y, B, C, [A|T]). \end{aligned}$$

The "*quote*" in the grammar rule is a command to DCSG-converter. It directs the converter to insert the following strings as it is. Namely, "$not\ member(B, T, _)$," is inserted as a Prolog clause in the DCSG-conversion. The last argument of *path* is substituted by a list of nodes which are already in the path and not to be used twice as a relay node in the process of finding path.

The following goal successully finds all pathes from the node 1 to the node 2 in the circuit *ca*49. The last argument of *path* is substituted by "[2]" which shows the ending node 2 must not be used as a relay node.

$$? -\ \ ca49(CT), subset(path(X, 1, 2, [2]), CT, _).$$

$X = [b1];$
$X = [r1, 3, r2];$
$X = [r1, 3, r5, 4, r4];$
$X = [r3, 4, r4];$
$X = [r3, 4, r5, 3, r2];$
No

If the ending node is equal to the starting node, the goal enumelates all loops through the node as:

$$? -\ \ ca49(CT), subset(path(X, 1, 1, [\,]), CT, _).$$

$X = [r1, 3, r5, 4, r3];$
$X = [r1, 3, r5, 4, r4, 2, -b1];$
$X = [r1, 3, r2, 2, -b1];$
$X = [r1, 3, r2, 2, r4, 4, r3];$
$X = [r3, 4, r4, 2, -b1];$
...

4 Recursively Defined Circuits

4.1 Left Recursion

Recusively defined circuits were not appeared in the actual IC-designs[6], but the circuit theory has concepts of those circuits. When we define a series-parallel circuit, we meet a problem called left recursion. The following grammar rules define the circuit which consists only of series and parallel connections of resistors.

$$
\begin{aligned}
spCircuit(R, A, B) &\longrightarrow res(R, A, B).\\
spCircuit(sr(X, Y), A, C) &\longrightarrow spCircuit(X, A, B),\\
&\qquad spCircuit(Y, B, C),\\
&\qquad not\ anyElm(_, B, _).\\
spCircuit(pr(X, Y), A, B) &\longrightarrow spCircuit(X, A, B),\\
&\qquad spCircuit(Y, A, B).
\end{aligned}
$$

The first rule defines a single resistor R as a series-parallel circuit $spCircuit(R, A, B)$. The second rule defines a series connection of series-parallel circuits as a series-parallel circuit $spCircuit(sr(X, Y), A, C)$. Here, $sr(X, Y)$ is the name given to the circuit. The grammar rule has a condition "$not\ anyElm(_, B, _)$" that the central node B of the series connection must not be connected to other elements. The third rule defines a parallel connection of series-parallel circuits as a series-parallel circuit $spCircuit(pr(X, Y), A, B)$. "$pr(X, Y)$" is the name given to the circuit.

In order to identify a series-parallel circuit connected to nodes 1 and 2 in Figure 2, we attempt the following goal:

$$?-\ \ ca37(CT), subset(spCircuit(X, 1, 2), CT, _).$$

But the second goal loops. The top-down mechanism of DCSG decomposes the starting symbol into terminal symbols iteratively until it generates a set of terminal symbols which coincides with the object data set. As the series-parallel connection is defined by left recursive rules, the system infinitely decomposes the starting symbol with the same symbol when the generated elements do not coincide.

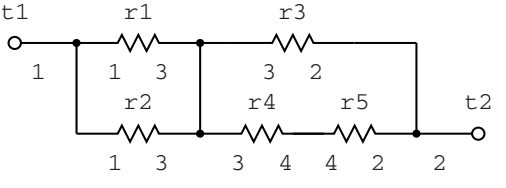


Fig. 2. Circuit ca37

4.2 Top-Down Parsing of Series-Parallel Circuits

In top-down parsing, we must avoid grammar rules with left recursion. In ordinary context-free grammars, this can be done by introducing additional non-terminal symbols to change rules into Greibach normal form[1]. Each grammar rule in the normal form first generates a terminal symbol then generate non-terminal symbols. This means that each terminal symbol is a component of some non-terminal symbols with other non-terminal symbols. So, we try to define the non-terminal symbol "series-parallel circuit" with a terminal symbol "resistor" and other circuits. If a resistor is found in a series-parallel circuit as a terminal symbol, the remaining circuit must be identified some non-terminal

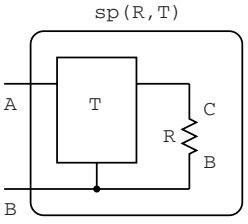


Fig. 3. Series-parallel circuit defined by three-terminals circuit

symbols. This can be realized by introducing a concept of three-terminals circuit as a non-terminal symbol.

Assume any resistor circuit connected to nodes A and B. If we remove a resistor R connected to node B from the circuit, the remaining circuit becomes a three-terminals circuit T as shown in Figure 3. If the original circuit is a series-parallel circuit, the three terminal circuit has one of the following structures shown in Figure 4. Figure 4-(A) is the simplest case which has no element in the three-terminal circuit. Figure 4-(B) has a series-parallel circuit X connected to the removed resistor in series. Figure 4-(C) has a series-parallel circuit X connected to the removed resistor in parallel.

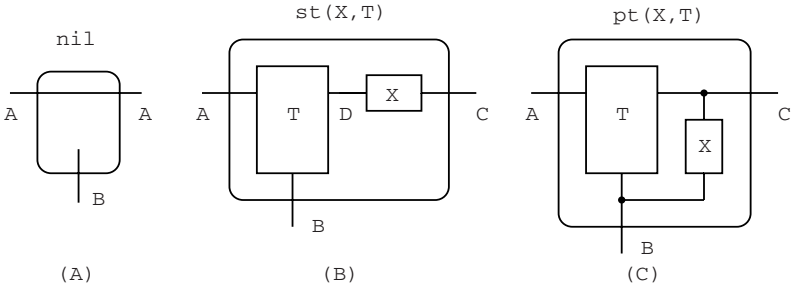


Fig. 4. Circuits with three terminals

Now we can write grammar rules which can execute top-down parsing of series-parallel circuits as:

$$
\begin{aligned}
&spCircuit(sp(R,T),A,B) \longrightarrow\ res(R,B,C).\\
&\qquad\qquad\qquad\qquad\quad\ tt(T,A,B,C),\\
&\qquad\qquad\qquad\qquad\quad\ not\ anyElm(_,C).\\
&tt(nil,A,B,A).\\
&tt(st(X,T),A,B,C) \longrightarrow\ spCircuit(X,C,D),\\
&\qquad\qquad\qquad\qquad\quad\ tt(T,A,B,D),\\
&\qquad\qquad\qquad\qquad\quad\ not\ anyElm(_,D).\\
&tt(pt(X,T),A,B,C) \longrightarrow\ spCircuit(X,B,C),\\
&\qquad\qquad\qquad\qquad\quad\ tt(T,A,B,C).
\end{aligned}
$$

The condition "*not anyElm*(_, C)" is added to the definition of *spCircuit*, because all elements connected to C except the resistor R must be identified as components of a three-terminal circuit T. The similar condition "*not anyElm*(_, D)" is added to the definition of $tt(st(X,T),A,B,C)$. Using these grammar rules, the circuit $ca37$ in Figure 2 is parsed as a series-parallel circuit:

$$?- \quad ca37(CT), subset(spCircuit(X,1,2), CT, _).$$

$$\begin{aligned} X = \quad & sp(r3, \\ & \quad pt(sp(r4, st(sp(r5, nil), nil)), \\ & \quad\quad st(sp(r1, pt(sp(r2, nil), nil)), nil))) \end{aligned}$$

5 Conclusions

The problems of finding all elements and finding routes in Section 3 show the advantage of topdown parsing than backward chaining. If the problems were formalized in backward chaining, looping problems due to using the same data repeatedly would occur.

In order to solve the problem of reft-recursion, we use the idea of Greibach normal form. As the normal form changes original grammar rules, the parse trees are different from original structures. Another method to solve this problem is to use bottom-up parsing. The bottom-up parsing corresponds to rewriting circuits into an equivalent one. This can be done by extending DCSG by introducing mechanism to write equivalent circuits. We are also extending DCSG by introducing semantic terms. Using the semantic terms, relations between circuit structures and functions are defined assuming the functions as their meaning.

References

1. Greibach, S.A.: A New Normal Form Theorem for Context-Free Phrase Structure Grammars. JACM 12, 42–52 (1965)
2. Pereira, F.C.N., Warren, D.H.D.: Definite Clause Grammars for Language Analysis. Artificial Intell. 13, 231–278 (1980)
3. Tanaka, T.: Definite Clause Set Grammars: A Formalism for Problem Solving. J. Logic Programming 10, 1–17 (1991)
4. Tanaka, T.: Parsing Circuit Topology in A Logic Grammar. IEEE-Trans. Knowledge and Data Eng. 5(2), 225–239 (1993)
5. Tanaka, T., Bartenstein, O.: DCSG-Converters in Yacc/Lex and Prolog. In: Proc. 12th International Conference on Applications of Prolog, pp. 44–49 (1999)
6. 101 Analog IC Designs: Interdesign Inc., Sunnyvale, CA (1976)

A Natural Deduction System for Annotated Predicate Logic

Seiki Akama[1], Kazumi Nakamtsu[2], and Jair Minoro Abe[3]

[1] Kawasaki, Japan
sub-akama@jcom.home.ne.jp
[2] University of Hyogo, Himeji, Japan
nakamatus@shse.u-hyogo.ac.jp
[3] Paulista University, Sao Paulo, Brazil
jairabe@uol.com.br

Abstract. Annotated logic is a paraconsistent logic for reasoning under incomplete and inconsistent information. This paper is a proof-theoretic investigation into annotated predicate logic by developing a natural calculus based on Prawitz style natural deduction system. We give a completeness proof.

Keywords: Annotated logics, natural deduction, completeness.

1 Introduction

Annotated logic, as proposed by Subrahmanian [13], is a paraconsistent logic for reasoning under incomplete and inconsistent information. By *paraconsistent logic*, we mean a logic for inconsistent but non-trivial theories. Annotated logic was originally devised in connection wih *paraconsistent logic programming*, but it is an interesting non-classical logics. Formal studies of annotated logic can be found in the papers by da Costa, Abe, Akama, and others; see da Costa et al. [9], [10], Abe [1], Abe and Akama [4], [5].

One of the intriguing research subjects is a proof-theoretic investigation. Although annotated logic was axiomatized by Hilbert formulation, Gentzen systems are also promising from the practical and formal point of view. This paper is a proof-theoretic investigation into annotated predicate logic by developing a natural calculus based on Prawitz style natural deduction system. We give a completeness proof by means of Henkin-like method.

2 Annotated Logic $Q\tau$

Recently, there is a growing interest for computer scientists in various non-classical logics, because classical logic is not suitable for a realistic representation of real world or common-sense reasoning. One of the challenging topics lies in the treatment of incomplete and inconsistent information. Annotated

B. Apolloni et al. (Eds.): KES 2007/ WIRN 2007, Part II, LNAI 4693, pp. 861–868, 2007.
© Springer-Verlag Berlin Heidelberg 2007

logic can be viewed as a new non-classical logic invented by computer scientist Subrahmanian [13] who proposed annotated logics as a foundation for paraconsistent logic programming for knowledge representation; also see Blair and Subrahmanian [7].

Now, we give the definitions of some notions used in paraconsistent logics. Let T be a theory whose underlying logic is $\mathcal{L}$. T is called *inconsistent* when it has the theorems of the form A and $\neg A$ (the negation of A). If T is not inconsistent, it is called *consistent*. T is called *trivial* if one can deduce from it any formula of its underlying language $\mathcal{L}$. Otherwise, T is said to be *non-trivial*. In trivial theories, the extensions of the concepts of formulas and theorems coincide. Thus, paraconsistent logic is a logic that can be used as the basis for inconsistent but non-trivial theories. A theory is called *paraconsistent* if its underlying logic is a paraconsistent logic. In paraconsistent logic, both A and $\neg A$ may be true for some A. Similarly, we can construct a *paracomplete logic* in which both A and $\neg A$ can be false for some A.

The symbol $\tau = \langle |\tau|, \leq, \sim \rangle$ indicates some fixed finite lattice called *lattice of truth-values*. We use the symbol $\leq$ to denote the ordering in which τ is a complete lattice, $\bot$ and $\top$ to denote, respectively, the bottom element and top element of τ. Also, $\wedge$ and $\vee$ indicate, respectively, the greatest lower bound and the least upper bound operators with respect to subsets of $|\tau|$. We also fix an operator $\sim : |\tau| \to |\tau|$, giving the "meaning" of the negation of the system $Q\tau$. The language $\mathcal{L}$ of $Q\tau$ has the following (denumerable) primitive symbols. Connectives: $\neg$ (negation), $\wedge$ (conjunction), $\vee$ (disjunction) and $\to$ (implication), quantifiers: $\forall$ (universal quantifier) and $\exists$ (existential quantifier), individual variables: $x, y, z, \cdots$, individual constants: $a, b, c, \cdots$, function symbols with arity $n > 0$: $f, g, h, \cdots$, annotated predicate symbols: $P_\lambda, Q_\mu, \cdots$, where $P, Q, \cdots$ are ordinary predicate symbols with arity n and $\lambda, \mu, \cdots \in |\tau|$, auxiliary symbols: (,).

Terms are defined as usual. *Formulas* are now defined as follows. (1) If P_λ is an annotated predicate symbol of arity n and $t_1, \cdots, t_n$ are terms, then $P_\lambda(t_1, \cdots, t_n)$ is a formula called an *annotated atom*. (2) A and B are formulas, then $\neg A, A \wedge B, A \vee B, A \to B$ are formulas (parentheses may be used to avoid ambiguities in the formula). (3) Only those expressions are formulas that are determined to be so by means of conditions (1) and (2).

$P_\lambda(t_1, \cdots, t_n)$ can be read "it is believed that $P(t_1, \cdots, t_n)$'s truth-value is at least λ". Let A be a formula. Then, $\neg^0 A$ is A, $\neg^1 A$ is $\neg A$, and $\neg^k A$ is $\neg(\neg^{k-1} A)$, where $k \geq 1$ is the natural number. The convention is also used for $\sim$. If $P_\lambda(t_1, \cdots, t_n)$ is an annotated atom, then the formula of the form $\neg^k P_\lambda(t_1, \cdots, t_n)$ is called a *hyper-literal*. A formula which is not a hyper-literal is called a *complex formula.*

Definition 1. *Let A and B be formulas. Then we put*

$$\begin{array}{ll} \textit{(equivalence)} & A \leftrightarrow B =_{def} (A \to B) \wedge (B \to A), \\ \textit{(strong negation)} & \neg_* A =_{def} A \to (A \to A) \wedge \neg(A \to A). \end{array}$$

The postulates (axiom schemata and primitive rules of inference) of $Q\tau$ are described below (cf Abe [1]). Here, A, B, and C are formulas whatsoever, F and

G are complex formulas, P is a predicate symbol, and λ, μ, λ_j are annotated constants.

Postulates for $Q\tau$

$(\rightarrow_1)$ $A \rightarrow (B \rightarrow A)$
$(\rightarrow_2)$ $(A \rightarrow (B \rightarrow C)) \rightarrow ((A \rightarrow B) \rightarrow (A \rightarrow C))$
$(\rightarrow_3)$ $((A \rightarrow B) \rightarrow A) \rightarrow A$
$(\rightarrow_4)$ $A, A \rightarrow B / B$
$(\wedge_1)$ $A \wedge B \rightarrow A$
$(\wedge_2)$ $A \wedge B \rightarrow B$
$(\wedge_3)$ $A \rightarrow (B \rightarrow (A \wedge B))$
$(\vee_1)$ $A \rightarrow A \vee B$
$(\vee_2)$ $B \rightarrow A \vee B$
$(\vee_3)$ $(A \rightarrow C) \rightarrow ((B \rightarrow C) \rightarrow ((A \vee B) \rightarrow C))$
$(\neg_1)$ $(F \rightarrow G) \rightarrow ((F \rightarrow \neg G) \rightarrow \neg F)$
$(\neg_2)$ $F \rightarrow (\neg F \rightarrow A)$
$(\neg_3)$ $F \vee \neg F$
(τ_1) $P_\perp(t_1, \cdots, t_n)$
(τ_2) $\neg^k P_\lambda(t_1, \cdots, t_n) \leftrightarrow \neg^{k-1} P_{\sim\lambda}(t_1, \cdots, t_n)$, where $k \geq 1$
(τ_3) $P_\lambda(t_1, \cdots, t_n) \rightarrow P_\mu(t_1, \cdots, t_n)$, where $\lambda \geq \mu$
(τ_4) $P_{\lambda_1}(t_1, \cdots, t_n) \wedge ... \wedge P_{\lambda_m}(t_1, \cdots, t_n) \rightarrow P_\lambda(t_1, \cdots, t_n)$, where $\lambda = \bigvee_{i=1}^{m} \lambda_i$
$(\forall_1)$ $\forall x A(x) \rightarrow A(t)$
$(\forall_2)$ $A \rightarrow B(x) / A \rightarrow \forall x B(x)$
$(\exists_1)$ $A(t) \rightarrow \exists x A(x)$
$(\exists_2)$ $A(x) \rightarrow B / \exists x A(x) \rightarrow B$

Note that $Q\tau$ is based on the positive classical predicate logic, we need $(\rightarrow_3)$ known as *Pierce's law* to get classical predicate logic.

Theorem 1. *In $Q\tau$, all valid formulas of the classical positive predicate logic are valid.*

Theorem 2. *In $Q\tau$, $\neg_*$ has all properties of the classical negation. For instance, we have the following. Here, $\vdash$ denotes the provability relation in $Q\tau$.*

$\vdash A \vee \neg_* A$, $\vdash \neg_*(A \wedge \neg_* A)$, $\vdash (A \rightarrow B) \rightarrow ((A \rightarrow \neg_* B) \rightarrow \neg_* A)$,
$\vdash A \rightarrow \neg_* \neg_* A$, $\vdash \neg_* A \rightarrow (A \rightarrow B)$.

We next give a semantics for $Q\tau$. An *interpretation* I for $\mathcal{L}$ consists of a non-empty set called the domain $dom(I)$ together with

1. Function α_I that maps constants of $\mathcal{L}$ to $dom(I)$.
2. Function β_I that assigns to each function symbol f of arity n in $\mathcal{L}$ a function from $(dom(I))^n$ to $dom(I)$.
3. Function γ_I that assigns to each predicate symbol of arity n in a function from $(dom(I))^n$ to $|\tau|$.

A *variable assignment* v for $\mathcal{L}$ with respect to I is a function from the set of variables of $\mathcal{L}$ to $dom(I)$. The denotation $d_{I,v}(t)$ of a term t with respect to an interpretation I and variable assignment v is defined as follows:

1. If t is a constant symbol, then $d_{I,v}(t) = \alpha_I(t)$.
2. If t is a variable, then $d_{I,v}(t) = v(t)$.
3. If t is of the form $f(t_1, \cdots, t_n)$, then $d_{I,v}(t) = \beta_I(f)(d_{I,v}(t_1), \cdots, d_{I,v}(t_n))$.

Let I and v be an interpretation for $\mathcal{L}$ and a variable assignment with respect to I. Let P is an atom, and F, G, H be any formula. We write $I, v \models A$ to state that a formula is true under I and v. Here are the definitions of $\models$:

1. $I, v \models P_\lambda(t_1, \cdots, t_n)$ iff $\gamma_I(P)(d_{I,v}(t_1), \cdots, d_{I,v}(t_n)) \geq \mu$.
2. $I, v \models \neg^k P_\lambda(t_1, \cdots, t_n)$ iff $I, v \models \neg^{k-1} P_{\sim\lambda}(t_1, \cdots, t_n)$
3. $I, v \models F \wedge G$ iff $I, v \models F$ and $I, v \models G$.
4. $I, v \models F \vee G$ iff $I, v \models F$ or $I, v \models G$.
5. $I, v \models F \rightarrow G$ iff $I, v \not\models F$ or $I, v \models G$.
6. $I, v \models \neg F$ iff $I, v \not\models F$, where F is not a hyper-literal.
7. $I, v \models \forall x H$ iff for all variable assignments v' such that for all variable y different from x, $v(y) = v'(y)$, we have $I, v' \models H$.
8. $I, v \models \exists x H$ iff for some variable assignments v' such that for all variable y different from x, $v(y) = v'(y)$, we have $I, v' \models H$.

We write $I \models H$ iff for all variable assignments v associated with I, $I, v \models H$. Let Γ be a set of formulas and H be a formula. We say that H is *valid* in $Q\tau$ iff for every interpretation I, $I \models H$. We say that H is a *semantic consequence* of Γ, writen $\Gamma \models H$ iff for any interpretation I such that $I \models G$ for any $G \in \Gamma$, $I \models F$. $Q\tau$ is shown to be sound and complete with this semantics as proved in da Costa et al. [10] and Abe [1], using the Hilbert system.

3 Natural Deduction System $NQ\tau$

Hilbert system is usually used when we want to define a new logical system. Unfortunately, Hilbert system is not practical. Thus, a logician prefer to other calculi. One of such proof methods is *natural deduction*. In this paper, by a natural calculus we mean a version of natural deduction system. It is known that a natural deduction system can be formalized by introduction and elimination rules. Sometimes, axioms may be used. For the terminology, the reader is referred to Prawitz [12]. In addition to standard rules for classical predicate logic, the natural calculus $NQ\tau$ needs extra rules for annotation.

Natural Deduction System $NQ\tau$

(Axiom)

$$(\tau_1)\frac{}{P_\perp(t_1, ...t_n)}$$

(Rules)

$$(\wedge I)\frac{A \quad B}{A \wedge B} \qquad (\wedge E)\frac{A \wedge B}{A} \quad \frac{A \wedge B}{B} \qquad (\vee I)\frac{A}{A \vee B} \quad \frac{B}{A \vee B} \qquad (\vee E)\frac{A \vee B \quad \overset{[A]}{C} \quad \overset{[B]}{C}}{C}$$

$$(\rightarrow I)\frac{\overset{[A]}{B}}{A \rightarrow B} \qquad (\rightarrow E)\frac{A \quad A \rightarrow B}{B} \qquad (P)\frac{\overset{[A \rightarrow B]}{A}}{A} \qquad (\neg I)\frac{\overset{[F]}{G \wedge \neg G}}{\neg G}$$

$$(\neg\neg E)\frac{\neg\neg F}{F} \qquad (\neg E)\frac{F \quad \neg F}{A} \qquad (\forall I)\frac{A(a)}{\forall x A(x)} \qquad (\forall E)\frac{\forall x A(x)}{A(t)}$$

$$(\exists I)\frac{A(t)}{\exists x A(x)} \quad (\exists E)\frac{\exists x A(x) \quad \begin{matrix}[A(a)]\\ B\end{matrix}}{B} \quad (\tau_2)\frac{\neg^k P_\lambda(t_1,\cdots,t_n)}{\neg^{k-1} P_{\sim\lambda}(t_1,\cdots,t_n)} \quad \frac{\neg^{k-1} P_{\sim\lambda}(t_1,\cdots,t_n)}{\neg^k P_\lambda(t_1,\cdots,t_n)}$$

$$(\tau_3)\frac{P_\lambda(t_1,\cdots,t_n) \;\; \lambda \geq \mu}{P_\mu(t_1,\cdots,t_n)} \quad (\tau_4)\frac{P_{\lambda_1}(t_1,\cdots,t_n)\cdots P_{\lambda_m}(t_1,\cdots,t_n) \;\; \lambda = \bigvee_{i=1}^{m} \lambda_i}{P_\lambda(t_1,\cdots,t_n)}$$

Here, in quantifier rules t is an arbitrary term and a is a term subject to the variable condition. For non-atomic level, $Q\tau$ is positive classical predicate logic. Thus, we need (P) a la Currry. In place of $(\neg\neg E)$, we can add the axiom of the form: $F \vee \neg F$.

4 Completeness Proof

The main result of this paper is to give a completeness proof of $Q\tau$ using the natural calculus $NQ\tau$. Although it is possible to have a purely syntactic proof, we here employ the semantics given above. It is here not difficult to show the equivalence of Hilbert and natural deduction versions of $Q\tau$.

Theorem 3. *(Soundness) If* $\vdash A$, *then* $\models A$.

Proof. It suffices to check that the axiom is valid and that each rule preserves validity. For classical rules, proofs are straightforward. The axiom (τ_1) is trivially valid via the semantics. We need to check (P). For (P): Suppose that $\models (A \to B) \to A$ and that $\not\models A$. From the former, we have that $\not\models A \to B$. This implies that $\models A$ and $\not\models B$. But, $\models A$ contradicts the assumption. Thus, we can conclude that there is no iterpretation I such that $I \models (A \to B) \to A$ and $I \not\models A$. For (τ_2): Trivial by definition of a hyper-literal. For (τ_3): Immediate from the property of lattice ordering. For (τ_4): Suppose that $\models P_{\lambda_1}(t_1,\cdots,t_n),\cdots,\models P_{\lambda_m}(t_1,\cdots,t_n)$ and $\lambda = \bigvee_{i=1}^{m} \lambda_i$ hold but $\not\models P_\lambda(t_1,\cdots,t_n)$. Now, pick up the supuremum annotation λ_i $(1 \leq i \leq n)$. Then, $\lambda = \lambda_i$. This implies that $\not\models P_{\lambda_i}(t_1,\cdots,t_n)$, contradicting the initial assumption.

For the completeness, we use the concept of *$Q\tau$-saturated set* of formulas. We say that a set of formulas Γ is *maximal non-trivial* iff it is trivial and $A \in \Gamma$ or $\neg A \in \Gamma$ for every formula A.

Definition 2. *A $Q\tau$-saturated set is a set of formulas Γ satisfying the following:*

(i) Γ is maximal non-trivial.
(ii) If $\Gamma \vdash A$ then $A \in \Gamma$.
(iii) If $\exists x A(x) \in \Gamma$, then $A(a) \in \Gamma$ for some constant a.

Definition 3. *A non-trivial set of formulas Γ can be extended to the $Q\tau$-saturated set of formulas with respect to the set of all formulas Form in $Q\tau$.*

Proof. We construct the sequence $\Gamma_0 = \Gamma, \Gamma_1, \cdots, \Gamma_n, \cdots$. Here, we expand the language $\mathcal{L}$ with the set $\mathcal{C}$ of constants new in $\mathcal{L}$ and set $\mathcal{L}' = \mathcal{L} \cup \mathcal{C}$. Let $A_1, A_n, \cdots$ be an enumeration of formulas of $Q\tau$. First, we check (i) in definition 6. Consider the case that n is even. Then, Γ_{i+1} is defined as follows:

$$\Gamma_{i+1} = \begin{cases} \Gamma \cup \{+A_{i+1}\} & \text{if } \Gamma_i \cup \{A_{i+1}\} \text{ is non-trivial} \\ \Gamma_i & \text{otherwise} \end{cases}$$

Here, each set of the sequence $\Gamma_0, \Gamma_1, \cdots$ is non-trivial and this is a non-decreasing sequence of sets such that $\Gamma_0 \subseteq \Gamma_1 \subseteq \cdots$. Then, we set:

$$\Gamma^* = \bigcup_{i=0}^{\infty} \Gamma_i$$

We can see that Γ^* is a maximally non-trivial set containing Γ. In fact, each finite subset of Γ^* must be non-trivial since Γ_i is non-trivial. It follows that Γ^* itself is non-trivial.

Now, we claim that Γ^* is a maximal non-trivial set. For suppose $A \notin \Gamma^*$ and $\neg A \notin \Gamma^*$. As A is a formula in $Q\tau$, it must appear in the above enumeration, say as A_k. If $\Gamma \cup \{A_k\}$ were non-trivial, then our construction would guarantee that $A_{k+1} \in \Gamma_{k+1}$, and hence that $A_k \in \Gamma^*$. From $A_k \notin \Gamma^*$, we have that $\Gamma_k \cup \{A_k\}$ is trivial. Thus, $\Gamma^* \cup \{A_k\}$ is trivial. It follows that Γ^* is a maximal non-trivial set. As $\Gamma \subseteq \Gamma_i$ $(i \in \omega)$, the following holds.

$$\Gamma \subseteq \Gamma^* = \bigcup_{i=0}^{\infty} \Gamma_i$$

On the other hand, suppose that Γ^* is trivial. Then, $\Gamma^* = Form$ follows. Thus, we have that $P_\lambda(t_1, \cdots, t_n), \neg_* p_\lambda(t_1, \cdots, t_n) \in \Gamma^*$. As $\mid \tau \mid$ is finite, any application of natural deduction rules has only finite number of premises. Thus, there are $n, m < \omega$ such that $P_\lambda(t_1, \cdots, t_n) \in \Gamma_n$ and $\neg_* P_\lambda(t_1, \cdots, t_n) \in \Gamma_m$. Therefore, $P_\lambda(t_1, \cdots, t_n), \neg_* P_\lambda(t_1, \cdots, t_n) \in \Gamma_{n_0}$, where $n_0 = max(n, m)$. Thus, Γ_{n_0} is trivial and we reach a contradiction. (ii) is straightforward by $(\rightarrow I)$ and maximality of Γ^*. For (iii), consider the enumeration in which n is odd. Let $\exists x A(x)$ be the first formula of the enumeration of existential formulas not considered before satisfying $\Gamma_n \vdash \exists x A(x)$. If we assume that a is a constant not in Γ_n, then $\Gamma_{n+1} = \Gamma_n \cup \{A(a)\}$. Otherwise $\Gamma_{n+1} = \Gamma_n$. Now, suppose $\exists x A(x) \in \Gamma^*, A(a) \notin \Gamma^*$. By construction, $\Gamma^* \vdash A(a) \rightarrow B$. By $(\exists E)$, $\Gamma^* \vdash \exists x A(x) \rightarrow B$. A contradiction. Thus, (iii) is verified. It is therefore shown that Γ^* is $Q\tau$-saturated.

Lemma 1. *Let Γ be a $Q\tau$-saturated set of formulas. Then, the following hold.*

(1) $A \wedge B \in \Gamma$ iff $A \in \Gamma$ and $B \in \Gamma$.
(2) $A \vee B \in \Gamma$ iff $A \in \Gamma$ or $B \in \Gamma$.
(3) $A \rightarrow B \in \Gamma$ iff $A \notin \Gamma$ and $B \in \Gamma$.
(4) $\neg_* A \in \Gamma$ iff $A \notin \Gamma$.

(5) $\neg^k P_\lambda(t_1, \cdots, t_n) \in \Gamma$ *iff* $\neg^{k-1} P_{\sim\lambda}(t_1, \cdots, t_n) \in \Gamma$.
(6) If $P_\lambda(t_1, \cdots, t_n), P_\mu(t_1, \cdots, t_n) \in \Gamma$, *then* $P_\theta(t_1, \cdots, t_n) \in \Gamma$, *where* $\theta = max(\lambda, \mu)$.
(7) $\exists x A(x) \in \Gamma$ *iff* $A(a) \in \Gamma$ *for some* a.
(8) $\forall x A(x) \in \Gamma$ *iff* $A(a) \in \Gamma$ *for all* a.

Proof. (1)-(4) are the same as classical cases. (5) is trivial by the definition of $Q\tau$-saturated set. For (6), suppose $P_\lambda(t_1, \cdots, t_n), P_\mu(t_1, \cdots, t_n) \in \Gamma$. By (1), we have $P_\lambda(t_1, \cdots, t_n) \wedge P_\mu(t_1, \cdots, t_n) \in \Gamma$. But, by (τ_4), $P_\theta(t_1, \cdots, t_n) \in \Gamma$, where $\theta = max(\lambda, \mu)$. For (7), one direction is immediate from the definition of $Q\tau$-saturated set. The other direction can be obtained by $(\exists I)$. (8) is treated similarly. We finish to check the $Q\tau$-saturatedness.

Definition 4. *Let the language of FN be* $\mathcal{L}$, $\mathcal{L}'$ *be* $\mathcal{L} \cup \mathcal{C}'$, *and* $Var_\mathcal{L}$ *and* $\mathcal{C}_\mathcal{L}$ *be the set of variables and of constants in* $\mathcal{L}$. *Then, a canonical interpretation* I^* *for a* $Q\tau$*-saturated set* Γ *of formulas is defined as follows:*

1. $dom(I^*) = \mathcal{C}_{\mathcal{L}'}$
2. $\alpha_{I^*}, \beta_{I^*}, \gamma_{I^*}$ *are defined above.*
3. *The variable assignment* v^* *is a function from* $Var_{\mathcal{L}'}$ *to* $dom(I^*)$.
4. $d_{I^*,v^*}(t)$ *is defined above.*
5. χ_Γ *is a charcteristic function from Form to* $\mathbf{2} = \{0, 1\}$.

Let *Atom* be a set of annotated atom. Then, we define the function $I_*^{v_*}$ from *Atom* to $|\tau|$ satisfying $I_*^{v_*}(P(t_1, \cdots, t_n)) = \bigvee\{\mu \in |\tau| \colon P_\mu(t_1, \cdots, t_n) \in \Gamma\}$. Next, we define the function $Val_{I_*}^{v_*}$ from *Form* to $\mathbf{2}$. It suffices to show that $Val_{I_*}^{v_*} = \chi_\Gamma$. Assume $P_\mu(t_1, \cdots, t_n) \in \Gamma$. Then, we have that $\chi_\Gamma(P_\mu(t_1, \cdots, t_n)) = 1$. On the other hand, it is clear that $\gamma_{I_*}(P)(d_{I_*,v_*}(t_1), \cdots, d_{I_*,v_*}(t_n)) \geq \mu$. So, $Val_{I_*}^{v_*}(P_\mu(t_1, \cdots, t_n)) = 1$. If $P_\mu(t_1, \cdots, t_n) \notin \Gamma$, then $\chi_\Gamma(P_\mu(t_1, \cdots, t_n)) = 0$. It is also not the case that $\gamma_{I_*}(P)(d_{I_*,v_*}(t_1), \cdots, d_{I_*,v_*}(t_n)) \geq \mu$, we have that $P_{\gamma_{I_*}(P)(d_{I_*,v_*}(t_1), \cdots, d_{I_*,v_*}(t_n))}(t_1, \cdots, t_n) \in \Gamma$, which is a contradiction. Thus, it is not the case that $\gamma_{I_*}(P)(d_{I_*,v_*}(t_1), \cdots, d_{I_*,v_*}(t_n)) \geq \mu$. Hence, we conclude that $Val_{I_*}^{v_*}(P_\mu(t_1, \cdots, t_n)) = 0$.

By Lemma 1 (5), $\neg^k P_\mu(t_1, \cdots, t_n) \in \Gamma$ iff $\neg^{k-1} P_{\sim\mu}(t_1, \cdots, t_n) \in \Gamma$. Thus, $\chi_\Gamma(\neg^k P_\mu(t_1, \cdots, t_n)) = \chi_\Gamma(\neg^{k-1} P_{\sim\mu}(t_1, \cdots, t_n))$ holds. We have to prove that $Val_{I_*}^{v_*}(\neg^k P_\mu(t_1, \cdots, t_n)) = \chi_\Gamma(\neg^k P_\mu(t_1, {}^c dots, t_n))$ by induction on k. If $k = 0$, then the proposition is obvious. Here, assume that it holds for $k-1 (k \geq 1)$. Then, $\chi_\Gamma(\neg^k P_\mu(t_1, \cdots, t_n)) = \chi_\Gamma(\neg^{k-1} P_{\sim\mu}(t_1, \cdots, t_n)) = Val_{I_*}^{v_*}(\neg^{k-1} P_{\sim\mu}(t_1, \cdots, t_n))$ $= Val_{I_*}^{v_*}(\neg^k P_\mu(t_1, \cdots, t_n))$.

Theorem 4. *Let* Γ *be* $Q\tau$*-saturated set of formulas. Then, we have:*

$\chi_\Gamma(A) = 1$ *iff* $A \in \Gamma$ *and* $\chi_\Gamma(A) = 0$ *iff* $A \notin \Gamma$.

Proof. For annotated atom and hyper-literal, it is clear from the above definition. The cases that A is any formula can be easily proved. If A is of the form $\neg B$, where B is a complex formula. So, $\chi_\Gamma(B) = Val_{I^*}^{v_*}(B)$. If $A \in \Gamma$, then $B \notin \Gamma$. Then, we have that $\chi_\Gamma(A) = 0$ and $\chi_\Gamma(B) = 1$. But, $Val_{I_*}^{v_*}(A) = 1 - Val_{I_*}^{v_*}(B)$. Thus, $Val_{I_*}^{v_*}(A) = 0$ follows. The other cases present no difficulty.

Theorem 5. *(completeness) If* $\models A$, *then* $\vdash A$.

5 Conclusion

We have presented a natural calculus $NQ\tau$ for annotated predicate logic $Q\tau$ with a completeness proof. Our proof uses Henkin method based on infinite construction of $Q\tau$-saturated set of formulas. Several topics remain to be worked out. First, we should be able to prove normal form theorem giving a syntactic completeness proof. Second, a natural deduction system could be extended with the rules of comprehension schema to formalize naive set theory,. Third, it would be interesting to present a completeness proof of annotated logics with infinite lattices. In the previous works on annotated logics, completeness proofs rely on finite lattices or infinite lattices with the *finite annotation property* (cf. [9]). However, for infinite lattice of truth-values, some form of the so-called ω-*rule* is needed to establish completeness even for propositional lelvel. The fourth issue is a formulation of $Q\tau$ based on *sequent calculus*. If the cut-elimination theorem can be proved, another syntactic completeness proof is available.

References

1. Abe, J.M.: On the Foundations of Annotated Logics (Portuguese). Ph.D. Thesis, University of São Paulo, Brazil (1992)
2. Abe, J.M.: Some recent applications of paraconsistent systems to AI. Logique et Analyse 157, 83–96 (1997) (published in 2000)
3. Abe, J.M.: Curry algebras $P\tau$. Logique et Analyse 161-162-163, 5–15 (1998) (published in 2001)
4. Abe, J.M., Akama, S.: Annotated logics $Q\tau$ and ultraproducts. Logique et Analyse 160, 335–343 (1997) (published in 2000)
5. Abe, J.M., Akama, S.: On some aspects of decidability of annotated systems. In: Arabnia, H.R. (ed.) Proc. of the International Conference on Artificial Intelligence, pp. 789–795 (2001)
6. Akama, S., Abe, J.M.: Many-valued and annotated modal logics. In: Proc. of the IEEE 28th International Symposium on Multiple-Valued Logic, Fukuoka, Japan, pp. 114–119. IEEE Computer Society Press, Los Alamitos (1998)
7. Blair, H., Subrahmanian, V.: Paraconsistent logic programming. Theoretical Computer Science 68, 135–154 (1989)
8. da Costa, N.: On the theory of inconsistent formal system. Notre Dame Journal of Formal Logic 15, 443–454 (1974)
9. da Costa, N., Subrahmanian, V., Vago, C.: The paraconsistent logic Pτ. Zeitschrift für mathematische Logik und Grundlagen der Mathematik 37, 139–148 (1991)
10. da Costa, N., Abe, J., Subrahmanian, V.: Remarks on annotated logic. Zeitschrift für mathematische Logik und Grundlagen der Mathematik 37, 561–570 (1991)
11. da Costa, N., Lu, J., Subrahmanian, V., Henschen, L.: Automatic theorem proving in paraconsistent logics: foundations and experiment results. In: Proc. of the 10th CADE, pp. 72–86. Springer, Berlin (1989)
12. Prawitz, D.: Natural Deduction: A Proof-Theoretical Study. Almqvist and Wiksell, Stockholm (1965)
13. Subrahmanian, V.: On the semantics of quantitative logic programs. In: Proc. of the 4th IEEE Symposium on Logic Programming, pp. 173–182. IEEE Computer Society Press, Los Alamitos (1987)

An Intelligent Coordinated Traffic Signal Control Based on EVALPSN

Kazumi Nakamatsu[1], Jair M. Abe[2], and Seiki Akama[3]

[1] University of Hyogo.
Himeji, Hyogo Japan
nakamatu@shse.u-hyogo.ac.jp
[2] Paulista University.
Sao Paulo, SP-Brazil
jairabe@uol.com.br
[3] Kawasaki.
Kanagawa, Japan
sub-akama@jcom.home.ne.jp

Abstract. We introduce a coordinated traffic signal control method based on EVALPSN. The simulation results of the EVALPSN coordinated traffic signal control method are shown with comparing to conventional ones.

Keywords: EVALPSN(Extended Vector Annotated Logic Program with Strong Negation), traffic signal control, paraconsistent annotated logic program, defeasible deontic reasoning.

1 Introduction

We have proposed a paraconsistent annotated logic program called EVALPSN (Extended Vector Annotated Logic Program) that can deal with defeasible deontic reasoning and paraconsistency [1]. Some applications of EVALPSN to various kinds of control and safety verification, robot action control, discrete event control, and safety verification for railway interlocking and air traffic control have been introduced [3,4].

Traffic jam caused by inappropriate traffic signal control is a serious environmental issue that we have to overcome. Conventionally, traffic signal control is divided into three kinds, single, coordinated and broad area ones according to the number of traffic signals to be controlled. In single traffic signal control each traffic light is controlled independently. On the other hand, in coordinated one a series of traffic lights on the same route are controlled. Recently an optimizing method for traffic signal control by Genetic Algorithm has been proposed and applied practically. However, the method takes a long time to compute optimal solutions and is not so appropriate for real-time control. We have already proposed an intelligent real-time single traffic signal control system based on EVALPSN and shown that it could reduce the traffic density 10% to 15% by simulation [2]. In this paper, we extend the idea of EVALPSN single traffic signal

B. Apolloni et al. (Eds.): KES 2007/ WIRN 2007, Part II, LNAI 4693, pp. 869–876, 2007.
© Springer-Verlag Berlin Heidelberg 2007

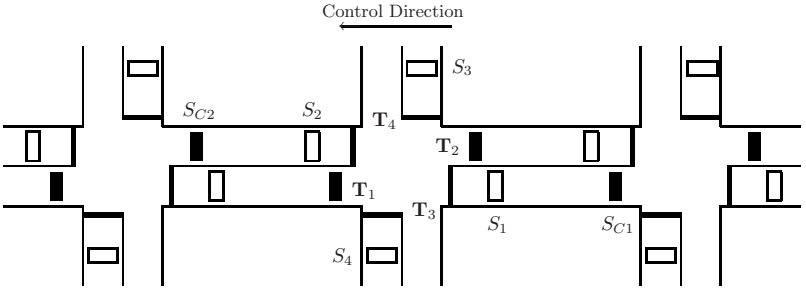


Fig. 1. Intersections in Japan

control to coordinated one and propose an EVALPSN based coordinated traffic signal control system.

2 Traffic Signal Control in EVALPSN

First of all we suppose a basic traffic signal control policy that traffic lights treating a larger amount of traffic should have the priority to be green until reducing the traffic density. Based on this control policy the traffic signal is controlled in real-time by defeasible deontic reasoning in EVALPSN.

Suppose that you are waiting for the traffic light turning from red to green at an intersection. Then, you must have a demand for turning the traffic light from red to green in your mind. On the other hand, if you are driving through the intersection with green you must have a demand for keeping the traffic light green. The first demand can be regarded as permission for turning the light from red to green and the second one can be regarded as forbiddance from turning the traffic light from green to yellow or red. Then, there is a conflict between the permission and forbiddance. The basic idea of EVALPSN traffic signal control is to resolve such conflicts by EVALPSN defeasible deontic reasoning. We will formalize the traffic signal control in defeasible deontic formulas firstly and translate them into EVALPSN later.

2.1 Single Traffic Signal Control

We take a series of three typical intersections in Japan as described in Fig.1. In single traffic signal control, we focus on only the middle intersection among the three ones in Fig.1 as the object of the single traffic signal control. We assume that: the middle intersection has four traffic lights $T_{1,2,3,4}$, each one has three kinds of displays, green, yellow and red, and they have the following signal chart,

$$\begin{array}{ll} T_{1,2} & \rightarrow \text{red} \quad \rightarrow \text{red} \quad \rightarrow \text{all red} \rightarrow \text{green} \rightarrow \text{yellow} \rightarrow \text{all red} \rightarrow \\ T_{3,4} & \rightarrow \text{green} \rightarrow \text{yellow} \rightarrow \text{all red} \rightarrow \text{red} \quad \rightarrow \text{red} \quad \rightarrow \text{all red} \rightarrow \end{array}$$

the intersection has four sensors $S_{1,2,3,4}$ to detect traffic density or flowing traffic, which are described as white boxes in Fig.1; yellow and all red terms are fixed

and only green term is variably controlled according to the traffic detected by the sensors; the minimum and maximum lengths of green term are given in advance, that is to say, green term must be controlled between the minimum and maximum terms.

For example, if the sensor S_1 detects flowing traffic, the green term of the traffic light T_1 should be extended within the range between the minimum and maximum terms. In the EVALPSN traffic signal control, the flowing traffic or traffic density of each road connected to the intersection are regarded as the forbiddance or permission for turning the object traffic light from green to yellow, and the conflict between the permission and forbiddance should be resolved by EVALPSN defeasible deontic reasoning. As the first step, we will formalize the traffic signal control in defeasible deontic formulas and translate them into EVALPSN.

Defeasible Deontic Rules for Single Traffic Signal Control. Firstly, we introduce defeasible reasoning briefly. The details are found in [5]. Basically, two kinds of rules, a strict rule $A \rightarrow p$ and a defeasible rule $B \Rightarrow q$ are used, where A,B and p,q are called the prerequisites and consequents of the rules $A \rightarrow p$ and $B \Rightarrow q$, respectively. Defeasible reasoning is carried out based on the superiority relations between defeasible rules. Strict rules are used for deriving facts and superior to defeasible rules. Suppose that there are conflicting defeasible rules: **R1**: $a \Rightarrow p$ and **R2**: $b \Rightarrow \neg p$. For example, if there is the superiority relation **R1** $<$ **R2**, and both a and b hold, then the rule **R2** defeats the rule **R1** and only the consequent $\neg p$ of the defeasible rule **R2** is derived.

Now we will start formalizing the single traffic signal control for the middle intersection in defeasible deontic formulas. We suppose that the traffic lights $T_{1,2}$ and $T_{3,4}$ are red and green, respectively. Then, the following conditions should be considered as strict or defeasible rules.

1. If one of the sensors $S_{1,2}$ detects traffic density more than the criterion at the time t, the permission for turning the traffic lights $T_{3,4}$ from green to yellow should be derived. This condition is represented by the defeasible rule,

$$S^{rg}_{1,2}(t) \wedge T_{1,2}(r,t) \wedge T_{3,4}(g,t) \Rightarrow \neg \bigcirc \neg T_{3,4}(y,t), \tag{1}$$

where, $S^{rg}_{1,2}(t)$ denotes that one of the sensors $S_{1,2}$ detects the traffic density over the criterion at the time t; $T_{1,2}(r,t)$ and $T_{3,4}(g,t)$ denote that the traffic lights $T_{1,2}$ and $T_{3,4}$ are red and green at the time t, respectively; and the symbol $\bigcirc$ is a modal operator to denote obligation, thus, the symbols $\neg \bigcirc \neg$ and $\bigcirc \neg$ denote permission and forbiddance, respectively.

2. On the other hand, if one of the sensors $S_{3,4}$ detects the flowing traffic over the criterion at the time t, the forbiddance from turning the traffic lights $T_{3,4}$ from green to yellow should be derived. Then, we have the defeasible rule conflicting to the defeasible rule (1),

$$S^{rg}_{3,4}(t) \wedge T_{1,2}(r,t) \wedge T_{3,4}(g,t) \Rightarrow \bigcirc \neg T_{3,4}(y,t). \tag{2}$$

We also have to define the superiority relation ($<$) between the defeasible rules (1) and (2). As we assume that maintaining the current signal state is superior to turning it to the next signal state, the forbiddance from turning the traffic lights $T_{3,4}$ from green to yellow is prior to the permission for it, that is to say, the defeasible rule (2) is considered to be superior to the defeasible rule (1).

We need more traffic signal control rules. If we take only the defeasible rules (1) and (2) into account, we might have an extreme situation such that the green term of $T_{1,2}$ is 2 hours but that of $T_{3,4}$ is 1 minute,. In order to avoid such an extreme control, we have to control green term in an appropriate range, which are represented by the following definite rules.

Minimum Green Term Rule. Each green term of the traffic lights $T_{1,2,3,4}$ should be guaranteed its minimum length. Let $MIN_i(g,t)$, $(i \in \{1,2,3,4\})$ denote that the green time of the traffic light T_i has not passed its minimum term at the time t. Then we have a rule : if the green time is shorter than its minimum term at the time t, it is forbidden from the traffic light turning from green to yellow. This rule can be represented by the strict rules,

$$MIN_{3,4}(g,t) \wedge T_{3,4}(g,t) \rightarrow \bigcirc \neg T_{3,4}(y,t). \tag{3}$$

Maximum Green Term Rule. Each green term of the traffic lights $T_{1,2,3,4}$ also should have its maximum term. Let $MAX_i(g,t)$, $(i \in \{1,2,3,4\})$ denotes that the green time of the traffic light T_i has already passed its maximum term at the time t. Then we have a rule : if the green time is longer than its maximum term at the time t, it is definitely permitted for the traffic light turning from green to yellow. This rule can be represented by the strict rules,

$$MAX_{3,4}(g,t) \wedge T_{3,4}(g,t) \rightarrow \neg \bigcirc \neg T_{3,4}(y,t). \tag{4}$$

There are conflicting defeasible rules (1) and (2) whose consequents are permission and forbiddance. As the defeasible reasoning result, if the permission is derived at the time t, the traffic light has to be turned from green to yellow at the next time $t+1$. On the other hand, if the forbiddance is defeasibly derived at the time t, the signal has to keep the current state at the next time $t+1$. These conditions can be represented by the following strict rules :

$$T_{3,4}(g,t) \wedge \neg \bigcirc \neg T_{3,4}(y,t) \rightarrow \bigcirc T_{3,4}(y,t+1), \tag{5}$$

$$T_{3,4}(g,t) \wedge \bigcirc \neg T_{3,4}(y,t) \rightarrow \bigcirc T_{3,4}(g,t+1). \tag{6}$$

Moreover, we need some more strict rules for synchronizing all the traffic lights. For example, if it is obligatory that the traffic light T_1 is red at the time t, the traffic light T_2 also must be red at the same time, and if it is obligatory that the traffic light T_3 is green at the time t, the traffic light T_1 must be red at the same time. Such synchronization is represented by the following strict rules,

$$\bigcirc T_1(r,t) \rightarrow \bigcirc T_2(r,t) \quad \text{and} \quad \bigcirc T_3(g,t) \rightarrow \bigcirc T_1(r,t) \tag{7}$$

EVALPSN for Single Traffic Signal Control. Now, we will translate the strict and defeasible rules into EVALPSN. The strict and defeasible rules can be

translated by formalizing their semantics in EVALPSN. The defeasible rules (1) and (2) are translated into the EVALPSN clauses,

$$S^{rg}{}_{1,2}(t):[(2,0),\alpha] \wedge T_{1,2}(r,t):[(2,0),\alpha] \wedge T_{3,4}(g,t):[(2,0),\alpha] \wedge \\ \sim MIN_{3,4}(g,t):[(2,0),\alpha] \wedge \sim S^{rg}{}_3(t):[(2,0),\alpha] \wedge \sim S^{rg}{}_4(t):[(2,0),\alpha] \\ \rightarrow T_{3,4}(y,t):[(0,1),\gamma], \tag{8}$$

$$S^{rg}{}_{3,4}(t):[(2,0),\alpha] \wedge T_{1,2}(r,t):[(2,0),\alpha] \wedge T_{3,4}(g,t):[(2,0),\alpha] \wedge \\ \sim MAX_{3,4}(g,t):[(2,0),\alpha] \rightarrow T_{3,4}(y,t):[(0,1),\beta]. \tag{9}$$

The EVALPSN clause (8) represents that: *if* the traffic sensors S_1 or S_2 have detected the traffic density over the criterion, the minimum green term has already passed, and the traffic sensors neither S_3 nor S_4 has detected the flowing traffic over the criteria at the time t under the condition that the traffic lights $T_{3,4}$ are green, *then* the traffic lights $T_{3,4}$ are permitted for turning to yellow. On the other hand, the EVALPSN clause (9) represents that: *if* the traffic sensors S_3 or S_4 have detected the flowing traffic over the criterion, and the maximum green term has not passed yet under the same condition that the traffic lights $T_{3,4}$ are green, *then* the traffic lights $T_{3,4}$ are forbidden for turning to yellow.

Similarly, the strict rules (3), (4), (5) and (6), are also translated into the EVALP clauses,

$$MIN_{3,4}(g,t):[(2,0),\alpha] \wedge T_{3,4}(g,t):[(2,0),\alpha] \rightarrow T_{3,4}(y,t):[(0,2),\beta], \tag{10}$$
$$MAX_{3,4}(g,t):[(2,0),\alpha] \wedge T_{3,4}(g,t):[(2,0),\alpha] \rightarrow T_{3,4}(y,t):[(0,2),\gamma], \tag{11}$$
$$T_{3,4}(g,t):[(2,0),\alpha] \wedge T_{3,4}(y,t):[(0,1),\gamma] \rightarrow T_{3,4}(y,t+1):[(2,0),\beta], \tag{12}$$
$$T_{3,4}(g,t):[(2,0),\alpha] \wedge T_{3,4}(y,t):[(0,1),\beta] \rightarrow T_{3,4}(g,t+1):[(2,0),\beta]. \tag{13}$$

Example 1. Suppose that the traffic lights $T_{1,2}$ are red and the traffic lights $T_{3,4}$ are green, furthermore, it has already passed the minimum green term then.

Case-1. If one of the sensors $S_{1,2}$ detects the traffic density over the criterion at the time t and the rest of the sensors do not react then, the EVALPSN clause (8) is fired and the permission $T_{3,4}(y,t):[(0,1),\gamma]$ is derived, furthermore, the EVALPSN clause (12) is also fired and the obligation $T_{3,4}(y,t+1):[(2,0),\beta]$ is also derived.

Case-2. If both the sensors S_1 and S_3 detect the traffic density and flowing traffic over the criteria at the time t, respectively, and the rest of the sensors do not detect then, the EVALPSN clause (9) is fired and the forbiddance $T_{3,4}(y,t):[(0,1),\beta]$ is derived, furthermore, the EVALPSN clause (13) is also fired and the obligation $T_{3,4}(g,t+1):[(2,0),\beta]$ is also derived.

2.2 Coordinated Traffic Signal Control

We show the basic idea of the EVALPSN coordinated traffic signal control with taking the same intersections in Fig.1 as an example. Here we focus on only the traffic from the right to the left and its converse direction as the object of the

EVALPSN coordinated traffic signal control. In conventional coodinated traffic signal control, not only each term of green, yellow and red but also a timelag called an *off set* between two neighbor traffic lights are controlled, and they should be preinstalled. On the other hand, in EVALPSN coordinated traffic signal control, only the green term of each coordinated traffic light should be controlled by EVALPSN defeasible deontic reasoning based on traffic amount information, which is detected by traffic sensors attached to not only the object intersection but also its both neighbor intersections. Therefore, when we consider the EVALPSN coordinated traffic signal control in terms of the middle intersection, the extra traffic sensors $Sc_{1,2}$ described by black boxes in Fig.1 attached to the right and left intersections should be taken into account to detect the traffic flowing into the middle intersection. Moreover, we note that when the traffic information detected by the remote sensors $Sc_{1,2}$ is processed, appropriate timelags between the detection and processing of the traffic information should be taken into account according to the distances between the object intersection and the sensors $Sc_{1,2}$.

Defeasible Deontic Rules for Coordinated Traffic Signal Control. We assume the same condition as the single traffic signal control, that is to say the traffic lights $T_{1,2}$ and $T_{3,4}$ are red and green, respectively. Then, we have the following conditions in defeasible deontic formulas.

3. If one of the sensors Sc_1 or Sc_2 detects flowing traffic over the criteria at the time t, the permission for turning the traffic lights $T_{3,4}$ from green to yellow should be derived. Then, we have the defeasible rule,

$$Sc^{rg}_{1,2}(t) \wedge T_{1,2}(r,t) \wedge T_{3,4}(g,t) \Rightarrow \neg \bigcirc \neg T_{3,4}(y,t), \tag{14}$$

where $Sc^{rg}_{1,2}(t)$ denotes that one of the sensors $Sc_{1,2}$ detects the flowing traffic over the criterion.

We also need to consider the superiority relation between the conflicting defeasible rules (2) and (14). As the coordinately controlled route traffic from the right to the left or its converse are regarded to be superior to other traffic, the permission for turning the traffic lights $T_{3,4}$ from green to yellow has a prior to the forbiddance from it, that is to say, the defeasible rule (14) is considered to be superior to the defeasible rule (2), i.e.,$(2) < (14)$.

EVALPSN for Coordinated Traffic Signal Control. Taking the superiority relations among the defeasible rules (1), (2) and (14), i.e., $(1) < (2) < (14)$ into account, those defeasible rules are retranslated into the EVALPSN clauses,

$$\begin{aligned} &{S^{rg}}_{1,2}(t) : [(2,0),\alpha] \wedge T_{1,2}(r,t) : [(2,0),\alpha] \wedge T_{3,4}(g,t) : [(2,0),\alpha] \wedge \\ &\sim MIN_{3,4}(g,t) : [(2,0),\alpha] \wedge \sim {S^{rg}}_3(t) : [(2,0),\alpha] \wedge \\ &\sim {S^{rg}}_4(t) : [(2,0),\alpha] \rightarrow T_{3,4}(y,t) : [(0,1),\gamma], \end{aligned} \tag{15}$$

$$\begin{aligned} &{S^{rg}}_{3,4}(t) : [(2,0),\alpha] \wedge T_{1,2}(r,t) : [(2,0),\alpha] \wedge T_{3,4}(g,t) : [(2,0),\alpha] \wedge \\ &\sim MAX_{3,4}(g,t) : [(2,0),\alpha] \wedge \sim {Sc^{rg}}_1(t) : [(2,0),\alpha] \wedge \sim {Sc^{rg}}_2(t) : [(2,0),\alpha] \\ &\rightarrow T_{3,4}(y,t) : [(0,1),\beta], \end{aligned} \tag{16}$$

$$Sc^{rg}{}_{1,2}(t):[(2,0),\alpha] \wedge T_{1,2}(r,t):[(2,0),\alpha] \wedge T_{3,4}(g,t):[(2,0),\alpha] \wedge \\ \sim MIN_{3,4}(g,t):[(2,0),\alpha] \rightarrow T_{3,4}(y,t):[(0,1),\gamma], \quad (17)$$

As well as the EVALPSN single traffic signal control the coordinated one also has **Minimum** and **Maximum Green Term Rules**, therefore, we also have the EVALPSN clauses (10), (11), (12), (13) as the translation of those rules.

Example 2. Suppose the same conditions as **Example 1**.

Case-3. If one of the sensors $Sc_{1,2}$ and the sensor S_4 detect the flowing traffic over the criteria at the time t, respectively, the EVALPSN clause (17) is fired and the permission $T_{3,4}(y,t):[(0,1),\gamma]$ is derived, furthermore, the EVALPSN clause (12) is also fired and the obligation $T_{3,4}(y,t+1):[(2,0),\beta]$ is also derived.

Case-4. If all the sensors $S_{1,2,3,4}$ and $Sc_{1,2}$ detect the flowing traffic or traffic density over the criteria at the time t, respectively, the EVALPSN clause (17) is fired and the permission $T_{3,4}(y,t):[(0,1),\gamma]$ is derived, furthermore, the EVALPSN clause (12) is also fired and the obligation $T_{3,4}(y,t+1):[(2,0),\beta]$ is also derived.

Case-5. If the sensors $S_{1,2}$ and $Sc_{1,2}$ detect the traffic density and flowing traffic over the criteria at the time t, respectively, both the EVALPSN clauses (15) and (17) are fired and the same permission $T_{3,4}(y,t):[(0,1),\gamma]$ is derived, furthermore, the EVALPSN clause (12) is also fired and the obligation $T_{3,4}(y,t+1):[(2,0),\beta]$ is also derived.

3 Simulation

We present the simulation results of the four kinds of traffic signal control, conventional (fixed time) single, coordinated, EVALPSN single, and EVALPSN coordinated ones for the intersections in Fig.1 by the cellular automaton simulation method. We have assumed the following conditions for simulation: the unit time called "step" is defined in the simulation system, which is the time that a car travels one cell to the next cell, and there are 5000 steps/hour in fact; each distance between the three intersections is the same 20 cells; cars are flowing into all the intersections from each road in 10% probability except for the traffic from the right road, which is 15%, where 15% probability means 15 cars/100 steps appear at the road; for fixed-time traffic signal control, the green term is 60 steps, the yellow term is 3 steps, and the off-set in the coordinated traffic signal control is 24 steps; for EVALPSN traffic signal control, the green term is between 14(min) and 60(max) steps, and the yellow term is fixed 5 steps.

We took the simulation data of 50 cars that are chosen at random among all the cars flowing into the intersections from the right and traveling to the left. All the four kinds of simulation have been carried out during 5000 steps each and repeated ten times. The simulation data in Tab.1 show the average values of the ten times simulation results, where "total car" denotes the total number of cars that flowed into all the three intersections from any road; "stop step" denotes the total number of steps that the 50 cars stopped; "travel step" denotes the total number of steps

Table 1. Simulation Results

	total car	stop step	travel step	travel car
Fixed-time Single	3804	66	177	495
Fixed-time Coordeinated	3823	65	173	500
EVALPSN Single	4149	43	152	577
EVALPSN Coordinated	4169	34	122	607

for the 50 cars having traveled from the right to the left; "travel car" denotes the total number of cars that traveled from the right up to the left.

This simulation results clearly say that: EVALPSN traffic signal control shows more efficient results than conventional fixed time one in both single and coordinated traffic signal control methods, and especially EVALPSN coordinated control is most efficient among them.

4 Conclusion

We have introduced a coordinated traffic signal control method based on an annotated logic program EVALPSN and its computer simulation results by the cellular automaton method. The EVALPSN traffic control is a sensor based real-time one and more efficient than conventional ones from the simulation results; however, if it is implemented practically, lots of sensor installation are required and too much cost; EVALPSN control can be implemented in both existent software and hardware such as PLC(Programmable Logic Controller), although it has not been addressed in this paper; since each traffic light is controlled independently as one autonomous agent in the EVALPSN coordinated traffic signal control, it is more fault tolerant than conventional coordinated one.

References

1. Nakamatsu, K., Abe, J.M., Suzuki, A.: Annotated Semantics for Defeasible Deontic Reasoning. In: Ziarko, W., Yao, Y. (eds.) RSCTC 2000. LNCS (LNAI), vol. 2005, pp. 470–478. Springer, Heidelberg (2001)
2. Nakamatsu, K., Seno, T., Abe, J.M., Suzuki, A.: Intelligent Real-time Traffic Signal Control Based on a Paraconsistent Logic Program EVALPSN. In: Wang, G., Liu, Q., Yao, Y., Skowron, A. (eds.) RSFDGrC 2003. LNCS (LNAI), vol. 2639, pp. 719–723. Springer, Heidelberg (2003)
3. Nakamatsu, K., Chung, S-L., Komaba, H., Suzuki, A.: A Discrete Event Control Based on EVALPSN Stable Model Computation. In: Ślęzak, D., Wang, G., Szczuka, M., Düntsch, I., Yao, Y. (eds.) RSFDGrC 2005. LNCS (LNAI), vol. 3641, pp. 671–681. Springer, Heidelberg (2005)
4. Nakamatsu, K., Mita, Y., Shibata, T.: An Intelligent Action Control System Based on Extended Vector Annotated Logic Program and its Hardware Implementation. Intelligent Automation and Soft Computing 13, 289–304 (2007)
5. Nute, D.: Basic Defeasible Logic. In: del Cerro, L.F., Penttonen, M. (eds.) Intensional Logics for Programming, pp. 126–154. Oxford Science Publications, Oxford (1992)

Intelligent Design of Diagnosable Systems: A Case Study of Semiconductor Manufacturing Machines

YuanLin Wen[1], ShengLuen Chung[2], LiDer Jeng[3], and MuDer Jeng[1,*]

[1] Dept. of Electrical Engineering, National Taiwan Ocean University,
Keelung 202, Taiwan
{d88530004,jeng}@mail.ntou.edu.tw
[2] Dept. of Electrical Engineering, National Taiwan University of Science and Technology
Taipei 106, Taiwan
slchung@mail.ntust.edu.tw
[3] Dept. of Electronic Engineering, Chung-Yuan Christian University,
Chung-Li 320, Taiwan
lider@cycu.edu.tw

Abstract. This paper presents an intelligent approach using Petri nets for designing diagnosable discrete event systems such as complex semiconductor manufacturing machines. The concept is based on diagnosability analysis and enhancement. We use a real-world Metal-Organic Vapor Phase Epitaxy (MOVPE) system to illustrate that our proposed approach is practically useful.

Keywords: Petri Nets, Semiconductor Manufacturing, Fault diagnosis.

1 Introduction

Traditionally, diagnosis has been studied in an ad hoc manner using techniques [3] such as heuristic search, expert system, and fault tree. Sampath et al. [3] first studied the problem in the domain of the discrete event system (DES) theory and proposed the concept of DES-based diagnosability [8]. The merit of their approach is that a model-based, systematic methodology is proposed to detect a specific type of failure in finite delay after its occurrence. In addition, they present a necessary and sufficient condition for checking diagnosability. One of the challenges of diagnosis is that not all events are observable. It is difficult to differentiate between a normal and a faulty event trace within a bounded delay from the time of the occurrence of a fault. Informally speaking, it is necessary to check all cycles in the state space to see if there are two cycles that have the same observable event traces. If yes, the system is non-diagnosable. Otherwise, it is diagnosable. Thus, the computational cost for checking diagnosability using the methods of Sampath and her followers [1] are significantly high. This causes their practical uses limited. As will be described below, we will adopt a more efficient method that does not enumerate the state space for checking diagnosability.

* Corresponding author.

B. Apolloni et al. (Eds.): KES 2007/ WIRN 2007, Part II, LNAI 4693, pp. 877–884, 2007.
© Springer-Verlag Berlin Heidelberg 2007

The remaining paper is organized as follows: Section 2 presents related basic concepts. Section 3 presents our proposed methodology. A case study is shown in Section 4. Finally, concluding remarks are presented in Section 5.

2 Basic Concepts

This section briefly summarizes Petri nets and diagnosability concepts as follows:

We consider a system modeled by a Petri net, $G = (P, T, I, O, M_o)$. The set of places P denotes conditions. The set of transitions T, denoting events, is partitioned into the set of observable transitions T_o and the set of unobservable transitions T_{uo}, i.e., $T = T_o \cup T_{uo}$ and $T_o \cap T_{uo} = \varnothing$. Labels, denoting events, are associated with transitions only. Failures are labels associated to unobservable transitions. A label associated with an unobservable transition is called an unobservable label, denoting an unobservable event. To focus on non-trivial diagnosis, it is assumed that failures are unobservable events.

We define diagnosability as follows [6]:

Definition 1. It is said that the system G is *diagnosable* if, whenever any failure transition occurs, its failure type is detected after any finite firings of transitions from the occurrence.

As mentioned before, the detection of diagnosability requires the computation of all cycles in the state space. These cycles are implied by Petri net structural objects called T-invariants, where a *T-invariant* is a vector X of non-negative integers denoting transition firing counts of the implied cycle such that $A^T X = 0$ where A is the incidence matrix, i.e., a matrix notation of the Petri net. For our problem, we only need to consider elementary cycles implied by so-called *minimal T-invariants* defined as follows: a T-invariant X_1 is minimal if there exists no other T-invariant X_2 such that $X_2 \leq X_1$. It is well-known that there exists a fast algorithm [7] to obtain all minimal T-invariants.

Therefore, the following property [4], [5] is useful for checking if a system is diagnosable:

Property 1. The system modeled by a Petri net G is diagnosable if there do not exist two T-invariants in G with the same observable labels.

From Property 1, we have the following property:

Property 2. A system modeled by a Petri net G is diagnosable if for any two minimal T-invariants $X_i = (x_{i1}, x_{i2}, \ldots . v_{i|T|})^T$ and $X_j = (x_{j1}, x_{j2}, \ldots . v_{j|T|})^T$, $X_i \cdot V \neq X_j \cdot V$ where $V = (v_1, v_2, \ldots . v_{|T|})$ is the numerical encoding of the observable label mapping vector from transitions to their corresponding labels as follows: assume that for any label into which a set U of transitions is mapped, the total number of transition firings of U in any T-invariant does not exceed k. Let $d = \lceil k/10 \rceil$. Then we use 10^0, 10^d, 10^{2d}, 10^{3d} … to represent each different label, respectively. Note that $X_i \cdot V = X_j \cdot V$ implies that two elementary cycles have the same observable event count, which may mean the same observable event trace. The following example illustrates the above concepts.

3 Proposed Design Methodology

In the proposed methodology, we assume that all available sensors, the locations in the target system where these sensors can be added, and all possible readings of these sensors are known beforehand by domain knowledge. We assume that unobservable labels remain unobservable no matter what sensors are added so we do not consider changing unobservable labels. The above description can be formalized using the following algorithm:

Algorithm *Diagnosable_System_Design*
Input: A Petri-net-modeled system, the set S of all available sensors, all sensor locations of S, all sensor readings of S;
Output: A diagnosabiliy-enhanced system, the set P of sensors added;
Step 1: Let $P = \varnothing$.
Step 2: Let $\tilde{X}$ = the set of all minimal T-invariants.
Step 3: Let V = the numerical encoding of the labels.
Step 4: If there exist two minimal T-invariants $X_i, X_j \in \tilde{X}$ such that $X_i \cdot V = X_j \cdot V$, go to Step 6.
Step 5: Otherwise, return. /* The system is diagnosable.*/
Step 6: Let $S' = S$.
Step 7: If $S' = \varnothing$, then return. /* The diagnosability cannot be enhanced further*/
Step 8: Otherwise, select a sensor s in S', and change the labels.
Step 9: Let V = the numerical encoding of the new labels.
Step 10: If $X_i \cdot V = X_j \cdot V$, then remove s from S′ and go to Step 7.
Step 11: Remove s from S, add s to P, and go to Step 4.
End of Algorithm

4 A Case Study: Metal-Organic Vapor Phase Epitaxy (MOVPE) System

In this section, we consider a real-world MOVPE [2] system as shown in Fig. 1. Due to space limitation, we will use its pumping and control subsystem in Fig. 2 to illustrate the approach. The same principle can be applied to the remaining part of the system. The pumping and control subsystem consists of three components. Fig. 3 depicts their Petri nets models, whose operations are explained as follows:

1) The VALVE model has four places: SC and SO represent the STUCK-CLOSE and STUCK-OPEN statuses of the VALVE, respectively; VC and VO represent the VALVE-CLOSE and VALVE-OPEN statuses, respectively.
2) The PUMP model has two places: POFF and PON represent the PUMP-OFF and PUMP-ON status of the PUMP, respectively.
3) The CONTROLLER has four places: C1 represents both the VALVE and PUMP are off. C2 represents that the VALVE opens first, and the PUMP is in the off status. C3 represents that the VALVE opens and the PUMP turns on. C4 represents that the VALVE closes first, and the PUMP is still in the on status.

The transitions of VALVE, PUMP, and CONTROLLER are mapped into 6 types of events; four are normal: OPEN_VALVE, CLOSE_VALVE, START_PUMP, and STOP_PUMP. The other two failure events are represented as valve STUCK_CLOSE (failure type f_1) and valve STUCK_OPEN (failure type f_2), respectively. The transitions and their mapped events are show in Table 1. Note that we assume that there are only two types of failure events.

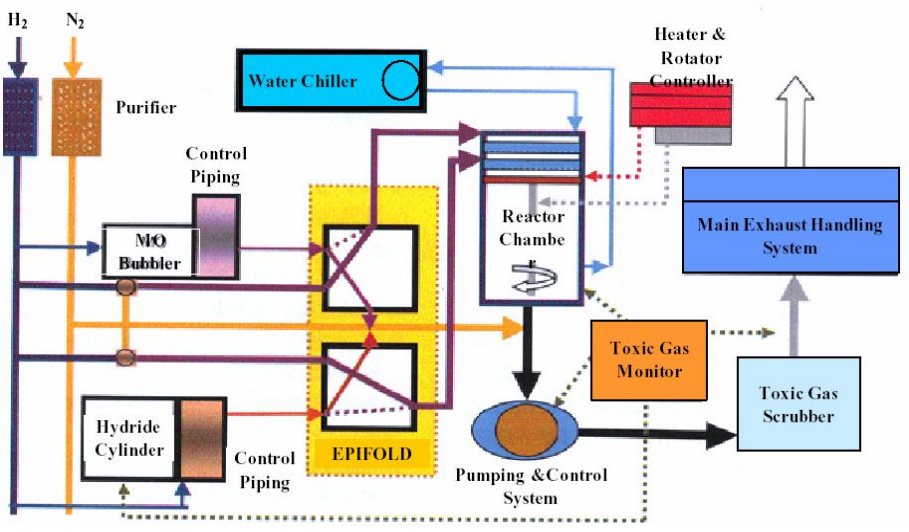


Fig. 1. An MOVPE System

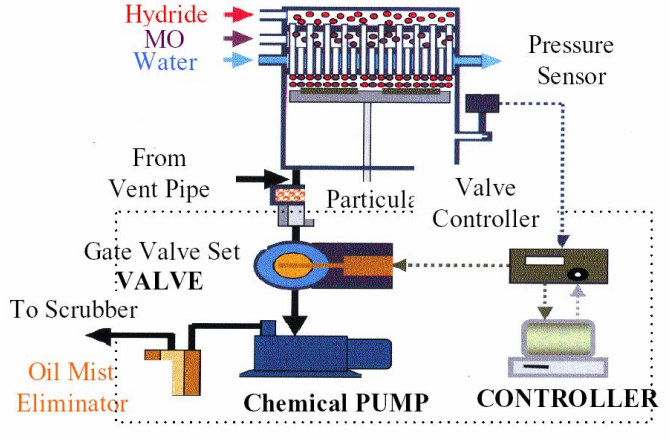
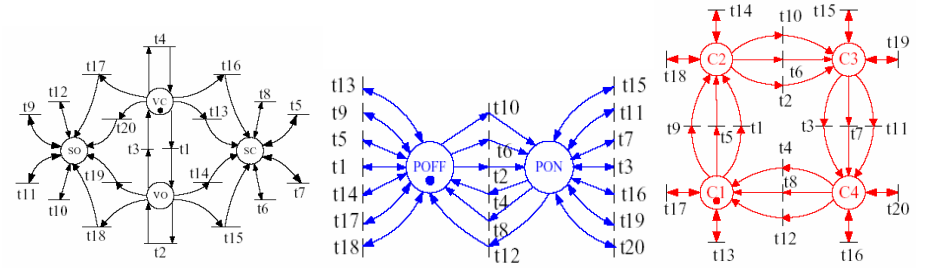


Fig. 2. The pumping and control subsystem

After each component is modeled, the complete system model G as shown in Fig. 4 can be built by merging the common transitions of the component models.



(a) VALVE (b) PUMP (c) CONTROLLER

Fig. 3. The models of the (a) VALVE (b) PUMP, and (c) CONTORLLER of Fig. 2

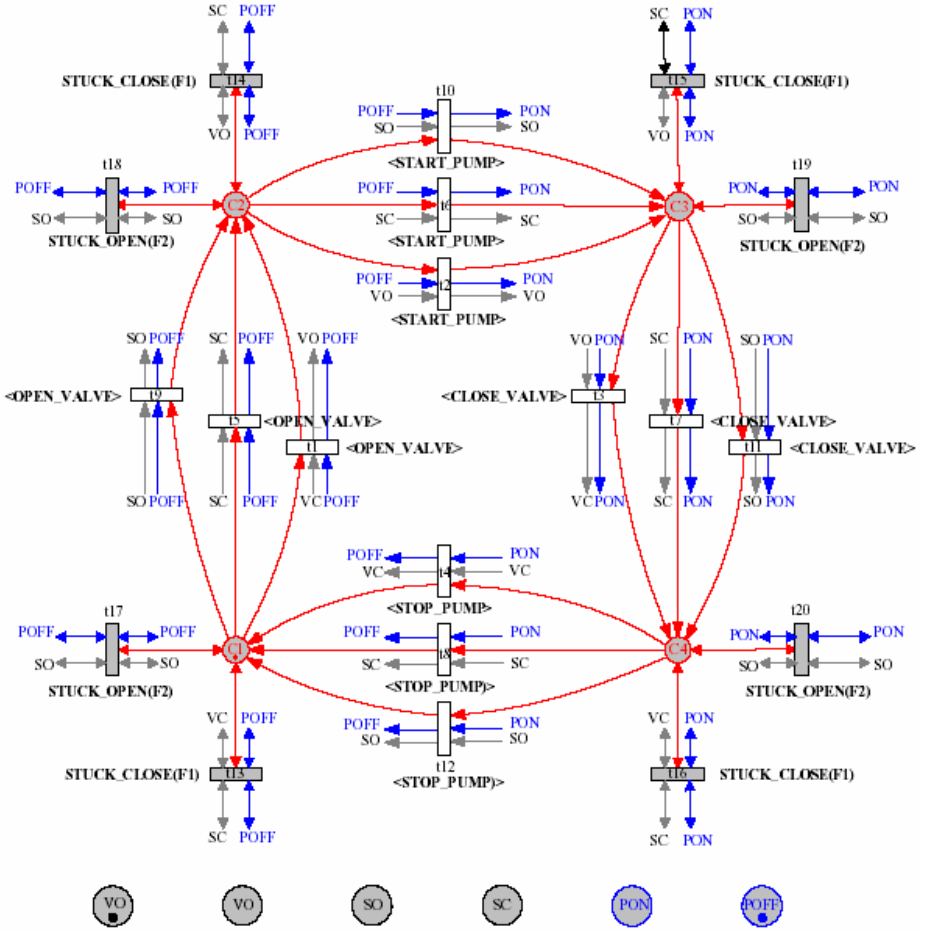


Fig. 4. The complete Petri net model of the pumping and control system

Table 1. Transitions and their mapped events

T	Event	Observability	T	Event	Observability
t1	OPEN_VALVE	observable	t11	CLOSE_VALVE	observable
t2	START_PUMP	observable	t12	STOP_PUMP	observable
t3	CLOSE_VALVE	observable	t13	STUCK_CLOSE (f_1)	unobservable
t4	STOP_PUMP	observable	t14	STUCK_CLOSE (f_1)	unobservable
t5	OPEN_VALVE	observable	t15	STUCK_CLOSE (f_1)	unobservable
t6	START_PUMP	observable	t16	STUCK_CLOSE(f_1)	unobservable
t7	CLOSE_VALVE	observable	t17	STUCK_OPEN (f_2)	unobservable
t8	STOP_PUMP	observable	t18	STUCK_OPEN (f_2)	unobservable
t9	OPEN_VALVE	observable	t19	STUCK_OPEN (f_2)	unobservable
t10	START_PUMP	observable	t20	STUCK_OPEN (f_2)	unobservable

Table 2. Available sensors and their locations in the target system

No	Failures	Available Sensors	Location
1	STUCK_CLOSE(f_1),STUCK_OPEN(f_2)	Pressure (S_p)	reactor chamber
2	STUCK_CLOSE(f_1),STUCK_OPEN(f_2)	Flow (S_f)	Output of the chemical pump
3	STUCK_CLOSE(f_1),STUCK_OPEN(f_2)	Temperature (S_t)	reactor chamber
4	STUCK_CLOSE(f_1),STUCK_OPEN(f_2)	Concentrations (S_c)	reactor chamber

Table 3. All sensor readings

Trans.	Event	Pressure sensor S_p	Flow sensor S_f	Temperature sensor S_t	Concentrations sensor S_c
t1	OPEN_VALVE	P(H to H)	F(L to H)	T(H to H)	C(H to H)
t2	START_PUMP	P(H to L)	F(H to H)	T(H to L)	C(H to L)
t3	CLOSE_VALVE	P(L to H)	F(H to L)	T(L to H)	C(L to H)
t4	STOP_PUMP	P(H to H)	F(L to L)	T(H to H)	C(H to H)
t5	OPEN_VALVE	P(H to H)	F(L to L)	T(H to H)	C(H to H)
t6	START_PUMP	P(H to H)	F(L to L)	T(H to H)	C(H to H)
t7	CLOSE_VALVE	P(H to H)	F(L to L)	T(H to H)	C(H to H)
t8	STOP_PUMP	P(H to H)	F(L to L)	T(H to H)	C(H to H)
t9	OPEN_VALVE	P(H to H)	F(L to H)	T(H to H)	C(H to H)
t10	START_PUMP	P(H to L)	F(H to H)	T(H to L)	C(H to L)
t11	CLOSE_VALVE	P(L to H)	F(H to H)	T(L to H)	C(L to H)
t12	STOP_PUMP	P(H to H)	F(H to L)	T(H to H)	C(H to H)

Assume that with domain knowledge, we obtain the information about all available sensors and the locations as shown in Table 2, and the sensor readings as shown in Table 3, where, for example, P(H to H) means that the reading of the pressure sensor changes from High to High.

We apply Algorithm *Diagnosable_System_Design* to the model of Fig. 5. The details are described as follows:

Steps 1-2: The model has three minimal T-invariants $X_1=(111100000000)^T$, $X_2=(000011110000)^T$, and $X_3=(000011110000)^T$. X_2 implies STUCK_CLOSE, i.e., f_1 occurs. X_3 implies STUCK_OPEN, i.e., f_2 occurs.

Step 3: <OPEN_VALVE>, <START_PUMP>, <CLOSE_VALVE>, and <STOP_PUMP> are encoded as 10^0, 10^1, 10^2, and 10^3, respectively. Therefore, $V=(10^0,10^1,10^2,10^310^0,10^1,10^2,10^310^0,10^1,10^2,10^3)$.

Step 4: $S'=\{S_p, S_f, S_t, S_c\}$.

Step 5: There exist two minimal T-invariants, $X_1=(111100000000)^T$, and $X_2=(000011110000)^T$ such that $X_1\cdot V=1\cdot10^0+1\cdot10^1+1\cdot10^2+1\cdot10^3=X_2\cdot V$. That is, they have the same observable event count that means <OPEN_VALVE>, <START_PUMP>, <CLOSE_VALVE>, and <STOP_PUMP> occurs 1, 1, 1, and 1 time, respectively. Go to step 7.

Step 7: $S'=\{S_p, S_f, S_t, S_c\}\neq\varnothing$ so go to the next step.

Step 8: Select S_p in S' and change the labels according to Table 3. The new labels are shown in Table 4.

Step 9: The numerical encoding of the new labels after S_p is selected is shown in Table 4.

Table 4. The new labels and their numerical encoding encoding after S_p is selected

T	Event	v_i
t1	<OPEN_VALVE,P(H to H)>	10^0
t2	<START_PUMP,P(H to L)>	10^1
t3	<CLOSE_VALVE,P(L to H)>	10^2
t4	<STOP_PUMP,P(H to H)>	10^3
t5	<OPEN_VALVE,P(H to H)>	10^0
t6	<START_PUMP,P(H to H)>	10^4
t7	<CLOSE_VALVE,P(H to H)>	10^5
t8	<STOP_PUMP,P(H to H)>	10^3
t9	<OPEN_VALVE,P(H to H)>	10^0
t10	<START_PUMP,P(H to L)>	10^1
t11	<CLOSE_VALVE,P(L to H)>	10^2
t12	<STOP_PUMP,P(H to H)>	10^3

Table 5. The new labels and their numerical after S_p, and S_f are selected

T	Event	v_i
t1	<OPEN_VALVE,P(H to H),F(L to H)>	10^0
t2	<START_PUMP,P(H to L),F(H to H)>	10^1
t3	<CLOSE_VALVE,P(L to H),F(H to L)>	10^2
t4	<STOP_PUMP,P(H to H),F(L to L)>	10^3
t5	<OPEN_VALVE,P(H to H),F(L to L)>	10^4
t6	<START_PUMP,P(H to H),F(L to L)>	10^5
t7	<CLOSE_VALVE,P(H to H),F(L to L)>	10^6
t8	<STOP_PUMP,P(H to H),F(L to L)>	10^3
t9	<OPEN_VALVE,P(H to H),F(L to H)>	10^0
t10	<START_PUMP,P(H to L),F(H to H)>	10^1
t11	<CLOSE_VALVE,P(L to H),F(H to	10^7
t12	<STOP_PUMP,P(H to H),F(H to L)>	10^8

Step 10: $X_1\cdot V=1\cdot10^0+1\cdot10^1+1\cdot10^2+1\cdot10^3=1111$, and $X_2\cdot V=1\cdot10^0+1\cdot10^4+1\cdot10^5+1\cdot10^3=111001$. Hence, $X_1\cdot V\neq X_2\cdot V$. Go to the next step.

Step 11: Remove S_p from S, and add S_p to P. Thus, $P=\{S_p\}$ and $S=\{S_f, S_t, S_c\}$. Go to Step 4.

Step 4: $S'=\{S_f, S_t, S_c\}$.

Step 5: According to the numerical encoding V in Table 4, we obtain two minimal T-invariants, $X_1=(111100000000)^T$ and $X_3=(000011110000)^T$ such that $X_1\cdot V=X_3\cdot V=1\cdot10^0+1\cdot10^1+1\cdot10^2+1\cdot10^3=1111$, meaning <OPEN_VALVE, P(H to H)>, <START_PUMP,P(H to L)>, <CLOSE_VALVE, P(L to H)>, and <STOP_PUMP, P(H to H)> occurs 1, 1, 1, and 1 time, respectively. Go to Step 7.

Step 7: $S'=\{S_f, S_t, S_c\}\neq\varnothing$. Go to the next step.

Step 8: Select S_f in S' and change the labels according to Table 3. The new labels are shown in Table 5.
Step 9: The numerical encoding of the new labels is shown in Table 5.
Step 10: $X_1 \cdot V = 1*10^0+1*10^1+1*10^2+1*10^3=1111$, and $X_3 \cdot V=1*10^0+1*10^1+1*10^7+1*10^8 = 110000011$. Hence, $X_1 \cdot V \neq X_3 \cdot V$. Go to the next step.
Step 11: Remove S_f from S, and add S_f to P. Thus, $P=\{S_p, S_f\}$ and $S=\{S_t, S_c\}$.Go to Step 4.
Step 4: $S'=\{S_t, S_c\}$.
Step 5: According to the numerical encoding V in Table 5, after S_p and S_f are added, there does not exist any two minimal T-invariants X_i and X_j such that $X_i \cdot V=X_j \cdot V$. Go to Step 6
Step 6: Return and the system is diagnosable after S_p and S_f are added.

Finally, we obtain the observable event counts and their corresponding events, which can be used for online fault diagnosis in the future.

5 Conclusions

In this paper, a novel approach based on Petri nets for designing diagnosable semiconductor manufacturing machines has been presented. The approach is under the assumption that the costs of sensors are not considered, comparing to the costs of machines. This is well justified in semiconductor manufacturing.

References

1. Jiang, S., Huang, Z., Chandra, V., Kumar, R.: A Polynomial Algorithm for Testing Diagnosability of Discrete-Event Systems. IEEE Trans. on Automatic Control 46(8), 1318–1321 (2001)
2. Lin, W.: Introduction to Metal-Organic Vapor Phase Epitaxy System and Technology. Instruments Today 24(2), 13–24 (2002)
3. Sampath, M., Lafortune, S., Sinnamohideen, K., Teneketzis, D.: Diagnosability of Discrete Event Systems. IEEE Trans. on Automatic Control 40(9), 1555–1557 (1995)
4. Wen, Y.L., Jeng, M.D., Huang, Y.S.: Diagnosability of Semiconductor Manufacturing Equipment. Material Science Forum 505, 1135–1140 (2006)
5. Wen, Y.L., Jeng, M.D., Jeng, L.D., Fan, P.S.: An Intelligent Technique Based on Petri Nets for Diagnosability Enhancement of Discrete Event Systems. In: Gabrys, B., Howlett, R.J., Jain, L.C. (eds.) KES 2006. LNCS (LNAI), vol. 4252, pp. 879–887. Springer, Heidelberg (2006)
6. Ushio, T., Onishi, I., Okuda, K.: Fault Detection Based on Petri Net Models with Faulty Behaviors. In: Proc. of the 1998 IEEE Int. Con. on Systems, Man, and Cybernetics, pp. 113–118. IEEE Computer Society Press, Los Alamitos (1998)
7. Martinez, J., Silva, M.: A simple and fast algorithm to obtain all invariants of a generalized Petri nets. In: Proc. of Second European Workshop on Application and Theory of Petri Nets. Informatik-Fachberichte, vol. 52, pp. 301–310 (1982)
8. Chung, S.-L.: Diagnosing PN-based models with partial observable transitions. Int. J. Computer Integrated Manufacturing 18(2–3), 158–169 (2005)

Knowledge Assessment Based on Many-Valued Logic

Sylvia Encheva[1] and Sharil Tumin[2]

[1] Stord/Haugesund University College, Bjørnsonsg. 45, 5528 Haugesund, Norway
sbe@hsh.no, mzs@hsh.no
[2] University of Bergen, IT-Dept., P.O. Box 7800, 5020 Bergen, Norway
edpst@it.uib.no

Abstract. This paper describes application of six-valued logic in an assessment sub-system of an intelligent tutoring system. This allows the system to handle situations with inconsistent and incomplete input. Special attention is paid to the decision making rules an intelligent agent is applying for assessing students' understanding of new terms. If an assessment is not in anyway connected with grading, i.e. a student will neither loose no gain by providing honest answers, we assume that there is no need for penalty rules deterring system misuse. Instead we implement rules that distinguish between students' hesitation in the process of giving an answer and lack of knowledge.

Keywords: Web-based assessment, learning.

1 Introduction

Most automated assessment tests are based on binary logic, i.e. an answer chosen by a student is either correct or incorrect and that implies possesion or lack of knowledge respectively. The assigned score in a case of an incorrect answer can be either a zero, which does not distinguish an incorrect answer from a missing answer, or a negative score. Penalty marking is applied for preventing guessing. However, according to [12] one should ask whether 'educated' guess is something to be discouraged since professional people use educated guesses in all fields of activity and by [23] 30% of the students choose the correct answer for wrong reasons.

Some systems reward partially correct answers in a sense that a part of an answer is correct and another part is either missing or incorrect. The assigned score is related to the significance of a missing part or an incorrect part of the answer.

However different, the above mentioned approaches have one thing in comman, they all use binary logic while making conclusions about student' knowledge. Thus any of the above mentioned approaches does not treat incomplete or inconsistent information.

In this paper we discuss how to assess students' understanding of new terms, shortly after these terms have been introduced in a subject. Application of a

B. Apolloni et al. (Eds.): KES 2007/ WIRN 2007, Part II, LNAI 4693, pp. 885–892, 2007.
© Springer-Verlag Berlin Heidelberg 2007

six-valued logic allows the system to handle situations with inconsistent and/or incomplete input. Special attention is paid to the decision making rules an intelligent agent is applying for assessing students' understanding of new terms. If an assessment is not in anyway connected with grading, i.e. a student will neither loose no gain by providing honest answers, we assume that there is no need for penalty rules deterring system misuse. Instead we implement rules that distinguish between students' hesitation in the process of giving an answer and lack of knowledge. Such rules are used only for the intelligent assessment sub-system. This work presents futher development of a response-driven Web-based assessment system enhancing learning [5]. The original system contains tests assessing recall of facts, high-level thinking, asking students to evaluate consequences and draw conclusions, and various quizzes. The system responds to students' needs as they progress through the system. This is incorporated by providing a structured learning system environment, suitable for offering automatic help through logic-based and qualitative reasoning mechanisms and guiding students in their navigation.

The rest of the paper is organized as follows. Related work and statements from many-valued logic may be found in Section 2. The main results of the paper are placed in Section 3. The system architecture is described in Section 4. The paper ends with a conclusion in Section 5.

2 Background

A three-valued logic, known as Kleene's logic is developed in [14] and has three truth values, truth, unknown and false, where unknown indicates a state of partial vagueness. These truth values represent the states of a world that does not change.

A brief overview of a six-valued logic, which is a generalized Kleene's logic, has been first presented in [16]. The six-valued logic was described in more detail in [10]. In [6] this logic is further developed by assigning probability estimates to formulas instead of non-classical truth values.

The six-valued logic distinguishes two types of unknown knowledge values - permanently or eternally unknown value and a value representing current lack of knowledge about a state [9].

Two kinds of negation, weak and strong negation are discussed in [24]. Weak negation or negation-as-failure refers to cases when it cannot be proved that a sentence is true. Strong negation or constructable falsity is used when the falsity of a sentence is directly established.

Very interesting application of paraconsistent logic is presented in [1], [2], [17] and [18]. A process of decision making, based on a paraconsistent annotated logic, is presented in [3]. The results are obtained applying paraconsistent annotated logic techniques, the para-analyzer algorithm and the baricenter analysis method.

A level-based instruction model is proposed in [19]. A model for student knowledge diagnosis through adaptive testing is presented in [11]. An approach for

integrating intelligent agents, user models, and automatic content categorization in a virtual environment is presented in [22].

Let P be a non-empty ordered set. If $sup\{x, y\}$ and $inf\{x, y\}$ exist for all $x, y \in P$, then P is called a *lattice* [4]. In a lattice illustrating partial ordering of knowledge values, the logical conjunction is identified with the meet operation and the logical disjunction with the join operation.

3 The Scenario

In this scenario we consider a multiple-choice test (MCT) assessing student's understanding of new terms. The test consists of two questions. According to the result of a MCT, understanding of a term is achieved if a student gives a correct answer to each of the two questions about that term. Such tests are placed after a new term has been introduced in the theoretical part of the system.

Modus ponens

$$\frac{(P, P \rightarrow Q)}{Q}$$

can be applied if there is no doubt about the truth-status of P. Most tests are based on the understanding that a 'correct answer to one question about a term' (P) implies 'understanding of that term' (Q).

In our view a single question is not enough, since a student can get a correct answer by guessing. To minimize the amount of opportunities for such occurrences we propose use of two questions.

What should an intelligent agent do in a situation where one of the questions receives a correct answer and the other one receives a wrong answer or is not answered? The binary logic does not provide a model for sets like {correct answer, incorrect answer} or {correct answer, no answer is provided}. Such a situation can be resolved by applying six-valued logic. Based on the theory of six-valued logic in [6], [7], and [8] we propose the following:

- Two correct answers imply understanding of that particular term. The process of questioning is terminated.
- One correct answer and one unanswered question imply some doubt about the student's understanding of that particular term. The system first provides additional explanations and then suggests to the student to answer one new question taken from the database.
- One correct answer and one incorrect answer imply doubt about the student's understanding of that particular term. The system first provides additional explanations and then suggests to the student to answer two questions where one new question is taken from the database and the other question is taken from the first trial and has received an incorrect answer.
- Two unanswered questions imply uncertainty about the student's understanding of that particular term. The system first provides additional explanations and then suggests two new questions taken from the database.

– One incorrect answer and one unanswered question imply doubt about the student's understanding of that particular term. The system first provides additional explanations and then suggests to the student to answer the same questions.
– Two incorrect answers imply lack of understanding of that particular term. The system first provides additional explanations and then suggests to the student to answer the same questions plus one new question taken from the database.

A lattice showing a partial ordering of the above described cases is presented in Fig. 1. The knowledge lattice illustrates how the truth value of a formula that has a temporary truth value can be changed as more knowledge becomes available.

If the second set of responces contains an incorrect answer and/or unanswered questions the system advises the student to work more with the originally provided learning materials and terminates the automated questioning process. We believe that several rounds of questioning would make the learning process time consuming for the student and thus disturb the learning flow.

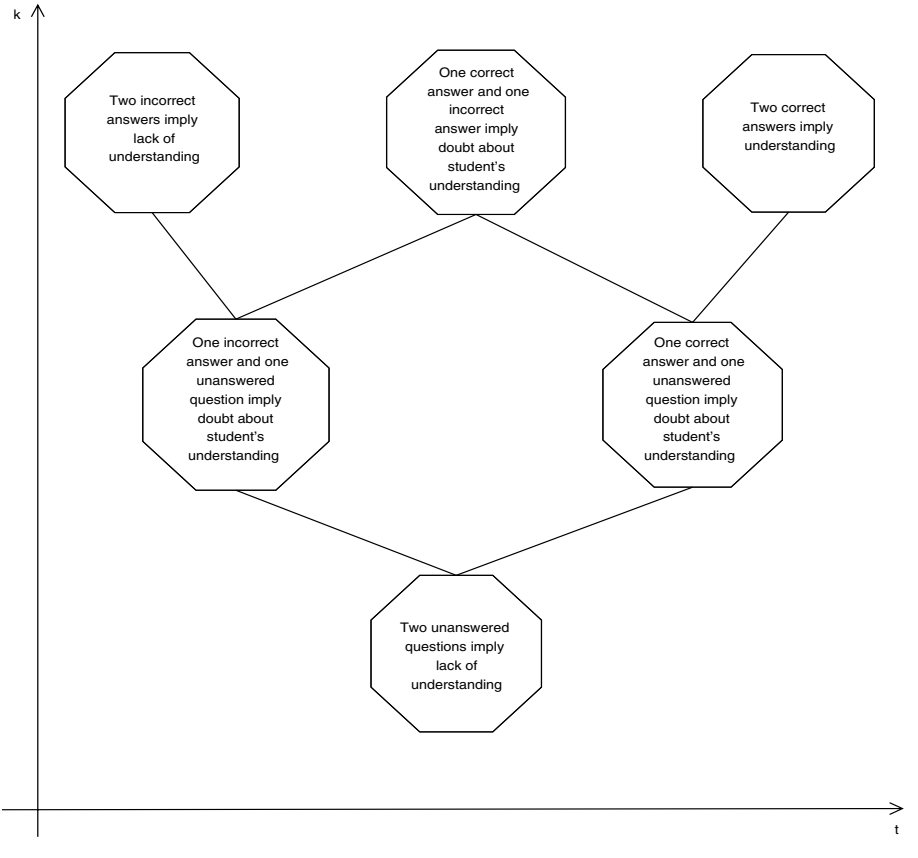


Fig. 1. Knowledge lattice

However, the student can start a new assessment of his/her understanding of that particular term at any time he/she wants.

4 System Architecture

In our earlier work [5] relations between wrong answers and understanding status of a student were implemented in tables of MCTs' distracters and their implications. These static bindings were done early at the implementation stage by the expert human tutor based on her knowledge and teaching experience of the subject matter. Each MCT was expertly crafted as to build tables of inferences to be used by intelligent diagnostic agents.

By employing six-valued logic in an assessment sub-system, we are able to postpone these binding between students' responses to tests and their understanding status at a later stage. This greatly reduces the difficulty of making/choosing appropriate distracters for the MCT. Instead of a wrong answer leading to a consequence that was hard coded in an inference table as in [5], truth value/consequence relations among the six truth values as discussed in [6], [7], and [8] are used to control the next step in the process of a student learning activities. The complexity of managing inference tables is reduced to only managing six inferences for each test.

The system is implemented using the so-called LAMP Web server infrastructure and deployment paradigm. It is a combination of free software tools of an Apache Web server, a database server and a scripting programming platform on a Linux operating environment.

Behind this traditional three-tiers Web deployment is a service support subsystem. Communication framework based on XML-RPC is used to connect the Web application middle-ware and the intelligent assessment/diagnostic system together. The separation of these two units made it possible to modularly design and implement the system as loosely couple independent sub-systems.

An authenticated user receives a unique session key that is used to identify user in the system for that particular session. This session key is saved in the user Web browser cookie. All sessions' dynamic data stored in the database are indexed using this session key. A session key is used to index state variables assigned to a user for that particular session. The session key cookie and dynamic data stored are used together to keep user's states of interaction with the system within otherwise a stateless HTTP protocol.

The dynamic page publisher compiles a page to be presented to the user from a template file in relation to the user response, current state variables and activities history. A template file contains the static declarations of a document. The variables in a particular template files are given values by the dynamic page publisher module during the production of an HTML document. The resulting HTML document is sent back to the user Web browser. This module also acts as a handler when a user requests a page or sends a form back to the Web server.

Each learning unit is atomic, self-contained and reusable. The dynamic page publisher makes use of these learning units to provide students with a dynamic

and personalized learning material as a direct reaction to students' interaction to the system.

The user authenticator authenticates a user during login and creates initial session contact in the system if the user provides correct credentials. It is realized by creating a unique session key and saving that key in the database of the server and a cookie in the client. The module also provides user authorization during an active user's session. The user authenticator module is responsible for session cleanup at user's logoff.

The users stack profiler keeps track of user activities history in a stack like data structure in the database. Each event, like for example response/result of a test or a change of learning flow after following a hint given by the system, is stored in the database. This module provides the percepts to the intelligent modules of the software agents sub-system. The users stack profiler communicates directly with the agents by sending messages over the XML-RPC communication channel. By using some common data stored in the database, the users stack profiler indirectly affects the behavior of the user's agents and vise verse.

The application middleware and the software agents run independently of each other. As such, they can be situated on different servers. The middleware implement the Web side of the system while the software agents implement the decision side of users learning process. Given a certain response to a particular test at a particular user state, what best action can be taken to increase the probability that the user will learn a particular unit of knowledge? This decision is done by the intelligent diagnostics agent.

The intelligent assessment agent is a special agent that implements six-valued logic table of inference for testing students' understanding of a particular basic term by using a pair of questions after the term has been introduced. The agent does an early diagnostic about absorption of knowledge. A response given by a particular student from a test will give the system an indication about the state of learning of a particular term. Each of the six different truth values of a response triggers different rule-based reaction as discussed in Section 3. This agent helps to implement a part of an intelligent tutoring system, which differ from the intelligent diagnostics agent. The intelligent assessment agent facilitates students' early absorption/assimilation of new terms.

5 Conclusion

This paper presents an intelligent sub-system assessing students initial understanding of new terms. The decision making process is based on a six-valued logic. Our motivation for employing six-valued logic is that this way the system will provide better personalized help to the students and course builders will receive more detailed information about the effectiveness of their learning materials. We believe that assessment rules for students' understanding of new terms, shortly after they have been introduced, should differ from the ones used for marking students' achievements in a subject.

References

1. Abe, J.M., Prado, J.C.A., Nakamatsu, K.: Paraconsistent Artificial Neural Network: Applicability in Computer Analysis of Speech Productions. In: Gabrys, B., Howlett, R.J., Jain, L.C. (eds.) KES 2006. LNCS (LNAI), vol. 4252, pp. 844–850. Springer, Heidelberg (2006)
2. Abe, J.M., Torres, L.G., Nakamatsu, K., Kondo, M.: Intelligent Paraconsistent Logic Controller and Autonomous Mobile Robot Emmy II. In: Gabrys, B., Howlett, R.J., Jain, L.C. (eds.) KES 2006. LNCS (LNAI), vol. 4252, pp. 851–857. Springer, Heidelberg (2006)
3. de Carvalho, F.R., Brunstein, I., Abe, J.M.: Decision Making based on Paraconsistent Annotated Logic. In: Nakamatsu, K., Abe, J.M. (eds.) Frontiers in Artificial Intelligence and Applications, vol. 132, pp. 55–62 (2005)
4. Davey, B.A., Priestley, H.A.: Introduction to lattices and order. Cambridge University Press, Cambridge (2005)
5. Encheva, S., Tumin, S.: Responce-Driven Web-Based Assessment System. In: Khosla, R., Howlett, R.J., Jain, L.C. (eds.) KES 2005. LNCS (LNAI), vol. 3683, pp. 749–755. Springer, Heidelberg (2005)
6. Fitting, M.: Kleene's Logic, Generalized. Journal of Logic and Computation 1(6), 797–810 (1991)
7. Fitting, M.: Kleene's three-valued logics and their children. Fundamenta Informaticae 20, 113–131 (1994)
8. Fitting, M.: Tableaus for many-valued modal logic. Studia Logica 55, 63–87 (1995)
9. Garcia, O.N., Moussavi, M.: A Six-Valued Logic for Representing Incomplete Knowledge. In: ISMVL. Proceedings of the 20th International Symposium on Multiple-Valued Logic, Charlotte, NC, USA, May 1990, pp. 110–114. IEEE Computer Society Press, Los Alamitos (1990)
10. Garcia-Duque, J., Lopez-Nores, M., Pazos-Arias, J., Fernandez-Vilas, A., Diaz-Redondo, R., Gil-Solla, A., Blanco-Fernandez, Y., Ramos-Cabrer, M.: A Six-valued Logic to Reason about Uncertainty and Inconsistency in Requirements Specifications. Journal of Logic and Computation 16(2), 227–255 (2006)
11. Guzmàn, E., Conejo, R.: A model for student knowledge diagnosis through adaptive testing. In: Lester, J.C., Vicari, R.M., Paraguaçu, F. (eds.) ITS 2004. LNCS, vol. 3220, pp. 12–21. Springer, Heidelberg (2004)
12. Johnstone, A.: Effective Practice in Objective Assessment, `www.physsci.heacademy.ac.uk/Publications/PracticeGuide/`
13. Kacprzyk, J., Roubens, M.: Non Conventional Preference Relations in Decision Making. LNMES, vol. 301. Springer, Heidelberg (1988)
14. Kleene, S.: Introduction to Metamathematics. D. Van Nostrand Co., Inc., New York (1952)
15. Lukasiewicz, J.: On Three-Valued Logic. Ruch Filozoficzny 5 (1920) English translation in Borkowski, L. (ed.) Jan Lukasiewicz: Selected Works, 1970. North Holland, Amsterdam (1920)
16. Moussavi, M., Garcia, O.N.: A Six-Valued Logic and its application to artificial intelligence. In: Proceedings of the Fift Southeastern Logic Symposium (1989)
17. Nakamatsu, K., Ishikawa, R., Suzuki, A.: A Paraconsistent Logic Program Based Control for a Discrete Event Cat and Mouse. In: Negoita, M.G., Howlett, R.J., Jain, L.C. (eds.) KES 2004. LNCS (LNAI), vol. 3214, pp. 954–960. Springer, Heidelberg (2004)

18. Nakamatsu, K., Akama, S., JAbe, J.M.: An Intelligent Safety Verification Based on a Paraconsistent Logic Program. In: Khosla, R., Howlett, R.J., Jain, L.C. (eds.) KES 2005. LNCS (LNAI), vol. 3682, pp. 708–715. Springer, Heidelberg (2005)
19. Park, C., Kim, M.: Development of a Level-Based Instruction Model in Web-Based Education. In: Luo, Y. (ed.) CDVE 2004. LNCS, vol. 3190, pp. 215–221. Springer, Heidelberg (2004)
20. Priest, G.: An Introduction to Non-Classical Logic, Cambridge (2001)
21. Prior, A.N.: A Statement of Temporal Realism. In: Copeland, B.J. (ed.) Logic and Reality: Essays on the Legacy of Arthur Prior, Clarendon Press, Oxford (1996)
22. Santos, C.T., Osòrio, F.S.: Integrating intelligent agents, user models, and automatic content categorization in virtual environment. In: Lester, J.C., Vicari, R.M., Paraguaçu, F. (eds.) ITS 2004. LNCS, vol. 3220, pp. 128–139. Springer, Heidelberg (2004)
23. Tamir, P.: Justifying the selection of answers in multiple-choice items. International Journal of Science Education 12, 563–573 (1990)
24. Wagner, G.: Vivid Logic. LNCS, vol. 764. Springer, Heidelberg (1994)

Monadic Curry Algebras Qτ

Jair Minoro Abe[1], Seiki Akama[2], and Kazumi Nakamatsu[3]

[1] Graduate Program in Production Engineering, ICET – Paulista University
R. Dr. Bacelar, 1212, CEP 04026-002 São Paulo – SP – Brazil
jairabe@uol.com.br
[2] Kawasaki – Japan
sub-akama@jcom.home.ne.jp
[3] School of Human Science and Environment/H.S.E. – University of Hyogo – Japan
nakamatu@shse.u-hyogo.ac.jp

Abstract. In this paper we present an algebraic version of the annotated predicate logics Qτ [1] by using the concept of Curry Algebras [5]. The algebraic structure obtained is called Curry Algebras Qτ. We study some basic properties of these algebras. The main ideas involved lean heavily on the works by P. Halmos and N.C.A. Da Costa [9], [5].

Keywords: Curry algebras, monadic algebras, paraconsistent annotated logic, algebraic logic.

1 Introduction

Paraconsistent annotated logic is a much discussed subject nowadays. Very significant applications were obtained in AI, Automation and Robotics, as well as in Computer Science. So it became convenient to study from foundational point of view by some authors: N.C.A. da Costa, J.M. Abe, S. Akama, among others. This logic has proved be powerful tool to deal with inconsistencies and paracompleteness in a non-trivial manner. It is to observed that such concepts are more and more common in several contexts in AI, Robotics, and other fields of applications.

Roughly speaking, logics that serve as underlying logic of inconsistent but non-trivial theories are called *paraconsistent logics*. And, logics that serve as underlying logics of theories in which there are propositions p and $\neg p$ (the negation of p) both false are called *paracomplete logics*. Logics that are both, paraconsistent and paracomplete, are called *non-alethic logics*.

One question arises naturally: algebraic version of these logics.

In this paper we introduce the concept of monadic Curry algebras Qτ that algebraises the annotated systems Qt extensively studied by [1]. Some basic initial results are presented.

2 The Annotated Logics Pτ

In this section we summarize the annotated propositional logics Pτ [1], [7]. Throughout this work, $\tau = < |\tau|, \leq >$ will be some arbitrary, but fixed finite lattice of

B. Apolloni et al. (Eds.): KES 2007/ WIRN 2007, Part II, LNAI 4693, pp. 893–900, 2007.
© Springer-Verlag Berlin Heidelberg 2007

truth-values. The least element of τ is denoted by $\bot$, while its greatest element by $\top$. There is a fixed unary operator $\sim$: $|\tau| \to |\tau|$ which constitutes the "meaning" of our negation. $\vee$ and $\wedge$ denote, respectively, the least upper bound and the greatest lower bound operators of τ.

The language of Pτ is composed by the following primitive symbols.

1. Propositional symbols *p, q,* ...
2. Each member of τ is an annotational constant: μ, λ, ...
3. The connectives $\neg$, $\wedge$, $\vee$, and $\to$
4. Auxiliary symbols: parentheses.

Formulas are defined as follows:

1. If p is a propositional symbol and λ is an annotational constant, then p_λ is an (atomic) formula.
2. If A and B are formulas, then $(\neg A)$, $(A \wedge B)$, $(A \vee B)$, and $(A \to B)$ are formulas.

Among several intuitive readings, the atomic formula p_λ can be read as *it is believed that p's truth-value is at least* λ.

We introduce some definitions.

1. The symbol of equivalence '$\leftrightarrow$' is introduced as usual.
2. Let A be a formula. $\sim A = A \to ((A \to A) \wedge \neg(A \to A))$ is called *strong negation of A*.
3. $\underbrace{\neg \ldots \neg}_{k-times} p_\lambda = \neg^k p_\lambda$ is called a *hyper-literal*. A formula other than hyper-literals is called a *complex formula*.

Definition 2.1. An *interpretation I* is a function I: $P \to \tau$. To each interpretation I, we associate a valuation V_I: $F \to \mathbf{2}$, where P is the set of propositional symbols, F is the set of all formulas, and $\mathbf{2} = \{0, 1\}$, defined as follows:

1. If p_λ is an atomic formula, then:
a) $V_I(p_\lambda) = 1$ iff $I(A) \geq \lambda$
b) $V_I(p_\lambda) = 0$ otherwise.
c) $V_I(\neg^k p_\lambda) = V_I(\neg^{k-1} p_{\sim\lambda})$, $k \geq 1$
2. If A and B are formulas, then:
a) $V_I(A \wedge B) = 1$ iff $V_I(A) = V_I(B) = 1$
b) $V_I(A \vee B) = 1$ iff $V_I(A) = 1$ or $V_I(B) = 1$
c) $V_I(A \to B) = 1$ iff $V_I(A) = 0$ or $V_I(B) = 1$
3. If A is a complex formula, then $V_I(\neg A) = 1 - V_I(A)$.

Now we present an axiomatization of Pτ. A, B and C are formulas whatsoever, F and G are complex formulas, p is a propositional symbol, and μ, μ_j, $1 \leq j \leq n$, are annotational constants.

$(\to_1)$ $A \to (B \to A))$
$(\to_2)$ $(A \to (B \to C)) \to ((A \to B) \to (A \to C))$
$(\to_3)$ $((A \to B) \to A) \to A$
$(\to_4)$ $A,\ A \to B / B$ (Modus Ponens, abbreviated by MP)
$(\wedge_1)$ $(A \wedge B) \to A$

$(\wedge_2)$ $(A \wedge B) \to B$
$(\wedge_3)$ $A \to (B \to (A \wedge B))$
$(\vee_1)$ $A \to (A \vee B)$
$(\vee_2)$ $B \to (A \vee B)$
$(\vee_3)$ $(A \to C) \to ((B \to C) \to ((A \vee B) \to C))$
$(\neg_1)$ $(F \to G) \to ((F \to \neg G) \to \neg F)$
$(\neg_2)$ $F \to (\neg F \to A)$
$(\neg_3)$ $F \vee \neg F$
(τ_1) $p_\perp$
(τ_2) $(\neg^k p_\mu) \leftrightarrow (\neg^{k-1} p_{\sim\mu})$, $k \geq 1$
(τ_3) $p_\mu \to p_\lambda$, where $\mu \geq \lambda$
(τ_4) $p_{\mu 1} \wedge p_{\mu 2} \wedge \ldots \wedge p_{\mu n} \to p_\mu$, where $\mu = \bigvee_{j=1}^{n} \mu_j$

In [1], [7] were proved that the axiomatization above is sound and complete with respect to the semantic discussed.

Some remarkable results of the systems Pτ are: (i) The set of formulas together with the connectives $\neg$, $\wedge$, $\vee$, and $\to$ has all properties of the Classical Propositional logic (see theorem 2.3 below). (ii) The set of all complex formulas with the connectives $\neg$, $\wedge$, $\vee$, and $\to$ has also all properties of the Classical Propositional logic. So, the inconsistencies appear among hyper-literals, or equivalently, among atomic formulas. This is a nice property, so in applications, for instance in databases, the conflicting informations are among hyper-literals; hardly we have inconsistencies among complex formulas. (iii) Pτ is a paraconsistent logic and, in general, paracomplete and non-alethic logics.

Definition 2.2. In Pτ: we define $A \leq B$ by setting $\vdash A \to B$ and $A \equiv B$ by setting $A \leq B$ and $B \leq A$.

Theorem 2.1. $\leq$ is a quasi-order and $\equiv$ is an equivalence relation.

Theorem 2.2. In general, in Pτ, $\neg$ is not compatible with the equivalence relation $\equiv$, and we have $\vdash (A \wedge \neg A) \to B$.

Theorem 2.3. The strong negation possesses all properties of the classical negation; for instance, let A, B, and C be formulas whatsoever.

$\vdash (A \to B) \to ((A \to \neg B) \to \neg A)$; $\vdash A \to (\neg A \to B)$; $\vdash (A \wedge \neg A) \to B$; $\vdash A \to \neg\neg A$; $\vdash \neg\neg A \to A$.

3 Pre-algebraic Structures

In order to obtain algebraic versions of majority of logical systems the procedure is the following: we define an appropriate equivalence relation in the set of formulas (v.g., identifying equivalent formulas in classical propositional logic), in a such way that the primitive connectives are compatible with the equivalence relation, i.e., a congruence. The resulting quotient system is the algebraic structure linked with the correspondent logical system. By this process, Boolean algebra constitutes the

algebraic version of the classical propositional logic, Heyting algebra constitutes algebraic version of some intuitionistic propositional logics, and so on.

However, in some non-classical logics, it is not always clear what "appropriate equivalence relation" can be; also it can happen the non existence of any significant equivalence relation among formulas of the calculi. This occurs, for instance, with some paraconsistent systems (see, v.g. [11]). Indeed, as pointed by [8], even for classical logics, it may not always be convenient to apply these ideas.

Definition 3.1. A system $(R, \equiv, \leq)$ is called a *Curry pre-ordered system* if

1. $\equiv$ is an equivalence relation on R
2. $x \leq x$
3. $x \leq y$ and $y \leq z$ imply $x \leq z$
4. $x \leq y$, $x' \equiv x$ and $y' \equiv y$ imply $x' \leq y'$
5. $(R, \equiv, \leq)$ is called a *pre-lattice* if $(R, \equiv, \leq)$ is a Curry pre-ordered system and
6. $\inf\{x, y\} \neq \emptyset$ and $\sup\{x, y\} \neq \emptyset$.
7. $(R, \equiv, \leq, \rightarrow)$ is called an *implicative pre-lattice* if it is a pre-lattice and
8. $x \wedge (x \rightarrow y) \leq y$
9. $x \wedge y \leq z$ iff $x \leq y \rightarrow z$
10. $(R, \equiv, \leq, \rightarrow)$ is called *classic* if it is an implicative pre-lattice and
11. $(x \rightarrow y) \rightarrow x \leq x$ (Peirce's law).

4 Curry Algebras Pτ

The algebraic structures considered here are those in [2]. Let us consider a non-empty set S and a finite lattice (in the usual sense) $\tau = (|\tau|, \leq)$ together with an operator $\sim$: $|\tau| \rightarrow |\tau|$. Let S^* be se set of all pairs (p, λ) where $p \in S$ and $\lambda \in |\tau|$. Consider now the set $S^* \cup \{\neg, \wedge, \vee, \rightarrow\}$. Let S^{**} be the smallest algebraic structure freely generated by the set $S^* \cup \{\neg, \wedge, \vee, \rightarrow\}$ (we do not write the details; it follows usual method to obtain free structures). Elements of S^{**} are classified in two categories: *hyper-literal elements* are of the form $\neg^k(p, \lambda)$ and *complex elements* are the remaining elements of S^{**}.

Now we introduce the concept of a Curry algebra Pτ.

Definition 4.1. A *Curry algebra* Pτ (abbreviated by Pτ-algebra) is a structure $R\tau = (R, (|\tau|, \leq, \sim), \equiv, \leq, \rightarrow, \neg)$ and, for $p \in R$, $a \in R^{**}$, and x, y complex elements of R^{**}:

1. R^{**} is a classical implicative lattice with a greatest element **1**.
2. $\neg$ is a unary operator $\neg$: $R^{**} \rightarrow R^{**}$.
3. $x \rightarrow y \leq (x \rightarrow \neg y) \rightarrow \neg x$
4. $x \leq \neg x \rightarrow a$
5. $(p, \perp) \equiv \mathbf{1}$ ($\perp$ is the bottom element of τ)
6. $x \vee \neg x \equiv \mathbf{1}$
7. $\neg^k(p, \lambda) \equiv \neg^{k-1}(p, \sim\lambda)$, $k \geq 1$
8. $\mu \leq \lambda$ then $(p, \mu) \leq (p, \lambda)$
9. $(p, \lambda_1) \wedge (p, \lambda_2) \wedge \ldots \wedge (p, \lambda_n) \leq (p, \lambda)$, where $\lambda = \bigvee_{i=1}^{n} \lambda_i$

Theorem 4.1. A Pτ-algebra is distributive and has a greatest element, as well as a first element.

Definition 4.2. Let x be an element of a Pτ-algebra. We put $x^* = x \to ((x \to x) \wedge \neg(x \to x))$.

Theorem 4.2. In a Pτ-algebra, x^* is a Boolean complement of x; so $x \vee x^* \equiv 1$ and $x \wedge x^* \equiv 0$. Moreover, in a Pτ-algebra, the structure composed by the underlying set and by operations $\wedge$, $\vee$, and * is a (pre) Boolean algebra. If we pass to the quotient by the basic relation $\equiv$, we obtain a Boolean algebra in the usual sense.

Definition 4.3. Let $(R, (|\tau|, \leq, \sim), \equiv, \leq, \to, \neg)$ be a Pτ-algebra, and $(R, (|\tau|, \leq, \sim), \equiv, \leq, \to, ^*)$ the Boolean algebra obtained as in the above theorem. Any Boolean algebra that is isomorphic to the quotient algebra of $(R, (|\tau|, \leq, \sim), \equiv, \leq, \to, ^*)$ by $\equiv$ is called Boolean algebra *associated with the* Pτ-algebra.

Hence, we have the following representation theorems for Pτ-algebras.

Theorem 4.3. Any Pτ-algebra is associated with a field of sets. Moreover, any Pτ-algebra is associated with the field of sets simultaneously open and closed of a totally disconnected compact Hausdorff space.

This is not the only way of extracting Boolean algebras out of Pτ-algebras. There is another natural Boolean algebra associated to a Pτ-algebra.

Definition 4.4. Let $(R, (|\tau|, \leq, \sim), \equiv, \leq, \to, \neg)$ be a Pτ-algebra. By RC we indicate the set of all complex elements of $(R, (|\tau|, \leq, \sim), \equiv, \leq, \to, \neg)$. Then the structure $(RC, (|\tau|, \leq, \sim), \equiv, \leq, \to, \neg)$ constitutes a (pre) Boolean algebra which we call it as *Boolean algebra c-associated with* the Pτ-algebra $(R, (|\tau|, \leq, \sim), \equiv, \leq, \to, \neg)$.

Thus, we obtain a second representation theorem for Pτ-algebras.

Theorem 4.4. Any Pτ-algebra is *c-associated* with a field of sets. Moreover, any Pτ-algebra is *c-associated* with the field of sets simultaneously open and closed of a totally disconnected compact Hausdorff space.

5 Monadic Curry Algebras Qτ

In this section we introduce the monadic Curry algebras Qτ.

Definition 5.1. Let A be a Pτ-algebra. Let $\exists$ and $\forall$ be operators on A. $(\exists, \forall)$ is called a quantifier on A if

1. $\exists 0 \equiv 0$
2. $x \leq \exists x$
3. $\exists(x \vee y) \equiv \exists x \vee \exists y$
4. $\exists\exists x \equiv \exists x$
5. $\exists(\exists x)^* \equiv (\exists x)^*$
6. $\exists(x \wedge \exists y) \equiv \exists x \wedge \exists y$

7. $\forall 1 \equiv 1$
8. $\forall x \leq x$
9. $\forall(x \vee y) \equiv \forall x \vee \forall y$
10. $\forall\forall x \equiv \forall x$
11. $\forall(\forall x)^* \equiv (\forall x)^*$

We suppose in the above definition that, if $x \equiv y$, then $\exists x \equiv \exists y$ and $\forall x \equiv \forall y$. $\exists$ is called existential quantifier on A and $\forall$ is called universal quantifier on A. The pair $<A, (\exists, \forall)>$ is called a monadic Curry algebra Qτ or a Qτ-monadic algebra (or Qτ-algebra).

Let C be a Pτ-algebra and $A = \{(x_1, x_2, \ldots, x_n) | \, x_i \in C, i = 1, 2, \ldots, n\}$. Let us suppose that if $x, y \in A$, then $x \vee y \in A$ and $\neg x \in A$, and if $(x_1, x_2, \ldots, x_n), (y_1, y_2, \ldots, y_n) \in A$ we define $(x_1, x_2, \ldots, x_n) \equiv (y_1, y_2, \ldots, y_n)$ iff $x_i \equiv y_i$, $i = 1, 2, \ldots, n$. Also, we put $\overline{x} = \bigvee_{i=1}^{n} x_i$, $\underline{x} = \bigwedge_{i=1}^{n} x_i$, and we assume that $(\underbrace{\overline{x}, \ldots, \overline{x}}_{n-times})$, $(\underbrace{\underline{x}, \ldots, \underline{x}}_{n-times}) \in A$. If we define $\exists x = (\overline{x}, \ldots, \overline{x})$ and $\forall x = (\underline{x}, \ldots, \underline{x})$, then A is a Qτ-monadic algebra.

A more useful example of Qτ-monadic algebras is the following. Let C be a Pτ-algebra, a set $K \neq \emptyset$, and C^K is the set of all functions of K into C. Let A be the set such that: (i) A is a Pτ-algebra with respect the pointwise operations, and (ii) if $x \in A$, then the range of x has a supremum $\overline{x}$ and a infimum $\underline{x}$ in C, and the functions that take the value $\overline{x}$ at each point of K and $\underline{x}$ at each point of K are in A. If $\exists x$ and $\forall x$ are defined to be those functions, then A becomes a Qτ-monadic algebra. Every Qτ-monadic algebra obtained in this way is called a C-valued functional algebra with domain K.

6 Qτ-Monadic Algebras and Monadic Algebras

Now we discuss some aspects of the algebraic structures originated in our discussion. In fact, the matter is very rich, but due limitations of this papers we'll concern only on some of them. As in paragraph 4, given a Pτ-algebra, it generates (at least) two (classical) Boolean algebras: associated to and c-associated to. These two Boolean algebras generate two correspondent (classical, in the sense of [9]) monadic algebras. Each one of these extensions, namely Qτ-monadic associated Pτ-algebra and Qτ-monadic c-associated Pτ-algebra, has a representation theorem in the following form.

In the sequel we use the definitions and conventions of paragraph 4.

Theorem 6.1. In a Qτ-algebra, the structure composed by the underlying set and by operations $\wedge$, $\vee$, *, $\exists$, and $\forall$ is a (pre) monadic algebra. If we pass to the quotient by the basic relation $\equiv$, we obtain a monadic algebra in the usual sense [9].

Definition 6.1. Let $(R, (|\tau|, \leq, \sim), \equiv, \leq, \rightarrow, \neg, \exists, \forall)$ be a Qτ-algebra, and $(R, (|\tau|, \leq, \sim), \equiv, \leq, \rightarrow, *, \exists, \forall)$ the monadic algebra obtained as in the above theorem. Any monadic algebra that is isomorphic to the quotient algebra of $(R, (|\tau|, \leq, \sim), \equiv, \leq, \rightarrow, *, \exists, \forall)$ by $\equiv$ is called monadic algebra *associated with the* Qτ-algebra.

Hence, we can establish the following representation theorems for Qτ-algebras.

Theorem 6.2. If C is a Qτ-algebra, then for its associated monadic algebra A, there exists a set X and there exits a Boolean algebra B, such that (i) A is isomorphic to a B-valued functional algebra A' with domain X, and (ii) for every element p of A' there exists a point x in X with $p(x) = \exists p(x)$.

There exists a second way of extracting monadic algebras out of Qτ-algebras. There is another natural monadic algebra associated to a Qτ-algebra.

Definition 6.2. Let $(R, (|\tau|, \leq, \sim), \equiv, \leq, \rightarrow, \neg, \exists, \forall)$ be a Qτ-algebra. By RC we indicate the set of all complex elements of $(R, (|\tau|, \leq, \sim), \equiv, \leq, \rightarrow, \neg, \exists, \forall)$. Then the structure $(RC, (|\tau|, \leq, \sim), \equiv, \leq, \rightarrow, \neg, \exists, \forall)$ constitutes a (pre) monadic algebra which we call it as *monadic algebra c-associated with* the Qτ-algebra $(R, (|\tau|, \leq, \sim), \equiv, \leq, \rightarrow, \neg, \exists, \forall)$.

Thus, we obtain a second representation theorem for Qτ-algebras.

Theorem 6.3. If C is a Qτ-algebra, then for its c-associated monadic algebra A, there exists a set X and there exits a Boolean algebra B, such that (i) A is isomorphic to a B-valued functional algebra A' with domain X, and (ii) for every element p of A' there exists a point x in X with $p(x) = \exists p(x)$.

Theorems 6.2 and 6.3 show us that Qτ-algebras constitute interesting generalizations of the concept of monadic algebras. We finish this work with some open problems. How many non-isomorphic monadic algebras associated with a Qτ-algebra are there? And, how many non-isomorphic monadic algebras c-associated with a Qτ-algebra are there? Establish connections between associated and c-associated monadic Qτ-algebras. All these questions are to be considered in forthcoming papers.

References

1. Abe, J.M.: Fundamentos da lógica anotada (Foundations of annotated logics), in Portuguese. Ph.D. thesis, University of São Paulo, São Paulo, p. 135 (1992)
2. Abe, J.M.: Curry algebras Pt. Logique et Analyse 161-162-163, 5–15 (1998)
3. Abe, J.M., Akama, S.: Annotated logics Qt and ultraproducts. Logique et Analyse 160, 335–343 (1997)
4. Abe, J.M., Akama, S.: On some aspects of decidability of annotated systems. In: Arabnia, H.R. (ed.) Proc. of the International Conference on Artificial Intelligence, IC-AI'2001, vol. II, pp. 789–795. CSREA Press, Las Vegas, Nevada, USA (2001)
5. Barros, C.M., da Costa, N.C.A., Abe, J.M.: Tópico de teoria dos sistemas ordenados: sistemas de Curry, Coleção Documentos, Série Lógica e Teoria da Ciência, IEA-USP, 20, vol. II, p. 132 (1995)
6. Curry, H.B.: Foundations of Mathematical Logic. Dover, New York (1977)
7. Da Costa, N.C.A., Abe, J.M., Subrahmanian, V.S.: Remarks on annotated logics. Zeitschrift f. math. Logik und Grundlagen d. Math. 37, 561–570 (1991)

8. Eytan, M.: Tableaux of Hintikka et Tout ça: un Point de Vue Algebrique. Math. Sci. Humaines 48, 21–27 (1975)
9. Halmos, P.R.: Algebraic Logic, Chelsea Publishing Co. Chelsea Publishing Co., New York (1962)
10. Kleene, S.C.: Introduction to Metamathematics. Van Nostrand, Princeton (1952)
11. Mortensen, C.: Every quotient algebra for C1 is trivial. Notre Dame J. of Formal Logic 21, 694–700 (1977)

A Method for Visualising Possible Contexts

Shigeki Amitani and Ernest Edmonds

Creativity & Cognition Studios,
Australasian CRC for Interaction Design,
University of Technology, Sydney
PO Box 123 Broadway NSW 2007, Australia

Abstract. This paper presents a method for visualising possible sequences of information. Possible sequences of information segments, called contexts, scenarios, narratives, or storytellings, have been used as a tool for exploring and stimulating thinking about possible events, assumptions relating to these events, and courses of actions, i.e. chances for a broad range of information designers from public audiences searching on the internet to analysts and policy makers. Thanks to the search technology, it has become easier to obtain information that a person is looking for, however, it is a laborious task to grab an overview of information space so that an information designer can find contextually relevant information pieces and sequence them into contextually meaningful ways. In this research, we design and develop a series of generative systems that visualises possible contexts.

Keywords: Generative systems, context, scenario generation.

1 Introduction

We are currently developing a system that visualises possible sequences of information out of existing information in order to increase "chances" for information designers to discover possible contexts that otherwise could be missed.

The concept of the system derived from Knowledge Liquidization & Crystallization [1] [2]. Figure 1 shows expected interactions between users and the system. First, information artefacts (existing ones and/or new pieces of information) are collected and stored (left in Figure 1). The system (top in Figure 1) segments the input information artefacts into small semantic units, then it generates restructure them into possible meaningful sequences of information shown in the right of Figure 1.

These outputs work in two ways such as:

1. As final products that a user can enjoy
2. As draft materials that a user can modify (at the centre of 1)
   A generative system visualises possible contexts, i.e. possible sequences of information as "stimulants" [3] regarding to a core theme that a user inputs.

Suppose you are investigating "public art" on the Internet to write, for example, a report. There is a plenty of information regarding to interactive art,

B. Apolloni et al. (Eds.): KES 2007/ WIRN 2007, Part II, LNAI 4693, pp. 901–908, 2007.
© Springer-Verlag Berlin Heidelberg 2007

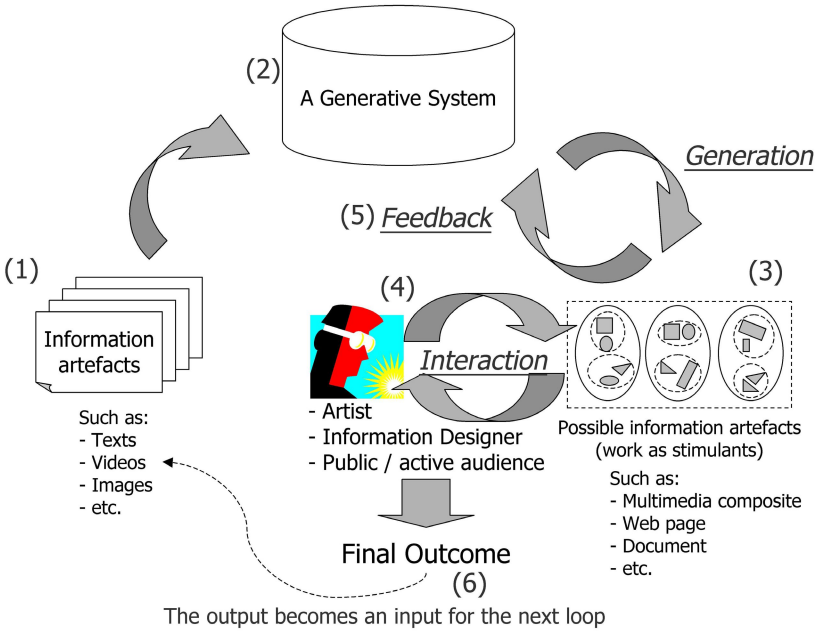


Fig. 1. How a Generative System Works

however, if you are writing a report, you have to decide from which viewpoint you are analysing public art (e.g. history of public art, types of public art, examples of public art, etc.), and also you have to compare which viewpoints yield interesting contexts for the report.

For this aim, you can repeatedly use a search engine to retrieve relevant information. As is often the case, enormous amount of web pages are retrieved. Then you use more keywords to improve the search result.

Problems here are: (1) the search results are limited by viewpoints that you have already in your mind; and (2) the search results with different keywords (viewpoints) actually have to be compared by jumping around multiple pages.

In this research, we propose a method to visualise possible sequences of information to challenge these problems.

2 Related Work

This research takes the knowledge creation viewpoint. Hori et al. have proposed a framework of the process of knowledge creation called Knowledge Liquidization & Knowledge Crystallization in order to fill a gap between theoretical frameworks and human practices [1]. Knowledge Liquidization means dissolving knowledge into small granularity that have a core grounding to the real world and that preserve the local semantic relationships around the core. Knowledge Crystallization

means restructuring the relationships among the granular units in accordance with a current context.

Suppose you are writing a paper, you refer to a number of relevant books and papers. You do not merely pile what is written in them as solid blocks, but you extract relevant parts (such as a paragraph, sentences, etc., i.e. semantic segments) from the books and the papers and fuse them into your paper. This is the process of knowledge creation, and the Knowledge Liquidization & Knowledge Crystallization are correspondent to this process.

They also proposed a conceptual system called Knowledge Nebula Crystallizer that supports the process of Knowledge Liquidization and Knowledge Crystallization [1]. The essential functions are: (1) dissolving information into small semantic segments (corresponding to Knowledge Liquidization); and (2) restructuring relationships among these semantic segments (corresponding to Knowledge Crystallization). Actual systems have been implemented and applied to several domains (e.g., [2] [4] [5]).

Akaishi et al. [6] have developed "Topic Tracer" that arranges segmented documents relevant to an input keyword (= a context) along with their time line. This is useful visualisation when a user knows in which context the user should analyse existing documents. This research challenges the case where a user explores possible contexts.

3 Design Rationales for Generative Systems

In the process of knowledge creation, it is possible to assume:

1. You have a certain topic in your mind
2. You are looking for contextually relevant pieces of information
3. You are trying to sequence the information pieces in a contextually correct way

In order for the generative system to support this process, the interaction should be:

1. A user enters or choose keywords
2. The system returns sequences of information pieces that provide reasonably well-connected contexts as stimulants

Following computations support the interaction:

1. Segmenting information into pieces
2. Restructuring segments into sequences
3. Representing the generated sequences

In the following section, we discuss necessary computations.

3.1 Segmenting Information into Pieces

We adopted the criteria called Term Dependency (TD) and Term Attractiveness (TA) [6]. The ideas of these criteria are:

- If the value of Term Attractiveness of a term in a document is increasing as the document proceeds, then the term is becoming a main topic in this part of the document
- If the value of Term Attractiveness of the main topic term decreases, then the topic has changed to another

The topic terms are defined as words that have the highest value of Term Attractiveness as a document proceeds. Segmentation is conducted at the points where the value of Term Attractiveness of a main topic term decreases. Term Dependency is defined as follows:

$$td(t_1, t_2) = \frac{NumOfSentences(t_1 \cap t_2)}{NumOfSentences(t_1)}$$

where t_1 and t_2 are terms in a given document d, $td(t_1, t_2)$ is the value of Term Dependency of t_1 on t_2 in the document d, and $NumOfSentences(t)$ is the number of sentences that includes t. That is, Term Dependency is a conditional probability that t_2 is included in a sentence that also includes t_1.

Term Attractiveness of a term t in a document d denoted as $attr_d(t)$ is defined as sum of the values of Term Dependency on the term t:

$$attr(t_1) = \sum_t td(t, t_1)$$

By definition, the value of Term Attractiveness of a term t changes as sentences proceed. So the value of Term Attractiveness of a term t_j up to i th sentence is denoted as $attr_i(t_j)$.

When $attr_i(t_j)$ is the maximum value among all the attractiveness values, then the term t_j is regarded as a main topic term. Segmentation is conducted when another term becomes main topic term. As a result of this process, the system produces segments each of which is tagged with its main topic term.

3.2 Restructuring Segments into Sequences

We extend the idea of the Term Dependency to Main Topic Dependency. Main Topic Dependency is defined as follows:

$$mtd(mt_1, mt_2) = \frac{NumOfDocuments(mt_1 \cap mt_2)}{NumOfDocuments(mt_1)}$$

where mt_1 and mt_2 are main topic terms in given document set D, $mtd(mt_1, mt_2)$ is the value of Term Dependency of mt_1 on mt_2 in D, and $NumOfDocuments(mt)$ is the number of documents that include mt. That is, Main Topic Dependency is a conditional probability that mt_2 is included in a document that also includes mt_1.

This value is used for sequencing the segments. If a keyword that a user enter is contained in some segments, then they are retrieved as first segments of sequences. Sequences are generated based on the value of the Main Topic Dependency. Segments with the highest values of the Main Topic Dependency are retrieved and connected to the first segments. Then this process is repeated to make longer sequences.

3.3 Representing the Generated Sequences

The generated sequences are arranged in a two-dimensional space. A number of studies that retrieves information pieces containing designated keywords have been done. The retrieved are usually arranged close to each other to show they are relevant.

This interaction, as implemented in most of other applications, provides an opportunity for users to explore information space, however, it does not support how the retrieved information are inter-related with each other, hence it does not support the process of restructuring existing information so that the user can produce new information artefacts. The user has to actually remember what he/she read in order to articulate the relationships among them.

We propose a representation shown in Figure 2. Each context is represented as a sequence of generated segments. If a document is shared by two contexts, as circled in Figure 2, that means that two contexts are inter-related with each other. We expect that crossing multiple contexts via the document yields a new context.

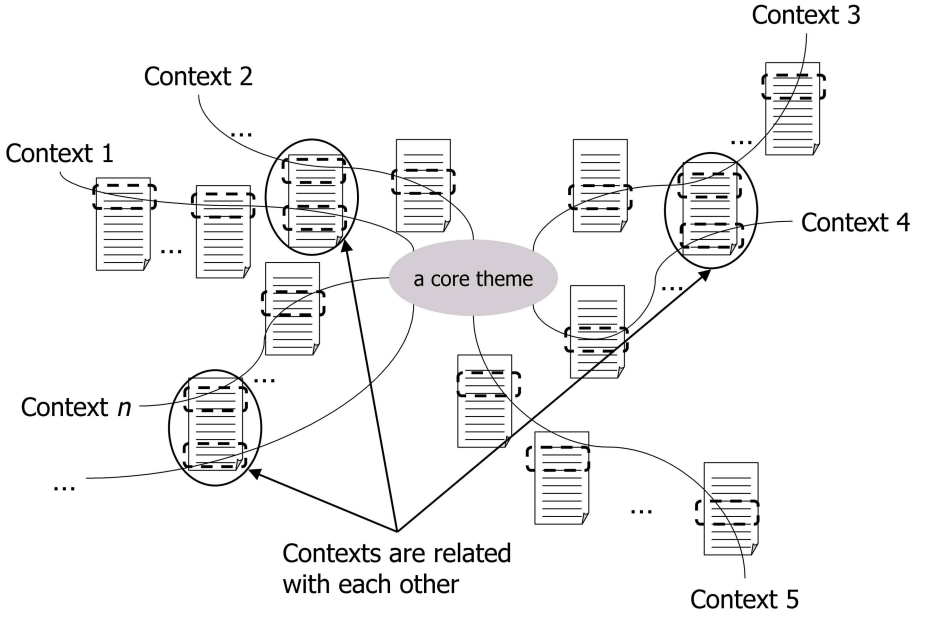


Fig. 2. Each one have to be read to know which is what

4 An Application: Generative Website

We are currently working on the Generative Website Project supported by the Australasian CRC for Interaction Design (ACID). This is a collaborative work with the International Federation of Arts Councils and Culture Agencies (IFACCA). They are renewing their web site[1] in order to provide more dynamic contents to audiences of the web site. The generative website is expected to be a place for followings:

- Audiences of the web site can post their articles, comments, and video clips
- The web site provides dynamically woven sequences of information so that audiences can grab the overview of the contents

We have implemented the algorithms listed above and applied to actual information. The web site of IFACCA has a number of PDF files which are mainly composed of reports from art conferences and news articles. As a first test of the prototype system, text data from 10 PDF files was excerpted and converted into 10 text files. The system segments them into 57 segments.

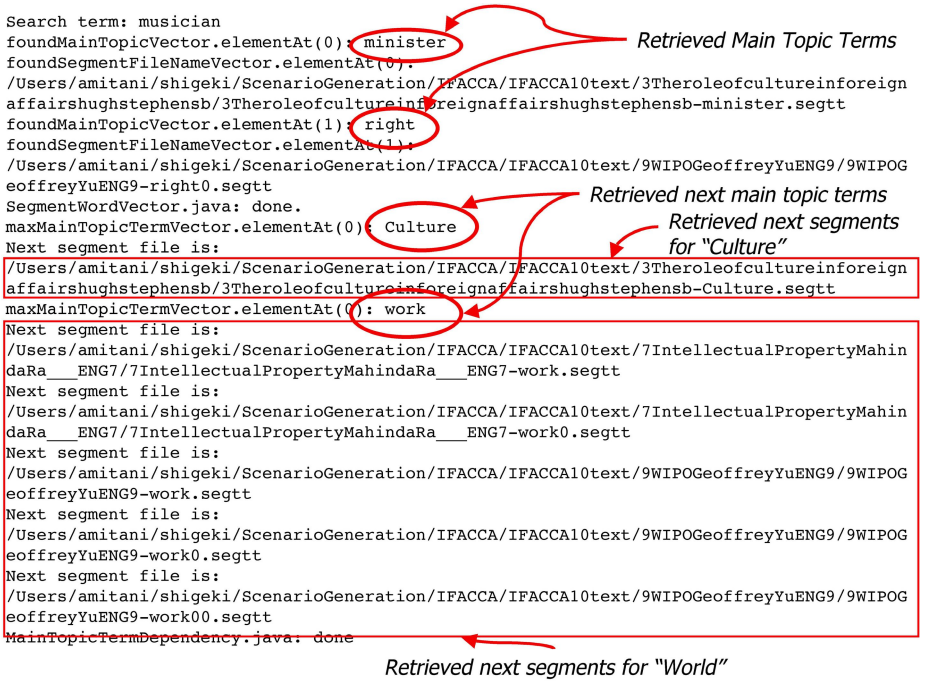


Fig. 3. A Result

[1] http://www.ifacca.org/

A scenario here is that a user is searching "a story" about "musician" on the IFACCA web site. The result is shown in Figure 3. Note that at the time of writing this paper, the system does not have a graphical user interface. The image of this result is shown in Figure 4.

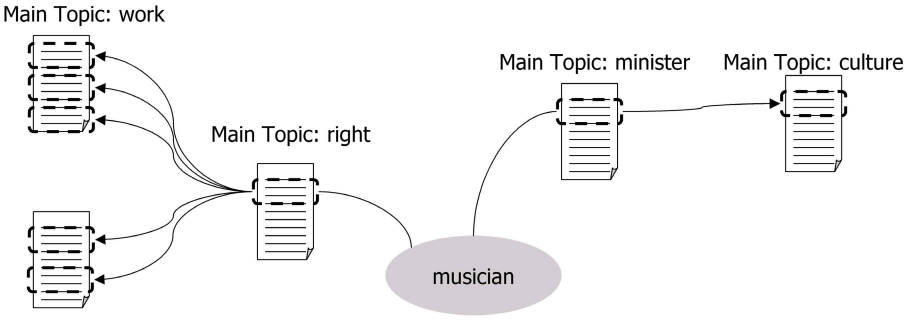


Fig. 4. An Image of the Result

The user inputs a keyword "musician", then the system returns two segments that contain the keyword. The main topics of the retrieved segments are "minister" and "right". Then the system searches other main topic terms "Culture" and "work" on which the two retrieved main topic terms are dependent most. The generated sequences show that this web site discusses "musicians" in two contexts[2]

- Government's actions toward cultural affairs
- Intellectual property issues

Although this result is preliminary that has only two steps and has not been evaluated, our impression is that the system has a potential to connect segments in meaningful ways.

5 Conclusion

In this paper, we presented a method of for visualising possible contexts out of existing information. Main Topic Term Dependency was proposed as criteria for sequencing information segments as a consistent context. Also we presented a prototype system with a preliminary result.

At the time when this paper is written, as the algorithms for the system has just been implemented, it does not have a graphical representation[3]. The

[2] Due to the page limitation, the generated texts in this preliminary test are available at: http://shigekifactory.com/research/GenerativeSystem/PreliminaryTestResult.html

[3] At the time when this camera-ready version is written, a prototype system has been developed. However, describing the prototype and its interactions are not explained in this paper as it is beyond the accepted contents.

interface will be developed for actual user studies. Also we are going to evaluate the algorithms and results through user studies to see the effectiveness of the system. We recognise that this algorithm is still argueable and that other sequencing algorithms (e.g. [7]) could produce semantically more meaningful sequences. These are the future works.

The algorithm presented in this paper is for text-based information, however, we are going to extend this algorithm for non-text-based information such as videos and music.

Acknowledgements

This research/work was partly conducted within the Australasian CRC for Interaction Design, which is established and supported under the Australian Government's Cooperative Research Centres Programme. We are also grateful for the support of Australian Centre for the Moving Image.

References

1. Hori, K., Nakakoji, K., Yamamoto, Y., Ostwald, J.: Organic perspecives of knowledge management: Knowledge evolution through a cycle of knowledge liquidization and crystallization. Jounal of Universal Computer Science 10(3), 252–261 (2004)
2. Amitani, S., Hori, K.: Knowledge nebula crystallizer for knowledge lequidization & crystallization - from a theory to a methodology of knowledge management. In: Proceedings of Expertise In Design, Design Thinking Research Symposium, vol. 6 (2003)
3. Edmonds, E.: Artists augmented by agents (invited speech). In: Proceedings of the 5th international conference on Intelligent user interfaces, New Orleans, Louisiana, United States, pp. 68–73 (2000)
4. Amitani, S.: A Method and a System for Supporting the Process of Knowledge Creation. PhD thesis, Department of Advanced Interdisciplinary Studies University of Tokyo (2004)
5. Amitani, S., Edmonds, E.: Knowledge nebula crystallizer for time-based information. In: Proceedings of Creativity & Cognition 2005, London, England, April 12-16, 2005 (2005)
6. Akaishi, M., Hori, K., Satoh, K.: Topic tracer: A visualization tool for quick reference of stories embedded in document set. In: Keim, D.A., Mansmann, F., Schneidewind, J., Ziegler, H., Tominski, C., Abello, J., Ham, F.v., Schumann, H., Jern, M., Franz, J. (eds.) International Conference on Information Visualisation 2006, London, UK, pp. 101–106 (2006)
7. Nishihara, Y., Sunayama, W., Yachida, M.: Information acquiring support system based on keywords' continuity and informational difficulty. In: International Conference on Human-Computer Interaction, Las Vegas, US (2005)

An Interface for Medical Diagnosis Support

Akinori Abe[1,2], Norihiro Hagita[1,3], Michiko Furutani[1], Yoshiyuki Furutani[1], and Rumiko Matsuoka[1]

[1] International Research and Educational Institute for Integrated Medical Science (IREIIMS), Tokyo Women's Medical University
8-1 Kawada-cho, Shinjuku-ku, Tokyo 162-8666 Japan
[2] ATR Knowledge Science Laboratories
[3] ATR Intelligent Robotics and Communication Laboratories
2-2-2, Hikaridai, Seika-cho, Soraku-gun, Kyoto 619-0288 Japan
ave@ultimaVI.arc.net.my, hagita@atr.jp,
{michi,yoshi,rumiko}@imcir.twmu.ac.jp

Abstract. In this paper, we propose an interactive interface to show results by C4.5 on a web browser, where the physicians can easily check the correspondence data on decision trees. With the interface, physicians can easily confirm or correct clinical data and analyzed results. We also propose a web-based interface that can estimate health levels of unknown data, and that can be used for medical diagnosis support. We show an interactive procedure by using the web-based interface and the possibility of chance discovery process where we can discover hidden or rare but very important relationships in a medical diagnosis support.

Keywords: Medical diagnosis, data mining, interface, chance discovery.

1 Introduction

Medical science as well as clinical diagnosis and treatment has progressed rapidly in recent years. In addition, clinical data that should be dealt with has been increased and become complicated. Therefore, we proposed a framework, called **Cyber Integrated Medical Infrastructure** (CIMI), of integrated management of clinical data on computer networks consisting of a database with a visualizer, a knowledge base, an inference component for diagnoses, and a data analysis (learning or data mining) component, which are connected to each other in the network Abe et al., 2007. In addition, the generated and collected data can be used for medical training of students or physicians as well as medical diagnoses. In this framework, medical information (e.g. clinical data) is analyzed or data mined to build a knowledge base for predicting all possible diseases and to support medical diagnosis. We previously applied C4.5 Quinlan, 1993 to analyze the clinical data and showed the possibility for a chance discovery procedure Abe, Kogure and Hagita, 2003, Abe et al., 2007. Various researches (e.g. Tsumoto, 2004) showed the importance of data mining of medical data to discover rare or unknown rules for a medical diagnosis. Though we could show the importance of data mining technique in medical data analysis, the result was

B. Apolloni et al. (Eds.): KES 2007/ WIRN 2007, Part II, LNAI 4693, pp. 909–916, 2007.
© Springer-Verlag Berlin Heidelberg 2007

not easy to be understood by general physicians. Therefore, it is necessary to develop a understandable interface for general physicians.

In this paper, as a first step, we propose an interactive interface to show results by C4.5 on a web browser, where the user can check the correspondence data on decision trees. We also propose a web-based interface that can show health levels of unknown data, and that can be used for a medical diagnosis support. Then we show the merit of using the web-based interface to manage and analyze large scaled clinical data. In addition, we show the possibility of chance discovery on the interface, where hidden or rare but very important relationships can be discovered.

In fact, powerful interactive data mining tools such as Clementine have been developed, where usual interactivity was developed for the purpose of the gradual reconstruction of decision trees for obtaining better decision or controlling users' focus on decision trees. Of course, construction of better decision trees is important in data mining process. However our main purpose is to discover rare or exceptional relationships. In addition, we also aim to discover human related errors (e.g. mis-writing) in the data. Therefore, our interactivity is mainly developed from the viewpoint of effective interface for discovering rare or exceptional data. No other data mining tools except KeyGraph Ohsawa, Benson and Yachida, 1998 explicitly have this type of objective. KeyGraph can suggest rare or hidden events by its visualization, but cannot suggest data that might have errors.

In Section 2, we illustrate medical data analyses in CIMI. In Section 3, we propose an interactive medical diagnosis support by using a web-based interface. In addition, we overview the possibility of chance discovery process in a medical diagnosis support.

2 Medical Data Analysis

In this section, we illustrate data mining of medial data. For the database in CIMI, medical data includes patients' clinical and personal data. Currently, more than 130 items are included in the clinical data. These items are categorized into 5 categories, which are 1) liver, pancreas, and kidney test data (Total protein, albumin, serum protein fraction-α1-globulin), 2) metabolic function test data (Na, K, Ferritin, total acid phosphatase), 3) general urine test data (Urobilinogen, urine acetone) 4) blood and immunity test data (Mycoplasma pneumoniae antibody, cellular immunity), and 5)tumor markers (Immunosuppressive acidic protein, Sialyl Le X-i antigen, urine β2-microglobulin). In addition, health levels are assigned to clinical data for each person. Health levels that express the health status of patients are defined based on "Tumor stage" Kobayashi and Kawakubo, 1994 and modified by Matsuoka[1]. Currently, the health levels are given by physicians with reference to clinical data and clinical interviews.

[1] Kobayashi classified health levels into 5 classes (1,2,3,4,5). In addition, Matsuoka classified them into 8 classes (1,2,3,4a,4b,4c,5a,5b), with which we can more precisely review patients' status.

As pointed out above, medical science as well as clinical diagnosis and treatment has progressed rapidly in recent years. Accordingly, medical data that should be analyzed has been increasing. Of course, such data are also intentionally collected for medical research purposes. We are now collecting various types of clinical data that are obtained in blood and urine tests. Currently, more than 130 items, including tumor markers, are included in the clinical data. Previously, such data was analyzed by physicians. However, due to the enormous increase in the amount, complexity and variety of the data, it is rather difficult, even for specialized physicians, to handle them. Therefore, automatic analysis is necessary to deal with this problem. In the field of artificial intelligence, the data mining technique is widely applied to analyze numerous and complex data to discover the features of the data. We applied C4.5 Quinlan, 1993, which is a machine learning tool that generates a decision tree, to the medical data of around 1200 individuals[2]. We apply C4.5 because it generates logical formulae and does not require any background knowledge. Since we try to discover novel, hidden, or rare relationships as well as well-known relationships, it is rather difficult to prepare background knowledge and sometimes improper background knowledge leads to incorrect results. In addition, we applied a statistical method to the data and compared the results. Statistical methods cannot generate logical formulae, but they can generate images of data relationships and the results are intuitive and understandable. Thus the generated decision tree can be regarded as logical formulae representing the relationships between clinical data and health levels.

In fact, when we apply C4.5 to the clinical data, we can obtain decision tree. For instance, we can obtain decision tree such as below (part of the generated decision tree).

```
ICTP <= 5.8
|  CYFRA <= 2.1
|  |  TK Activity <= 5
|  |  |  Leucocyte image:Neutro <= 40
|  |  |  |  SCC Antigen <= 0.9 : 4a.......
```

The shape of the generated decision tree has a tree or hierarchical pattern, and does not seem like a logical formula. However, it contains logical interpretations. For instance, the above decision tree can be interpreted as the following formula (if-then style).

If ICTP $\leq$ 5.8, CYFRA $\leq$ 2.1, ..., and SCC Antigen $\leq$ 0.9 then health level = 4a.

Thus, if sets of the generated decision trees are confirmed by physicians, we can use them as sets of rules for a medical diagnosis support. However, the most serious problem is that since ordinal physicians are not familiar with decision tree they cannot interpret the meaning of decision trees. We can obtain more complicated relationships than those shown above. Therefore, we need to prepare an understandable interface for them. Even for researchers who are familiar with

[2] Some data are taken from the same person on different days.

the decision tree approach, it would be better to provide understandable interface to conduct chance discovery Ohsawa and McBurney, 2003.

3 Medical Diagnosis Support

3.1 Clinical Data Analysis Interface

In the previous section, we illustrated data mining of clinical data. We applied C4.5 to obtain relationships between health levels and medical data. At the end of the previous section, we pointed out that for those who are not familiar with decision tree, it is not easy to understand or comprehend decision tree. In addition, we sometimes need to check which patients are on each trees and leaves. Therefore, we have developed an interface that can show who are on each trees and leaves. Results by C4.5 can be displayed on web browsers and leaves have link anchors that can show various information on data in the decision tree. The current interface is shown in Fig. 1. If we click link anchors on leaves, another browser will be coming up that has more detailed information (the center browser in Fig. 1). In the second (center) browser, we can check a range of each items relating to the health level. For instance, for health level 5a, the range for *TK Activity* is less than 5.0[3]. We can also check who are on the leaf that has a certain health level (e.g. "5a") in the top of the second browser. In addition, we can check individual data by following links in the top of the second browser. Thus we can review results by C4.5 in the organic view and we can check each value relating to certain health levels in detail.

The other merit of using the interface is that we can confirm or check the clinical data. For instance, in Fig. 2, differently classified data are shown in different colour (red: 2012,2006-06-01,3). In the case, a datum that has health level 3 is on the leaf of health level 4b. If we encounter differently classified data, by checking the data on the browser, we can confirm whether the health levels are correct or not. When the health level is incorrect, we can check individual data to change the health level. On the other hand, when the health level is correct, we should check the clinical data. This procedure sometimes works as a correction of error in data recording or writing. In addition, when both data and health level are correct, we might be able to discover hidden factors that have serious effects on patients health. For instance, for the data of "2012,2006-06-01,3" we can see several out-ranged data, if we check each data, we can determine if he/she is in health level 3 or 4b. If he/she should be categorized to 3, data that do not appear on decision trees might give positive effects on patient, which we can call chances and we can add new rule as discovered in this process. On the other hand, if a person who has health level 5a is classified to 4a, and if he/she

[3] Of course, the other factors also affect the health level. In fact, in the case Fig. 1, the range of *CA72-4* is more than 4.0. If the range of the data is outside of the standard value range, the item is shown in red color. The standard value is obtained by referring to the data library of Mitsubishi Chemical Medience Corporation (`http://www.medience.co.jp/`).

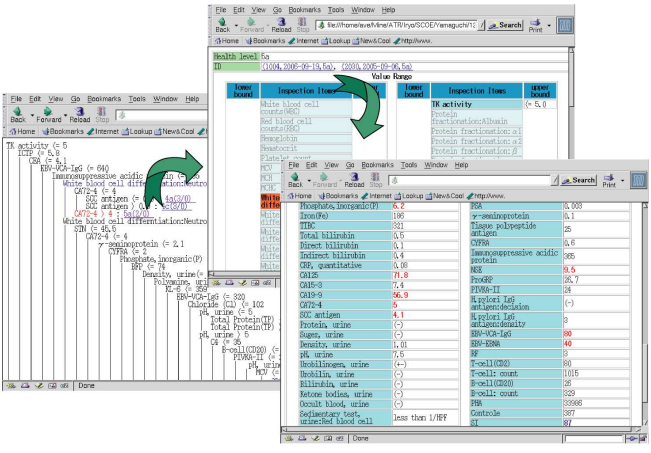


Fig. 1. Interactive interface of clinical data analysis

should be categorized to 5a, data that do not appear on decision trees might give negative effects on patient, which we can call chances and we can add new rule as discovered in this process. Of course, in the process, we can discover wrong classifications. Thus we can check the data from the various viewpoint as well as analyzing the clinical data on the web interface.

3.2 Medical Diagnosis Support Interface

In the previous section, we illustrated the developed web interface for analysis of clinical data. The same interface can be used for a medical diagnosis support, because in the above application, all clinical data are checked to determine decision trees that can explain them. The same procedure can be applied to determine health levels of clinical data. Thus for a medical diagnosis support, we mean that the system can estimate patients' health levels by applying results of known patients' data analyses by C4.5. Since we are planing to develop an integrated medical diagnosis system involving a data analyzer and a decision maker in medical cases.

Figure 3 shows the first browser that shows estimated health levels of each patients. To generate health levels, we applied the results of the 1200 individuals' clinical data analysis by C4.5. As pointed out above, we can adopt the rules shown in the decision trees for the estimation. In this framework, we directly use the decision trees for the estimation. That is, health levels are calculated by applying each data to the decision tree. As shown in Fig. 3, a list of estimated health levels for each patient are shown in the left window. In addition, the applied decision trees are shown in the right window. When we click an ID information, a tree that contains the ID's data appears (moves) in the top of the right window. The following procedure is the same as that shown above. We can check the data on leaves to confirm or change health levels.

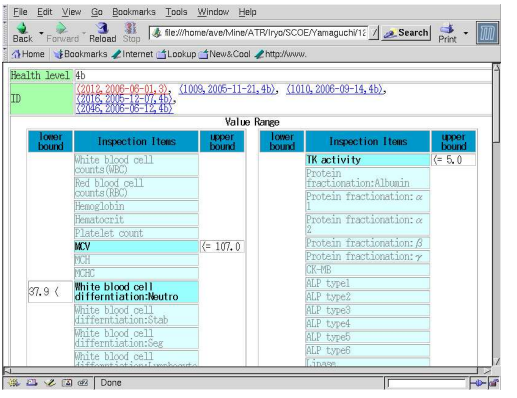


Fig. 2. Interactive interface of clinical data analysis

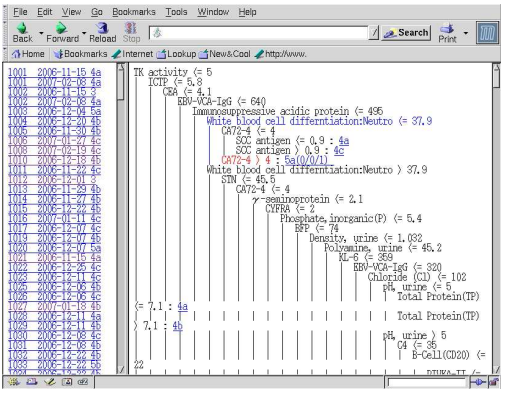


Fig. 3. Interactive interface for medical diagnosis support

For Fig. 4, we can review a certain person whose health level is estimated as 5a. In the right window, we can check the selected data to confirm if he/she is in the health level of 5a. Similarly, we can view all the estimated health levels at a glance and check doubtful or uncertain person's data. In fact, since the quality of the estimation depends upon the analysis of 1200 samples by C4.5, currently the system will not always generate correct estimations. However, from the viewpoint of accessibility and handling of large and complicated clinical data, the system provides a flexible and understandable environment to conduct chance discovery as well as medical diagnosis. For instance, if we focus on wrongly estimated data, we can sometimes discover hidden factors that have serious or positive effects on patients' health. In some cases, physicians focus on special data in their diagnoses. However, since statistical analysis such as C4.5 treat all data equally, it is difficult to do such unbalanced analysis. If we use this

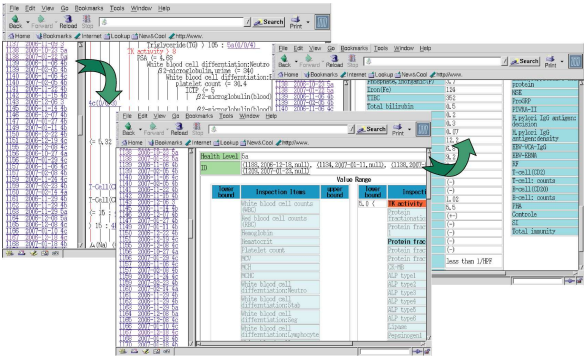


Fig. 4. Interactive interface for medical diagnosis support

type of interface, we can handle such an unbalanced data set after statistical analyses. Thus unfortunately the current system only shows estimated health levels as suggestions. If physicians are doubtful about the result, they can check the original data in detail. In this procedure, chance discovery can be conducted. After chance discovery, we can tuneup the analysis system. For AI researchers, since decision trees have a structure, the interface is very easy to understand and intuitive.

To obtain exact health levels is not the main aim of this paper, though of course it is very important for medical diagnosis. However, in the future, we need to develop or improve the estimation quality. If we can collect large amount of clinical data to analyze and tuneup the data mining engine, the estimation quality will be improved. As shown in the above, the estimation quality can be improved by using the developed interface, because it can support chance discovery procedure.

4 Conclusions

In this paper, we proposed an interactive interface for a medical diagnosis support. First, we showed a clinical data analysis interface where individual data are placed on decision trees and we can review individual data by following links on a web browser. We pointed out that by using the interface, we can check individual data to change or confirm health levels. Second, we showed a medical diagnosis support interface where individuals' health levels are estimated (calculated). Since it uses the same interface as a clinical data analysis interface, it can also change or confirm health levels. In addition, we pointed out that during interaction, we might be able to discover hidden factors that might have serious or positive effects on patients health. In Abe et al., 2007, for chance discovery we proposed an integrated data mining approach, where data are categorized into small number of items and each small sized data are analyzed to discover less influential data. In this framework, we can check an individual data that might have exceptional data to discover rare or hidden relationships.

Though we found the proposed system is very useful to discover hidden, rare or unusual data and relationships, it has yet not checked by physicians. It need to be checked by them to improve the interface according to their usage.

In this paper, we applied C4.5 as a data mining tool. We did not modify the quality of C4.5, but we added a pre-processing and a post-processing phase that is supported by the interface. The pre-processing usually works as data formatting and cleaning. This is an usual case in data mining. Furthermore, the post-processing is important to conduct chance discovery where we might discover hidden or rare relationship. Since the interface designed for general usage, we can apply it to another or improved data mining method.

Our future work is to develop an intuitive visualization of decision trees. As mentioned above, the output from C4.5 is not easy to understand for those who are not familiar with it. In the future, we need to develop an understandable interface for those who are not familiar with C4.5.

Acknowledgments

This research was supported in part by the Program for Promoting the Establishment of Strategic Research Centers, Special Coordination Funds for Promoting Science and Technology, Ministry of Education, Culture, Sports, Science and Technology (Japan).

References

[Abe, Kogure and Hagita, 2003]Abe, A., Kogure, K., Hagita, N.: Discovery of Hidden Relations from Medical Data. In: Proc. of HCI2003 3rd. Int'l Workshop on Chance Discovery, pp. 37–43 (2003)

[Abe et al., 2007]Abe, A., Hagita, N., Furutani, M., Furutani, Y., Matsuoka, R.: Possibility of Integrated Data Mining of Clinical Data. Post-Proc. of CODATA06, Data Science Journal (2007)

[Ohsawa, Benson and Yachida, 1998]Ohsawa, Y., Benson, N.E., Yachida, M.: KeyGraph: Automatic Indexing by Co-occurrence Graph based on Building Construction Metaphor. In: Proc. of IEEE ADL'98, pp. 12–18 (1998)

[Ohsawa and McBurney, 2003]Osawa, Y., McBurney, P. (eds.): Chance Discovery. Springer, Heidelberg (2003)

[Tsumoto, 2004]Tsumoto, S.: Mining Diagnostic Rules from Clinical Databases Using Rough Sets and Medical Diagnostic Model. Information Sciences 162(2), 65–80 (2004)

[Kobayashi and Kawakubo, 1994]Kobayashi, T., Kawakubo, T.: Prospective Investigation of Tumor Markers and Risk Assessment in Early Cancer Screening. Cancer 73(7), 1946–1953 (1994)

[Quinlan, 1993]Quinlan, J.R.: C4.5: Programs for Machine Learning. Morgan Kaufman, San Francisco (1993)

Creating Chances Through Cognitive Niche Construction

The Role of Affordances

Lorenzo Magnani

Department of Philosophy and Computational Philosophy Laboratory,
University of Pavia, Pavia, Italy,
and Department of Philosophy, Sun Yat-sen University,
Guangzhou (Canton), P.R. China
lmagnani@unipv.it

Abstract. As a matter of fact, humans continuously delegate and distribute cognitive functions to the environment to lessen their limits. They build models, representations, and other various mediating structures, that are considered to aid thought. In doing these, humans are engaged in a process of cognitive niche construction. In this sense, I argue that a cognitive niche emerges from a network of continuous interplays between individuals and the environment, in which people alter and modify the environment by mimetically externalizing fleeting thoughts, private ideas, etc., into external supports. Through mimetic activities humans create *external semiotic anchors* that are the result of a process in which concepts, ideas, and thoughts are projected onto external structures. Once concepts and thoughts are externalized and projected, new *chances* and ways of inferring come up from the blend. For cognitive *niche construction* may also contribute to make available a great portion of knowledge that otherwise would remain simply unexpressed or unreachable.

Keywords: Chance, Cognitive Niche, Abduction, Affordances.

1 Introduction

As a matter of fact, humans continuously delegate and distribute cognitive functions to the environment to lessen their limits. They build models, representations, and other various mediating structures, that are thought to be good to think. The aim of this paper is to shed light on these *designing* activities. In the first part of the paper I will argue that these designing activities are closely related to the process of niche construction. I will point out that in building various mediating structures, such as models or representations, humans alter the environment and thus create *cognitive niches*.

In dealing with the exploitation of cognitive resources and chances embedded in the environment, the notion of *affordance*, originally proposed by Gibson [1] to illustrate the hybrid character of visual perception, can be extremely relevant.

B. Apolloni et al. (Eds.): KES 2007/ WIRN 2007, Part II, LNAI 4693, pp. 917–925, 2007.
© Springer-Verlag Berlin Heidelberg 2007

The analysis of the concept of affordance also provides an alternative account about the role of external – also artifactual – objects and devices. Artifactual cognitive objects and devices extend, modify, or substitute "natural" affordances actively providing humans and many animals with new opportunities for action [2]. In order to solve various controversies on the concept of affordance, I will take advantage of some useful insights that come from the study on *abduction* [3]. Abduction may fruitfully describe all those human and animal hypothetical inferences that are operated through actions which consist in smart manipulations to both detect new affordances and to create manufactured external objects that offer new affordances.

2 Humans as Chance Seekers

2.1 Incomplete Information and Human Cognition

Humans usually make decisions and solve problems relying on incomplete information [4]. Having incomplete information means that 1) our deliberations and decisions are never *the best* possible answer, but they are at least *satisficing*; 2) our conclusions are always *withdrawable* (i.e. questionable, or never final). That is, once we get more information about a certain situation we can always revise our previous decisions and think of alternative pathways that we could not "see" before; 3) a great part of our job is devoted to elaborating conjectures or hypotheses in order to obtain more adequate information. Making conjectures is essentially an *act* that in most cases consist in manipulating our problem, and the representation we have of it, so that we may eventually acquire/create more "valuable" knowledge resources. Conjectures can be either the fruit of an abductive selection in a set of pre-stored hypotheses or the creation of new ones, like in scientific discovery. To make conjectures humans often need more evidence/data: in many cases this further cognitive action is the only way to simply make possible (or at least enhance) a thought to "hypotheses" which are hard to successfully perform. In this sense, humans can be considered *chance seekers*, because they are continuously engaged in a process of building up and then extracting latent possibilities to uncover new valuable information and knowledge.

The idea I will try to deepen in the course of this paper is the following: as chance seekers, humans are *ecological engineers*. That is: humans like other creatures do not simply live their environment, but they actively shape and change it looking for suitable chances. In doing so, they construct *cognitive niches* [5,6,7] through which the offerings provided by the environment in terms of cognitive possibilities are appropriately selected and/or manufactured to enhance their fitness as chance seekers. Hence, this ecological approach aims at understanding cognitive systems in terms of their *environmental situatedness* [8,9]. Within this framework, chances are that "information" which is not stored internally in memory or already available in an external reserve but that has to be "extracted" and then *picked up* upon occasion.

As I will show in the following two sections, "the active interrogation of the environment" is also at the root of the evolution of our organism and its cognitive

system. I will also describe this ecological activity by using the notion of abduction [3] and its semiotic dimension. (I have already treated the relationships between abduction and chance in [8,10]).

2.2 Cognitive Niche Construction and Human Cognition as a Chance-Seeker System

It is well-known that one of the main forces that shape the process of adaptation is natural selection. That is, the evolution of organisms can be viewed as the result of a selective pressure that renders them well–suited to their environments. Adaptation is therefore considered as a sort of *top-down process* that goes from the environment to the living creature [11]. In contrast to that, a small fraction of evolutionary biologists have recently tried to provide an alternative theoretical framework by emphasizing the role of niche construction [12] [13].

According to this view, the environment is a sort of "global market" that provides living creatures with unlimited possibilities. Indeed, not all the possibilities that the environment offers can be exploited by the human and non-human animals that act on it. For instance, the environment provides organisms with water to swim in, air to fly in, flat surfaces to walk on, and so on. However, no creatures are fully able to take advantage of all of them. Therefore, all organisms try to modify their surroundings in order to better exploit those elements that suit them and eliminate or mitigate the effect of the negative ones. This process of *environmental selection* [14] allows living creatures to build and shape the so-called "ecological niches". An ecological niche can be defined as a "setting of environmental features that are suitable for an animal" [1]. It differs from the notion of habitat in the sense that the niche describes *how* an organism lives its environment, whereas habitat simply describes *where* an organism lives. In any ecological niche, the selective pressure of the *local* environment is drastically modified by organisms in order to lessen the negative impacts of all those elements which they are not suited to. Indeed, this does not mean that natural selection is somehow halted.

My contention is that the notion of niche construction can be also usefully applied to human cognition. More precisely, I claim that cognitive niche construction can be considered as one of the most distinctive traits of human cognition, i.e. humans construct "cognitive niches".[1] It emerges from a network of continuous interplay between individuals and the environment, in which they more or less tacitly manipulate what is occurring outside at the level of the various structures of the environment in a way that is suited to them. Accordingly, we may argue that the creation of cognitive niches is *the* way cognition evolves, and humans can be considered as ecological cognitive engineers.

[1] If we also recognize in animals, like many ethologists do, a kind of nonlinguistic thinking activity basically model-based (i.e. devoid of the cognitive functions provided by human language), their ecological niches can be called "cognitive", when for example complicate animal artifacts like landmarks of caches for food are fruit of "flexible" and learned thinking activities which indeed cannot be entirely connected with innate endowments [15].

Recent studies on distributed cognition seem to support my claim [16,17,18,19]. According to this approach, cognitive activities like, for instance, problem solving or decision-making, cannot only be regarded as internal processes that occur within the isolated brain. Through the process of niche creation humans extend their minds into the material world, exploiting various external resources. For "external resources" I mean everything that is not inside the human brain, and that could be of some help in the process of deciding, thinking about, or using something. Therefore, external resources can be artifacts, tools, objects, and so on. They are property of individuals in so far as those individuals are embedded in given settings or environments [20]. Something important must still be added, and it deals with the notion of representation: the traditional notion of representation as a kind of abstract mental structure is old-fashioned and misleading [21]. If some cognitive performances can be viewed as the result of a smart interplay between humans and the environment, the representation of a problem is partly internal but it also depends on the smart interplay between the individual and the environment.

As I have already said, an alternative definition of the ecological niche that I find appealing in treating our problem has been provided by Gibson [1]: he pointed out that a niche can be seen as a set of *affordances*. My contention is that the notion of affordance may help provide sound answers to the various questions that come up with the problem of ecological niches. The notion of affordance is fundamental for two reasons. First of all, it defines the nature of the relationship between an agent and its environment, and the mutuality between them. Second, this notion may provide a general framework to illustrate humans as chance seekers.

3 Affordances as Eco-Cognitive Interactional Chances

As I have illustrated in the first part of this paper, humans and some animals manipulate and distribute cognitive meanings after having delegated them to suitable environmental supports. The activity of cognitive niche construction reveals something important about the human and animal cognitive system. As already mentioned, human cognition can be better understood in terms of its environmental situatedness. This means humans do not retain in their memory an explicit and complete representation of the environment and its variables, but they actively manipulate it by picking up information and resources upon occasion. Information and resources are not only given, but they are actively sought and even manufactured. In this sense, I consider human cognition as a chance-seeker system. In my terminology, chances are not simply information, but they are "affordances", namely, *environmental anchors* that allow us to better exploit external resources.

One of the most disturbing problems with the notion of affordance is that any examples provide different, and sometimes ambiguous insights on it. This fact makes very hard to give a conceptual account of it. That is to say, when making examples everybody grasps the meaning, but as soon as one tries to conceptualize

it the clear idea one got from it immediately disappears. Therefore, I hope to go back to examples from abstraction without loosing the intuitive simplicity that such examples provide to the intuitive notion.

Gibson defines "affordance" as what the environment offers, provides, or furnishes. For instance, a chair affords an opportunity for sitting, air breathing, water swimming, stairs climbing, and so on. Gibson did not only provide clear examples, but also a list of definitions that may contribute to generating possible misunderstanding: 1) affordances are opportunities for action; 2) affordances are the values and meaning of things which can be directly perceived; 3) affordances are ecological facts; 4) affordances imply the mutuality of perceiver and environment.

I contend that the Gibsonian ecological perspective originally achieves two important results. First of all, human and animal agencies are somehow hybrid, in the sense that they strongly rely on the environment and on what it offers. Secondly, Gibson provides a general framework about how organisms directly perceive objects and their affordances, as behavioral and cognitive chances. His hypothesis is highly stimulating: "[...] the perceiving of an affordance is not a process of perceiving a value-free physical object [...] it is a process of perceiving a value-rich ecological object", and then, "physics may be value free, but ecology is not" [1, p. 140]. These two issues are related, although some authors seem to have disregarded their complementary nature.

We can find that a very important aspect that is also disregarded in literature is the dynamic one, related to designing affordances with respect to their evolutionary framework: human and non-human animals can "modify" or "create" affordances by manipulating their cognitive niches. Moreover, it is obvious to note that human, biological bodies themselves evolve: and so we can guess that even the more basic and wired perceptive affordances available to our ancestors were very different from the present ones. Of course different affordances can be detected in children, and in the whole realm of animals.

Moreover, organisms need to become *attuned* to the relevant offered features and much of the cognitive tools built to reach this target are a result of evolution and of merely wired and embodied perceptual capacities like imagistic, empathetic, trial and error, and analogical devices. These capabilities, that in our epistemological perspective have to be considered "cognitive" even if instinctual, can be seen as devices of organisms that provide potential implicit abductive powers: they can provide an overall appraisal of the situation at hand and so orient action, and can be seen as forms of pseudo-explanation of what is occurring over there, as emerging in that material contact with the environment granted in the perceptual interplay. It is through this embodied process that affordances can arise as chances both in wild and artificially modified niches. Indeed, humans and animals make available – create – new affordances through the manipulation of the environment and the construction of artifacts; moreover these artifacts are often fruit of high-level – not merely reflex-based, instinctual – cognitive plastic endowments.

3.1 Affordances and Abduction

Let us consider the case of a chair that affords sitting. Now, my point is that we should distinguish between two cases: in the first one, the cues we come up with (flatness, robustness, rigidity) are *highly diagnostic* chances to know whether or not we can sit down on it, whereas in the second case we eventually decide to sit down, but we do not have any precise clue about. How many things are there that are flat, but one cannot sit down on? A nail head is flat, but it is not useful for sitting. This example introduces two important elements: firstly, finding/constructing affordances deals with a (semiotic) inferential activity; secondly, it distinguishes between an affordance and the information that specify it that only arise in the *eco-cognitive interaction* between environment and organisms. I maintain that the notion of abduction can clarify this puzzling problem.

The term "highly diagnostic" explicitly refers to the abductive framework. Abduction is a process of *inferring* certain facts and/or laws and hypotheses that render some sentences plausible, that *explain* or *discover* some (eventually new) phenomenon or observation. The distinction between theoretical and manipulative abduction extends the application of that concept beyond a sentential dimension. From Peirce's philosophical point of view, all thinking is in signs, and signs can be icons, indices or symbols. Moreover, all inference is a form of sign activity, where the word sign includes "feeling, image, conception, and other representation" [22, 5.283], and, in Kantian words, all synthetic forms of cognition. That is, a considerable part of the thinking activity is "model-based". Of course model-based reasoning acquires its peculiar creative relevance when embedded in abductive processes, so that we can individuate a *model-based abduction*, a considerable part of the "performances that involve sign activities are abductions" (*cit.*). [22, 5.283]. In the case of diagnostic reasoning, a physician detects various symptoms (that are signs or clues), for instance, cough, chest pain, and fever, *then* he/she may infer that it is a case of pneumonia.

The original Gibsonian notion of affordance deals with those situations in which the "perceptual" signs and clues we can detect prompt or suggest a certain action rather than others.[2] They are already available and belong to the normality of the adaptation of an organism to a given ecological niche. Nevertheless, if we acknowledge that environments and organisms' instinctual and cognitive plastic endowments change, we may argue that affordances can be related to the variable (degree of) *abductivity* of a configuration of signs: *a chair affords sitting* in the sense that the action of sitting is a result of a sign activity in which we perceive some physical properties (flatness, rigidity, etc.), and

[2] In the original Gibsonian view the notion of affordance is strictly referred to proximal and immediate perceptual chances, which do not involve higher cognitive functions such as memory and schemata, but which are merely "picked up" by a stationary or moving observer. In this paper I maintain that perceiving affordances also involves evolutionary changes and the role of more sophisticated and plastic cognitive capacities. In an analogous perspective Norman [2] and Zhang and Patel [19] expand the notion into a framework for design and for distributed cognition.

therefore we can ordinarily "infer" (in Peircean sense) that a possible way to cope with a chair is sitting on it. So to say, in most cases it is a spontaneous abduction to find affordances because this chance is already present in the perceptual and cognitive endowments of human and non-human animals.

I maintain that describing affordances that way may clarify some puzzling themes proposed by Gibson, especially the claim concerning the fact that we directly perceive affordances and that the value and meaning of a thing is clear on first glance. As I have just said, organisms have at their disposal a standard endowment of affordances (for instance through their wired sensory system, which is the only cognitive system "available" in the case of simple organisms), but at the same time they can extend and modify the range of what can afford them through the appropriate cognitive abductive skills (more or less sophisticated). As maintained by several authors (for example cf. [23,3]), what we see is a result of an embodied cognitive abductive process. For example, people are adept at imposing order on various, even ambiguous, stimuli [3, p. 107]. Roughly speaking, we may say that what we *see* is what our visual apparatus can, so to say, "explain". It is worth noting that this process happens almost simultaneously without any further mediation. Perceiving affordances has something in common with it. Visual perception is indeed a more automatic and "instinctual" activity, I have already said that Peirce claimed to be essentially abductive, even if not propositional. Indeed he considers inferential any cognitive activity whatever, not only conscious abstract thought: he also includes perceptual knowledge and subconscious cognitive activity.

We also have to remember that environments change and so the perceptive capacities when enriched through new or higher-level cognitive skills, which go beyond the ones granted by the merely instinctual levels. This dynamics explains the fact that if affordances are usually stabilized this does not mean they cannot be modified and changed and that new ones can be formed. First of all, affordances appear durable in human and animal behavior, like a kind of habit, as Peirce would say [22, 2.170]. For instance, that a chair affords sitting is a fair example of what I am talking about. This deals with what we may call *stabilized affordances*. That is, affordances that we have experienced as highly successful. Once evolutionarily formed, or created/discovered through cognition, they are stored in embodied or explicit cognitive libraries and retrieved upon occasion. Not only can they be a suitable source of new chances, through analogy. We may have very different objects that equally afford sitting. For instance, a chair has four legs, a back, and it also stands on its own. The affordances exhibited by a traditional chair may be an analogical source and transferred to different new artifacts that present the affordance of a chair for sitting down (and that to some extent can still be described as a chair). Consider, for instance, the variety of objects that afford sitting without having four legs or even a back. Let us consider a stool: it does not have even a back or, in some cases, it has only one leg or just a pedestal, but it affords sitting as well as a chair.

Second, affordances are also subjected to changes and modifications. Some of them can be discarded, because new artifacts are invented with more powerful

ones. Consider, for instance, the case of blackboards. Progressively, teachers and instructors have partly replaced them with new affordances brought about by various tools, for example, slide presentations. In some cases, the affordances of blackboards have been totally re-directed or re-used to more specific purposes. For instance, one may say that a logical theorem is still easier to be explained and understood by using a blackboard, because of its affordances that give a temporal, sequential, and at the same time global perceptual depiction to the matter. In this perspective we acknowledge that artifacts like computational programs are tools for thoughts as is language: tools for exploring, expanding, and manipulating our own minds. The *information artifacts* [24] or *cognitive artifacts* [25] which represent the external multiple tools - communication, context shifting, computational devices expressly constructed to the aim of creating opportunities and risks, like KeyGraph, etc. - recently analyzed by the researchers in the field of chance discovery [26], are artifacts as *chance seeking prostheses* offered to humans to enhance their cognitive capabilities.

4 Conclusion

In this paper I have argued that cognitive niche construction is one of the most distinctive traits of human cognition. Humans and many other organisms continuously manipulate the environment in order to exploit its offerings. In doing this, they are engaged in a process of altering or even creating external structures to lessen and overcome their limits. New ways of coping with the environment, through both evolution and cultural evolution, namely affordances, are thus created. In dealing with cognitive niche construction, I have argued that abduction can fruitfully shed light on various aspects of building affordances.

References

1. Gibson, J.J.: The Ecological Approach to Visual Perception. Houghton Mifflin, Boston, MA (1979)
2. Norman, D.: The Design of Everyday Things. Addison Wesley, New York (1988)
3. Magnani, L.: Abduction, Reason, and Science. Processes of Discovery and Explanation. Kluwer Academic/Plenum Publishers, New York (2001)
4. Simon, H.: A behavioral model of rational choice. The Quarterly Journal of Economics 69, 99–118 (1955)
5. Tooby, J., DeVore, I.: The reconstruction of hominid behavioral evolution through strategic modeling. In: Kinzey, W.G. (ed.) Primate Models of Hominid Behavior, pp. 183–237. Suny Press, Albany (1987)
6. Pinker, S.: How the Mind Works. Norton, New York (1997)
7. Pinker, S.: Language as an adaptation to the cognitive niche. In: Christiansen, M.H., Kirby, S. (eds.) Language Evolution, Oxford University Press, Oxford (2003)
8. Magnani, L.: Chance discovery and the disembodiment of mind. In: Oehlmann, R., Abe, A., Ohsawa, Y. (eds.) Proceedings of the Workshop on Chance Discovery: from Data Interaction to Scenario Creation. International Conference on Machine Learning (ICML 2005), pp. 53–59 (2005)

9. Clancey, W.J.: Situated Cognition: on Human Knowledge and Computer Representations. Cambridge University Press, Cambridge (1997)
10. Magnani, L.: Prefiguring ethical chances: the role of moral mediators. In: Oshawa, Y., Tsumoto, S. (eds.) Chance Discoveries in Real World Decision Making: Data-based Interaction of Human and Artificial Intelligence, pp. 205–229. Springer, Berlin (2006)
11. Godfrey-Smith, P.: Complexity and the Function of Mind in Nature. Cambridge University Press, Cambridge (1998)
12. Laland, K., Odling-Smee, J., Feldman, M.W.: Niche construction, biological evolution and cultural change. Behavioral and Brain Sciences 23(1), 131–175 (2000)
13. Laland, K.N., Odling-Smee, F.J., Feldman, M.W.: Cultural niche construction and human evolution. Journal of Evolutionary Biology 14, 22–33 (2001)
14. Odling-Smee, J.J.: The Role of Behavior in Evolution. Cambridge University Press, Cambridge (1988)
15. Magnani, L.: Animal abduction. From mindless organisms to artifactual mediators. In: Magnani, L., Li, P. (eds.) Model-Based Reasoning in Science, Technology, and Medicine, Springer, Heidelberg (forthcoming, 2007)
16. Hutchins, E.: Cognition in the Wild. MIT Press, Cambridge, MA (1995)
17. Clark, A., Chalmers, D.J.: The extended mind. Analysis 58, 10–23 (1998)
18. Wilson, R.A.: Boundaries of the Mind. Cambridge University Press, Cambridge (2004)
19. Zhang, J., Patel, V.L.: Distributed cognition, representation, and affordance. Cognition & Pragmatics 2(14), 333–341
20. Wilson, R.A.: Collective memory, group minds, and the extended mind thesis. Cognitive Process 6(4), 429–457 (2005)
21. Gatti, A., Magnani, L.: On the representational role of the environment and on the cognitive nature of manipulations. In: Magnani, L., Dossena, R. (eds.) Computing, Philosophy and Cognition, pp. 227–242. College Pubblications, London (2005)
22. Peirce, C.S.: Collected Papers of Charles Sanders Peirce, vol. 1-6, Hartshorne, C., Weiss, P. (eds.), vol. 7-8, Burks, A.W.(ed.) Harvard University Press, Cambridge, MA (1931-1958)
23. Hoffman, D.D.: Visual Intelligence: How We Create What We See. Norton, New York (1998)
24. Amitani, S., Hori, K.: A method and a system for supporting the process of chance discovery. In: Abe, A., Oehlmann, R. (eds.) The First European Workshop on Chance Discovery, Valencia, pp. 62–71 (2004)
25. Shibata, H., Hori, K.: Towards and integrated environment for writing. In: Abe, A., Oehlmann, R. (eds.) The First European Workshop on Chance Discovery, Valencia, pp. 232–241 (2004)
26. Oshawa, Y., McBurney, P. (eds.): Chance Discovery. Springer, Heidelberg (2003)

Discovering Color Semantics as a Chance for Developing Cross-Cultural Design Frameworks

Gyoung Soon Choi[1], Ruediger Oehlmann[2], Hilary Dalke[1], and David Cottington[1]

[1] Kingston University London, Faculty of Art, Design and Architecture,
Kingston upon Thames, KT1 2QJ, UK
[2] Kingston University London, Faculty of Computing, Information Systems and Mathematics,
Kingston upon Thames, KT1 2EE, UK
R.Oehlmann@Kingston.ac.uk

Abstract. Chance Discovery aims at determining a rare event as a chance in the sense of an opportunity or risk for future decision making. Such opportunities or risks are of particular importance for a Web designer who designs for a cross-cultural audience. An on-going research program aims at developing a new method of deriving opportunities and risks for design frameworks from the semantic and context analysis of the culturally constrained use of color. As part of this program, this paper describes an experiment where two groups of Korean and English subjects were presented with colors that are preferred in these cultures. For each color the subjects had to write brief text about their feelings and associations. Then results confirmed the prediction that Korean subjects form more associations related to emotions and nature, while English subjects form more associations with artificial concept categories, such as man-made objects. These differences will inform the next step of the research program that is concerned with the discovery of chances from context analysis.

Keywords: Discovering Color Semantics as a Chance, Cross-Cultural Design Framework, Semantics and Context of Color.

1 Introduction

Chance Discovery is concerned with the identification and the management of rare, but significant events, such as potential risks or opportunities, in some domain or application [1]. Recognizing the significance of such chances is of particular importance for designers of cross-cultural websites.

Users from around world will use websites differently [2]. There should be a difference between a site that is designed for domestic users only and a site that addresses potential customers from other cultures. For instance, Hu et. al. [3] have demonstrated that subjects in China and Japan perceive Web sites differently from subjects in the UK.

Therefore achieving social objectives in different cultural interactions based on the understanding of cultural differences [4] plays a crucial role in the design of e-commerce Web sites. As markets become increasingly global, organizations not

B. Apolloni et al. (Eds.): KES 2007/ WIRN 2007, Part II, LNAI 4693, pp. 926–933, 2007.
© Springer-Verlag Berlin Heidelberg 2007

only need to address the needs in different cultures but also communicate their objectives in different cultural interactions. This communication has a nonverbal component, in addition to the usual verbal communication.

There are many words to express human emotions and thoughts and in a similar way, colors have been used [5]. Colors are one form of communication. If commercial Web sites want to communicate such non-verbal messages, color becomes a crucial issue.

Typically, designers rely on design frameworks, which enable the designer to make decisions for the use of color. Such design frameworks include guidelines for suitable color, font, or style combinations. Unfortunately typical frameworks offer no cross-cultural dimension. However, color experience can be well specified in interactive design if there are semantic dimensions, which are to some extent consistent between cultures. Moreover, a design framework that can account for cultural differences in terms of color preferences, semantics and context would allow the designer to target audience from another culture more effectively [6]. Therefore, the authors have started a research program with the objective to develop such a cross-cultural framework that is capable to help designers in England to develop Web sites that are attractive to Koreans and vice versa. The resulting design framework will guide designers of commercial websites with worldwide audience to create culturally specific, effective and significant colour schemes in order to improve their users' experience.

This paper describes the second experiment of the research program. The crucial question for this stage is how opportunities and risks for the intended design framework can be derived from the semantic analysis of associations between color and meaning. Based on the preference ratings in a previous experiment, the most preferred colors in both groups were selected and presented in a mixed format with a random sequence of Korean and English preferred colors. Subjects were asked to write brief texts that describe their semantic associations with the presented colors. Based on a general understanding of the English and Korean culture, it was predicted that Korean subjects form more associations related to emotions and nature, while English subjects form more associations with artificial concept categories, such as man-made objects.

The next section will give an outline of the research program. The subsequent section will describe the experiment on color semantics and the final section will relate to the experimental results to the program and will give some indication of the planned next stage.

2 Chances for Design Frameworks: A Research Program

This program aims at identifying chances for the development of design frameworks. The underlying method is to investigate the relationships between colour design and presentational context of commercial websites that intend to communicate cross-cultural meaning. It is planned to develop an interactive design framework based on the investigation that will guide designers of commercial websites to create culturally specific, effective and significant colour schemes in order to improve their users' experience. Fig. 1 shows the key relationships in this research program. They form a triangle between meaning, colour and context.

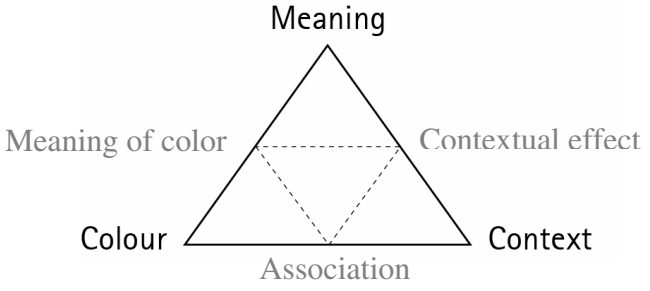


Fig. 1. The key relationships between color context and meaning

Exploring general meaning of colour is very ambiguous but when it is associated with a particular context, it constrains the set of possible alternative meanings. 'Meaning' like 'emotion', is a relational or process concept [7]. Language signs have certain meanings that they are used in certain situations and consistently produce certain behaviours and they occur in predictable associations with other signs.

In this sense the interactive triangle will form and identify relationship between colour, meaning and contextual effects. However there is another factor which affects these relationships and that is culture.

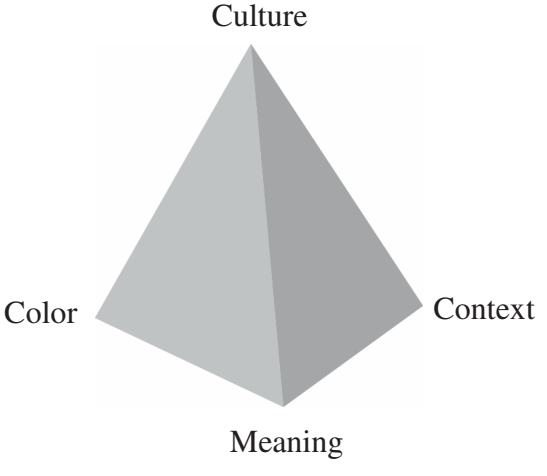


Fig. 2. The relationship between colour, meaning, context and culture

As shown in Fig. 2, colour, meaning, context and culture form a pyramid structure. This relational structure is planned to be investigated further with controlled experiments. The experiments are conducted in parallel in UK and Korea and will identify cross-cultural differences and their colour perceptions. The investigation is based on the five-stage approach described below. It should however be noted that

the aim is not just to identify opportunities of an effective design framework but also risks in the sense of identifying design decisions, which a designer for a cross-cultural context should avoid.

Stage 1: Exploration of meaning of color using Q-Methodology
This involves an exploration of meaning in a sequence of three experiments. In the first experiment, English and Korean subjects gave scores that indicated their preferred colors. In the next experiment, a group of English subjects and a group of Korean subjects is presented with a combined set of English and Korean preferred colors from the first experiment. The subjects have to characterize the meanings of the colors in a brief text. This experiment is reported in this paper. The meaning derived from the experiment is then used in a third experiment that uses Q-methodology. In the Q-Methodology, subjects are presented with a set of statements on a stimulus board. Then the subjects are asked to place a word or color card on those statements. For instance, the general attribution of blue may be sky and its concreted meaning may be happy feeling.

Stage 2: Investigating the relationship between meaning and context of color by discourse analysis
This is another explorative analysis using different methods, such as discourse analysis or content analysis, as alternative way to measure the response of feeling that is evoked by viewing a website. The purpose of this stage is to investigate how experience affects the association of color with a contextual situation.

This can be seen as an 'Interpretative act of an individual' [8]. For example, in Fig. 3 the items are all of the same yellow color. But using different context, it shows how these form different meanings with different context. The first image is a signal in a computer program to attract users' attention, the second sign is associated with a smile or a happy feeling and the third warns of radiation dangers. This is an example of how meaning is changed by different context.



Fig. 3. Relationship between Meaning and Context

Stage 3: Generating and testing a model based on those investigations
This stage involves generating and testing a model in controlled experiments based on those previous investigations. The experiments will involve Web sites that will be constructed according to the resulting model and ones that is live on the Internet.

Stage 4: Creating a design framework for web application based on previous investigations
Based on the results of the previous stages interactive/dynamic guidelines will be constructed. These can then be used by designers of commercial website who create designs based on color, meaning and context. This design framework also can be used to evaluate existing commercial websites.

The next section will describe the second experiment in this program.

3 Experiment

The main objective of this experiment was to investigate the meaning that is associated with a given color. Based on the preference ratings in the previous experiment, the most preferred colors in both groups were selected and presented in a mixed format with a random sequence of Korean and English preferred colors. After the presentation of a single color subjects were asked to write down their feelings or items they associate with that color.

3.1 Method

3.1.1 Design
The experiment was conducted in a mixed design. In this experiment, 10 colors were presented to two groups of 12 subjects each. One group consisted of Korean subjects and the other group of English subjects.

3.1.2 Subjects
Subjects in the Korean group were non-designers, who were born and educated in Korea and now live in Korea. Subjects in the British group were non-designers who were born and educated in UK and now live in UK. In both groups the cultural origin criterion also applied to parents and grandparents, the age ranged from 31 years to 50 years (means 38.3 and SD 4.8). The groups were matched according to age and gender.

3.1.3 Materials
As shown in Fig. 4, web pages were designed to show 10 different colors which were presented in sequence. Each page showed just one color that fills nearly the entire page in front of a small gray (50% of Black) background (based on Latin-square). The sequence of colors was formed by selecting from a previous experiment the 5 most preferred colors by Koreans and the 5 most preferred colors by English subjects. The subjects in this previous experiment differed from the subjects in the current experiment. The sequence of 10 selected colors were randomly generated to avoid after-effects and shown at; http://www.ygassociates.com/gsoonchoi/colours2.

The 10 selected colors were Red (FF0505), Dark Blue (163876), Dark Green (006666), Blue (6484BD), Dark Brown (5D0105), Bright Blue (00B2CC), Pink (F4358D), Bright Green (00ED2A), White (FFFFFF), Yellow (F9EA02). For the color blindness test Ishihara's standardized figures were used [9].

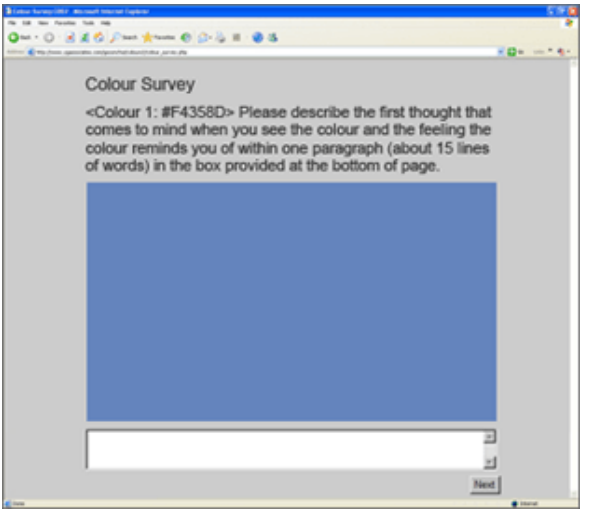


Fig. 4. The experimental website

3.1.4 Apparatus

The web pages were displayed on a Sony VAIO VGN-S28LP Laptop Computer, which was set up in a black box. The box was sufficiently large to include the computer and the head of the subjects. In this way subject's perceptions were excluded from any visual effects outside of the computer screen. Measurements of the black box were 65cm width X 40cm height X 85cm depth. The same set-up was used for all subjects in Korea and in the UK.

3.1.5 Procedure

Each subject was tested for color blindness before participating in the experiment to ensure they have normal color vision. Participants were required to touch the back of the box with the back of their head. This resulted in a fixed distance between computer screen and head of the subject. The computer screen was set up in an angle of 90 degree to the base and raised to the participant's eye level. The computer was covered with the black box described above. Then the designed web pages were presented in Korean to Korean subjects and in English to English subjects. After presenting each page, subjects were asked to write down what thoughts they had and what they feel when they see the color.

3.2 Results

Korean subjects described in Korean what meaning they associated with the presented color. The generated text file was then translated into English. Afterwards all the obtained texts were analyzed using the qualitative research software N-Vivo [10]. As part of the analysis, the descriptors used in the text were used to generate keywords.

For example, one of description for color 'Red' in English group, *'Fields covered with poppies'* could create keywords such as 'Poppy fields' and these keywords could be grouped in the categories warm emotion, cold emotion, spiritual, physical/material,

nature, artificial, objects and others. The descriptors used in each category were counted separately for the Korean and the English group. The results are summarized in Fig. 5, which shows that Korean subjects form more associations related to emotions and nature, while English subjects form more associations with artificial concept categories, such as man-made objects.

Qualitative differences of the colors 'red', 'navy blue', 'dark green', 'bright green', 'white', 'yellow' were found in relation to cross-cultural differences between English and Korean subjects. For example, the semantic descriptors of the color 'red' used by Korean subjects were for instance 'hot summer day', 'well matured Sharon fruits', 'weather between spring and summer' and 'sun', which were categorized as warm emotion and nature. In contrast, English subjects used descriptors such as 'poppy's field', 'plastic', 'football team' and 'modern design', which are a wide range of artificial concepts and objects.

For the color 'yellow', English subjects used descriptors such as 'children's painting', 'washing line', 'envelope' and 'ceiling' which are associations with objects they possess (immobility), whereas Korean subjects used descriptors such as 'Forsythia flower', 'picnic in spring', 'raining' and 'spring comes' which are associations with nature and related to activities (mobility).

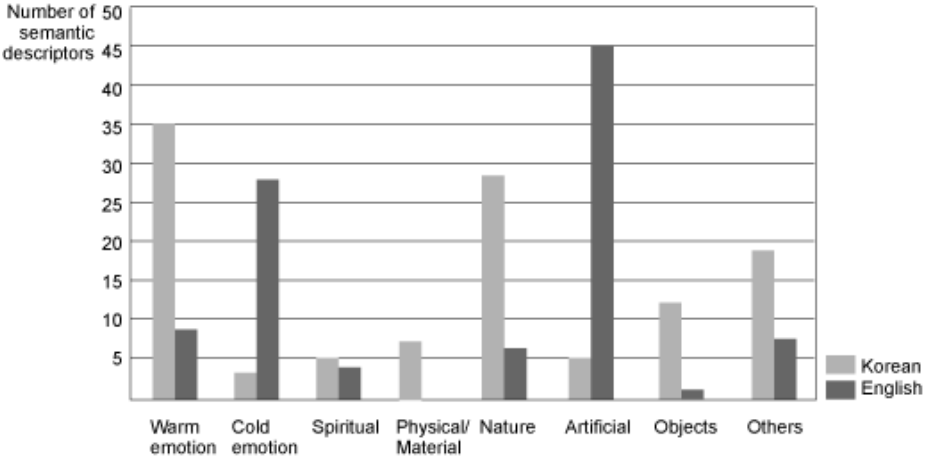


Fig. 5. Categorization of semantic descriptors used by Korean and English subjects

4 Discussion

The results confirmed the prediction that Korean subjects form more associations related to emotions and nature, while English subjects form more associations with artificial concept categories, such as man-made objects. This suggests potential chances for design frameworks, in that for instance appealing to nature-related aspects of a product or service appears to be more suitable for Korean consumers than for English.

However, it also showed that the same subject formed different associations for the same color. Therefore a design framework that just derived chances from these semantic results would be ambiguous. This ambiguity can be resolved by presenting colors in different context and analyzing how this context constraints the semantic interpretation of the color. Hence it can be argued that in order to identify chances for a cross-cultural design framework all three dimensions of cross-culture, semantic and context need to be considered.

Acknowledgments. The first author gratefully acknowledges the support of the Arts and Humanities Research Council in the UK.

References

1. Ohsawa, Y.: Modeling the Process of Chance Discovery. In: Ohsawa, Y., McBurney, P. (eds.) Chance Discovery, Springer, Heidelberg (2003)
2. Nielsen, J.: Designing Web Usability, New Riders, USA (2000)
3. Hu, J., Shima, K., Oehlmann, R., Zhao, J., Takemura, Y., Matsumoto, K.: An Empirical Study of Audience Impressions of B2C Web Pages in Japan, China and the UK. Electronic Commerce Research and Applications (Journal) 3(2), 176–189 (2006)
4. Ascalon, M., Schleicher, D., Born, M.: Cross-cultural social intelligence: An assessment for employees Working in Cross-National Contexts. Cross Cultural Management: An International Journal (submitted, 2006)
5. Kuno, N.: FORMs Inc.: Colors in context. Graphic-Sha, Japan (1999)
6. Wright, B., Rainwater, L.: The meaning of Color. The Journal of General Psychology 67, 89–99 (1962)
7. Osgood, C.E., Suci, G.J., Tannenbaum, P.H.: The measurement of meaning. The University of Illinois, USA (1957)
8. Kelly, G.A.: The Psychology of Personal Constructs, vol. 1, 2, New York (1955)
9. Ishihara, S.: Ishihara's Tests for Color Deficiency. Kanehara Trading Inc., Tokyo, Japan (2001)
10. Richards, L.: Using N-Vivo in qualitative research. SAGE Publications Ltd., London

Evaluating a Constructive Meta-learning Algorithm for a Rule Evaluation Support Method Based on Objective Indices

Hidenao Abe[1], Shusaku Tsumoto[1], Miho Ohsaki[2], and Takahira Yamaguchi[3]

[1] Department of Medical Informatics, Shimane University, School of Medicine
89-1 Enya-cho, Izumo, Shimane 693-8501, Japan
abe@med.shimane-u.ac.jp, tsumoto@computer.org
[2] Faculty of Engineering, Doshisha University
1-3 Tataramiyakodani, Kyo-Tanabe, Kyoto 610-0321, Japan
mohsaki@mail.doshisha.ac.jp
[3] Faculty of Science and Technology, Keio University
3-14-1 Hiyoshi, Kohoku Yokohama, Kanagawa 223-8522, Japan
yamaguti@ae.keio.ac.jp

Abstract. In this paper, we present an evaluation of learning algorithms to select proper ones in a rule evaluation support tool for post-processing of mined results. Post-processing of mined results is one of the key processes in the data mining process. However, it is difficult for human experts to completely evaluate several thousand of rules from a large dataset with noises. To reduce the costs in such a rule evaluation task, we have developed a rule evaluation support method with rule evaluation models, which learn from objective indices for mined classification rules and evaluations by a human expert for each rule. To enhance the adaptability of rule evaluation models, we introduced a constructive meta-learning system to choose proper learning algorithms. Then, we performed the case study on the meningitis data mining as an actual problem. The obtained results demonstrate the applicability of the proposed rule evaluation support method.

Keywords: Data Mining, Post-processing, Rule Evaluation Support, Objective Rule Evaluation Index, Constructive Meta-learning.

1 Introduction

In recent years, an enormous number of data are stored in information systems used in natural science, social science, and business domains. Valuable knowledge has become increasingly accessible as a result of the development of information technology. Besides, data mining techniques combine different types of technologies such as database technologies, statistical methods, and machine learning methods. Then, data mining has been well known for utilizing data stored on database systems. In particular, if-then rules, which are outputs of data mining

B. Apolloni et al. (Eds.): KES 2007/ WIRN 2007, Part II, LNAI 4693, pp. 934–941, 2007.
© Springer-Verlag Berlin Heidelberg 2007

produced by rule induction algorithms, are considered to be highly usable and readable. However, for large datasets that have hundreds of attributes and that include noise, the process often generates several thousand rules. From such a large rule set, it is difficult for human experts to find out valuable knowledge, which is rarely included in the rule set.

To support such a rule selection, several studies have been conducted using objective rule evaluation indices such as recall, precision, and other interestingness measures (hereinafter, referred to as "objective indices") [1,2,3]. Furthermore, it is difficult to estimate the criterion of a human expert using a single objective rule evaluation index, because his/her subjective criterion, such as interestingness or importance for his/her purpose, is influenced by the amount of his/her knowledge and/or the passage of time. In addition, rule selection methods have never explicitly re-used the history of each rule evaluation, such as focused items and relationships between items, which is only stored in the mind of the human expert.

With respect to the above-mentioned issues, we have developed an adaptive rule evaluation support method for human experts with rule evaluation models. This method predicts the criteria of the experts based on objective indices by re-using the results of the evaluations by human experts.

First, in Section 2, we describe our rule evaluation support method by constructing rule evaluation models based on objective indices. After implementing the method as a tool, we carried out a performance comparison of learning algorithms for constructing rule evaluation models in Section 3. Finally, we conclude the present paper in Section 4.

2 Rule Evaluation Support with Rule Evaluation Model Based on Objective Indices

In practical data mining situations, a human expert repeatedly performes costly rule evaluation procedures. In these situations, useful experience related to each evaluation such as focused attributes, interesting combinations of antecedents, and valuable facts, are not explicitly used by any rule selection system, but rather are tacitly stored in the human expert. To solve these problems, we suggest a method to construct rule evaluation models based on objective rule evaluation indices as a way to describe the criteria of a human expert explicitly, re-using the human evaluations. We then implemented the method as a tool for rule evaluation support of human experts.

2.1 The Rule Evaluation Support Tool Based on Objective Indices

We considered the process of modeling rule evaluations of human experts as the process by which to clarify the relationships between the human evaluations and features of input if-then rules. Based on this consideration, we decided that the rule evaluation model construction process could be implemented as a learning task. Figure 1 shows the rule evaluation model construction process based on the re-use of human evaluations and objective indices for each mined rule.

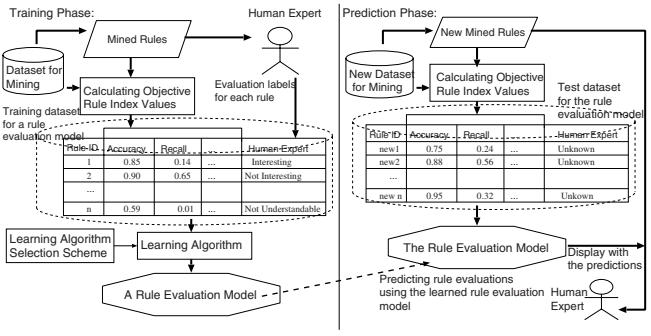


Fig. 1. Overview of the construction method of rule evaluation models

The inputs of the system are a mined rule set and evaluations by a human expert. The system then outputs the prediction of rule evaluation labels based on rule evaluation models such as linear expressions, decision trees, neural networks, and expressions that are more complex.

In the training phase, the attributes of a meta-level training data set are obtained by objective indices such as recall, precision and other rule evaluation values. The human evaluations for each rule are combined as classes of each instance. To obtain this data set, a human expert has to evaluate the entirety or a part of the input rules at least once. After obtaining the training data set, its rule evaluation model is constructed by using a learning algorithm. To construct more correct rule evaluation models for each execution, the training phase also has a function by which to select a proper learning algorithm.

At the prediction phase, a human expert receives predictions for new rules based on their objective index values. Since rule evaluation models are used for prediction, we need to choose a learning algorithm with a high accuracy, which is similar to the current classification problems.

2.2 Learning Algorithm Selection with Meta-learning

As shown in Figure 1, the training phase has a process to select a proper learning algorithm for constructing more correct rule evaluation models. Considering the current situation, this appears to be the best way to introduce a meta-learning scheme to this learning algorithm selection process.

Meta-learning algorithms are often used to enhance a classification task. One of the approaches integrates prepared base-level learning algorithms with a meta-strategy such as voting, selecting or meta-level learning. This approach is called "selective meta-learning". Besides, we have developed another approach of meta-learning scheme that constructs a proper learning algorithm for a given dataset using de-composed base-level learning algorithms. This approach is called "constructive meta-learning".

In the field of meta-learning, there are many studies about the selective algorithms. There are two approaches to selective scheme. The first includes bagging

[4] and boosting [5], combining base-level classifiers from multiple training data with different distributions. In these meta-learning schemes, we should select just one learning algorithm to learn base-level classifiers. The other approach includes voting, stacking [6] and cascading [7], which combines base-level classifiers from different learning algorithms. METAL [8] and IDA [9] are also selective meta-learning approaches, selecting a proper learning algorithm for the given data set with a heuristic score, which is called meta-knowledge.

The constructive meta-level processing scheme [10] uses the meta-learning approach, which controls the objective process with meta-knowledge. In this scheme, we construct a meta-knowledge, represented with method repositories. The meta-knowledge consists of information of functional parts, restrictions of combinations of each functional part, and methods by which to re-construct object algorithms with the functional parts.

3 Performance Comparisons of Learning Algorithms for Rule Model Construction

In this section, we present the results of empirical evaluations using the dataset obtained from the results of a meningitis data mining [11].

First, we obtained a dataset for each rule set from the above data mining results. The nine learning algorithms, including selective and constructive meta-learning algorithms, were then used to construct each rule evaluation model. Based on the experimental results, we evaluated the availability of each learning algorithm with respect to accuracies of rule evaluation models.

Since a rule evaluation model requires higher accuracy to support a human expert exactly in the proposed rule evaluation support method, we compared the predictive accuracies on the entire dataset and Leave-One-Out validation. The accuracy of a validation dataset D is calculated with correctly predicted instances $Correct(D)$ as $Acc(D) = (Correct(D)/|D|) \times 100$, where $|D|$ is the size of the dataset. The recalls of class i on a validation dataset are calculated using correctly predicted instances of the class $Correct(D_i)$ as $Recall(D_i) = (Correct(D_i)/|D_i|) \times 100$, where $|D_i|$ is the number of instances of class i. Furthermore, the precision of class i is calculated using the number of instances which are predicted i as $Precision(D_i) = (Correct(D_i)/Predicted(D_i)) \times 100$, where $Predicted(D_i)$ means the number of instances predicted as i by a rule evaluation model.

In order to construct a dataset to learn a rule evaluation model, the values of the objective indices have been calculated for each rule by considering 39 objective indices, which are investigated and re-formulated by Ohsaki et. al.[12]. Thus, each dataset for each rule set has the same number of instances as the rule set. Each instance has 40 attributes including those of the class.

We applied the nine learning algorithms to these datasets to compare their performances as a rule evaluation model construction method. We used the following learning algorithms from Weka [13]: J4.8 (a C4.5 decision tree learner) [14], a neural network learner with back propagation (BPNN)[15], support vector

machines (SVM)[1][16], classification via linear regressions (CLR)[2][17] and OneR [18]. In addition, we used the following selective meta-learning algorithms: Bagging [4], Boosting [5] and Stacking[3] [6]. As a constructive meta-learning scheme, we used CAMLET [10] to construct proper learning algorithms for each dataset.

In these experiments, we have not evaluated the computational costs of each of the learning algorithms.

3.1 Constructing Rule Evaluation Models for the Meningitis Data Mining Result

In this case study, we considered 244 rules mined from six datasets for six types of diagnostic problems, as shown in Table 1. In these datasets, the appearances of meningitis patients were considered as attributes, and the diagnosis of each patient was considered as a class. Each rule set was mined with its proper rule induction algorithm composed by a constructive meta-learning system called CAMLET [11]. For each rule, we assigned a label (I: Interesting, NI: Not-Interesting, or NU: Not-Understandable) according to evaluation comments provided by a medical expert.

Table 1. Description of the meningitis datasets and the results of data mining

Dataset	#Attributes	#Class	#Mined rules	#'I'	#'NI'	#'NU'
Diag	29	6	53	15	38	0
C_Course	40	12	22	3	18	1
Culture+diag	31	12	57	7	48	2
Diag2	29	2	35	8	27	0
Course	40	2	53	12	38	3
Cult_find	29	2	24	3	18	3
TOTAL	—	—	244	48	187	9

A proper learning algorithm to construct the meningitis rule evaluation model. Abe and Yamaguchi have developed a constructive meta-learning system called CAMLET to choose a proper learning algorithm for a given dataset with a machine learning method repository. To implement the method repository, we first identified each functional part, called a method, from the following eight learning algorithms: Version Space [19], AQ15 [20], Classifier Systems [21], Neural Network, ID3 [22], C4.5, Bagging and Boosting. With the method repository, CAMLET constructs a proper learning algorithm for a given dataset, searching the possible learning algorithm specification space obtained by the method repository for the best learning algorithm with a Genetic Algorithm.

[1] A polynomial kernel function was used.

[2] We set up the elimination of collinear attributes and the model selection with greedy search based on the Akaike information metric.

[3] This stacking uses the other seven learning algorithms as base-level learners and J4.8 as a meta-level learner.

Since we have set up the initial population $\tau = 4$ and the number of refinements is $N = 100$[4], CAMLET searched up to 400 learning algorithms without any parameter optimization from 6,000 possible learning algorithms for the best learning algorithm. Figure 2 shows the algorithm constructed by CAMLET for the dataset of the meningitis data mining result.

This algorithm iterates boosting of the C4.5 decision tree for randomly split training datasets. Each classifier set generated by the C4.5 decision tree learner is reinforced with the method from the Classifier Systems. The learned committee then aggregates with the weighted voting from boosting.

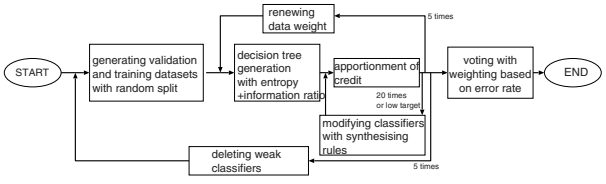


Fig. 2. The learning algorithm constructed by CAMLET for the dataset of the meningitis data mining result

Comparison of the classification performances. In this section, we present the results of accuracy comparison over the entire dataset, the recall of each class label, and the precisions thereof. Since Leave-One-Out holds just one test instance and the remaining as the training dataset repeatedly for each instance of a given dataset, we can evaluate the performance of a learning algorithm for a new dataset without any ambiguity.

The results of the performances of the five learning algorithms for the entire training dataset and the results of Leave-One-Out are shown in Table 2.

All of the Accuracies, Recalls of I and NI, and Precisions of I and NI are higher than those of the predicting default labels.

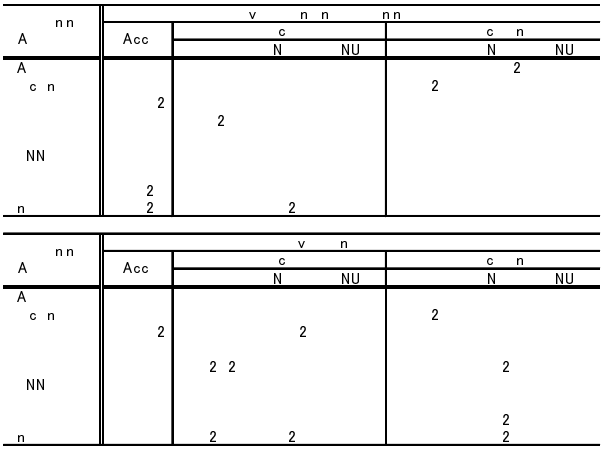
Compared to the accuracy of OneR, the other learning algorithms achieve equal or higher performances using combinations of multiple objective indices than by sorting with single objective index. For Recall values over class I, BPNN achieved the highest performance. The other algorithms exhibit lower performance than that of OneR, because they tend to learn classification patterns for the major class NI.

The accuracy of Leave-One-Out demonstrates the robustness of each learning algorithm. The Accuracy (%) of these learning algorithms ranges from 75.8% to 81.9%. However, these learning algorithms have not been able to classify the instances of class NU, because it is difficult to predict a minor class label in this dataset.

The learning algorithm constructed by CAMLET shows the second accuracy for the entire training dataset, compared with other learning algorithms.

[4] These parameters are decided based on the e well known heuristics of the Genetic Algorithm search.

Table 2. Accuracies (%), Recalls (%), and Precisions (%) of the five learning algorithms



Although Boosted J4.8 outperforms the entire training dataset compared to CAMLET, it has lower accuracy for LOO. This means that Boosted J4.8 has a problem of overfitting. Thus, CAMLET shows higher adaptability than the other selective meta-learning algorithms.

4 Conclusion

In the present paper, we have described the evaluation of the nine learning algorithms for a rule evaluation support method with rule evaluation models to predict evaluations for an if-then rule based on objective indices by re-using evaluations by a human expert.

Based on the performance comparison with the nine learning algorithms for the dataset of the meningitis data mining result, rule evaluation models have achieved higher accuracies for simply predicting each default class. For this dataset, the learning algorithm constructed by CAMLET provides higher accuracy with higher reliability than the other eight learning algorithms, including the three selective meta-learning algorithms.

In the future, we will improve the method repository of CAMLET to construct suitable learning algorithms for rule evaluation models. In addition, we will apply this rule evaluation support method to other datasets from various domains.

References

1. Hilderman, R.J., Hamilton, H.J.: Knowledge Discovery and Measure of Interest. Kluwer Academic Publishers, Dordrecht (2001)
2. Tan, P.N., Kumar, V., Srivastava, J.: Selecting the right interestingness measure for association patterns. In: Proceedings of International Conference on Knowledge Discovery and Data Mining KDD-2002, pp. 32–41 (2002)

3. Yao, Y.Y., Zhong, N.: An analysis of quantitative measures associated with rules. In: Zhong, N., Zhou, L. (eds.) Methodologies for Knowledge Discovery and Data Mining. LNCS (LNAI), vol. 1574, pp. 479–488. Springer, Heidelberg (1999)
4. Breiman, L.: Bagging predictors. Machine Learning 24, 123–140 (1996)
5. Freund, Y., Schapire, R.E.: Experiments with a new boosting algorithm. In: Thirteenth International Conference on Machine Learning, pp. 148–156 (1996)
6. Wolpert, D.: Stacked generalization. Neural Network 5, 241–260 (1992)
7. Gama, J., Brazdil, P.: Cascade generalization. Machine Learning 41, 315–343 (2000)
8. Metal (2002), http://www.metal-kdd.org/
9. Bernstein, A., Provost, F.: An intelligent assistant for knowledge discovery process. In: Proceedings IJCAI 2001 Workshop on Wrappers for Performance Enhancement in KDD (2001)
10. Abe, H., Yamaguchi, T.: Constructive meta-learning with machine learning method repositories. In: Orchard, B., Yang, C., Ali, M. (eds.) IEA/AIE 2004. LNCS (LNAI), vol. 3029, pp. 502–511. Springer, Heidelberg (2004)
11. Hatazawa, H., Negishi, N., Suyama, A., Tsumoto, S., Yamaguchi, T.: Knowledge discovery support from a meningoencephalitis database using an automatic composition tool for inductive applications. In: Terano, T., Chen, A.L.P. (eds.) PAKDD 2000. LNCS, vol. 1805, pp. 28–33. Springer, Heidelberg (2000)
12. Ohsaki, M., Kitaguchi, S., Kume, S., Yokoi, H., Yamaguchi, T.: Evaluation of rule interestingness measures with a clinical dataset on hepatitis. In: Boulicaut, J.-F., Esposito, F., Giannotti, F., Pedreschi, D. (eds.) PKDD 2004. LNCS (LNAI), vol. 3202, pp. 362–373. Springer, Heidelberg (2004)
13. Witten, I.H., Frank, E.: Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations. Morgan Kaufmann, San Francisco (2000)
14. Quinlan, J.R.: Programs for Machine Learning. Morgan Kaufmann, San Francisco (1993)
15. Hinton, G.E.: Learning distributed representations of concepts. In: Proceedings of 8th Annual Conference of the Cognitive Science Society (1986)
16. Platt, J.: Fast training of support vector machines using sequential minimal optimization. In: Burges, C., Smola, A. (eds.) Advances in Kernel Methods - Support Vector Learning, pp. 185–208. MIT Press, Cambridge (1999)
17. Frank, E., Wang, Y., Inglis, S., Holmes, G., Witten, I.H.: Using model trees for classification. Machine Learning 32, 63–76 (1998)
18. Holte, R.C.: Very simple classification rules perform well on most commonly used datasets. Machine Learning 11, 63–91 (1993)
19. Mitchell, T.M.: Generalization as search. Artificial Intelligence 18, 203–226 (1982)
20. Michalski, R., Mozetic, I., Hong, J., Lavrac, N.: The AQ15 inductive learning system: An over view and experiments. Reports of Machine Learning and Inference Laboratory MLI-86-6, George Mason University (1986)
21. Booker, L.B., Holland, J.H., Goldberg, D.E.: Classifier systems and genetic algorithms. Artificial Intelligence 40, 235–282 (1989)
22. Quinlan, J.R.: Induction of decision tree. Machine Learning 1, 81–106 (1986)

Information Literacy and Everyday Life Risks

Yumiko Nara

The University of the Air
2-11 Wakaba, Mihama-ku Chiba City 261-8586 Japan
narayumi@u-air.ac.jp

Abstracts. Living is inevitably accompanied by risks, and the environment surrounding the life system in present day society is constantly changing. The facts that risks are becoming further diversified and are keeping pace with the increase in hazards, there is a pressing need to acquire techniques to counter risks. This study aims to examine the significance and contents of information literacy to perceive risks and cope with them.

Keywords: everyday life risk, information, literacy, risk management.

1 Introduction: Everyday Life Risks, Perception and Countermeasures

Chance Discovery aims at identifying and managing rare, but significant events, such as potential risks or opportunities, in some domain and application [Ohsawa, 2003]. As the risks of daily life become increasingly diverse and widespread, receipt and provision of information are essential for enabling each individual to obtain chances, i.e., to perceive and counter such risks. This paper discusses everyday life risks, focusing on effective receipt and provision of risk information and the definition of literacy in the entire human social life, from the viewpoint that the individual is the risk-information processing system.

The concept and definition of risk vary depending on the field of study and the researcher. Risks are understood as 'possibilities,' 'uncertainties,' 'deviations of expectations and results,' 'probabilities,' and 'expected losses,' which are contained in hazards, as well as obstacles and other undesirable phenomena which jeopardize the safety of life, health, assets, socioeconomic activities, and the natural environment. Despite such variations, the true nature of risks is invariably understood as the 'occurrence of undesirable phenomena' that may inflict 'harmful influences.' Thus, everyday life risks shall herein be defined as the 'possibility of the occurrence of phenomena that will eventually result in inflicting harmful influences upon the life system.'

Daily life risks can be classified as material risks, human risks, and indemnity-liable risks, depending on the type of exposure. They can also be classified by the conditions that give rise to risks, such as risks related to highly advanced science and technology, environmental problems, consumers' lives and products, health and medical problems, and disasters.

B. Apolloni et al. (Eds.): KES 2007/ WIRN 2007, Part II, LNAI 4693, pp. 942–949, 2007.
© Springer-Verlag Berlin Heidelberg 2007

The individual perceives and counters everyday life risks in order to minimize the harmful influences which may be sustained, and to maintain life stability. In other words, risk perception and countermeasures are performed as risk management in the daily life routine management cycle.

Specifically, the individual (1) recognizes risks by finding the risks (the exposures that the individual is involved with shall be examined for possibilities of occurrence of phenomena that may provoke harmful influences); (2) measures the risks (the frequency and intensity of the risks that exposure may sustain shall be evaluated); and (3) counters the risks by selecting and executing measures (the measures to counter the risks that have been found and measured shall be planned, selected, and executed). The individual subsequently reevaluates risk management and feeds it back into the risk-finding stage.

2 Ambiguity and Difficulty in Risk Perception and Countermeasures

Everyday life risks involve harmful influences upon life. Accordingly, it is considered essential that the individual makes scientifically reasonable efforts to perceive and minimize all possibilities of risk occurrence. Nevertheless, it sometimes happens that what is perceived as a risk by one individual in a certain situation, is not perceived as a risk, or is even ignored despite being recognized as a risk, by another individual in a different situation. In this chapter, ambiguity and difficulty in daily life risk perception and countermeasures are discussed.

2.1 Ambiguity in Risk Perception: Discrepancy Between Objective and Subjective Risks

In the case of individuals from the general public, risk perception is synonymous with understanding uncertain phenomena. This understanding is based on an integrated set of criteria, including frequency and intensity, anxiety and fear, optimism, benefits, and acceptability, from subjective viewpoints. Therefore, risks are perceived differently depending on each individual's limitation of information-processing ability, surrounding situations, and attributes, such as past risk career, age, sex, occupation, and academic background.

On the other hand, in the case of risk assessment and management organizations, including experts, companies, and administrative organizations, risks are analytically perceived based on estimations of probability, quantitative measurements of losses, and calculations of costs and benefits. Accordingly, there is a great discrepancy in risk perception between individuals and organizations. Risks perceived by the former are subjective risks, while those perceived by the latter are objective risks.

Risks that are ranked according to comprehensive quantification and comparison using their degree of influences upon human health, the ecosystem, and life quality as indices, are comparative risks. Comparative risks are also classified as objective comparative risks and subjective comparative risks, and it has been pointed out that these also involve discrepancies. The former is analyzed scientifically by experts, based on statistical data and theoretical estimation, while the latter is judged by

individuals and is based on their personal risk perception. For example, a report has been done on the results of investigating the ranks of thirty risk items, including smoking, drinking, automobiles, electric power, swimming, atomic power, and antibiotics, using a risk-ranking method for comparing objective and subjective comparative risks. According to the results, the order of objective comparative risks (subjective comparative risks are shown in parentheses) of the seven items are: 1st (8th), 2nd (21st), 3rd (7th), 5th (25th), 7th (30th), 20th (1st), and 27th (4th), which apparently reveals a discrepancy between objective and subjective comparative risks [Kobayashi, 2000]. Such a discrepancy is called the risk perception bias, which has been pointed out by many psychologists [Hirose, 1993; Okamoto, 1995].

2.2 Difficulty in Risk Perception and Countermeasures: Context Shift and Information

Just as risk perception is not uniform due to the various attributes of each individual, the behavioral patterns of countermeasures for accepting, avoiding, and mitigating risks vary depending on the attributes of each individual. Such variations are considered to be attributable to the influence of the context of each individual.

The term context refers to situations and circumstances surrounding a matter or incident. The context of the individual is the state of his/her integrated circumstances, which is expressed by a function using his/her knowledge, sense of values, past risk career, age, occupation, academic background, annual income, family makeup, residential conditions, and current situation as variables. This serves as the initial condition that determines the direction of an individual's behavior (i.e. perception and countermeasures) in dealing with various phenomena. Individuals of the same culture and environment may react differently to the same phenomenon, i.e. some consider it a risk and some do not; some take action to mitigate the impact and some do not. Such variations in risk perception are attributable to the difference of the context.

In principle, everyday life risks can be mitigated by increasing the redundancy of the life system. This can be achieved by preparing various control measures within the life system in consideration of the degree of diversity of external disturbances and other factors that impede the life system. This requires cost investment. In the case of the individual, such costs include financial costs (money) and non-financial costs (labor, mental strength, adjustment of human relations, and time).

When there is a budgetary ceiling for investment of life resources, the costs shall be prioritized. In the case of persons, for example, central interests vary depending on their family cycle. Individuals in a period of family growth concentrate their life interest on their jobs, education of their children, and housing loan reimbursement, while those in a period of family maturity concentrate on health and retirement. As their interests are directed to their life activities, and resources are invested accordingly, it is a matter of low priority to turn their interests and resources to latent risks with unknown occurrence probability. In addition, it is unlikely that all risks can be kept under equally constant alert; if an individual directs his/her interests and resources towards one particular risk, the perception of other risks is lowered and resources are not invested.

Thus, a spiral phenomenon is considered to be occurring when individuals actually deal with risks. The finite resources in the realities of life impose conditions upon

relations, interests, norms, and values. This becomes the basis for determining risk perception and countermeasures, although theoretically it is considered essential that the prevalent value of life, i.e. 'safety and security of life,' determines the prevalent norm that 'risks should be mitigated,' based on which interests, relations, and resources shall be determined, in that order. It is a prerequisite for each individual to have a certain context for noticing and taking interest in a risk and understanding its impact on his/her own life, and for determining a countermeasure, preparing necessary techniques and resources, and taking action against the risk.

Incidentally, people seriously practiced precautions immediately after the Hanshin-Awaji Earthquake Disaster, but they have become less desperate as time passes [Nara, 1999]. This shows that the context has shifted from earthquake risk perception and countermeasures to engagement in daily life activities. It must be noted, however, that elder people and people who sustained considerable earthquake damage continued to harbor feelings of high anxiety towards earthquake risks and still practice precautions. Such individuals are considered to have retained the context for being ready to perceive and counter earthquake risks.

The flow of everyday life risk perception and countermeasures is as follows: notice a risk → take interest in the risk → identify the risk → acquire the necessary techniques and resources for countering the risk → take actual countermeasures. To shift the context for guiding people to follow the stages in the order shown by the arrows above, external stimulation must be introduced. The term context shift means the creation of clues for changing the interrelationship between life resources and other relations based on new patterns of life value, which are created by external and internal stimulation. When no external stimulation is introduced, people will stick to the existing context. They will impose conditions upon their life interests in line with existing resource distributions. Based on this, they will impose conditions upon the patterns of their life's norms and values, which will eventually become customary and fixed. This flow costs the least. In this case, the context shift can be promoted by introducing outside stimulation, which is no less than the acceptance of everyday life risk information.

3 Literacy and Interaction of the Life Subject, Risk Information, and the Information Provider

In the previous chapter, the author discussed the difficulty in achieving scientifically reasonable risk perception and countermeasures. Even if an individual who is jeopardized does not perceive a risk, or neglects to take countermeasures despite recognition of a risk, this behavior has been chosen at his/her own discretion. Nevertheless, if it is objectively apparent that the risk is jeopardizing the safety and security of his/her life, or that his/her ignorance or negligence in taking countermeasures results in sharing or shuffling off the risk to family members or any third person, the situation cannot be overlooked. In this case, someone must stimulate the context shift by providing the individual with information about the nature of the risk, along with effective countermeasures. In this chapter, risk information provision in consideration of the standpoints of both the receiver and provider, is discussed.

3.1 Receipt and Dispatch of Life Risk Information

There are two cases of interaction between risk information and the individual. One is without the mediation of a third party that provides information, and the other is with the apparent mediation by a third party that warns against risks by providing information.

In the case of the former, the individual obtains risk information at the primary level based mainly on images obtained by the five senses through past experiences with risks, the disaster subculture, and other social and cultural frameworks, and by perceiving the frequent vibrations of buildings and the increase in traffic volume in his/her residential area. Based on such information, the individual perceives the risk himself/herself.

In the case of the latter, when there is a provider of risk information, there are two stages of information provision depending on the processing level. In the first stage, information at the primary and secondary level is provided. The primary risk information is analyzed and evaluated by experts, and processed into secondary information, based on which a more objective and direct risk message is created. Such a message shall be dispatched to the individual to promote proper understanding and countermeasures for the risk. The provider of the risk information at the secondary level is mainly risk management and control organizations, including the central and municipal governments, various research institutions, and private-sector companies.

In the second stage, interpretation is applied and new information added to the primary and secondary information, based on which more sophisticated information is created and dispatched. Specifically, a risk message at the tertiary information level will be delivered to the individual through newspapers and other mass media channels.

Upon delivery, the information shall be either accepted or rejected by each individual. Information that has been accepted shall be analyzed and evaluated, and the individual perceives and counters the risk. This individual behavior of risk perception and countermeasures will not only affect and change the risk, but will also be fed back to the risk management and control organizations and the mass media, which are the information providers.

3.2 Literacy of the Information Provider

As discussed so far, risk perception and countermeasures vary depending on the individual. What is primarily important for the information provider is knowing the characteristics, attributes, and situations related to an individual's risk perception. Without understanding these factors, the provider will fail to stimulate the context shift and change the individual's behavior, no matter how scientifically accurate the information is. Unidirectional information provision that neglects such important factors will merely result in discordant perception and will adversely affect individuals' countermeasures for risks.

In addition, the risk information provider should be well aware of the possibility that they may have a biased perception of risks because of their expertise. Because their views of risks lean toward technical issues in their respective specialized fields and there is a tendency to attach too much importance to occurrence probability,

experts are apt to overlook the fact that the significance of risks is determined within the relationship between individuals and the society where the risks actually take shape. Risks such as traffic accidents and fire incidents may have relatively low frequency and intensity of occurrence when evaluated at the level of entirety; however, such risks are regarded to be very rare and considerable at the individual level. Information will not be accepted by the receiver unless it is dispatched in compliance with the needs, anxieties, and sensibilities of individuals.

Selection of the optimum medium for providing information is also important. Although optimum medium differs depending on the content of messages and the attributes of receivers, there are two media that are considered to be effective. Mass communication is effective for providing warning information because of its capacity to present problems and provide a huge volume of information, while person-to-person communication is considered to be effective for providing information for establishing agreement because of its superiority in interaction and information density.

One more important issue is to assure the receiver of the reliability of the dispatched information, which is the provider's responsibility. According to the results of investigations conducted, the general public has low confidence in companies and administrative organizations (central government, ministries, agencies, and municipal governments), but high confidence in experts from universities, research institutions, international organizations, and NGOs, who have expert knowledge and who take a neutral standpoint [Kinoshita, 1997; Murayama, et al., 1999].Accordingly, mediation by a third party should be added to the options for means of dispatching information.

3.3 Literacy of the Information Receiver

The individual is required to acquire literacy for perceiving and countering daily life risks. Literacy for risk perception (finding and evaluating risks) can be acquired by increasing the understanding schema and improving the content, while literacy for risk countermeasures (selecting and executing risk countermeasures) can be acquired by increasing the behavioral scripts and improving the content quality. In the case of risk perception and countermeasures, acquisition of literacy is synonymous with acquisition of scientific knowledge of diversified risks.

The standard of scientific knowledge is classified as three levels: (1) cultural scientific knowledge (understanding the meaning of scientific technical terms which general literate adults know); (2) functional scientific knowledge (properly using scientific technical terms); and (3) true scientific knowledge (thinking analytically and deductively). Risk experts analyze and evaluate risks at the level of true scientific knowledge, which can hardly be expected of the general public. It is not peculiar that the general public does not have the ability to process primary information into secondary information that is scientifically reasonable.

Accordingly, the risk information provider must make the effort to provide information that is acceptable for the general public, while the individual must make the effort to acquire scientific knowledge and then acquire the additional ability to examine the information they receive. Even if it is almost impossible for the individual to have highly sophisticated knowledge equivalent to that of experts,

he/she should acquire techniques to critically examine the risk information provided. This is called media literacy. In particular, it is quite likely that messages of risk information are created using special techniques for enhancing risk perception and countermeasures, which will be discussed in the next subchapter. Each individual must evaluate information based on awareness of the possible use of such techniques.

3.4 Techniques for Stimulating Risk Perception and Countermeasures

Various effective techniques, mainly in the field of social psychology, have been studied for persuading people. Needless to say, such techniques should not be abused or misused. There are cases in which techniques of persuasive communication need to be employed in order for risk information to be effectively accepted by the receiver. Examples of messages using such techniques include warning signs for preventing misuse of products, and evacuation recommendations in the event of disasters, both of which are to protect people from risks.

As an effective technique of persuasive communication regarding risk information, bilateral communication (which provides both supporting arguments and counter-arguments for the information provider's opinion) is preferred to unilateral communication (which provides only supporting arguments for the information provider's opinion). Bilateral communication is particularly advantageous when the receiver (1) disagrees with the information provider's opinion; (2) has high educational standards; (3) already has sufficient information and knowledge of the topic of the persuasive communication; and (4) is likely to be informed of opposing opinions.

In addition, fear-inciting communication (which incites the feeling of fear by informing the receiver of possible physical danger) is often used in persuasive communication. It must be noted, however, that the correlation between the degree of fear incited by the information and the effect of the fear on actually changing the individual's behavior varies depending on the receiver's ability to counter risks. For instance, it has been pointed out that when the feeling of fear is provoked among people whose ability to counter risks is low, they will experience perception discord, which will eventually discourage them from taking countermeasures.

Furthermore, the risk information provider can also use the framing effect, which takes place when people perceive risks (the effect that the same phenomenon is accepted differently depending on the expression, i.e. by changing the frames). The provider can intentionally set up a frame to guide the receiver towards the choice of behavior that the provider desires.

4 Conclusion: Formation of the Subject of the Life System and Risk Information

If the author continues to conduct research on everyday life risk perception and countermeasures, and risk information, it may be able to unfold discussions in two directions. One direction is to discuss the possibility that the information on everyday life risks accepted by the receiver will change the value of life after going through the process of context shift at the individual level. This will eventually affect the value of

the social system that is the external environment. The other direction is to identify the mechanisms and solutions for cases in which a certain risk inflicts discord among several individuals (e.g. construction of a waste disposal facility).

There will be other opportunities to discuss these topics; however, there is only one goal to consider regardless of the direction of the discussions: the expectation that the subject of the life system can be formed by providing the individuals with information for perceiving and countering everyday life risks. The author believes that reiterative exchanges of risk information shall form collective risk perception based on comprehensive and adjustable decision-making, which is not only scientifically reasonable but also free from individual life values, and that risk countermeasures shall eventually be reinforced as a whole.

Acknowledgement

This study was supported by the Japan Society for the Promotion of Science (JSPS), Grant-in-Aid for Scientific Research (B).

References

1. Hirose, H.: Risk Perception. Journal of The Society for Risk Analysis 5(1) (1993)
2. Kinoshita, T.: Symbiosis between Science and Technology and Humans: Concepts and Techniques of Risk Communication, and Nature, Society and Culture as the Environment, Kyoto University Press (1997)
3. Kobayashi, S., Kanda, R.: Comparative Risks, Handbook of Risk Research. In: The Society for Risk Analysis, Japan Section, TBS Britannica (2000)
4. Murayama, R., Uchiyama, I., Sazuka, H., Hirose, H., et al.: Survey of Environmental Risk Levels (2). Collection of Papers of The Society for Risk Analysis 12 (1999)
5. Nara, Y.: Time-series Variations and Effectiveness of Earthquake Disaster Prevention Measures in Daily Lives, Disaster Prevention Measures and Communities: Consumers' Cooperative's Report on the 3rd Investigation of the Hanshin-Awaji Earthquake Disaster, Consumers' Cooperative Research Organization of Co-op Kobe (1999)
6. Ohsawa, Y.: Modeling the Process of Chance Discovery. In: Ohsawa, Y., McBurney, P. (eds.) Chance Discovery: Foundations and Applications, Springer, Heidelberg (2003)
7. Okamoto, K.: Introduction to Risk Psychology: Human Errors and Risk Images. Science (1995)

Logic of Discovery in Uncertain Situations– Deciding Algorithms

V. Rybakov

Department of Computing and Mathematics,
Manchester Metropolitan University,
John Dalton Building, Chester Street, Manchester M1 5GD, U.K.
V.Rybakov@mmu.ac.uk

Abstract. We study[1] a logic LDU (logic of Discovery in Uncertain Situations) generated in a semantic way as the set of all formulas valid in Krike/Hintikka models, which are models of linear discrete time with time clusters imitating possible uncertain states. The *possibility of discovery* and *uncertain necessity of discovery* are modeled by modal operations. The logic LDU differs from all standard normal and non-normal modal logics because the modalities ate not mutually expressible in standard way. We discuss properties of this logic, i.e. study its fragments, compare LDU with well known modal logics and study the main question about decidability of this logic. We propose an algorithm recognizing theorems of LDU (so we show that LDU is decidable), which is based on verification of validity of special normal reduced forms of rules in models of quadratic polynomial size in the testing rules.

Keywords: Modal logic, logic of discovery, Kripke/Hintikka models, inference rules, rules in normal reduced form.

1 Introduction

Non-classical logics form a modern branch of mathematical logic, which has diversified application in CS (Computer Science) and AI (Artificial Intelligence). Technique of non-classical logic has been involved to cover various aspects of problems for computerized modeling of intelligent behavior. A significant part of this technique is based at modal-type logics, which were introduced at the beginning of previous century. Major part of applications was focused to reasoning about knowledge, time and computation (cf., for instance, Goldblatt [8], van Benthem [22]). Semantic tools for modal logic are often based on Kripke/Hinttikka models and temporal algebras (cf. Thomason [21]). Since then, the branch of modal logics formed a solid direction in contemporary mathematical logic (cf., for instance, Goldblatt [9]). In our paper we attempt to apply techniques borrowed from modal logics to analysis of logic for chance and discovery.

[1] This research is supported by Engineering and Physical Sciences Research Council (EPSRC), U.K., grant EP/F014406/1.

B. Apolloni et al. (Eds.): KES 2007/ WIRN 2007, Part II, LNAI 4693, pp. 950–958, 2007.
© Springer-Verlag Berlin Heidelberg 2007

Chance Discovery (CD) (cf. Ohsawa and McBurney [13]) is a contemporary direction in AI and CS which analyzes events with uncertain information, incomplete past data, chance events, which are typically rare or hard to find. A chance is defined as some event which is significant for decision-making in a specified domain, Chance Discovery is defined by Y.Ohsawa as learning or explaining a chance event. The aims of CD as a discipline within AI are the determination of methods for discovering various chance events. It is an intriguing question what could be a logic of CD, which basic logical laws hold for CD and how they can be implemented. There are various viewpoints what could be meant by a logic of discovery and whether this logic exists at all (cf. M.Alai [1], M.Alani and Gino Tarozzi [2]). The logic of discovery has a solid prehistory, maybe, starting from the monograph Logic of Discovery and Logic of Discourse by Jaakko Hintikka and Fernand Vandamme [11]. This logic has interdisciplinary direction and is influenced by various ideas coming from researchers with distinct background. In particular, the modeling of environmental decision support systems been undertaken (cf. Cortés M., Sánchez-Marré and L. Ceccaroni [6], N.M. Avouris [4]), applications of the fuzzy logic were discovered (cf. A.J. Bugarin and S. Barro [5], D. Dubois and H. Prade [7]), tools involved in sematic web and multi-agent systems has been established (cf. F. Harmelem and I. Horrocks [10], J. Hendler [12], K. Arisha et.al [3]).

This paper studies a variation of logic of CD, – the logic of discovery in uncertain situations LDU, which is based on a modal logic different from all standard normal and non-normal modal logics (it is actually a special bi-modal logic). The semantic basis for this logic is formed by Kripke/Hintikka models generated by linear discrete time with time points being some clusters of states - uncertain situations. The possibility of discovering, – the *chance of discovery* – is modeled by modal operation $\Diamond$ (with meaning *possible*, and it is the standard modality), and the *necessity of discovering* is imitated by non-classical operation $\Box$ – *weak necessity* – with meaning that the fact necessary can be discovered in any time point in any time cluster. Actually, time points model discrete linear time flow, and each of them contains of states in current time point (all available databases, data structures etc.). The statement $\Diamond A$ says that the assertion A holds in one of states in future time points, that the *chance to discover A* is possible in current time point. $\Box A$ means that the assertion A holds in a state of any future time point, i.e. that the *chance to discover A* always holds in future.

We discuss the language and rules for computing truth values of formulas in suggested language in special Kripke/Hintikka models, define formally the logic LDU – logic of discovery in uncertain situations and compere fragments of the logic with standard modal logics. The basic question we are dealing with is construction of an algorithm recognizing theorems of LDU and checking satisfiability of formulas. We propose an algorithm (so we show that LDU is decidable) based on reduction of formulas to rules and converting rules in special reduced normal form, and, then, on checking validity of such rules in models of square polynomial size. The technique which we use is borrowed from our research (cf. [14] - [20] devoted to study inference rules in non-classical logics. It turns out,

that these instruments work rather good and we achieve only square polynomial bound for models involved in the algorithm for the decision problem. At the end of the paper some open problems are formulated.

2 Notation, Preliminaries

The paper uses standard notation and known facts concerning modal, multimodal and temporal logics, and some familiarity with basic facts is assumed, though we give below all necessary definitions to follow the paper.

To formulate logic LDU we proceed by introduction a semantic motivation on choice its language and rules for computing truth values of formulas. The model of chance discovery, which we consider in this paper, is based on the linear frames $\mathcal{N}_C := \langle \bigcup_{i\in N} C(i), R\rangle$. Here N is the set of natural numbers, each $i \in N$ (any natural number i) is the time index for possible current events from now to future. Any $C(i)$ is a finite set of all possible states in the time point i. In our formalism, any $C(i)$ is simply a set of elements (worlds in terms of Kripke/Hintikka semantics). The time flow (a computational process, an evolving of a system, a discovery search, a verification) is modeled by the binary accessibility relation R in $\mathcal{N}_C := \langle \bigcup_{i\in N} C(i), R\rangle$, where for all elements a and b from $\bigcup_{i\in N} C(i)$,

$$aRb \Leftrightarrow [a \in C(i) \textit{ and } b \in C(j) \textit{ and } i \leq j] \textit{ or } [a, b \in C(i) \textit{ for some } i.]$$

Less formally, R imitates the discrete flow of time connecting states in time, so, aRb means that a and b are some states at the same time pint or the state b will be achieved after the time point where the state a will be accomplished. So, the flow of the time is supposed to be linear, which matches well with daily human experience. How we could model the chance of discovery in these frames? First of all, we need to indicate possible events in models based on frames $\mathcal{N}_C$. For this we consider a set P of facts, propositions, which may or may not happen, may be true or nor true in time flow. For each frame $\mathcal{N}_C$, we consider valuations V of P, which are mappings of P into the set of all subsets of the set $\bigcup_{i\in N} C(i)$, so in symbols,

$$\forall p \in P, V(p) \subseteq \bigcup_{i\in N} C(i).$$

If, for an element $a \in \bigcup_{i\in N} C(i)$, $a \in V(p)$ we say *the fact* $\mathbf{p}$ *is discovered in the state* $\mathbf{a}$.

For example, consider the frame $\mathcal{N}_C := \langle \bigcup_{i\in N}[2i, 2i+1], R\rangle$, with the valuation V of the letter p, where $V(p) := \{2i \mid i \in N\}$. Then the fact p is discoverable in any cluster $[2i, 2i+1]$, this follows from the fact that any cluster $[2i, 2i+1]$ consists of an even number $2i$ and followed odd number $2i+1$. Let the frame $\mathcal{N}_C$ above to model the parallel computing, where at each computational step i we achieve exactly two states $2i$ and $2i+1$ (numeration of which is not essential), but $2i$ always is the state, where the property p – to have consistent

outputs – holds. Then the frame $\mathcal{N}_C$ models the special parallel computing, where, at each computational step, some state with consistent outputs is discoverable.

The question is how we could analyze the possibility of discovering (*chance of discovering*) a fact in logical terms. A close analogy to approach from modal and temporal logic and usage Kripke/Hintikka models is immediately coming to mind. How to say that it is possible to discover a fact p in future research?

A way is to introduce the modal operation $\Diamond$ (possible), where $\Diamond p$ means p is possible. To describe possible logical dependence and laws we evidently need standard Boolean logical operations: $\wedge(and), \vee(or), \neg(not), \rightarrow$ $(implease)$. To analogy to standard modal logic it is relevant to consider the modal operation $\Box$ (necessary), where $\Box p$ means p is necessary. Still, all these operations can be applied to arbitrary formulas but not merely propositions p, this way we may generate the formulas of arbitrary structure.

For instance, the formula $\Diamond(p \vee q) \rightarrow (\Diamond p \vee \Diamond q)$ says us that if we can discover $p \vee q$ then we can discover p or can discover q. The formal definition of formulas is as follows:

(i) Any propositional letter from P is a well formed formula (wff).
(ii) If φ and ψ are wff's then $\varphi \wedge \psi$, $\varphi \vee \psi$, $\varphi \rightarrow \psi$ and $\neg\varphi$ are wffs also.
(iii) If φ is a wff then $\Diamond\varphi$ and $\Box\varphi$ are also wffs.

Now we define rules for computing truth values of formulas in models $\mathcal{N}_C$ with valuations V of propositions P. In the notation below $(\mathcal{N}_C, a) \Vdash_V \varphi$ is meant to say the formula φ in true at the state a in the model $\mathcal{N}_C$ w.r.t. valuation V. The rules are as follows:

$\forall p \in P,\ \forall a \in \mathcal{N}_C\ \ (\mathcal{N}_C, a) \Vdash_V p \Leftrightarrow a \in V(p)$;

$(\mathcal{N}_C, a) \Vdash_V \varphi \wedge \psi \Leftrightarrow [(\mathcal{N}_C, a) \Vdash_V \varphi$ and $[(\mathcal{N}_C, a) \Vdash_V \psi]$;

$(\mathcal{N}_C, a) \Vdash_V \varphi \vee \psi \Leftrightarrow [(\mathcal{N}_C, a) \Vdash_V \varphi$ or $[(\mathcal{N}_C, a) \Vdash_V \psi]$;

$(\mathcal{N}_C, a) \Vdash_V \varphi \rightarrow \psi \Leftrightarrow [not[(\mathcal{N}_C, a) \Vdash_V \varphi]$ or $[(\mathcal{N}_C, a) \Vdash_V \psi]$;

$(\mathcal{N}_C, a) \Vdash_V \neg\varphi \Leftrightarrow not[(\mathcal{N}_C, a) \Vdash_V \varphi]$;

$(\mathcal{N}_C, a) \Vdash_V \Diamond\varphi \Leftrightarrow \exists b \in \mathcal{N}_C[(aRb)$ and $(\mathcal{N}_C, b) \Vdash_V \varphi]$;

$(\mathcal{N}_C, a) \Vdash_V \Box\varphi \Leftrightarrow [a \in C(i) \Rightarrow [\forall j \geq i \exists b \in C(j)(\mathcal{N}_C, b) \Vdash_V \varphi]$;

The truth valuation for Boolean operations is quite standard, and the same for the operation *possible to discover* – $\Diamond$. Indeed, the statement $(\mathcal{N}_C, a) \Vdash_V \Diamond\varphi$ codes that the fact (the formula) φ may be discovered – there is a future state for the current time point where the fact φ is discovered (the formula φ is true). But for our interpretation of the operation $\Box$ we choose a non-standard way. This is because we wont to interpret $\Box\varphi$ as

it is necessary that the fact φ may be discovered in any time point.

So, what we suggest is a thinner treatment of the operation $\Box$ as a variant of *week necessity*. $\Box\varphi$ does not say that φ is always true, but it says that φ is discoverable in a certain state of any time point (so, it says that φ is necessary discoverable). This specification of $\Box$ makes the logic we offer to be different from all standard modal logic - to be a non-classical modal logic. Notice that the Kripke models we are using in this research are *special kind of standard Kripke models* for bi-modal logics.

Definition 1. *The logic LDU is the set of all formulas which are true in any state of any frame $\mathcal{N}_C$ w.r.t. any valuation.*

What is evident immediately is that the $\Diamond$-fragment of LDU is equal to $\Diamond$ fragment of the well known standard modal logic $S4.3$ - it follows directly from the finite model property of $S4.3$. The $\Box$-fragment of LDU is already not equal to the $\Box$-fragment of $S4.3$. Moreover, the lemma.

Lemma 1. *The following holds $\Box p \rightarrow p \notin LDU, \Box p \equiv \neg\Diamond\neg p \notin LDU$ and $\Diamond p \equiv \neg\Box\neg p \notin LDU$.*

immediately derivable from definitions. So, $\Box$-fragment of LDU is not a fragment of any reflexive modal logic.

What is more interesting, $\Box$ and $\Diamond$ are not mutual counterparts of each other, so LDU differs with any standard normal or not-normal modal logic (because $\Box p \equiv \neg\Diamond\neg p$ and $\Diamond p \equiv \neg\Box\neg p$ do not hold in LDU). It is interesting whether $\Box$ and $\Diamond$ may be mutually expressed by any other ways. At the same time.

Lemma 2. *The following holds*
$\Box(p \rightarrow q) \rightarrow (\Box p \rightarrow \Box q) \in LDU$,
$\Box p \rightarrow \Box\Box p \in LDU$,
$\Box(\Box p \rightarrow \Box q) \vee \Box(\Box q \rightarrow \Box p) \in LDU$,
$\varphi \in LDU \Rightarrow \Box\varphi \in LDU$.

So, $\Box$-fragment of LDU extends the fragment of the standard normal modal logic $K4.3$. But more interesting question is what may happen if we combine both modalities $\Box$ and $\Diamond$ in formulas, the language became to be much more flexible and expressible. For instance, in a frame $\mathcal{N}_C$ the truth of formula $\Diamond\Box\varphi$ w.r.t a valuation (labeling) V means that a time point i is possible, where, since i in any time cluster the property φ is discoverable. The truth of the formula $\Box p \wedge \Box\neg p \rightarrow \neg\Diamond q$ would mean that if in any time point both p and $\neg p$ are discoverable then in any length of computation q cannot be discovered. All such properties are difficult to be expressed in the standard interpretation of modal language.

Thus, the logic LDU is defined, and we would like to know which logical laws hold for it. Say, whether two formulas φ and ψ are equivalent, whether a formula φ is a theorem (a logical law for LDU), i.e. whether $A \in LDU$. These questions are considered below.

3 Results, Decidability

The prime question we are dealing with in the sequel is how, for any given formula φ, to determine whether φ is a theorem of LDU, – whether $\varphi \in LDU$. If an algorithm for solving this task is possible the logic is said to be decidable. Actually LDU is a special bimodal logic and we can borrow some evolved technique to approach this task. Several ways are possible and we apply a technique based on our own approach tested for several logics (cf. [14] – [20]). This approach uses a representation of formulas by rules, and converting rules in a special normal reduced form. Recall, a (sequential) rule is an expression

$$\mathbf{r} := \frac{\varphi_1(x_1,\ldots,x_n),\ldots,\varphi_m(x_1,\ldots,x_n)}{\psi(x_1,\ldots,x_n)},$$

where $\varphi_1(x_1,\ldots,x_n),\ldots,\varphi_m(x_1,\ldots,x_n)$ and $\psi(x_1,\ldots,x_n)$ are some formulas constructed out of letters $x_1,\ldots,x_n$. Letters $x_1,\ldots,x_n$ are called variables of $\mathbf{r}$.

A formula φ is **valid** in a frame $\mathcal{N}_C$ (notation $\mathcal{N}_C \Vdash \varphi$) if, for any valuation V of $Var(\varphi)$ and for any element a of $\mathcal{N}_C$, $(\mathcal{N}_C, a) \Vdash_V \varphi$. We use notation $\mathcal{N}_C \Vdash \varphi$ for this fact.

Definition 2. *A rule* $\mathbf{r}$ *is said to be* **valid** *in the Kripke model* $\langle \mathcal{N}_C, V\rangle$ *with the valuation* V *(we will use notation* $\mathcal{N}_C \Vdash_V \mathbf{r}$*) if* $\forall a((\mathcal{N}_C, a) \Vdash_V \bigwedge_{1\le i\le m} \varphi_i) \Rightarrow \forall a((\mathcal{N}_C, a) \Vdash_V \psi)$. *Otherwise we say* $\mathbf{r}$ *is* **refuted** *in* $\mathcal{N}_C$*, or* **refuted in** $\mathcal{N}_C$ **by** V*, and write* $\mathcal{N}_C \not\Vdash_V \mathbf{r}$.

A rule $\mathbf{r}$ is **valid** in a frame $\mathcal{N}_C$ (notation $\mathcal{N}_C \Vdash \mathbf{r}$) if, for any valuation V of $Var(\mathbf{r})$, $\mathcal{N}_C \Vdash_V \mathbf{r}$. A rule $\mathbf{r}$ is said to have the *reduced normal form* if $\mathbf{r} = \varepsilon_c/x_1$ where

$$\varepsilon_c := \bigvee_{1\le j\le m} \Big(\bigwedge_{1\le i,k\le n, i\ne k} [x_i^{t(j,i,0)} \wedge (\Box x_i)^{t(j,i,1)} \wedge (\Diamond x_i)^{t(j,i,2)}],$$

and x_s are certain letter (variables), $t(j,i,z), t(j,i,k,z) \in \{0,1\}$ and, for any formula α above, $\alpha^0 := \alpha$, $\alpha^1 := \neg\alpha$. We employ here both $\Diamond$ and $\Box$ because they are not mutually expressible in LDU.

For any formula φ we can convert it in the rule $x \to x/\varphi$ and employ technique of reduced normal forms as follows.

Definition 3. *Given a rule* $\mathbf{r_{nf}}$ *in the reduced normal form,* $\mathbf{r_{nf}}$ *is said to be a normal reduced form for a rule* $\mathbf{r}$ *iff, for any frame* $\mathcal{N}_C$, $\mathcal{N}_C \Vdash \mathbf{r} \Leftrightarrow \mathcal{N}_C \Vdash \mathbf{r_{nf}}$.

Based on proofs of Lemma 3.1.3 and Theorem 3.1.11 from [15], by similar technique, we obtain.

Theorem 1. *There exists an algorithm running in (single) exponential time, which, for any given rule* $\mathbf{r}$*, constructs its normal reduced form* $\mathbf{r_{nf}}$.

It is immediate to see that a formula φ is valid in a frame $\mathcal{N}_C$ iff the rule $x \to x/\varphi$ is valid in $\mathcal{N}_C$, so from Theorem 1 we obtain

Proposition 1. *A formula φ is a theorem of LDU iff the rule $(x \to x/\varphi)_{nf}$ is valid in any frame $\mathcal{N}_C$.*

Thus, to solve the question about decidability of LDU it is sufficient to find an algorithm recognizing rules in reduced normal form which are valid in all frames $\mathcal{N}_C$. To describe our algorithm we need the following special finite Kripke models. Take any frame $\mathcal{N}_N$ and some numbers n, m, where $m > n > 1$. The frame $\mathcal{N}_C(n,m)$ has the following structure: $\mathcal{N}_C(n,m) := \langle \bigwedge_{1 \le i \le m} C(i), R \rangle$, where R is the accessibility relation from $\mathcal{N}_C$ extended by pairs (x,y), where $x, y \in [n,m]$, so xRy holds for all such pairs. If given a valuation V of letters from a formula φ in $\mathcal{N}_C(n,m)$, the truth values of φ can be defined in elements of $\mathcal{N}_C(n,m)$ by the same rules as for frames $\mathcal{N}_C$ above (actually just in accordance with standard computing of truth values for modalities). Just in case we describe below steps for modalities.

$$(\mathcal{N}_C(n,m), a) \Vdash_V \Diamond\varphi \Leftrightarrow \exists b \in \mathcal{N}_C[(aRb) \text{ and } (\mathcal{N}_C(n,m), b) \Vdash_V \varphi];$$

$$(\mathcal{N}_C(n,m), a) \Vdash_V \Box\varphi \Leftrightarrow [(a \in C(i)) \& (i \le n) \&$$

$$(\forall j (m \ge j \ge i) \exists b \in C(j) (\mathcal{N}_C(n,m), b) \Vdash_V \varphi)] \text{ or}$$

$$[(a \in C(i)) \& (i > n) \& (\forall j (n \le j \le m) \exists c \in C(j) (\mathcal{N}_C(n,m), b) \Vdash_V \varphi)].$$

Using this modified Kripke structures $\mathcal{N}_C(n,m)$ we derive

Lemma 3. *A rule $\mathbf{r_{nf}}$ in the reduced normal form is refuted in a frame $\mathcal{N}_C$ w.r.t. a valuation V if and only if $\mathbf{r_{nf}}$ is refuted in a frame $\mathcal{N}_C(n,m)$ by a valuation V_1, where*

- *(i) The size of any cluster $C(i)$ in $\mathcal{N}_C(n,m)$ is linear in $\mathbf{r_{nf}}$;*
- *(ii) n and m are linear in $\mathbf{r_{nf}}$;*
- *(iii) The size of the frame $\mathcal{N}_C(n,m)$ is square polynomial in $\mathbf{r_{nf}}$.*

Combining Theorem 1, Proposition 1 and Lemma 3 we derive.

Theorem 2. *The logic LDU is decidable.*

The algorithm of verification a formula to be a theorem consists of verification validity of rules in reduced normal form in Kripke frames of size square polynomial in the size of reduced normal forms. The overall complexity includes as well the reduction of rules to normal reduced forms, but this complexity is single exponential.

Conclusion. The logic LDU is seemed to be rather interesting because it models situations resistant to describe in the standard modal language. LDU has an

epistemological value, because it attempts to formalize the notion of *chance discovery* in logical terms. We propose an algorithm for recognizing logical laws of LDU (so, we proved that LDU is decidable) but there is a good avenue for future research. Say, it would be interesting to find axiomatization for LDU, to clarify whether the operations $\Box$ and $\Diamond$ are not expressible at all (no formula $\varphi(p)$ constructed out of the letter p without usage of $\Box$, where $\Box p \equiv \varphi(p) \in LDU$, and vise versa for $\Diamond$). It would be interesting to clarify the role of LDU among bi-modal logics, and to look for implementations.

References

1. Alai, M.: Univesity of Urbiono, Manuscript (2006), `http://www.uniurb.it/Filosofia/isonomia/alai.PDF`
2. Alai, M., Tarozzi, G. (eds.): Karl Popper Philosopher of Science. Rubbettino Editore, Soveria Mannelli (2006)
3. Arisha, K., Ozcan, F., Ross, R., Subrahmanian, V.S., Eiter, T., Kraus, S.: Impact: A platform for collaborating agents. IEEE Intelligent Systems 14(2), 64–72 (1999)
4. Avouris, N.M.: Co-operation knowledge-based systems for environmental decision-support. Knowledge-Based Systems 8(1), 39–53 (1995)
5. Bugarin, A.J., Barro, S.: Fuzzy reasoning supported by Petri nets. IEEE Transactions on Fuzzy Systems 2(2), 135–150 (1994)
6. Cortés, U., Sánchez-Marré, M., Ceccaroni, L.: Artificial intelligence and environmental decision support systems. Applied Intelligence 13(1), 77–91 (2000)
7. Dubois, Prade, H.: The three semantics of fuzzy sets. Fuzzy Sets and Systems 90(2), 141–150 (1997)
8. Goldblatt, R.: Logics of Time and Computation. CSLI Lecture Notes, 7 (1992)
9. Goldblatt, R.: Mathematical Modal Logic: A View of its Evolution. J. Applied Logic 1(5-6), 309–392 (2003)
10. Harmelem, F., Horrocks, I.: The semantic web and its languagesfaqs on oil: The ontology inference layer. IEEE Intelligent Systems 15(6), 69–72 (2000)
11. Hintikka, J., Vandamme, F.: Logic of Discovery and Logic of Discourse. Springer, Heidelberg (1986)
12. Hendler, J.: Agents and the semantic web. IEEE Intelligent Systems 16(2), 30–37 (2001)
13. Ohsawa, Y., McBurney, P. (eds.): Chance Discovery (Advanced Information Processing). Springer, Heidelberg (2003)
14. Rybakov, V.V.: A Criterion for Admissibility of Rules in the Modal System $S4$ and the Intuitionistic Logic. Algebra and Logic 23(5), 369–384 (1984) (Engl. Translation)
15. Rybakov, V.V.: Admissible Logical Inference Rules. Studies in Logic and the Foundations of Mathematics, vol. 136. Elsevier Sci. Publ., North-Holland (1997)
16. Rybakov, V.V.: Construction of an Explicit Basis for Rules Admissible in Modal System S4. Mathematical Logic Quarterly 47(4), 441–451 (2001)
17. Rybakov, V.V.: Logical Consecutions in Intransitive Temporal Linear Logic of Finite Intervals. Journal of Logic Computation (Oxford Press, London) 15(5), 633–657 (2005)
18. Rybakov, V.V.: Logical Consecutions in Discrete Linear Temporal Logic. Journal of Symbolic Logic 70(4), 1137–1149 (2005)

19. Rybakov, V.V.: Linear Temporal Logic with Until and Before on Integer Numbers, Deciding Algorithms. In: Grigoriev, D., Harrison, J., Hirsch, E.A. (eds.) CSR 2006. LNCS, vol. 3967, pp. 322–334. Springer, Heidelberg (2006)
20. Rybakov, V.V.: Branching Time Logic PTLα with Operations Until and Since Based on Bundles of Integer Numbers. Logical Consecutions, Deciding Algorithms (submitted, 2006)
21. Thomason, S.K.: Semantic Analysis of Tense Logic. Journal of Symbolic Logic 37(1) (1972)
22. van Benthem, J.: The Logic of Time, Reidel, Dordrecht, Synthese Library, vol. 156 (1983)

RFID Tags Without Customers ID in Book Library for Detecting Latent Interest of User Group

Yukio Ohsawa[1], Takuma Hosoda[1], Takeshi Ui[2],
Misato Ueda[2], and Hirofumi Tanaka[2]

[1] The University of Tokyo
[2] Toppan Forms Co. Ltd.,
ohsawa@q.t.u-tokyo.ac.jp

Abstract. The RFID tags attached on items in a retail store are expected to come up to marketers' requirement to understand latent interest of customers. The constraint due to security, however, is forcing the difficulty in recording the ID of customers, i.e., data on the information to distinguish customers. In this paper, we apply a method of chance discovery to ID-less RFID tag data obtained in an experimental library. Here we show a viewpoint to regard the patterns obtained from RFID-tag data without each customer's ID as the log of the group behavior of customers. Using the data to visualize a scenario map showing the relations between items, we found (1) the behavior of each customer tends to correspond to a part of the map, and (2) the preference of customers converge to interests latently shared by the group, in the extension of each customer's behaviors when they pick items looking at the scenario map. These results imply the ID-less RFID tags can be useful for activating the market as a community of customers rather than individual customer's behavior.

Keywords: RFID tags, Marketing, Customer group, Chance Discovery.

1 Introduction

The interest of marketing researchers in RFID tags is being enhanced recently [1]. The ability to record the picking behaviors of customers in a retail store, before they decide to buy or not to buy, enables to understand their latent interests more deeply than the previous data on the position of sales (POS). That is, the information that a customer's interest in an item was strong enough to pick it but not enough to buy it has not been dealt with in the POS data, in spite of the apparent importance of purchase decision. For example, suppose there is a cheese which seems good to eat for customers of a supermarket. This cheese, however, costs 2500 JPY (ca 20 US dollars). Many customers walking around the store pick the cheese, but mostly give up buying when they look at the price label. As a result, this cheese hardly appears in the POS data. However, if we can record the picking behaviors of customers, the data includes useful information, e.g., 10 of 1000 customers picked the cheese but did not buy. Introducing the RFID tags, we may be able to tell it is meaningful to discount the price of this cheese.

B. Apolloni et al. (Eds.): KES 2007/ WIRN 2007, Part II, LNAI 4693, pp. 959–969, 2007.
© Springer-Verlag Berlin Heidelberg 2007

Motivated with this merit of RFID tags, the data from RFID tags were dealt with from the aspect of data-based marketing [2,3,4]. Applying the methods of data mining and compression, we are able to acquire deeper level knowledge about consumers and customers than when we did on POS data. This can be compared with our experience that users' log data in a shopping web site enabled to analyze the latent interest of people than by seeing just the purchase records in the same site.

However, customers ID, i.e., the information about each customer tends to be missed when RFID tags are attached to items in a retail store, for the reason of data security [5]. This has been regarded as an obstacle in predicting customer's behaviors on each item. In this paper, we consider the retailing circumstance where (customer-) ID-less RFID tags are used, and go further within this physical constraint. By applying the concept of chance discovery, we first show how the expensive cheese above can be meaningful even if we do not discount its price. Based on this idea, we expect the group behavior of customers recorded with ID-less RFID tags reflects their desire for new values. This expectation is validated by experiments in a book library.

2 Chance Discovery as a Highlighted Application of RFID Tag

2.1 20$ Cheese: A Chance Is a Supermarket

In chance discovery, a chance has meaning an event which plays a significant role in human decision making since 2000 [6]. For example, let us take the expensive cheese again. Suppose you often go to the supermarket close to your house, and mostly buy liquor. The supermarket keeper likes you to buy things to eat also, and you should eat rather than only drink. However, an established style of life is hard to change. One day, you find a very nice looking cheese close to the liquor shelf from which you usually buy beer and wine. The cheese looks good to eat with drinking wine, but... you pick up the cheese and return it to the shelf, because it is as expensive as 20$.

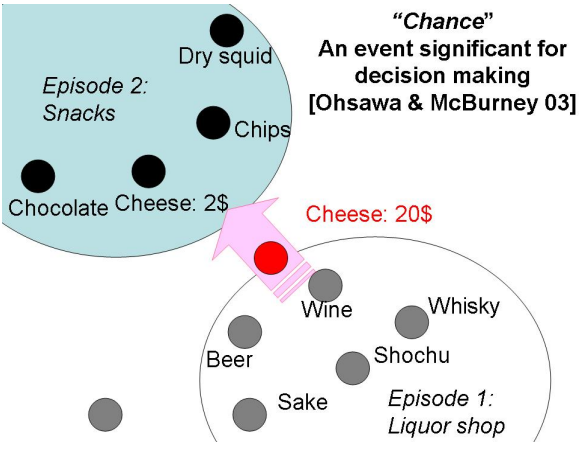


Fig. 1. Chance discovery in a supermarket: The expensive cheese plays a role

Yet, you do not go home but to the shelf of cheese and snacks, because you are now interested in something to eat which tastes like cheese. You learn to buy things to eat with things to drink since then. Although the 20$ cheese was not bought, it led the customer, you, to a new decision of purchase.

See figure 1 for the conceptual sketch of this scenario. The cheese of 20$ here is apparently a chance, because the event of picking the cheese made a strong influence on your decision. As well, the supermarket keeper may decide to keep the expensive cheese on the same shelf rather than to move it elsewhere or to discount its price. Changing the lifestyle of a customer makes more benefit than having the customer once buy an expensive item. Thus, we can expect to enable chance discovery more finely by introducing RFID tags data and visualizing the data in the way like the map in figure 1. We call this map a scenario map [7,8], because this map aids user in designing a valuable future scenario. This effect is expected to work more, if other customers come, see, touch and talk about the cheese as mentioned in the next section.

2.2 A Chance for a Community

Let us define a community as a group of people who do communication. When each member of the community embraces a scenario – a series of actions and events which occur under a coherent context - planned in the mind, he may talk to other members about the scenario, or realize his scenario by real actions. Another member who hears about his scenario may like to import it into her future action, if she feels empathy, as in figure 2. This effect is not rare, when we go and buy in a well-doing supermarket where you find other customers. Even though you do not talk to other customers, you can see their behaviors and this will be a non-verbal communication from which you can import other customers' scenarios into your new action. That is, the drift of your favor will trace just a short path connecting items e.g. beer, wine, and cheese of 20$. However, the short path of another customer will be combined to yours, and you will acquire a new scenario.

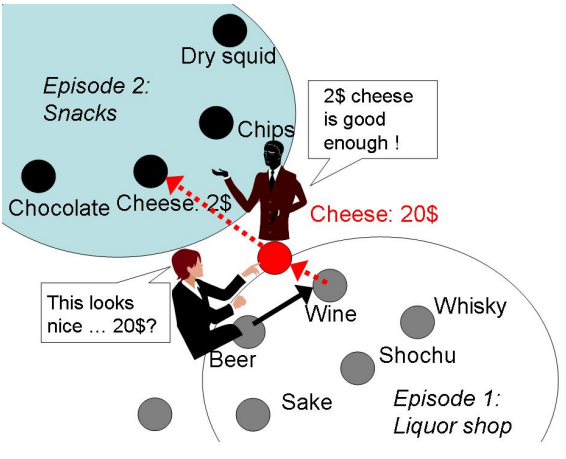


Fig. 2. She is just giving up the expensive cheese, but her colleague encourages her by presenting the choice of another scenario

On the basis of the discussion above, we take a hypothesis that the scenario map as in figure 1 or in figure 2 can be taken from the data on the behaviors of a community of people, rather than one person's history of actions. And, the definition of community can be relaxed to a group of people who do implicit communication, i.e., who see each other's behaviors. And, if one regards another person's choice to be similar to one's own recent choice, then will feel empathy with the person and import her scenario into one's own. For example, if we take customers' sales data as follows, we can a scenario map of as in figure 2. On this map, if the group of customers behave like a community, we can expect each customer will be interested in items in the extension of connected arrows of his/her recent purchase history.

On February 1, Mr. X bought the set {beer, wine, cheese 20$}
On February 2, Mr. X bought the set {wine, cheese 2$, chips}
On February 1, Ms. Y bought the set {beer}
On February 2, Ms. Y bought the set {beer, wine, cheese 2$}

3 Introducing RFID Tags in a Book Library

In this paper, we exemplify the data on customer's picking of books with RFID tags. The reason why we took a book library as the test bed in this study is three fold:

1) Each book has a high-dimensional and understandable value-set (see 3.2)
2) A large number of items (i.e. books) can be put on the shelves. We can store 50 books in a shelf of 1 meter width.
3) People often do pick-and-return behaviors when they touch a book shelf

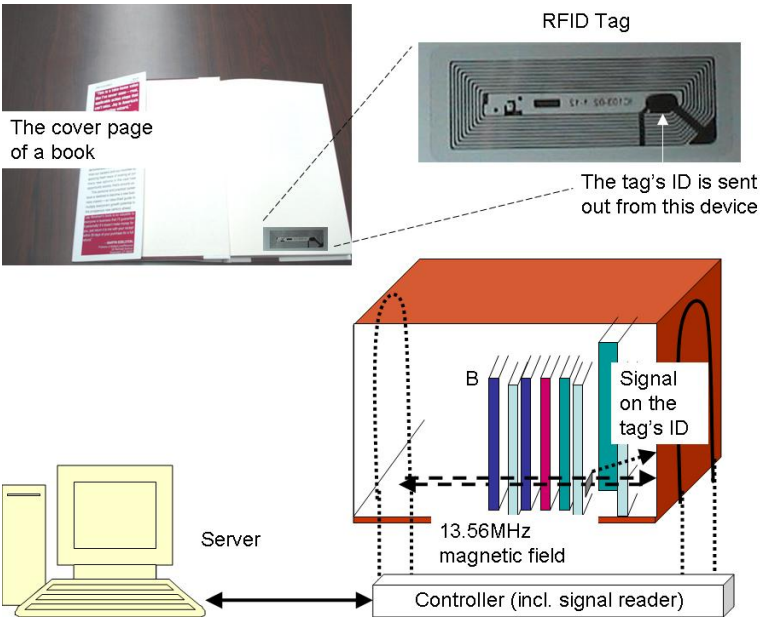


Fig. 3. The overview of the book library system with a RFID tag in each book

3.1 The RFID System

In our experimental library, one RFID tag is put on the back of the cover page of each book, similarly to the literature [9]. As in figure 3, the system here uses 13.56 MHz magnetic field to feed power to the passive RFID tags. The controller under the shelf manages this field, reads the data obtained from the tag-ID signals sent in from the RFID tags (attached to books), and forwards to the server. Via this server, the picking data of people who uses this library are recorded in the database. From this setting, we should keep in mind that (1) The reach length of 13.56MHz magnetic field is 30 cm to 50 cm. The reader judges the book left the shelf if the book has longer distance from the shelf than 50cm, but may take it as sitting still if a user just picks the book within 30cm off the shelf. (2) The switching of the magnetic field is managed by the controllers to be circulated throughout all shelves every 30 seconds. This means a picking of less than 30 seconds may be ignored. (3) The passive tags do not have power source in itself, so are light enough for the user to feel no gap of feeling from his/her daily experience with books.

The shelves in the library are separated as in figure 4. There are six stacks, and each stack has three level (upper, middle, and lower). Note the Chinese characters are used for representing upper, middle, and lower. Each level is separated into four shelf boxes (shelves, hereafter) A, B, C, and D. As a result, each shelf in the stack is called "1-A-Upper" ("Upper" is written in Chinese in the figures) if it is in the box A of the upper level of the first stack. The users of library are free to walk around, and can also take the seats nearby to read books they pick.



Fig. 4. The view of book shelves in the library

3.2 The Data Format of Users' Picking Books with RFID

The data as in table 1 is recorded as a result. Each line has information about the book picked/returned at each moment, such as the title, authors, and the belonging shelf. Note that there is no delimiters of baskets as POS data have, because we do not have information to distinguish users of the library. That is, each line of table 1 means a book picked or returned by "somebody" who visited.

Table 1. The data format of RFID tags on books in the library

Pick (1) and return(0)	Date and time	Shelf		ISBN	Title	Author
1	2006/11/6 14:32	4-A-下	4	4-7673-0803-8	小売・サービス業勝ち残る店は	安田龍平 編著
1	2006/11/6 14:33	2-A-上	177	4-532-13247-9	ゼミナール経営学入門	伊丹敬之, 加護野
1	2006/11/6 14:33	1-A-上	1351	4-478-50210-2	マーケティング原理	フィリップ・コトラー
0	2006/11/6 14:33	2-C-上	1570	4-8201-1736-X	組織設計のマネジメント	ジェイ Rガルブレ
1	2006/11/6 14:33	2-C-上	1570	4-8201-1736-X	組織設計のマネジメント	ジェイ Rガルブレ
1	2006/11/6 14:33	6-C-中	1151	4-502-59260-9	「人間力」で仕事が変わる	鈴野智子 著
0	2006/11/6 14:33	6-C-上	1977	4-901318-34-9	堀江貴文のお金をつかめ！	大谷晋郎&ホリエ
0	2006/11/6 14:34	1-A-上	1351	4-478-50210-2	マーケティング原理	フィリップ・コトラー

4 Experiment (I): Picking from Shelves and Data Visualization

4.1 The Setting of Experiment and Basic Data

We had 10 subject users who used the library for 120 minutes. They were instructed to read the books, and were encouraged to write down their impression of each book they pick. This was for having them keep a book more than 30 seconds, more than 50cm distance from the shelves. We also aimed to check the subjects' intension by this questionnaire, e.g., whether they meant to read them or were just attracted by the outlook of the cover page.

The 10 subjects were split into 3 groups, of 4 members, 3 members, and of 3 members. Subjects in each group were allowed to talk to each other, but no inter-group communications were allowed. This setting realized a realistic situation in a bookstore, where 2 to 3 people may come together and talk (although they should not).

See figure 5, to find the shelves in stack No.6 (the right part of the graph) was the most popular when we count the frequency for all the time. The shelves in stack No.3, and No.4 are much less. The most elementary data analysis may conclude books on these shelves should be removed. Books on 2-B might be also judges in the same way.

4.2 Movement from/to Shelves

In figure 6, we see the change of picked shelves by each 40 minutes of the 120 minutes of experiment. These graphs were drawn by the drawing function of Excel5.0. The horizontal axis means the stack number, and the vertical means the

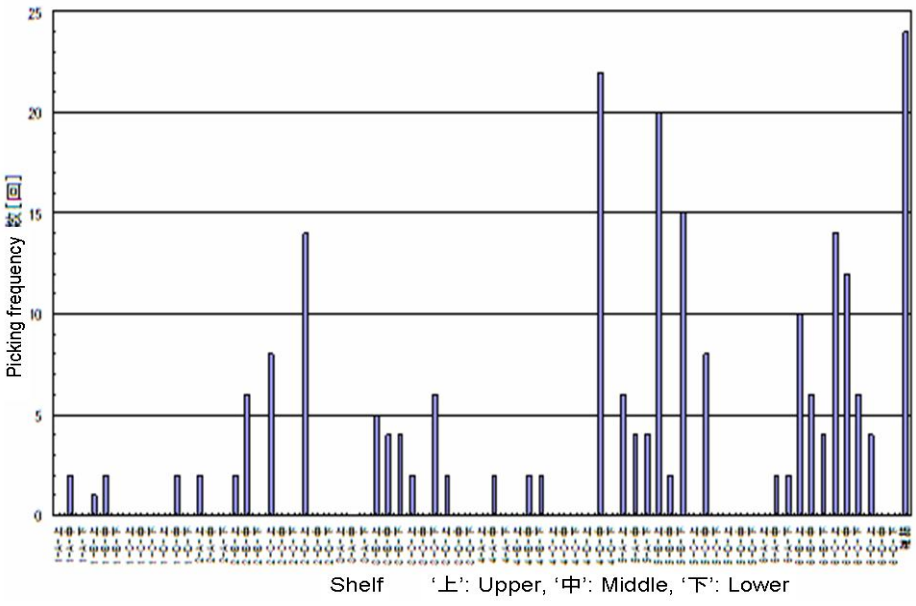


Fig. 5. The picking frequencies of book on each shelf

shelf box in each stack. The dots mean the shelves from/to which the subjects picked/returned books. The lines show the sequential occurrence of these shelves in the data as in Table 1. That is, a pair of dots connected by a line in figure 6 means that books in these shelves were touched in a short time (less than 30 seconds before or after the other). Figure 6 shows the three stages in the users' behavior:

Stage 1) Users started from stack No.6. Especially, they took books in shelf 6-C, and move to other shelves in a short time.

Stage 2) Users' attention changed to No.5. They moved further to the other shelves. Some seem to pick a few books from No.2

Stage 3) There is no more a central target such as No.5 and No.6 we found in the earlier stages. The subjects are checking books in the cluster of shelves including 2-C, 2-D, 3-B, 4-B, 4-D, and are still interested also in 5-C and 6-C.

5-A is playing the role of a bridge here, i.e., users move to form a new cluster after using 5-A. This shelf had books on marketing. Before this movement, the subjects' targets of picking were ranging across areas without a coherent context, because they just started from the shelves close to the place they started walking. After this, in stage 3), the subjects started to read from shelves relevant to the business methods based on established theories, such as marketing and strategic organizations.

Then we visualized the same data with KeyGraph [7,8,10]. KeyGraph shows islands (co-occurrence clusters of frequent events) and bridges (events which may be rare but co-occur with items in multiple islands). Here, the shelves in 2-B are found to bridge the starting shelves and the new cluster as in the upper area of figure 7. Items

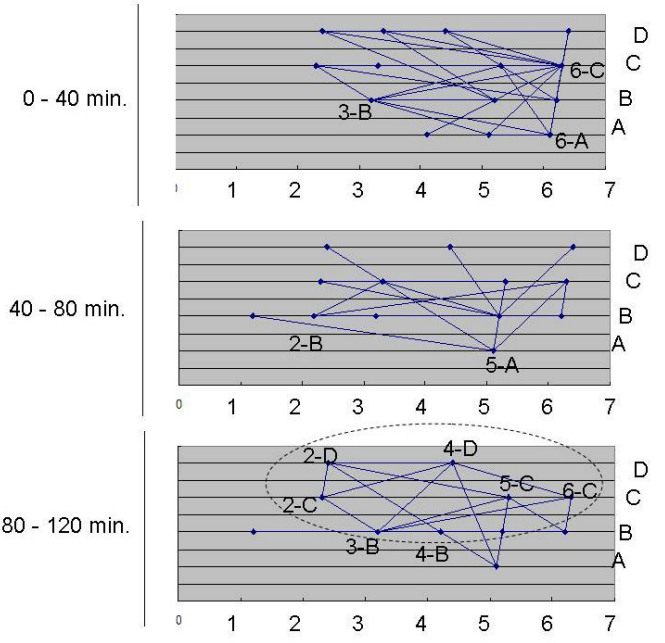


Fig. 6. The emergence of a shelf-cluster after 80 minutes

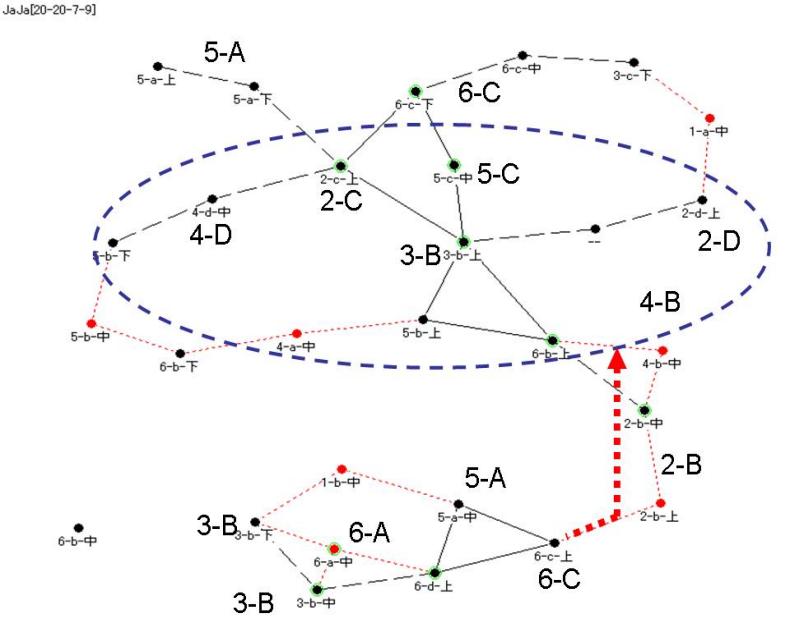


Fig. 7. The emergence of a shelf-cluster after 80 minutes, by KeyGraph

in this new cluster correspond to ones in the new cluster in figure 6. The 2-B-Upper shelf on the bridge, however, disappeared when the new cluster was formed according to figure 6. Thus 2-B-Upper was the least outstanding. On the other hand, as in figure 7, KeyGraph visualized 2-B including a scientific book titled "Complexity" which they picked few times and moved to the new cluster where science and business topics are merged. Hence the detection of bridging effect as the 20$ cheese played in figure 1 and 2, whether they are picked frequently as 5-A or rarely as 2-B.

5 Experiment (II): Picking of Books on KeyGraph

The values of the "title" variable in table 1 were taken for data in this section, and here we use only KeyGraph for visualization because it is desired to find items which do not outstand but may work as the 20$ cheese did. The scenario map resulted as in figure 8. Looking at this figure, the 10 subjects who did the picking in the previous section were engaged in another experiment here. Here, the 23 books appearing in the graph were put on a table, and each subject was instructed to pick and read as they like. We recorded the sequences of books that one read. Then, the next task of the subjects was to look at the graph and to choose favorite books.

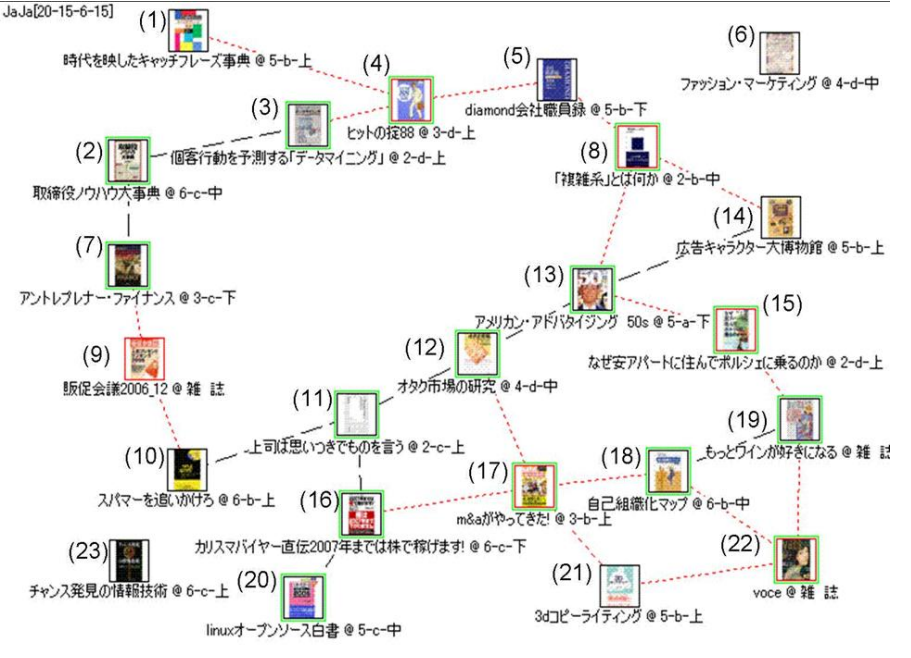


Fig. 8. KeyGraph visualizing the titles of picked books

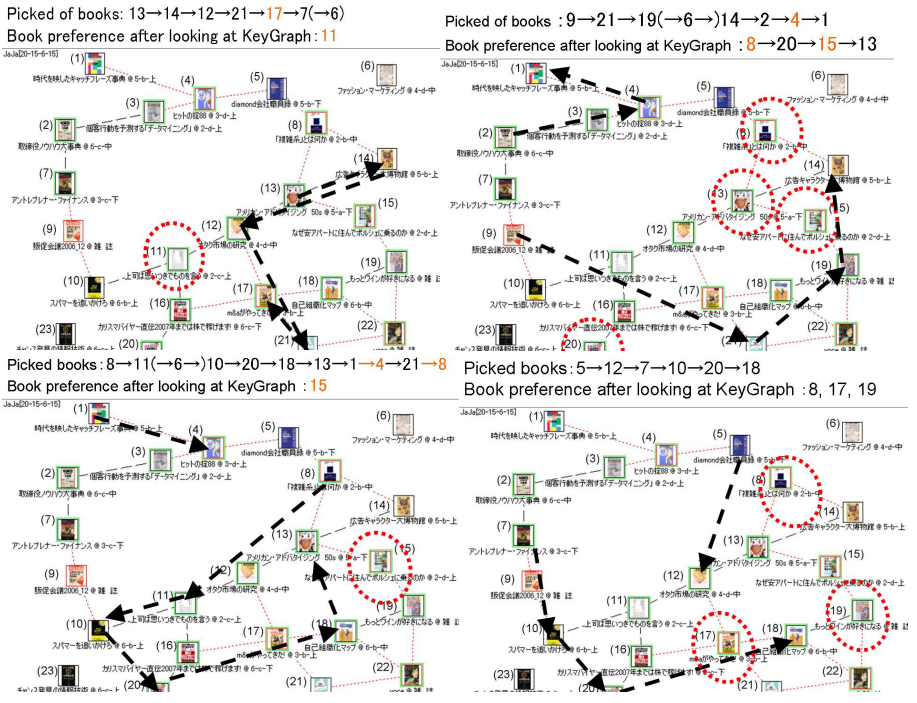


Fig. 9. The sequential order of booked picked, in the new experiment

The results were as shown in figure 9. Here, we attached the arrows onto the graph as follows: Item (6) and (23) in figure 8, isolated from other items, are deleted from the sequence of each subject. If two adjacent items A and B in the sequence are linked by a short path of less than for links in the graph, an arrow is attached from A to B.

Qualitatively, we find the following tendencies:

1) A connected sequence of arrows covers 2 to 6 among the 23 books in the graph.
2) The favorite books chosen looking at KeyGraph tend to be in the region in the extension of the arrow sequence. And, the subjects tend to nominate the items in the right-hand region of the graph, especially by high frequencies at 2-B, 3-B, and 5-A, that range close to the bridge in figure 6 and figure 7.

1) means the overall graph is the result of combining subjects' behaviors. The effect of KeyGraph shown to subjects, as 2), can be regarded as to remind them of books which affected their new choices in the group behavior with looking at other subjects and talking to each other.

6 Conclusions

An expectation of RFID tags applied to marketing have been to detect items that are touched but not bought, which may correspond to customers' latent interest which did not appear in sales data. On the other hand, we aimed in this paper to realize chance

discovery, which here means to detect items which are touched and influence the near-future behavior of customers. The experimental results in this paper show our successful progress in confirming that our aim comes true for RFID tags applied to libraries. Considering the findings here, we can conclude the RFID tags without the ID of customers are useful in navigating customers, because they organize human-group behaviors in the market not only on individual interest.

Acknowledgments. This study is under progress on the financial and environmental supports by Toppan Forms Co. Ltd. We appreciate for the assists of research by staffs of this company and Mr. Ken Hasegawa and Ms. Hikaru Kimura, students in the TMI department of The University of Tokyo. We also appreciate Dr. Eiji Murakami for constructive discussions about applications of RFID tags.

References

1. Lace, S.: Radio Frequency Identification Technology in Retail Briefing, for the 5 February 2004 summit at the National Consumer Council Prepared by Dr (2004), http://www.ncc.org.uk/technology/rfid.pdf
2. Gonzalez, H., Han, J., Li, X.: Mining compressed, commodity workflows from massive RFID data sets. In: Proceedings of the 15th ACM international conference on Information and knowledge management, pp. 162–171. ACM Press, New York (2006)
3. Murakami, E., Terano, T.: Fairy Wing:Distributed Information Service with RFID Tags. In: Kurumatani, K., Chen, S.-H., Ohuchi, A. (eds.) IJCAI-WS 2003 and MAMUS 2003. LNCS (LNAI), vol. 3012, pp. 174–189. Springer, Heidelberg (2004)
4. Hsu, H.-H., Cheng, Z., Huang, T., Han, Q.: Behavior Analysis with Combined RFID and Video Information. In: Ma, J., Jin, H., Yang, L.T., Tsai, J.J.-P. (eds.) UIC 2006. LNCS, vol. 4159, Springer, Heidelberg (2006)
5. Landwehr, C.E.: Conference Report on RFID Privacy Workshop, Concerns, Consensus, and Questions, IEEE Security and Privacy, March/April 2004, pp.34–36 (2004)
6. Ohsawa, Y., McBurney, P.: Chance Discovery. Springer, Heidelberg (2003)
7. Minami, T.: RFID in marketing Library marketing with RFID for supporting learning patrons. In: International Conf. on Multimedia and Information, Spain (November 2006)
8. Horie, K., Ohsawa, Y.: Product Designed on Scenario Maps Using Pictorial KeyGraph. WSEAS Transaction on Information Science and Application 3(7), 1324–1331 (2006)
9. Ohsawa, Y.: Scenario Maps on Situational Switch Model, Applied to Blood-Test Data for Hepatitis C Patients. In: Ohsawa, Y., Tsumoto, S. (eds.) Chance Discoveries in Real World Decision Making, pp. 69–82, 415–438. Springer, Heidelberg (2006)
10. Ohsawa, Y., Benson, N.E., Yachida, M.: KeyGraph: Automatic Indexing by Co-occurrence Graph based on Building Construction Metaphor. In: IEEE ADL'98. Proc. Advanced Digital Library Conference, pp. 12–18. IEEE Computer Society Press, Los Alamitos (1998)

Trigger to Switch Individual's Interest Toward Unconscious Preference

Yoshiharu Maeno[1], Yukio Ohsawa[2], and Takaichi Ito[3]

[1] Graduate School of Systems Management, Tsukuba University
maeno.yoshiharu@nifty.com
[2] School of Engineering, University of Tokyo
[3] Graduate School of Media and Governance, Keio University

Abstract. Unconscious preference is a potential trigger to switch an individual's interest toward new products, or unfamiliar persons. Its relationship to conscious preference sometimes appears vaguely in observation. We formulate a process to visualize such relationship and demonstrate a simple experiment to invent a scenario on a trigger to turn interests toward unconsicous preference.

Keywords: Human-interactive annealing, Scenario map, Unconscious preference.

1 Problem

It is important to understand individual's latent tendency in thinking and acting, to acquire opportunity and eliminate threat in marketing and communication. The latent tendency is secondary, hidden behind the primary tendency which is noticed by the outside observers easily. The latent tendency may, however, be indicated by the person's own stated words and interaction with the others implcitly. We call it unconscious preference in this paper. Unconscious preference is a potential trigger to turn the individual's interest toward new products and unfamiliar persons.

Various methods have been proposed to understand and to foresee individual consumer's behavior. Conjoint analysis is used to discover the optimal combination of factors which customers prefer [Green 1981]. Hierarchical Bayes model is employed to treat different nature of individual consumers. It is applied to the primary and secondary aspects of individual consumer's demand [Arora 1998]. On the other hand, a number of studies on latent class model, or latent trait model address modeling of latent variables and their statistical testing [Cheung 2004], [Zhang 2004]. These models, however, focus on model testing, rather than on hypothesis on a latent structure.

In this paper, we present a different approach to individual's unconscious preference. In a prior understanding, the existence and identity of unconscious preference is always missing. Its association to conscious preference, however, sometimes appears vaguely in observation. Various persons, objects, and their relationships are visualized on a scenario map to recognize individual's prior

B. Apolloni et al. (Eds.): KES 2007/ WIRN 2007, Part II, LNAI 4693, pp. 970–977, 2007.
© Springer-Verlag Berlin Heidelberg 2007

understanding, and cognition to observed preference. We believe that the gap between the prior understanding and cognition to observed preference is a clue to infer something beyond the conscious preference. The clue is grown up into a scenario on a trigger toward unconscious preference. This process is formulated into the human-interactive annealing process [Maeno 2007]. An experiment using questionnaire on art pieces are demonstrated to evaluate the quality of invented scenarios on individual subjects' trigger to switch their interests toward something unexpected.

2 Method

2.1 Process

The human-interactive annealing process aims at discovery from the understanding gap. The subjects, whose preference is of interest in the questionnaire on art pieces, become scenario map drawers and scenario inventors. The scenario map is a communication tool over which the subjects can point to drawn objects and talk with each other about their preference and the understanding gap. Group discussion and incorporation of the expert knowledge [Rush 1997] are beneficial to deepen interpretation of the scenario map. The scenario is not prediction, but a hypothetical action plan to propmt decision-making and to change the future. The understanding gap lies between the internal construct of the subjects' mind (human's work) and the external artifact supporting the subjects' decision making (computer's work) [Tory 2004]. The human's work includes prior understanding and circumstantial knowledge of the individual subject's own preference. The computer's work helps the subjects understand their cognition to observation (answers of the all subjects to the questionnaire).

The scenario map evolves in the iteration of subjects' interpretation (human's work) and computational processing (algorithm: computer's work) of the observation [Maeno 2006]. The process is illustrated in Fig.1 (5 scenario maps in 3 iterations). The map changes from a coarse-grained map to a fine-grained map as the iterations proceeds. Clusters in the map indicate preference discovered in the answers. Larger clusters are more commonly shared preference. Smaller clusters are more niche (less frequently encountered) preference. The clusters are connected with dummy events to form a single construct. The dummy event is a latent structure revealing a relationship from the primary (conscious) preference toward the secondary (unconscious) preference of a particular subject. The secondary (unconscious) preference may be something unexpected. The dummy event is, therefore, expected to work as an indicator of the understanding gap of the subject. The subjects acquire understanding and remove misunderstanding about their primary and secondary preference as the map evolves through the iteration of interpretation and algorithm. To acquire understanding means to move toward higher levels of understanding (Level 1: pre-structural, Level 2: uni-structural, Level 3: multi-structural, Level 4: relational, and Level 5: extended abstract [Kwok 2002]).

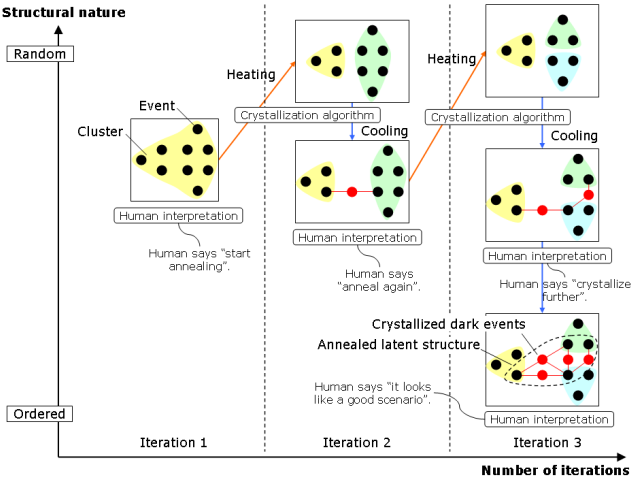


Fig. 1. Scenario map evolving in the iteration of interpretation and algorithm in the human-interactive annealing process (Copyright, IEEE 2007)

While annealing is executed with the modified peak temperature, heating step and cooling steps occur. In the heating step, up to the specified peak temperature, the number of clusters decreases. Dummy events disappear. Then, a cooling step comes, where cluster fragments are connected into a single construct with dummy events as temperature goes down. The number of dummy events between clusters increases. When the actors' degree of understanding is recognized to be high, the peak temperature shall be set higher. More complex higher-order latent structures shall be revealed. On the other hand, when the actors' degree of understanding is low, the peak temperature shall be set lower. The actor should try to learn from the basic lower-order structures. The iteration in the annealing process is terminated when the actors converges into complete understanding, resulting in many scenarios.

2.2 Algorithm

A basic structure found in the scenario map in Fig.1 is presented in Fig.2. Given a subject s_i, three structural components are drawn on a scenario map. The primary preference is a cluster of art pieces (c_{P}) which attracts the subject most strongly. It is drawn on the upper-left bright side. The secondary preference is a cluster of art pieces (c_{S}) which attracts the subject weakly. It is drawn on the lower-right dark side. The subject indicates preference to a few of the art pieces, but is indifferent to the others (this is why the cluster is on the dark side). The dummy event DE_i is a trigger event which stimulates the relationship between the primary and secondary preference. The dummy event is, in particular, interpreted as a switch to turn the subjct's interest toward unconscious preference (a clue to prompt the subject to buy something unfamiliar).

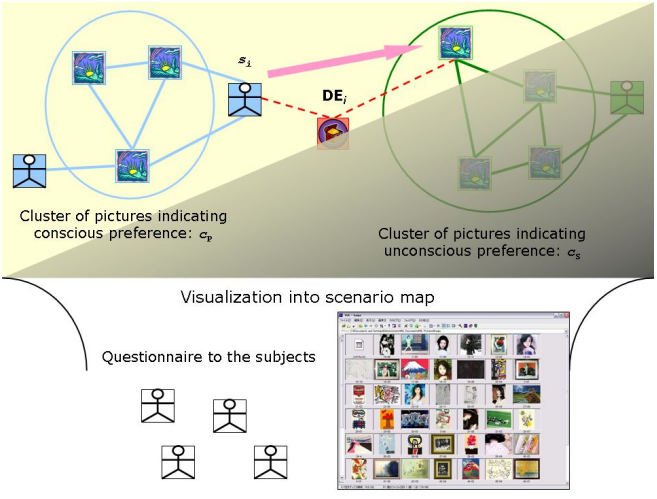


Fig. 2. Scenario map indicating a trigger to switch (dummy event) from conscious preference (upper-left bright side) toward unconscious preference (lower-right dark side)

The algorithm to draw a map like Fig.2 is presented below. Subjects' answers to the questionnaire is denoted by a market basket data (eq.(1)), which is equivalent to a bipartite graph structure. The market basket consists of a subject and a group of events. The set of events are those occurring simultaneously, spatially, or related strongly under a given subject. They are prefered art pieces in this case. The events are clustered into groups based on co-occurrence as a measure of similarity between events [Maeno 2006]. The co-occurrence is calculated by Jaccard coefficient in eq.(2). We can utilize various expertise on clustering. For example, k-medoids [Hastie 2001] or hierarchical clustering are simple, but efficient techniques. The k-medoids clustering is an EM algorithm similar to k-means algorithm for numerical data. Advanced algorithms for unsupervised learning like self-organization mapping can also be employed. A dummy event DE_i, representing a latent structure, is inserted into b_i, that results in b'_i in eq.(3) [Ohsawa 2005]. The index i can be used to identify the market basket where the corresponding dummy event was inserted.

$$b_i = s_i : \{e_{ij}\} \ (0 \leq i \leq B-1, 0 \leq j \leq |b_i|-1) \tag{1}$$

$$\mathrm{J}(e_i, e_j) = \frac{\mathrm{Freq}(e_i \cap e_j)}{\mathrm{Freq}(e_i \cup e_j)} \tag{2}$$

$$b'_i = s_i : \{e_{ij}\} \cup \mathrm{DE}_i, \ \mathrm{DE}_k = \mathrm{DE}_i \ (s_{\forall k} = s_i) \tag{3}$$

The clusters which the dummy event connect are derived with a measure in eq.(4). When individual subject appears only once in the market basket data, eq.(4) degenerates into a simpler form (eq.(5)). The primary and secondary

clusters for the dummy events are determined according to eq.(6) and eq.(7). The secondary cluster is either one that the dummy event is associated next to the primary cluster, or one that the dummy event is associated the least. The subject, its corresponding dummy events, the primary and secondary clusters are connected on a scenario map, resulting in a map like Fig.3.

$$\mathrm{C}(\mathrm{DE}_i, c_n) = \max_{e_m \in b_j \ (s_{\forall j} = s_i), c_n} \mathrm{J}(\mathrm{DE}_i, e_m) \tag{4}$$

$$\mathrm{C}(\mathrm{DE}_i, c_n) = \max_{e_m \in b_i, c_n} \frac{1}{\mathrm{Freq}(e_m)} \tag{5}$$

$$c_{\mathrm{P}}(i) = \arg \max_n \mathrm{C}(\mathrm{DE}_i, c_n) \tag{6}$$

$$c_{\mathrm{S}}(i) = \arg \max_{c_n \neq c_{\mathrm{P}}(i)} \mathrm{C}(\mathrm{DE}_i, c_n), \ c'_{\mathrm{S}}(i) = \arg \min_n \mathrm{C}(\mathrm{DE}_i, c_n) \tag{7}$$

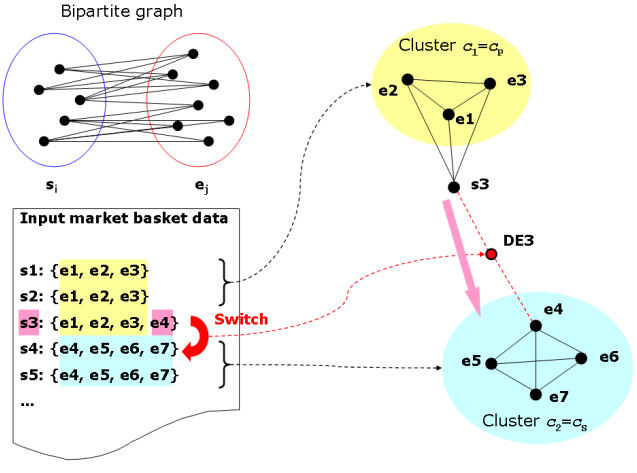


Fig. 3. Algorithm to convert a bipartite graph (corresponding market basket data) to a scenario map where dummy events are identified and placed to connect clusters

3 Experiment

The experiment uses the market basket data configured from the answers to the questionnaire on the preference of 20 subjects on 50 art pieces. The art pieces range widely from portraits and landscapes to abstract paintings. A few of subjects gathered, drew a scenario map using the process in 2.1, and discussed to invent scenarios on the primary and secondary preference of themselves and other subjects. The scenario map shown in Fig.4 is obtained as an output. In this case, art pieces are attached to the map to help the subjects understand intuitively. Most subjects have both primary and secondary preference. A few (**tkm** and **ymd** in the cluster 2) have multiple secondary preference. Some (**shimo** and

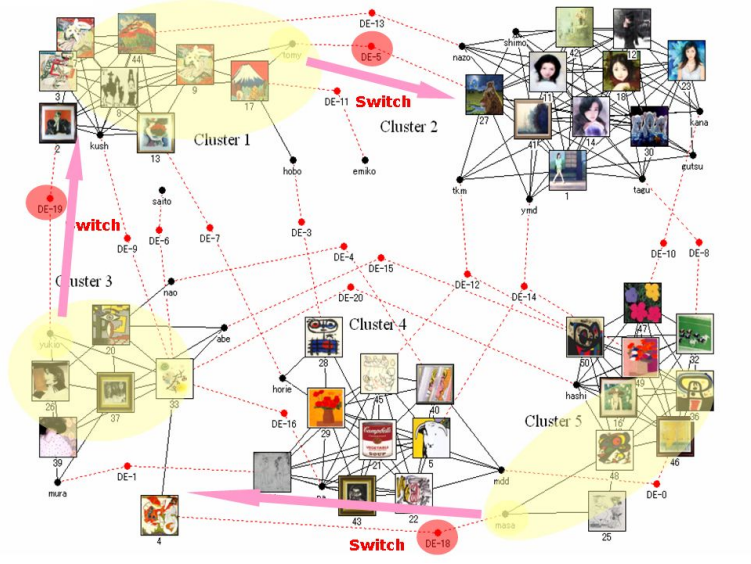


Fig. 4. Scenario map derived from 20 subjects' preference on 50 art pieces (Copyright, IPSJ 2007)

gutsu in the cluster 2) do not have secondary preference. Their preference is concentrated. The other (**saito** and **emiko**) do not have primary preference (do not belong to any clusters). Their preference is sparse and distributed.

At first, the participants of the discussion attempted to verbalize the preference represented by individual clusters. It is summarized in Table 1. Then, the participants tried to invent scenarios as many as possible. The participants focused on the identity of the triggers denoted by DE_i to switch the subjects' interests toward something unexpected. As a result, 18 scenarios were invented, and the identity of 12 triggers are inferred. Three examples are shown below. Corresponding triggers are labeled in red in Fig.4.

- **masa** in the cluster 5 encourages himself and succeeds. But he unconsciously wants to escape from the stress by mental healing (the identity of a trigger: DE_{18}).
- **yukio** in the cluster 3 is toward something genius. But he unconsciously wants to spiritual uplifting by risky adventures, or gambles (the identity of a trigger: DE_{19}).
- **tomy** in the cluster 1 sets a high goal and tries to grow up. But he unconsciously wants to be as he is by enjoying maniac hobbies, and enthusiasm (the identity of a trigger: DE_5).

Finally, the participants evaluated the scenarios to see if they are agreeable. Being agreeable means that an action plan can be proposed from the scenario or the trigger, which can not be easily noticed, but can be carried out practically.

Table 1. Interpretation of the clusters in Fig.4, verbalized by the subjects in a group discussion

Cluster No.	Interpretation of tasts indicated by the cluster	Keyword
1	colorful, oriental, symbolic	soul
2	transparent, sky-high, clear, calm	peace
3	wonder, paranormal, ghostly	genius
4	abnormal, grotesque	false genius
5	stirring, uplifting, boiling, radical	oppression

There are 3 viewpoints: proposeability (possibility to be a concrete action plan), unnoticeability (difficulty to notice), and growability (feasibility to carry out). If a scenario is good in the all 3 viewpoints, it is assigned True in agreeability. In some scenarios, triggers can not be clearly identified. The triggers inferred in the scenarios are also evaluated to see if they are agreeable. The evaluation results are summarized in Table 2. The ratio, 78%, and 83 %, is the average quality of the scenarios. After the experiment, the participants expressed an opinion that a process visualizing weak relationship like this is stimulus to invent something new but close to the present state. We believe that this can be a starting point to test the applicability of the process, although the evaluation is still subjective and preliminary.

Table 2. Summary of the agreeability evaluation to the invented scenarios by the subjects

Evaluation item	Score
Number of triggers whose identity is inferred (Y_1)	12
Number of triggers whose identity is agreeable (X_1)	10
Rate X_1/Y_1	**0.83**
Number of invented scenarios (Y_2)	18
Number of scenarios which are agreeable (X_2)	14
Rate X_2/Y_2	**0.78**

4 Discussion

Scenarios such as addressed in the human-interactive annealing here might be useful to stimulate brand switch in marketing or to extend belonging groups in communication. The invented scenarios may depend on specific characteristics of the art pieces, and subjects employed in the experiment. Questionnaire also affects the scenarios. Which art piece is suitable to your room? Which art piece do you like to buy ? These two questions may cause different responses. The action plan based on the scenarios is subject to the nature and change of the environment. Such sensitivity is essential to consumers, in particular, early adaptors, or trend leaders, in the post-modern marketing. We believe that it is more beneficial

to focus on such case-by-case scenarios with the aid of computers and experts in information gathering and analysis to accelerate the human-interactive annealing process, than to make an effort to figure out universal laws of human behaviors.

In this paper, we take a first step toward individual's latent tendency in thinking and acting. It is completely different from a conventional subliminal advertising, where something is input to individual's tendency in thinking and acting unconsciously. Our approach is to stimulate existing individual's latent tendency to prompt change. The individual's latent tendency itself may change by becoming aware of its nature. We believe that the human-interactive annealing process fosters understanding various visible, and invisible aspects of person's thinking and acting in marketing, communication, and many other applications.

References

[Arora 1998] Arora, N., Allenby, G.M., Ginter, J.L.: A hierarchical Bayes model of primary and secondary demand. Marketing Sci. 17, 29–44 (1998)

[Cheung 2004] Cheung, K.W., Tsui, K.C., Liu, J.: Extended latent class models for collaborative recommendation. IEEE Trans. Sys., Man & Cyber., Part A 34, 143–148 (2004)

[Green 1981] Green, P., Carroll, J., Goldberg, S.: A general approach to product design optimization via conjoint analysis. J. Marketing 43, 17–37 (1981)

[Hastie 2001] Hastie, T., Tibshirani, R., Friedman, J.: The elements of statistical learning: Data mining, inference, and prediction. Springer, Heidelberg (2001)

[Kwok 2002] Kwok, R., Ma, J., Zhou, D.: Improving group decision making: a fuzzy GSS approach. IEEE Trans. Sys., Man, & Cyber. Part C 32, 54–63 (2002)

[Maeno 2007] Maeno, Y., Ohsawa, Y.: Human-computer interactive annealing for discovering invisible dark events. IEEE Trans. Ind. Electron. 54, 1184–1192 (2007)

[Maeno 2006] Maeno, Y., Ohsawa, Y.: Stable deterministic crystallization for discovering hidden hubs. In: Proc. IEEE Int'l. Conf. Sys., Man & Cyber., Taipei, pp. 1393–1398. IEEE Computer Soceity Press, Los Alamitos (2006)

[Ohsawa 2005] Ohsawa, Y.: Data crystallization: chance discovery extended for dealing with unobservable events. New Math. and Nat. Comp. 1, 373–392 (2005)

[Rush 1997] Rush, R.: Elicitation of knowledge from multiple experts using network Interface. IEEE Trans. Knowl. and Data Eng. 9, 688–696 (1997)

[Tory 2004] Tory, M., Moeller, T.: Human factors in visualization research. IEEE Trans. Vis. and Comp. Graph. 10, 72–84 (2004)

[Zhang 2004] Zhang, N.L., Kocka, T.: Efficient learning of hierarchical latent class models. In: Proc. IEEE Int'l. Conf. Tools with Art. Intel., Boca Raton, pp. 585–593. IEEE Computer Society Press, Los Alamitos (2004)

Visualization of Similarities and Dissimilarities Between Rules Using Multidimensional Scaling

Shusaku Tsumoto and Shoji Hirano

Department of Medical Informatics,
Shimane University, School of Medicine,
Enya-cho Izumo City, Shimane 693-8501 Japan
tsumoto@computer.org, hirano@ieee.org

Abstract. This paper proposes a visualization approach to show the similar relations between rules by using multidimensional scaling, which assign a two-dimensional cartesian coordinate to each data point from the information about similiaries between this data and others data. As similarities, we adopt the average and absolute value of the difference of precision and recall. We evaluated this method on two medical data sets, whose experimental results show that knowledge useful for domain experts could be found.

Keywords: Rule Mining, Similarity, Visualization, MDS.

1 Introduction

One of the most important problems with rule induction methods is that it is very difficult for domain experts to check millions of rules generated from large datasets. Moreover, since the data collection is deeply dependent on domain knowledge, rules derived by datasets need deep interpretation made by domain experts. For example, Tsumoto and Ziarko reported the following case in analysis of a dataset on meningitis [1].

Even though the dataset is small, the number of records is 198, they obtained 136 rules with high confidence (more than 0.75) and support (more than 20). Here are the examples which are unexpected to domain experts.

```
1. [WBC  12000] & [Gender=Female] & [CSFcell  1000] => Virus meningitis
   (Accuracy: 0.97, Coverage: 0.55)
2. [Age > 40] & [WBC > 8000] => Bacterial meningitis
   (Accuracy: 0.80, Coverage: 0.58)
3. [WBC > 8000] & [Gender=Male] => Bacterial menigits
   (Accuracy: 0.78, Coverage: 0.58)
4. [Gender=Male] & [CSFcell>1000] => Bacterial meningitis
   (Accuracy: 0.77, Coverage: 0.73)
```

The factors in these rules unexpected to domain experts are gender and age, which have not been pointed out in the literature on meningitis [2].

B. Apolloni et al. (Eds.): KES 2007/ WIRN 2007, Part II, LNAI 4693, pp. 978–986, 2007.
© Springer-Verlag Berlin Heidelberg 2007

Since these detected patterns may strongly depend on the characteristics of data, Tsumoto and Ziarko searched for the hidden factors. For this analysis, several groupings of attributes are processed into the dataset.

The results obtained from the secondary analysis of processed data show that both $[Gender = male]$ and $[Age > 40]$ are closely related with chronic diseases, which is a risk factor of bacterial meningitis. The first attribute-value pair, $[Gender = male]$ is supported by 70 cases in total 198 records: 48 cases are bacterial meningitis, all of which suffered from chronic diseases (25 cases: diabetes mellitus, 17 cases: liver cirrhosis and 6 cases: chronic sinusitis.) On the other hand, $[Age > 40]$ is supported by 121 cases: 59 cases are bacterial meningitis, 45 cases of which suffered from chronic diseases (25 cases: diabetes mellitus, 17 cases: liver cirrhosis and 3 cases: chronic sinusitis.) Domain explanation was given as follows: chronic diseases, especially diabetes mellitus and liver cirrhosis degrade the host-defence to microorganisms as immunological deficiency and chronic sinusitis influences the membrane of brain through the cranial bone. Epidemiological studies show that women before 50 having mensturation suffer from such chronic diseases less than men.

This example illustrates that deep interpretation based on data and domain knowledge is very important for discovery of new knowledge. Especially, the above example shows the importance of similarities between rules. When one rule r_i has reasonable features and the other rule r_j with high similarity to r_i includes unexpected factors, the relations between these rules will become a trigger to the discovery of knowledge, especially about risk or chance.

In this paper, we propose a visualization approach to show the similarity relations between rules based on multidimensional scaling, which assign a two-dimensional cartesian coordinate to each data point from the information about similiaries between this data and others data. We evaluated this method on three medical data sets. Experimental results show that several knowledge useful for domain experts could be found.

2 Similarity of Rules

Let $sim(a, b)$ denote a similarity between objects a and b. Formally, similarity relation should hold the following relations:

1. An object a is similar to oneself: $sim(a, a)$.
2. If $sim(a, b)$, then $sim(b, a)$. (Symmetry)

It is notable that the second property is the principal axiom for the similarity measure. In this section, we define three types of similarity measures for rules which hold the above relations. As shown in the subsection 1, rules are composed of (1) relation between attribute-value pairs (proposition) and (2) values of probabilistic indices (and its supporting sets). Let us call the former component a **syntactic** part and the latter one a **semantic** part. Two similarities are based on the characteristics of these parts.

2.1 Syntactic Similarity

Syntatic similarity is defined as the similarity between conditional parts of the same target concept. In the example shown in Section 1, the following two rules have similar conditional parts:

```
R2. [Age > 40] & [WBC > 8000] => Bacterial meningitis
   (Accuracy: 0.80, Coverage: 0.58)

R3. [WBC > 8000] & [Gender=Male] => Bacterial menigits
   (Accuracy: 0.78, Coverage: 0.58)
```

The difference between these two rules are $[Age > 40]$ and $[Gender = Male]$. To measure the similarity between these two rules, we can apply several indices of two-way contigency tables. Table 1 gives a contingency table for two rules,

Table 1. Contingency Table for Similarity

		$Rule_j$		
		Observed	*Not Observed*	Total
$Rule_i$	*Observed*	a	b	$a+b$
	Not Observed	c	d	$c+d$
	Total	$a+c$	$b+d$	$a+b+c+d$

$Rule_i$ and $Rule_j$. The first cell a (the intersection of the first row and column) shows the number of matched attribute-value pairs.

2.2 Similarities

From this table, several kinds of similarity measures can be defined [3,4]. The best similarity measures in the statistical literature are seven measures shown in Table 2. It is notable that these indices satisfies the property on symmetry shown in the beginning of this section.

2.3 Semantic Similarity: Covering

The other similarity which can be defined from the definition of the rule is based on the *meaning* of the relations between formulas f_i and f_j from the viewpoint of set-theoretical point of view. Let us assume that we have two rules:

$$f_i \rightarrow D\ (\alpha_{f_i}(D), \kappa_{f_i}(D))$$
$$f_j \rightarrow D\ (\alpha_{f_j}(D), \kappa_{f_j}(D))$$

As shown in the last subsection, syntactic similarity is defined as $sim(f_i, f_j)$ from the viewpoint of syntactic representations. Since f_i and f_j have meanings

Table 2. Definition of Similarity Measures

(1) Matching Number	a
(2) Jaccard's coefficient	$a/(a+b+c)$
(3) χ^2-statistic	$N(ad-bc)^2/M$
(4) point correlation coefficient	$(ad-bc)/\sqrt{M}$
(5) Kulczynski	$\frac{1}{2}(\frac{a}{a+b}+\frac{a}{a+c})$
(6) Ochiai	$\frac{a}{\sqrt{(a+b)(a+c)}} = \sqrt{\alpha_R(D)\kappa_R(D)}$
(7) Simpson	$\frac{a}{min\{(a+b),(a+c)\}} = min(\alpha_R(D), \kappa_R(D))$

$N = a+b+c+d$, $M = (a+b)(b+c)(c+d)(d+a)$

(supporting sets), f_{iA} and f_{jA}, respectively, where A denotes the given attribute space. Then, we can define $sim(f_{iA}, f_{jA})$ by using a contingency table: Table 1 in the same way.

2.4 From Assymmetric Indices to Symmetric Ones

Since a similarity measure between two rules is symmetric, it may not capture assymetrical information between two rules. To extract simmilarities and dissimilarities information, we defined the following two measures:

$$average(R, D) = \frac{1}{2}(\frac{a}{a+b} + \frac{a}{a+c}), \quad and$$
$$difference(R, D) = \frac{1}{2}|\frac{a}{a+b} - \frac{a}{a+c}|$$

where the difference is equivalent to the dissimilarity. It is notable that the difference has a meaning if $a \neq 0$. If $a = 0$, the difference will be maximum. We assume that $difference(R, D)$ is equal to 1.0 if the intersection a is equal to 0.

3 Multidimensional Scaling

3.1 Problems with Clustering

After calculating the similarity relations among rules and deriving similarity matrix, patterns with respect to similarities will be investigated. Usually, data miners apply clustering methods to the similarity matrix obtained from given datasets [3]. However, clustering have several problems: most of clustering methods, hierarchical or nearnest neighbor based methods, forces grouping given examples, attributes or classes into one or several classses, with one dimension. This limitation is already given as a formal form of similarity function, $sim(a, b)$. As a mapping, similarity function is given as $O \times O \rightarrow R$, where O and R denote a set of objects and real number, respectively.

Since each object may have many properties, one dimensional analysis may not be sufficient for detecting the similarity between two objects. In these cases, the increase of the dimensionality gives a wider scope for reasoning about similarities. Multidimensional scaling (MDS) is one solution for dimensionality, which uses two-dimensional plane to visualize the similarity relations between objects.

3.2 How MDS Works

Metric MDS. The most important function of MDS is to recover cartesian coordinates (usually, two-dimensional) from a given similarity matrix. For recovery, we assume that a similarity is given as an inner product of two vectors for objects. Although we need three points to recover coordinates, but one point is fixed as the origin of the plane.

Let us assume that the coordinates of two objects x_i and x_j are given as: $(x_{i1}, x_{i2}, \cdots, x_{ip})$ and $(x_{j1}, x_{j2}, \cdots, x_{jp})$, where p is the number of dimension of the space. Let k denote the origin of the space $(0, 0, \cdots, 0)$. Then, here, we assume that the distance betweeen x_i and x_j d_{ij} is given as the formula of distance, such as Eucledian, Minkowski, and so on. MDS based on this assumption is called *metric MDS*. Then, the similarity between i and j s_{ij} is given as:

$$s_{ij} = d_{ik} d_{jk} \cos\theta = \sum_{m=1}^{p} x_{im} x_{jm}$$

From the triangle ijk, the following formula holds:

$$d_{ij}^2 = d_{ik}^2 + d_{jk}^2 - 2 d_{ik} d_{jk} \cos\theta.$$

Therefore, similarity should hold the following formula.

$$s_{ij} = \frac{d_{ik}^2 + d_{jk}^2 - d_{ij}^2}{2}$$

Since s_{ij} is given as $\sum_{m=1}^{p} x_{im} x_{jm}$, the similarity matrix for s_{ij} is given as:

$$\mathbf{Z} = \mathbf{X}\mathbf{X}^{\mathbf{T}},$$

where $\mathbf{X}^{\mathbf{T}}$ denotes the transposition matrix of $\mathbf{X}$. To obtain $\mathbf{X}$, we consider the minimization of an objective function Q defined as:

$$Q = \sum_i \sum_j \left(z_{ij} - \sum_{m=1}^{p} x_{im} x_{jm} \right)^2.$$

For this purpose, we apply *EckartandYoung* decomposition [5] in the following way. first, we calculate eigenvalues, denoted by $\lambda_1, \cdots, \lambda_p$, and eigenvectors of $\mathbf{Z}$, denoted by $\mathbf{v^1}, \cdots, \mathbf{v^p}$. Then by using a diagnoal matrix of eigenvalues, denoted by $\mathbf{\Lambda}$ and a matrix with eigenvectors, denoted by $\mathbf{Y}$, we obtain the following formula:

$$\mathbf{X} = \mathbf{Y}\mathbf{\Lambda}\mathbf{Y}^{\mathbf{T}},$$

where

$$\mathbf{\Lambda} = \begin{pmatrix} \lambda_1 & 0 & \ldots\ldots\ldots & 0 \\ 0 & \lambda_2 & 0 \ldots & \ldots & 0 \\ \ldots\ldots & & \lambda_i & \ldots\ldots\ldots & \\ 0 & 0 & \ldots & \lambda_{p-1} & 0 \\ 0 & 0 & \ldots & 0 & \lambda_p \end{pmatrix} \quad and \quad \mathbf{Y} = \left(\mathbf{v^1}, \mathbf{v^2}, \cdots, \mathbf{v^p}\right).$$

From this decomposition, we obtain $\mathbf{X}$ as $\mathbf{X} = \mathbf{Y\Lambda^{1/2}}$.

Nonmetric MDS. The above metric MDS can be applied to the case only when the difference between similarities has the meaning. In other words, the similarity index holds the property of interval calculus (interval scale). If the similarity index holds the property of order, we should not apply the above calculus to the similarity matrix, but we should apply nonmetric MDS method. Here, we will introduce Kruskal method, which is one of the most well-known nonmetric MDS method [6].

First, we calculate given similarities s_{ij} into distance data d^*_{ij} (dissimilarity). Next, we estimate the coordinates of x_i and x_j from the minimization of *Stress* function, defined as:

$$S = \sqrt{\frac{\sum\sum_{i<j}(d_{ij} - d^*_{ji})^2}{\sum\sum d^2_{ij}}},$$

where the distance d_{ij} is defined as a Minkowski distance:

$$d_{ij} = \left(\sum_{m=1}^{P} |x_{im} - x_{jm}|^t\right)^{1/t},$$

where t denotes the Minkowski constant. For the minimization of S, optimization methods, such as gradient method are applied and the dimensionality and t will be estimated. Since the similarity measures given above do not hold the property of distance (triangular inequality), we adopt nonmetric MDS method to visualize similarity relations.

3.3 Process for Visualization

Process for Visualization is given as follows:

1. Induce rules with accuracy and coverage from a given dataset.
2. Construct a Synctactic similarity matrix for rules for a target concept.
3. Construct a Dependency Matrix for rules for a target concept.
4. Construct Average and Difference matrices from a Dependence matrix.
5. Apply nonmetric MDS methods to Synctactic matrix.
6. Apply nonmetric MDS methods to Average and Difference matrices.

Average and Difference matrix is calculated as follows. Let $M_{dep}(D)$, $M_{avg}(D)$ and $M_{diff}(D)$ denote a dependency, average, difference matrix for a target concept D, respectively. From the definition of average and difference measure shown in 2.4, $M_{avg}(D)$ and $M_{diff}(D)$ are obtained as:

$$M_{avg}(D) = \frac{1}{2}(M_{dep}(D) + M_{dep}^{T}(D)), \tag{1}$$

$$M_{diff}(D) = \frac{1}{2}|M_{dep}(D) - M_{dep}^{T}(D)|, \tag{2}$$

$$\tag{3}$$

where $M_{dep}^{T}(D)$ denotes the transposition of $M_{dep}(D)$.

After visualization, differences between synctactic and semantic similarities of rules obtained from a dataset will be observed in a two or three-dimensional plane. The patterns shown in the plane include some manifestions of chances or unexpected discovery with respect to the differences between the information about attributes and that about supporting sets. Thus, this framework can be viewed as a type of tool for chance discovery from if-then rules.

4 Experimental Results

We applied the combination of rule induction and nonmetric MDS to a medical dataset on differential diagnosis of headache, which has 52119 esamples, 45 classes and 73 attributes. In these experiments, rule induction based on rough sets [7] is applied, where δ_α and δ_κ were set to 0.75 and 0.5 for rule induction, respectively. For similarity measures, we adopt Kulczynski's similarity and calculate similarity measures from accuracy and coverage of rules obtained from data.

Due to the limitation of space, we focus on the most interesting visualized patterns for each datasets. First, Figure 1, 2 and 3 show the pattern of synctatic, average and difference similarity of rules for headache *m.c.h.*, respectively. The

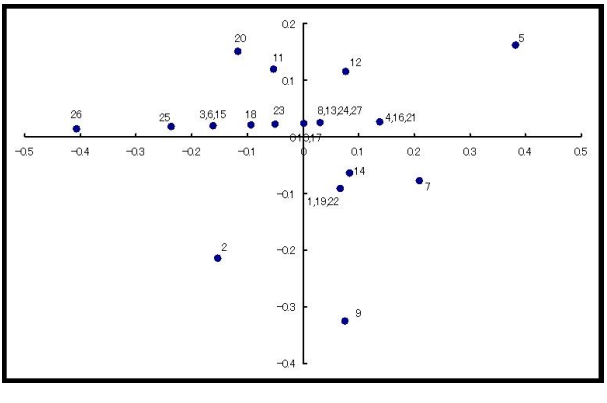


Fig. 1. Patterns of Similar Rules for *m.c.h.* (Synctactic Similarity: Headache)

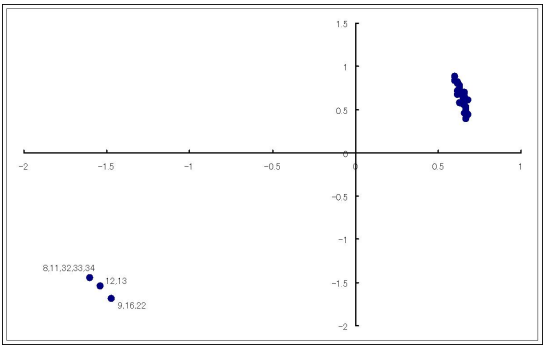


Fig. 2. Patterns of Similar Rules for *m.c.h.* (Semantic Similarity(Average): Headache)

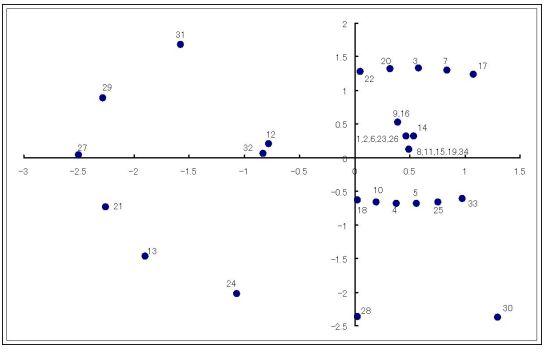


Fig. 3. Patterns of Similar Rules for *m.c.h.* (Semantic Similarity(Difference): Headache)

figures suggest two groups of the features of descriptors from average and three groups from different. While the descriptors in the right upper region in Figure 2 are those which are regularly used for the differential diagnosis of headache, the left lower group shows the descriptors used for special types of *m.c.h.*, which we call emotional-evoked *m.c.h.* and occupational *m.c.h.*. Thus, these patterns match with expert's knowledge.

However, Figure 3 gives a different pattern: two groups are scattered in different ways, which suggests that the left lower group in Figure 2 does not give overlapped intersections between rules, while the lower upper group has high overlapped region.

5 Conclusion

In this paper, we propose a visualization approach to show the similar relations between rules based on multidimensional scaling, which assign a two-dimensional

cartesian coordinate to each data point from the information about similiaries between this data and others data. As similarity for rules, we define three types of similarities: syntactic, semantic (covering based) and semantic (indice based). Syntactic similarity shows the difference in attribute-value pairs, semantic similarity gives the similarity of rules from the viewpoint of supporting sets. MDS assigns each rule into the point of two-dimensional plane with distance information, which is useful to capture the intuitive dissimilarities between rules. Since the indices for these measures may not hold the property of distance (transitivity), we adopt nonmetric MDS, which is based on the stress function.

Finally, we evaluated this method on a medical data set, whose experimental results show that several knowledge useful for domain experts could be found. This study is a preliminary study on the visualization of rules' similarity based on MDS. Further analysis of this method, such as studies on computational complexity, scalability will be made and reported in the near future.

Acknowledgement

This work was supported by the Grant-in-Aid for Scientific Research (13131208) on Priority Areas (No. 759) "Implementation of Active Mining in the Era of Information Flood" by the Ministry of Education, Science, Culture, Sports, Science and Technology of Japan.

References

1. Tsumoto, S., Ziarko, W.: The application of rough sets-based data mining technique to differential diagnosis of meningoenchepahlitis. In: Michalewicz, M., Raś, Z.W. (eds.) ISMIS 1996. LNCS, vol. 1079, pp. 438–447. Springer, Heidelberg (1996)
2. Adams, R., Victor, M.: Principles of Neurology, 5th edn. McGraw-Hill, New York (1993)
3. Everitt, B.S.: Cluster Analysis, 3rd edn. John Wiley & Son, London (1996)
4. Yao, Y., Zhong, N.: An analysis of quantitative measures associated with rules. In: Zhong, N., Zhou, L. (eds.) Methodologies for Knowledge Discovery and Data Mining. LNCS (LNAI), vol. 1574, pp. 479–488. Springer, Heidelberg (1999)
5. Eckart, C., Young, G.: Approximation of one matrix by another of lower rank. Psychometrika 1, 211–218 (1936)
6. Cox, T., Cox, M.: Multidimensional Scaling, 2nd edn. Chapman & Hall/CRC, Boca Raton (2000)
7. Tsumoto, S.: Automated induction of medical expert system rules from clinical databases based on rough set theory. Information Sciences 112, 67–84 (1998)

A Unified Approach to Web Usage Mining Based on Frequent Sequence Mining

Nobuhiro Inuzuka and Jun-ichi Hayakawa

Nagoya Institute of Technology,
Gokiso-cho Showa, Nagoya 466-8555, Japan
inuzuka@nitech.ac.jp, yoiko@phaser.eclom.nitech.ac.jp

Abstract. This paper gives a method to treat Web usage mining as request sequence mining. It provides a unified representation of three different types of source for Web usage mining. The source include an access log recorded in a Web server, the hypertext documents, and the hype-text structure or the graph of linkage. All of these information are represented as sequences. An efficient sequence mining *PrefixSpan* is inspected to mine with the unified information and an extension of *PrefixSpan* is proposed for our purpose. A preliminary experiment is also explained.

1 Introduction

Datamining gathers increasing interest in many areas. Its application to the World Wide Web becomes important because WWW has a large influence in our everyday life and a large part of industry is becoming to rely on it.

Mining on WWW is classified into three areas, (1)Web usage mining, (2)Web structure mining and (3) Web content mining. Web usage mining tries to find user's behaviour in Web accesses and is mostly approached from access log analysis. Web structure mining finds linkage structure of the complex Web structure, which often means the hypertext structure. The linkage analysis may find some community on Web. Web content mining aims automatic extraction of desirable information from Web. It studies to use the Web as a big dictionary or databases. It is mainly approached from the text analysis of the HTML documents through WWW.

This paper studies a method for Web usage mining. It provides a tool to designers of Web sites finding characteristic behaviour of users[7,8,9]. It mainly uses an access log of a Web server, the log which is the most important source of user's behaviour. A click stream analysis is a representative method of server-log analysis. It gives a path of users' access to reach some target pages, which Web-designers intend to show to users.

When we think from users' viewpoint it may be found that the access log analysis is not enough for Web-usage mining. User will explore WWW according to the contents which is provided as in the form of HTML documents. And the exploration is done by clicking the hyperlink in the hypertext space. Then we

B. Apolloni et al. (Eds.): KES 2007/ WIRN 2007, Part II, LNAI 4693, pp. 987–994, 2007.
© Springer-Verlag Berlin Heidelberg 2007

find that for the Web-usage mining we need not only the access log analysis but also the analysis of HTML documents and also the structure analysis or the hyperlink structure. Accordingly we need a method to analyze the three types of sources. This paper gives a unified formalism to combine the three information sources as a sequence database. Thereafter we will treat the mining problem as frequent sequence mining.

2 Frequent Sequence Mining

Sequence is a data type to treat a series of items. Some data, such as a corpus consisting of sentences of a language and DNA sequences, are treated as a set of sequences. Mining patterns for sequences also becomes important because of the large varieties of application areas[2,3].

Let I to be a set of items, of which sequences consist. We assume a linear order $\ll$ on I, which is used just for treating items in an algorithm. What we treat is a finite sequence $\langle e_1, \cdots, e_n \rangle$, where e_i's are subsets of I and are called elements of the sequence. For example $\langle e_1, e_2, e_3 \rangle = \langle \{a, c\}, \{b\}, \{b, d, e\} \rangle$ is a sequence which consists of its elements $e_1 = \{a, c\}$, $e_2 = \{b\}$ and $e_3 = \{b, d, e\}$. When e is an element of a sequence s, we write $e \in s$. When an item $i \in I$ is in an element in s, we write $i \lhd s$. The number of elements in a sequence s is denoted by $|s|$.

A sequence database is a set of sequences affixed with id numbers,

$$\{(1, s_1), (2, s_2), \ldots, (n, s_n)\}.$$

Sequence mining aims to find frequent patterns in a given sequence database. Patterns are given also by a sequence as a sub-sequence relation.

For a sequence $s = \langle e_1, \ldots, e_n \rangle$ and another sequence $\alpha = \langle \epsilon_1, \ldots, \epsilon_m \rangle$ (we use Greek letters to represent sequences for patterns), α is a subsequence of s and we write $\alpha \sqsubseteq s$ if there is an increasing sequence $j_1, \ldots, j_m$ $(1 \leq j_1 < \ldots < j_m \leq n)$, which satisfies $\epsilon_k \subseteq e_{j_k}$ for $k = 1, \ldots, m$. When $j_m = n$ α is an end-meet subsequence of s. For example $\langle \{a, c\}, \{f\} \rangle$ is a subsequence of $\langle \{a, b, c\}, \{d, e\}, \{f\}, \{h, i, j\} \rangle$ but is not an end-meet subsequence. $\langle \{a, c\}, \{f\}, \{h, i\} \rangle$ is an end-meet subsequence of it.

For the later sections we introduce concepts for sequences. At first we give two operations to extend a sequence pattern. For a sequence $\alpha = \langle e_1, \ldots, e_n \rangle$ and an item i, the representation α,i means the sequence $\langle e_1, \ldots, e_n, e_{n+1} \rangle$ where $e_{n+1} = \{i\}$. With this situation α_i means the sequence $\langle e_1, \ldots, e_{n-1}, e'_n \rangle$ where $e'_n = e_n \cup \{i\}$.

For sequence $s_1 = \langle e_1, \ldots, e_m \rangle$ and $s_2 = \langle \epsilon_1, \ldots, \epsilon_n \rangle$, and they satisfy,

$$i \ll i' \text{ for any } i \in e_m \text{ and } i' \in \epsilon_1$$

then $s_1 s_2$ means the sequence $\langle e_1, \ldots, e_{m-1}, e_m \cup \epsilon_1, \epsilon_2, \ldots, \epsilon_n \rangle$. When $s = s_1 s_2$ then s_1 is a prefix of s and s_2 is a postfix of s. When $s = s_1 s_2$ and α is an end-meet subsequence of s_1, s_2 is called a projection of s by α.

For example, when $s = \langle\{a,b,c\},\{d,e\},\{f\},\{h,i,j\}\rangle$, $s_1 = \langle\{a,b,c\},\{d\}\rangle$ and $s_2 = \langle\{e\},\{f\}\{h,i,j\}\rangle$, $s = s_1 s_2$ then s_1 is a prefix of s. We think of the alphabetic order as $\ll$ on items. For $\alpha = \langle\{b\},\{d\}\rangle$, it is an end-meet subsequence of s_1 and then s_2 is a projection of s by α.

For another example let think of $s = \langle\{a,b,c\},\{d,e\},\{f\},\{c,d,g\},\{h\}\rangle$ and $\alpha = \langle\{a\},\{d\}\rangle$. Then both of $\langle\{a,b,c\},\{d\}\rangle$ and $\langle\{a,b,c\},\{d,e\},\{f\},\{c,d\}\rangle$ are prefixs of s and α is end-meet subsequence of both of them. As a result s has two projection by α, $\langle\{e\},\{f\},\{c,d,g\},\{h\}\rangle$ and $\langle\{g\},\{h\}\rangle$.

For sequence DB D and a sequence pattern α the support of α in D, denoted by $support_D(\alpha)$, is the cardinality of $\{(i,s) \in D | \alpha \sqsubseteq s\}$.

For a given minimal support parameter $\min_{\sup}$, frequent sequence mining is to enumerate all of sequence pattern whose support are greater than $\min_{\sup}$. We call such sequence patterns frequent sequences.

2.1 PrefixSpan

PrefixSpan[4] is one of the most efficient algorithm for sequence mining. Which grows a frequent sequence simple to complex by adding items to the last of a sequence pattern which is already found frequent. The outline is in Table 1.

In the algorithm the idea of projection of a database is used. For a sequence database D and a sequence pattern α, the sub-database containing all sequences in D that containing α as their subsequence is called the projection of D by α and denoted by $D|_\alpha$. The algorithm in Table 1 obeys this notation although $D|_\alpha$ and $D|_{\alpha,i}$ can be seen simply as variables representing databases.

The most important principle in mining algorithms is the principle of Apriori, since it is adopted in Apriori algorithm[1]. It says that the more complex pattern the fewer it appears. *PrefixSpan* is also based on this principle: The frequency of a sequence pattern α is smaller than the frequency of a pattern which is a subsequence (or a prefix, which is used for a reason of efficiency) of α.

Although *PrefixSpan* algorithm is completely described in Table 1, as an implementation it has another description, where a projection of database is implemented by the following structure:

$$D|_\alpha = \{(id, s') \mid (id, s) \in D \text{ and } s' \text{ is a projection of } s \text{ by } \alpha\}$$

The support of α in D is the number of different id's in $D|_\alpha$. When using the structure above we can calculate $D|_{\alpha,i}$ and $D|_{\alpha_i}$ from $D|_\alpha$.

$$\begin{aligned} D|_{\alpha,i} = \{&(id, s') \mid (id, s) \in D|_\alpha \text{ for } s = \langle e_1, e_2, \ldots, e_n\rangle \\ &\text{and } s' \text{ is a projection of } \langle e_2, \ldots, e_n\rangle \text{ by } \langle\{i\}\rangle\} \\ D|_{\alpha_i} = \{&(id, s') \mid (id, s) \in D|_\alpha \text{ for } s = \langle e_1, e_2, \ldots, e_n\rangle \text{ and } i \in e_1 \\ &\text{and } s' = \langle e'_1, e_2, \ldots, e_n\rangle \text{ for } e'_1 = \{j \in e_1 | i \ll j\} \end{aligned}$$

3 Unifying Access Information in Sequences

We treat a log recorded in a Web server. It is a sequence of records each of which corresponds to a request from a Web client program, such as a browser, to the

Table 1. PrefixSpan algorithm

input: a sequence DB D and a minimum support threshold $\min_{\sup}$
output: all of frequent sequences

call *PrefixSpan*$(\langle\rangle, D)$;

procedure *PrefixSpan*$(\alpha, D|_\alpha)$:
for each $\beta = \alpha, i$ or $\beta = \alpha_i$ for $i \in I$ **do**
  if $support_{D|_\alpha}(\beta) \geq \min_{\sup}$ **then**
    write the sequence β and its support $support_{D|_\alpha}(\beta)$;
    $D|_\beta := \{(\mathrm{id}, s) \in D|_\alpha \mid \beta \sqsubseteq s\}$;
    call *PrefixSpan*$(\beta, D|_\beta)$;

server. Such a program requests with a url to the server and is responded with a data, such as an HTML document.

Let $R = (r_1, r_2, \ldots, r_n)$ to be a log, ie. a sequence of tuples of attribute values. Normally it includes attributes, such as *remote-host* which takes the name of client computer, *date* for a date and time of access by a client, *request* for a request url, and *referrer* which takes a url of a document which provides a link that caused this request. The *referrer* attribute takes a null value when the request is not happen from any link but by a directly given url or a bookmark. For an attribute a and a record r the value of a for r is denoted by $a(r)$. The set of all possible values of an attribute a is denoted by range_a.

If you see a tuple of attribute values as an element a log is already a sequence of elements and is in the form of sequence mining. However, it is not convenient for usage mining because the sequence includes whole information without any structure. The rest of this section gives a method to generate sequences, called access sequences, which represent user's behaviour naturally.

3.1 Generating Access Sequences

Access sequences can hold to unify information related on user's behaviour and factors which may affect user's decision. The outline of the procedure is as follows:

1. Generates a graph called a log-record graph, which represents a behaviour of a user.
2. Generates a graph called a url graph, which projects the user's behaviour to the hyperlink graph. Then adds access-related attributes to the log-record graph. The attributes connect user's behaviour and topological position of document accessed.
3. Adds document-related attributes to the graph.
4. Generates a set of access sequences.

Log-record graph. A log-record graph is a graph which represents behaviour of a user and is defined for a user. Unfortunately a user is not recorded in a server

log. We may roughly treat the *remote-host* attribute to give the information of users. Although this is not precise, we may assume it is reasonable because our target is not to detect a user but to understand user's behaviour.

For a log sequence $R = (r_1, \ldots, r_n)$ and a host u (which is intended to mean a user) we define a log sequence $R_u = (r_{i_1}, \ldots, r_{i_m})$, where $\{i_1, \ldots, i_m\} = \{i \in \{1, \ldots, n\} \mid \textit{remote-host}(r_i) = u\}$. Here we denote the set of records $\{r_{i_1}, \ldots, r_{i_m}\}$ also by R_u.

A log-record graph for a user u represents the relation of reference from a url to another. The reference relation is not linear because of reasons. First, the user may use the *back* button to step back to a document that includes a link to another document. Another possibility is using a frame structure to give a *menu*. It is also possible that a user may open a window or a tab for a document and leave open the referrer document. A menu and a document left open allow the user to step back the url of them and make the reference relation to be non-linear.

Then we define a graph called a log-record graph to represent a reference relation. A log-record graph $G_u^{\log} = (N_u^{\log}, E_u^{\log})$ is a directed graph defined by its nodes and edges. The set of nodes $N_u^{\log}$ is R_u and its set of edges $E_u^{\log} \subseteq R_u \times R_u$ is as follows.

$$\begin{aligned} E_u^{\log} = \{&(r_{i_j}, r_{i_k}) \in R_u \times R_u \mid \\ &\textit{referrer}(r_{i_k}) = \textit{request}(r_{i_j}) \ \wedge \ k > j \ \wedge \\ &\textit{request}(r_{i_l}) \neq \textit{request}(r_{i_j}) \text{ for any } l \in \{j+1, \ldots, k-1\}\} \end{aligned}$$

Url graph and access-related attributes. A log-record graph represents user behaviour from the aspect of reference relation, that is it represents paths of documents which a user passes through. When we would like to see topological positions of documents the following url graph is convenient.

For a log-record graph $G_u^{\log} = (N_u^{\log}, E_u^{\log})$ defined on a log-record sequence R_u for a user u, a url graph is a directed graph $G_u^{\text{url}} = (N_u^{\text{url}}, E_u^{\text{url}})$, where

$$\begin{aligned} N_u^{\text{url}} &= \{\textit{request}(r) \in v_{\textit{request}} \mid r \in R_u\} \text{ and} \\ E_u^{\text{url}} &= \{(\textit{request}(r), \textit{request}(r')) \in N_u^{\text{url}} \times N_u^{\text{url}} \mid (r, r') \in E_u^{\log}\}. \end{aligned}$$

A url graph is regarded as a projection of a log-record graph by *request*.

Url graph can also be considered as a part of hypertext graph at user's context. Based on this graph we may add access-related features to the log record. Here we give examples of attributes for $r \in R$, which are all Boolean features.

- $\textit{initial}(r) = \text{true} \Leftrightarrow$ there is no node $r' \in R$ s.t. $(r', r) \in E_u^{\log}$ for any u.
- $\textit{terminal}(r) = \text{true} \Leftrightarrow$ there is no node $r' \in R$ s.t. $(r, r') \in E_u^{\log}$ for any u.
- $\textit{branch}(r) = \text{true} \Leftrightarrow$ there are nodes r_1 and r_2 $(r_1, r_2 \in R)$ s.t. $(r, r_1), (r, r_2) \in E_u^{\text{url}}$ and $\textit{request}(r_1) \neq \textit{request}(r_2)$ for some u.

We call a node with a true initial attribute *an initial node* and a node a true terminal attribute *a terminal node*. We may construct other features, although our preliminary experiment does not use them.

- $choice(r) = \text{true} \Leftrightarrow$ there is a node r' s.t. $branch(r') = \text{true}$ and $(r', r) \in E_u^{\text{url}}$ for some u.
- $firstchoice(r) = \text{true} \Leftrightarrow choice(r) = \text{true}$ and there is no node r' s.t. (r'', r), $(r'', r') \in E_u^{\text{url}}$ and $date(r') < date(r)$ for any r'' and u.

Document-related attributes. Features related to document types, such as

- $image(r) = \text{true} \Leftrightarrow$ the source of $request(r)$ is an image,

and similarly $pdf(r)$ and $text(r)$ are added as attributes. Some document components included are also unified as attributes, such as,

- $has_table(r) = \text{true} \Leftrightarrow$ the source of $request(r)$ includes a `table` tag,

and also $has_image(r)$ and ones for other characteristic tags.

Access sequences. We have three different types of attributes, the original log attributes, access-related attributes and document-related attributes, on log-records. We denote the set of all these attributes by $\mathcal{A}$. With these attributes we would like to treat records as a sequence. For this purpose we propose the idea of access sequences.

For a log R an access sequence is a path $\langle r_1, \ldots, r_m \rangle$ of nodes or log-records in a log-record graph $G_u^{\log}$ of R from an initial node r_1 to a terminal node r_m. We treat attribute values of each record by their set as an element. An element is a subset of the set of all possible items or attribute values, the set which we denote by $I_{\mathcal{A}}$. Here we combine all values for all of attributes in $\mathcal{A}$

$$\begin{aligned} I_{\mathcal{A}} = \text{range}_{request} \oplus \text{range}_{date} \oplus \text{range}_{referrer} \oplus \ldots \\ \oplus \text{range}_{initial} \oplus \text{range}_{terminal} \oplus \ldots \\ \oplus \text{range}_{image} \oplus \ldots \oplus \text{range}_{has_image} \oplus \ldots, \end{aligned}$$

where $\oplus$ means the disjoint union. It distinguishes elements from different operands, for example, true for *initial* and true for *image* are different.

Accordingly we have a sequence $\langle e_1, \ldots, e_m \rangle$ from the access sequence $\langle r_1, \ldots$ $\ldots, r_m \rangle$, where $e_i = \{a(r_i) \mid a \in \mathcal{A}\} \subseteq I_{\mathcal{A}}$ for $i = 1, \ldots, m$. This sequence is also called an access sequence. We collect all of access sequences from a log. This is the proposing unified representation for Web usage mining.

4 Extension of *PrefixSpan* for Access Sequence Mining

Here we give a proposal of *PrefixSpan* algorithm in order to adopt it to access sequence mining. *PrefixSpan* is based on the subsequence relation, which is useful to find characteristic sequences. However, for the purpose of Web usage mining we should consider other patterns. An element in a sequence means an access. Then two elements appear in an access sequence in an order and the elements appear successively in the order is very different. The proposal is to give patterns for this distinction.

A pattern is a sequence of alternation of elements and connection symbols:

$$\langle \epsilon_1 \sigma_1 \epsilon_2 \sigma_2 \dots \sigma_{n-1} \epsilon_n \rangle,$$

where $\epsilon_1, \dots, \epsilon_n$ are elements and $\sigma_1, \dots, \sigma_n$ are connection symbols which we explain the following paragraph. When we restrict the connection symbols to ',' it represents the conventional sequence patterns.

In order to give a semantics for patterns, we assign relations to connection symbols. For a connection symbols σ we assume that there is a binary relation $r(\sigma) \subseteq N \times N$ on the set of natural numbers N. Then, a pattern $\alpha = \langle \epsilon_1 \sigma_1 \dots \sigma_{m-1} \epsilon_m \rangle$ matches a sequence $s = \langle e_1, \dots, e_n \rangle$ and we write $\alpha \preceq s$ if there is a sequence $j_1, \dots, j_m$ which satisfies (1) $\epsilon_k \subseteq e_{j_k}$ for $k = 1, \dots, m$ and (2) $j_k \, r(\sigma_k) \, j_{k+1}$ for $k = 1, \dots, m-1$.

As for Web mining we propose two connection symbols ',' and '-', where we give binary relations for them, $r(,) = \{(i,j) \in N \times N \mid i < j\}$ and $r(\text{-}) = \{(i,j) \in N \times N \mid j := i+1\}$. Then when we use only ',' for connection symbols sequence patterns have the conventional role. Elements connected by '-' mean that elements have to match successive elements in sequences.

With the introduction of the connection symbols *PrefixSpan* algorithm is easily extended by giving the projection of sequence DB. We introduced the two extension operations α,i and α_i in Section 2. Here we introduce extension operations for each of connection symbols. For a sequence pattern $\alpha = \langle \epsilon_1 \sigma_1 \dots \sigma_m \epsilon_m \rangle$ and an item i and a connection symbols σ, the extension $\alpha \sigma i$ means the sequence pattern $\langle \epsilon_1 \sigma_1 \dots \sigma_m \epsilon_m \sigma \{i\} \rangle$. *PrefixSpan* algorithm can also be extended with this extension. It only relies on the calculation of projection DB by the extended sequence patterns. The following is for the projection DB for α-i.

$$D|_{\alpha\text{-}i} = \{(id, s') \mid (id, s) \in D|_\alpha \text{ for } s = \langle e_1, e_2, e_3, \dots, e_n \rangle \\ \text{and } s' = \langle e_2', e_3, \dots, e_n \rangle \text{ for } e_2' = \{j \in e_2 | i \ll j\}$$

We may imagine further extension of connection symbols. While - means that the elements lie successively in sequence, we may give a connection symbols meaning that an element lies after two elements, and it is easily processed by the idea of projection.

5 A Preliminary Experiment

The generation method of access sequence and the extended version of *PrefixSpan* are implemented. With the implementation a preliminary experiment was done with an access log data, which is acquired from the KDD portal site, KDnuggets (`http://www.kdnuggets.com`). The site provides many resources for KDD study and also an access log of the server of KDnuggets itself.

We used the log on 16th November 2005. It includes 8,033,005 bytes and 39,974 records (each record corresponds a request). It includes for 900 users (although it does not accurately correspond 900 users). The generation method transforms this log to a set of access sequences. It included 18,897 access sequences and the

total number of elements was 47,067. The total number of items or attribute values was 241,601. The extended version of *PrefixSpan* extracted all of frequent sequences from this data in 10 minutes with $\min_{\sup}$of 1%. This is execution time is enough for practical use.

6 Conclusion

This paper proposed the unified representation of three types of information for Web usage mining as access sequences. The access sequences can be extended to consider other attributes which we mentioned in Section 3. The access sequences are treated by the sequence miner *PrefixSpan*, although it is only a choice and we should consider other algorithms and frameworks. The preliminary experiment showed that the access sequence generated and the extended version of *PrefixSpan* works in a practical time even with a smaller $\min_{\sup}$. The resulted sequences have not been analysed yet and we need to give an example to use this method for a scenario of Web design in a future study.

References

1. Agrawal, R., Srikant, R.: Fast Algorithms for Mining Association Rules. In: Proc. 20th Int'l Conf. Very Large Data Bases, pp. 487–499 (1994)
2. Agrawal, R., Srikant, R.: Mining Sequential Patterns. In: Proc. 11th Int'l Conf. Data Engineering, pp. 3–14 (1995)
3. Han, J., et al.: FreeSpan:Frequent Pattern-Projected Sequential Pattern Mining. In: Proc.6th ACM SIGKDD Int'l Conf. Knowledge Discovery and Data Mining, pp. 355–359. ACM Press, New York (2000)
4. Han, J., et al.: Mining Sequential Patterns by Pattern-Growth: The PrefixSpan Approach. IEEE Trans. Knowledge and Data Engineering 16(11), 1424–1440 (2004)
5. Tzvetkov, P.D., Yan, X.D., Han, J.D.: TSP: Mining Top-K Closed Sequential Patterns. In: Proc. 3rd IEEE Int'l Conf. on Data Mining, pp. 347–354 (2003)
6. Srikant, R., Agrawal, R.: Mining Sequential Patterns: Generalizations and Performance Improvements. In: Proc. 5th Int'l Conf. Extending Database Technology, pp. 3–17 (1996)
7. Chen, M., Park, J., Yu, P.: Data mining for path traversal patterns in a Web envirionment. In: Proc. 16th Int'l Conference on Distributed Computing Systems, pp. 385–392 (1996)
8. Srivastava, J., Cooley, R., Deshpande, M., Tan, P.: Web usage mining: Discovery and applications of usage pattern from web data. SIGKDD Explorations 1(2), 12–23 (2000)
9. Yan, T.W., Jacobsen, M., Garcia-Molina, H., Dayal, U.: From User Access Patterns to Dynamic Hypertext Linking. In: Proc. 5th Int'l World Wide Web Conference, pp. 1007–1014 (1996)

DACS Web Service

Kazuya Odagiri[1,3], Rihito Yaegashi[2], Masaharu Tadauchi[3], and Naohiro Ishii[1]

[1] Aichi Institute of Technology, 1-38-1 Higashiyamadouri Chikusa-ku
Nagoya-city Aichi, Japan
odagiri@toyota-ti.ac.jp,
ishii@aitech.ac.jp
[2] Shibaura Institute of Technology, 3-7-5 Toyosu Koutou-ku
Tokyo, Japan
rihito@sic.shibaura-it.ac.jp
[3] Toyota Technological Institute, 2-12-1 Hisakata Tenpaku-ku
Nagoya-city Aichi, Japan
tadauchi @toyota-ti.ac.jp

Abstract. As a communication mean, Personal Portal which is an entrance where each user can acquire his/her interested information on the network, is often used. We have problems on the existing Personal Portal described as follows. First, the program to realize the Personal Portal which is located on a specific Web Server, becomes large-scaled and complicated. Because Web pages as Personal Portal depend on that program, those Web pages can't always be useful for each user equally. Therefore, to solve these problems, we developed two kinds of new Web Service which are realized on the network introducing DACS(Destination Addressing Control System) Scheme, which is the network management scheme by communication control of client computers. By use of these Web Services, each user can create the customized Personal Portal, which displays the information by a user unit, for using in actual network easily and freely for oneself. In this paper, to deal with information not only by a user unit, but also by a group unit and by all users unit, DACS Web Service is proposed. DACS Web Service is realized by extending two kinds of functions of Web Service respectively and integrating these functions. As the result, each user can create the customized Web page as Personal Portal, which will be often useful in actual network.

Keywords: Personal Portal DACS Scheme Web Service CGI.

1 Introduction

Web pages are often used as communication means other than E-mails and telephones. As a communication mean, static Web pages are used. Static Web pages are often used as a communication mean for the unspecified number of users, and are unsuitable to communicate among respective individual users for respective purposes or interests. As another communication mean, Personal Portal realized by Web Service which can change the contents of Web page by a user unit dynamically, is

B. Apolloni et al. (Eds.): KES 2007/ WIRN 2007, Part II, LNAI 4693, pp. 995–1004, 2007.
© Springer-Verlag Berlin Heidelberg 2007

used. The word of ”Web Service” used in this paper, is different from ”Web Services” which is defined in W3C [1], and means the network service which is provided to users through Web Server. Moreover, though the word of “Portal” often indicates Web page for information searching [2] [3] such as Google and Yahoo, the word of “Personal Portal” used in this paper is different from “Portal”. “Personal Portal” indicates the entrance where each user can acquire his/her interested information on the network, and can display the different information for respective individual users on Web Browser dynamically by using the program such as CGI [4]. Therefore, Personal Portal is suitable to communicate among respective individual users for respective purposes or interests. To display the necessary information for individual user on Web Browser, that information is searched and extracted from databases on the network, and notified to each user by the program such as CGI. When each database is distributed on the network, the program to pick up the information from each database, will be large-scaled and complicated. Because the program is introduced by a system administrator and can't be changed by a user freely, Web page as Personal Portal is not always useful for each user. To solve this problem, new form of Personal Portal, which each user can create and customize freely and easily, is needed. To realize Personal Portal such as this, DACS Web Service, is proposed and examined. DACS Web Service is realized by extending two kinds of functions of Web Service, which is realized on the network introducing DACS (Destination Addressing Control System) Scheme. By using this DACS Web Service, each user can create and customize Personal Portal for oneself freely and easily regardless of information storage form such as database and document medium. For reference's sake, DACS Scheme is one of network management methods. The basic principle of DACS Scheme is that, through communication control of a client (a client computer) by a user unit, a whole network system is managed [5]. Moreover, for using in a real network, functional extension is done in the form of letting communication control by a user unit coexist with communication control by a client computer unit [6][7]. Then, two kinds of functions of Web Service, which is realized on the network introducing DACS Scheme, is described as follows. The first function of Web Service is that, data which is stored in database dispersed on the network, can be used efficiently [8]. The second function of Web Service is that, data which is stored in document medium as PDF file and simple text file, can be used efficiently. By using both functions, when each different user inputs same URL on Web Browser, the different information for each user is searched and extracted from database or document medium, and is displayed on Web Browser. By implementing various kinds of URL into a static HTML file, each user can create and customize Personal Portal freely and easily. That is, by letting these two kinds of functions of Web Service coexist, the service which a user can use information on the network regardless of information storage form is realized. However, in actual network, function of sending and accepting information not only to a specific user but also to grouped users and all users, is also necessary. By extending two kinds of functions of Web Service respectively, DACS Web Service with those functions is realized. DACS Web Service has the characteristic of using data stored in database or document medium with integrated interface for a user, and of sending and accepting information by a user unit, by a group unit and by all users unit.

2 Issue on Mechanism of Existing Personal Portal

So far, various researches have been performed on Personal Portal [9][10]. In addition, it is possible to realize Personal Portal by a commercial product. In both cases, basic mechanism, which is processed in the sequence from (1) to (9), is shown in Fig.1.

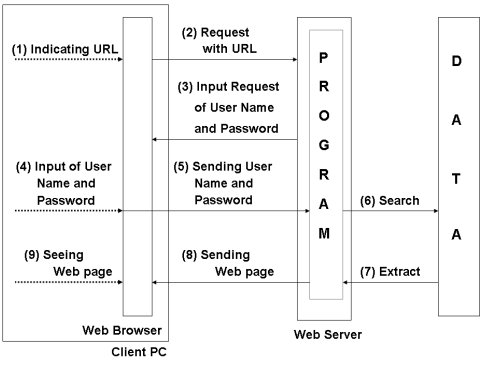


Fig. 1. Mechanism of Existing Personal Portal

A characteristic of this mechanism is the next two processing on Web Server.

(a) User authentication is performed by user information.
(b) Information related to user is searched and extracted from data which is accumulated beforehand.

Processing of (a) is performed after processing of (5) in Fig.1. This processing is necessary to perform processing of (b) and becomes essential so that Web Service premises anonymous user.

Web page as Personal Portal is generated by the program such as CGI on the Web Server. Because the program is introduced by system administrator and can't usually be changed by a user, the Personal Portal can't always be easy to use for each user or customize for personal use. In this paper, to overcome this problem, two kinds of functions of Web Service based on DACS Scheme, which make each user creates Personal Portal freely and easily, were developed. By using these functions, when each different user inputs same URL on Web Browser, the different information for each user is searched and extracted from database or document medium, and displayed on Web Browser. However, in these functions, it is possible only to send and accept information by a user unit. Because it is necessary to send and accept information by a group unit and by all users unit, these functions are insufficient. In addition, these two kinds of functions are independent with each other. So, to use in actual network, these two kinds of functions need to be integrated as one service, and integrated interface needs to be brought to each user. Therefore, after extending these two kinds of functions of Web Service to send and accept information not only by a user unit, but also by a group unit and by all users unit, DACS Web Service, is

proposed, which is realized as the result of having integrated these extended two kinds of functions of Web Service.

To explain DACS Web Service, in chapter 3, synopsis of DACS Scheme and two kinds of functions based on that DACS Scheme, system configuration by these functions to realize the customized Personal Portal, are shown. In chapter4, DACS Web Service is proposed. Especially, the part which is extended from conventional two kinds of functions of Web Service, are explained mainly.

3 DACS Scheme and Customized Personal Portal by User Unit

First, synopsis of DACS Scheme is explained. Fig.2 shows the basic principle of the network services by DACS Scheme. At the timing of the (a) or (b) as shown in the following, DACS rules (rules defined by user unit) are distributed from DACS Server to DACS Client.

(a) At the time of user's logging in the client
(b) At the time of a delivery indication from the system administrator

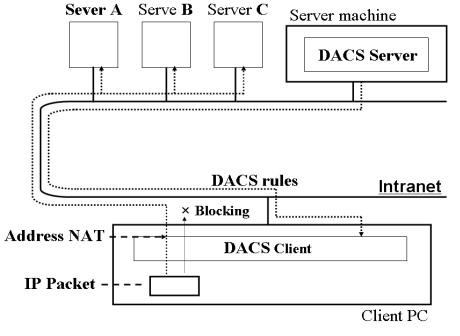


Fig. 2. Basic Principle of DACS Scheme

According to distributed DACS rules, DACS Client performs (1) or (2) operation as shown in the following. Then, communication control of the client is performed for every login user.

(1) Destination information on IP Packet, which is sent from application program, is changed.
(2) Packet from the client, which is sent from the application program to the outside of the client, is blocked.

An example of the case (1) is shown in Fig.2. In Fig.2, the system administrator can distribute a communication of the login user to the specified server among servers A, B or C. Moreover, the case (2) is added. When the system administrator wants to forbid user to use MUA (Mail User Agent), it will be performed by blocking IP Packet with the specific destination information.

Two kinds of functions of Web Services based on DACS Scheme are developed, here. At First, the function to use data from database is developed. To realize this

function, DACS Scheme needs to be extended, and the program on Web Server needs to be implemented in correspondence to the extended DACS Scheme as shown in Fig.3. In existing DACS Scheme, after a user's logging in a client (a), user name and IP address are sent to DACS SV (b). Then, DACS rules is sent back to DACS CL (c). In the extend DACS Scheme, moreover, user name and IP address are sent to the program on Web Server. A characteristic of extended DACS Scheme is that, the server side program on Web Server identifies the user by checking the login information and the source IP address from the client, and the processing of the program is different for each user. When each different user accesses the program with same URL, different information for each user can be searched and extracted from database, and be displayed on Web Browser. On extend DACS Scheme such as this, the processing from (1) to (7) is performed.

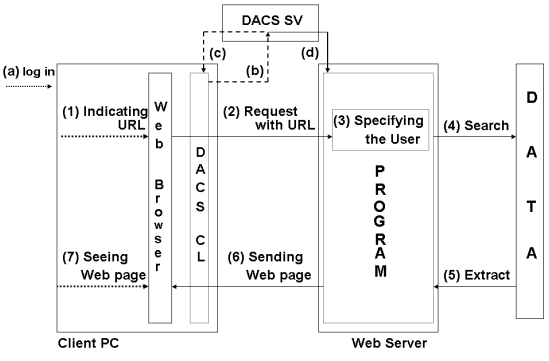


Fig. 3. Function Using Data from Database

Next, function to use data from document medium is developed for the respective user. In the network with DACS scheme, different IP address and TCP port can be assigned for one host name by a user unit. Therefore, different document medium with same file name on different Web Server can be referred for each user by inputting same URL to Web Browser. If this principle is combined with the function of virtual host which is equipped as Web Server, it is possible to use Web Server as shown in Fig.4. By the function of virtual host, multiple groups of socket (IP address and TCP port) can be assigned for one Web Server. The referred document can be changed every socket. First, in Document root of Web Server in Fig.4, directories (Dir A,B,C,D) are prepared for each user. By the function of virtual host, each directory is connected to each socket as one pare. By changing TCP port number (3000,3001,3002) for one IP address (192.168.1.1), sockets corresponding to each directory are prepared. Next, movement on this mechanism is described. One user inputs one URL to Web Browser. If the URL is inputted by User A, the file in Dir A which is connected to the socket (192.168.1.1:3000) is referred. Equally, if by User B, the file in Dir B which is connected to the socket (192.168.1.1:3001) is referred. If by User C, the file in Dir C which is connected to the socket (192.168.1.1:3002) is referred. If the document medium with same name exists in each directory (Dir A,B,C), each user can see different contents by inputting same URL to Web

Browser. At the point of user's view, the same function of Web Service on extend DACS Scheme is realized. For information sender, because it is possible to notify information to the specific user by uploading document medium to the predetermined directory, information usage becomes largely wide. Because information sender can describe the content of document medium easily and freely, it is possible to communicate the information with much expressive power and impact.

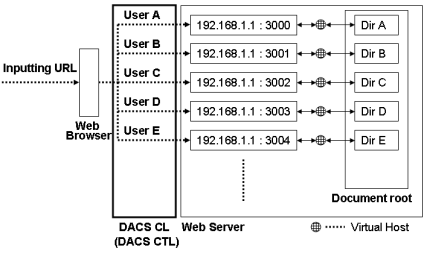


Fig. 4. Function Using Data from Document Medium

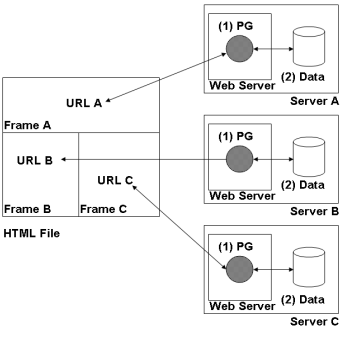


Fig. 5. System Configuration of Type 1

By using the function to use data from database, the system configuration of Type 1 is enabled as shown in Fig.5. First, in Fig.5, Server A C which has programs based on extended DACS Scheme (1) and data (2), respectively, are distributed on the network. In extended DACS Scheme, information related to user is displayed on Web Browser by only inputting the URL. The following mechanism is built based on this principle. One window of Web Browser is divided into some frames. For example, it is divided into three frames (Frame A-C). The static HTML file with each URL (URLA C) in each frame is created for displaying Web page as Personal Portal. The static HTML File is put on Web Server or on the client. When the static HTML file is opened through Web Browser, information extracted from each server is distributed on Web Browser. In extended DACS Scheme, URL corresponding to each server is only incorporated in the static HTML file. Thus, if the static Web page is created, user can create the customized Web page as Personal Portal easily.

The system configuration of Type 2 as shown in Fig.6 can be realized by using the function to use data from document medium. Each URL (URL A,B) is prepared. URL and the kind of information which can be seen after inputting the URL to Web Browser are notified to users. For example, URL for acquiring an issue in one classroom and URL for communicating from office to each individual user is enumerated as a kind of URL. By inputting URL to Web Browser, the file with same name (File Name A or B) is referred. The file is stored in each directory for each user. Each user can see the file which is stored in the directory for oneself. In other words, each user can see the information to oneself. Therefore, the customized Personal Portal with static HTML File same as Type 1 can be created.

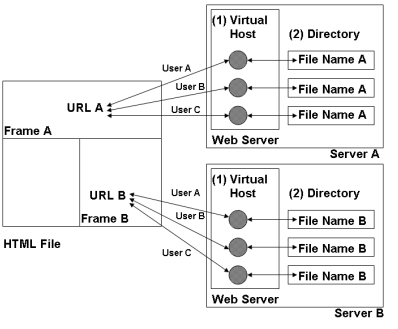


Fig. 6. System Configuration of Type 2

4 DACS Web Service

In chapter 3, synopsis of DACS Scheme was explained. In addition, two kinds of functions of Web Service which could be realized in DACS Scheme or extended DACS Scheme, and system configuration by these two kinds of functions to realize the customized Personal Portal, was explained. Because these functions of Web Service are used to send and accept information by a user unit, it is not enough to create the customized Personal Portal for using in the group actually. As explained in chapter 2, three functions for sending and accepting information by a user unit, by a group unit and by all user unit, are needed. Among these three functions, the function to send and accept information by a group unit and by all users unit doesn't exist. Therefore, in this chapter, those two kinds of functions of Web Service are integrated after extending the function to cover this insufficient point. DACS Web Service, which is realized by the integrative function, is proposed to realize Personal Portal for using in actual network.

In Fig.7, synopsis of DACS Web Service is shown. The function to use data from database of information system such as a system managing results for a student, is shown as Function α. The function to use data from document medium such as a simple text file and PDF file, is shown as Function β. After a user's inputting URL into Web Browser, communication control by DACS CL (DACS CTL) is performed. As the result, Function α or Function β is used. Because the function of either is automatically selected every each URL according to DACS rules, a user can use data

from information system or document medium dispersing on the network without being conscious of which function is used. In other words, a user can use information regardless of storage form and storage place of data freely and easily, if a user knows URL and the kind of information acquired by that URL. Even if whichever of Function α or Function β is used, data is displayed on Web Browser after inputting URL. Three kinds of data, which are data sent by a user unit (a) and by a group unit (b) and by all users unit, are displayed.

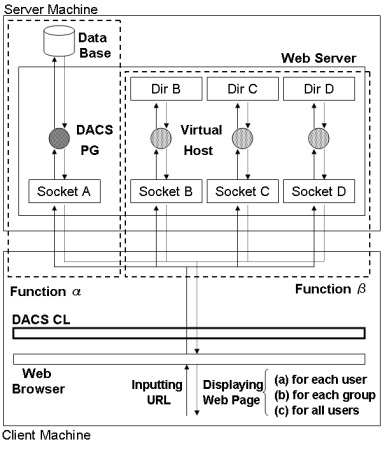


Fig. 7. DACS Web Service

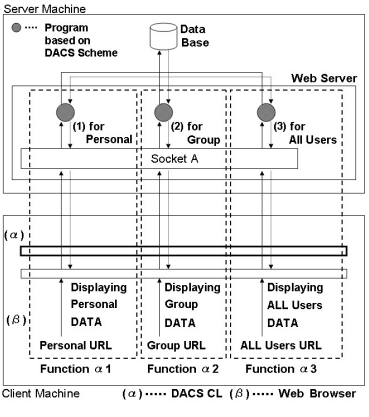


Fig. 8. Content of Function α

In Fig.8, the content of Function α is shown. Function α extends the function of using data form database by a user unit (Function α1 in Fig.8). After extension, data can be used by a group unit(Function α2 in Fig.8), and by all users unit (Function α3 in Fig.8). There are differences among Function α1, Function α2 and Function α3 in

the program extracting data of database for a request from Web Browser. In the program of Function α1, data is extracted by a user unit as shown by (1). In the program of Function α2, data is extracted by a group unit as shown by (2). In the program of Function α3, data is extracted by all users unit as shown by (3). In the existing function to use data from database as shown by (1), it is possible to specify which user sends communication through Web Browser. Therefore, the function is extended to set a correspondence list of user name and group name in DACS SV and send that correspondence list from DACS SV to the program of Function α2. As the result, because the program of Function α2 can know the group which a user belongs to , it is possible to extract information by a group unit. Even if a user belongs to multiple groups, it is possible to extract data of all groups. In addition, it is possible to extract data of a specific group by sending its group name as a parameter of URL. In the program of Function α3, data is extracted by all users unit. Because it is the function as normal Web Service not introducing DACS Scheme, it is realized without a technical problem normally.

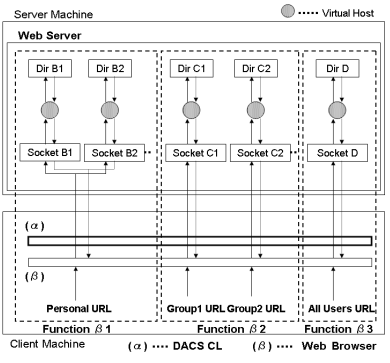


Fig. 9. Content of Function β

In Fig.9, the content of Function β is shown. In the Function β1, data of document medium is displayed dynamically by a user unit. The function by a group unit and the function by all users unit are realized by the same mechanism of Function β1. URL for each group (Group URL1,Group URL2) and URL for all users (All Users URL) correspond with each document medium, which is stored in each directory (Socket C1,Socket C2,Socket D), via each Virtual Host. To send information, only uploading a file as document medium into the predetermined directory (directory for each user, directory for each group, directory for all users) is needed. Information for each group can be accepted by use of URL for each Group. In addition, if a user's access is not permitted by DACS rules, it is not possible for the user not belonging to that group to access information by use of that URL. Information for all users can be accepted by use of URL for all users. By using DACS Web Service, not only information for each user but also information for each group and for all users, can be used from document medium.

5 Conclusion

In this paper, DACS Web Service was proposed to realize the customized Personal Portal for using in the group actually. DACS Web Service is realized by integrating two kinds of functions of Web Service after extending these functions respectively. It has the characteristic of using various kinds of data by a user unit, by a group unit, by all user unit, without distinguishing data stored in database from in document medium. By using DACS Web Service, data dispersing on the network in database or in document medium such as PDF file or text file, can be used as Personal Portal efficiently. From the view point of sending information in the group, it is possible to send information efficiently via DACS Web Service. The reason is why information sending can be realized through document medium. Conversely, from the view point of accepting and using information, data related to each user and to the group to, to all users, are acquired and used freely and easily.

References

1. http://www.w3.org/2002/ws/
2. Hartmann, J., Sure, Y.: An infrastructure for scalable, reliable semantic portals. IEEE Intelligent Systems 19(3), 58–65 (2004)
3. Lowe, N., Datta, A.: A New Technique for Rendering Complex Portals. IEEE Tran. on Visualization and Computer Graphics 11(1) (January-February 2005)
4. Robinson, D.: The WWW Common Gateway Interface Version 1.1. Internet Draft (1995)
5. Odagiri, K., Yaegashi, R., Tadauchi, M., Ishii, N.: Efficient Network Management System with DACS Scheme: Management with communication control. International Journal of Computer Science and Network Security 6(1), 30–36 (2006)
6. Odagiri, K., Yaegashi, R., Tadauchi, M., Ishii, N.: Efficient Network Management System with DACS Scheme. In: ICNS'06. Proc. of International Conference on Networking and Services, Silicon Valley, USA, July 16-19, 2006, IEEE Computer Society, Los Alamitos (2006)
7. Odagiri, K., Yaegashi, R., Tadauchi, M., Ishii, N.: New Network Management Scheme with Client's Communication Control. In: Gabrys, B., Howlett, R.J., Jain, L.C. (eds.) KES 2006. LNCS (LNAI), vol. 4252, pp. 379–386. Springer, Heidelberg (2006)
8. Odagiri, K., Yaegashi, R., Tadauchi, M., Ishii, N.: New Web Service Based on Extended DACS Scheme. International Journal of Computer Science and Network Security 6(3), 8–13 (2006)
9. Tushar, H.K.: Building enterprise portals: principles to practice. In: Proc. of International Conference on Software Engineering table of contents, pp. 623–633. ACM Press, New York (2002)
10. Hanakawa, N., Akazawa, Y., Mori, A., Maeda, T., Inoue, S., Tsutsui, S.: A Web-Based Integrated Education System for a Seamless Environment among Teachers, Students, and Administrators. Trans. of IEICE J88-D-I(2), 498–507 (2005)

Estimation of Initial Contour Based on Edge Background Subtraction for Self-affine Mapping System

Haruki Kawanaka[1], Hirofumi Kato[2], Fuminori Matsubara[2], Yuji Iwahori[2], and Robert J. Woodham[3]

[1] Faculty of Information Science and Technology, Aichi Prefectural University, Nagakute-cho, Aichi-gun 480-1198, Japan
kawanaka@ist.aichi-pu.ac.jp
[2] Faculty of Engineering, Chubu University, Matsumoto-cho 1200, Kasugai 487-8501, Japan
iwahori@cs.chubu.ac.jp
[3] Department of Computer Science and Technology, University of British Columbia, Vancouver, B.C. Canada V6T 1Z4
woodham@cs.ubc.ca

Abstract. Self-affine mapping system is a technique that can extract the contour on the basis of rough form. A self-similarity at the block unit in square is utilized. The initial contour of this technique is given as a simple form by hand, the result of background subtraction or frame subtraction and the selection from candidates of the plural outline form with degree of separation. However, there are some problems in case that the cost of manual operation is large, or the self-similarity does not hold when the object has the similar brightness as the background. So, in this paper, we propose a new method to extract the contour with self-affine mapping system from the rough form of the object obtained by using edge background subtraction and interpolation.

Keywords: Background subtraction, self-affine mapping system, self-similarity, object contour.

1 Introduction

An essential method, which automatically extracts the contour of the specified object from the image, is useful in the various fields such as image processing, searching and recognition.

As methods to extract the region by observing a contour, there are Interactive Graph Cuts[1] and Grab Cuts[2] which presume if a contour belongs to the object at each pixel. These techniques assumes that a probability density function of brightness is even for whole of the image for an object region and a background region respectively. Those functions are presumed by the histogram and mixture Gaussian distribution. If the value of probability density function at one pixel is high, the pixel is easily extracted as object region and become congruent with

B. Apolloni et al. (Eds.): KES 2007/ WIRN 2007, Part II, LNAI 4693, pp. 1005–1012, 2007.
© Springer-Verlag Berlin Heidelberg 2007

other pixels whose brightness resembles. In this way, the object form can be expressed in detail because it is distinguished at each pixel. However, some error occurs in case that the object has the same brightness to the background.

In addition, Active Contour Model[3][4][5] have been proposed. These methods can extract the object region from the rough outline form. An object contour is expressed by a curved line with parameters. That is, an energy minimization is applied to make the smooth curved line along some edge. However, it is difficult to express the detailed form due to the influence of smoothness constraint.

On the other hand, there is a Self-affine Mapping System (SMS)[6] as a technique of contour fitting. SMS can extract the actual contour on the basis of rough contour. To obtain a self-affine mapping of the object form, we find some part which is similar in each block unit of the initial contour. The self-mapping is repeated until it is converged to the actual contour form. It is necessary to give a rough contour of the object as an initial contour. Simple way is to set a simple form by hand. In this case, some cost of manual operation exists. As other methods to set the rough form, for example, we use a result of background subtraction or frame subtraction, and select from some candidates of plural outline form with degree of separation. However, if the brightness of an object has been similar to the background, large slips to the actual contour are occurred by SMS from the initial contour made by the background subtraction. It is much the same for the frame subtraction or the outline form selection with degree of separation. In these cases, it is not possible to presume a precise contour.

So, in this paper, we propose a new method to estimate the initial contour for SMS. The method introduces the region interpolation using edge background subtraction to estimate the contour. Based on this initial contour, the object region is automatically extracted by SMS.

2 Self-affine Mapping System

SMS is a contour fitting method with "self-similarity". The contour of any objects is able to closely resemble with combination in the straight line or the corner. Fig.1 shows the relationship of the self similarity around the contour between $\boldsymbol{C}_i (i = 1, \ldots, N)$ and $\boldsymbol{P}_i (i = 1, \ldots, N)$. Here, $\boldsymbol{C}_i$ and $\boldsymbol{P}_i$ represent the blocks of $N (N \geq 1)$ regions in the image. $\boldsymbol{C}_i$ is called child block and $\boldsymbol{P}_i$ is called parent block, where $\boldsymbol{X}$ represents the object region.

$\boldsymbol{g}(x) \in \boldsymbol{R}^1$ is called a shading image, and is given to all points $x \in \boldsymbol{G}$ of the n dimensional space $\boldsymbol{R}^n$ and its subspace $\boldsymbol{G} \subset \boldsymbol{R}^n$ as brightness value. $\boldsymbol{g}(x) = \{1, 0\}$ is called binary image. Let $\boldsymbol{G}$ be the screen. When the linear mapping $\{\boldsymbol{a}_i : \boldsymbol{P}_i \rightarrow \boldsymbol{C}_i\}$ exists for blocks $\boldsymbol{C}_i$ and $\boldsymbol{P}_i$ in a shading image and the mapping satisfies the next equation, $\boldsymbol{C}_i$ and $\boldsymbol{P}_i$ have a relationship of the similarity.

$$\forall \in \boldsymbol{P}_i, \boldsymbol{g}(x) = \boldsymbol{g}(\boldsymbol{a}_i(x)). \tag{1}$$

In this case, the image becomes a partially self similar and $\{\boldsymbol{P}_i, \boldsymbol{C}_i, \boldsymbol{a}_i\}$ is called the self similar model. A mapping to that itself $\{\boldsymbol{S} : \Omega \rightarrow \Omega\}$, which

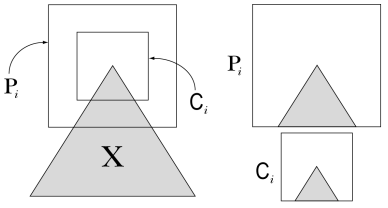


Fig. 1. Self similar model

is obtained by using self similar model $\{\boldsymbol{P}_i, \boldsymbol{C}_i, \boldsymbol{a}_i\}$, is called "self similar mapping". Ω is all subclasses of $\boldsymbol{G}$. The next equation shows a self-affine mapping (self similar mapping).

$$\boldsymbol{S}(\boldsymbol{X}) = \{\cup \boldsymbol{a}_i(\boldsymbol{P}_i \cap \boldsymbol{X})\} \cup \boldsymbol{\Gamma}. \tag{2}$$

To keep the state before the mapping of the part where $\boldsymbol{C}_i$ is not arranged, the next equation is applied.

$$\boldsymbol{\Gamma} = \boldsymbol{X}_0 \cap \{\cap(\boldsymbol{C}_i)^C\}. \tag{3}$$

Here, let $\boldsymbol{X}_0$ be early set of $\boldsymbol{X}$ and $(\boldsymbol{C}_i)^c$ be $\boldsymbol{C}_i$. Let mapping $\boldsymbol{a}_i$ be

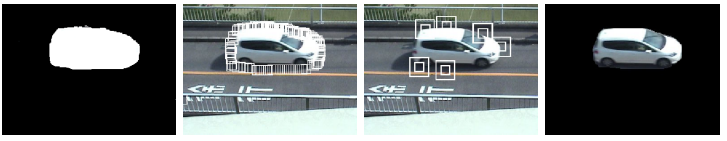
$$\boldsymbol{a}_i(x) = x/r + \boldsymbol{t}_i. \tag{4}$$

where $\boldsymbol{t}_i \in \boldsymbol{R}^n$ is a fixed vector and r is constant.

In SMS, child blocks are arranged around the initial contour, we search the similar part $\boldsymbol{P}_i$ in every block $\boldsymbol{C}_i$ that is shown in Fig.1 and calculate $\boldsymbol{a}_i$. We bring close the contour to the actual contour by using $\boldsymbol{S}(\boldsymbol{X})$ as the initial contour of the next iteration. The concrete algorithm is described as follows.

1. Child blocks $\boldsymbol{C}_i$ are arranged (Fig.2(b)) in order to cover around the rough contour of the object shown by a binary image (Fig.2(a)).
2. The parent blocks $\boldsymbol{P}_i$ are found so that self similar model $\{\boldsymbol{P}_i, \boldsymbol{C}_i, \boldsymbol{a}_i\}$. The each $\boldsymbol{P}_i$ is the similar relationship to each $\boldsymbol{C}_i$ by the block matching method (Fig.2(c)).
3. For the object region of the binary image, a self similar mapping is repeated with the specified number of times until it is converged (Fig.2(d)).

The bigger the block size of $\boldsymbol{C}_i$, the wider the width of the gap between the initial contour and the actual contour which we can adjust, but the lower the precision of the result contour. The smaller the block size of $\boldsymbol{C}_i$ conversely, the higher the accuracy of the result contour, but the narrower the width of the gap which we can adjust. Therefore, by changing the size of $\boldsymbol{C}_i$ from big ones gradually, precision can be increased.



(a) (b) (c) (d)

Fig. 2. Flow of self-affine mapping system. (a) Rough initial contour by hand, (b) Arrangement of child blocks $\boldsymbol{C}_i$, (c) Similar relationship of $\boldsymbol{C}_i$ and $\boldsymbol{P}_i$ (The number of blocks is decreased to make clearly similar relationship understandable of block), (d) Convergence result.

3 Generation of Initial Contour by Interpolation of Change Region with Edge Background Subtraction

The object region, where the brightness value of the background is close, cannot be extracted with only the background subtraction. When it is utilized as the initial contour of self-affine mapping system, it is not possible to closely resemble self-similarity. And at the initial value due to simple form, the contour extraction ends in the failure as shown in Fig.3.



(a) (b)

Fig. 3. Failure example. (a) Initial contour by simple form, (b) Result of SMS.

Some differences appear at the object region in the edge images of both the background and the present frame. And the edge of the present frame often includes the contour of the object. If this part is brought out, it is possible to extract the object region. So, we think the subtractions of the edges that are made from both the background image and the present frame image. Then, to extract the object region with higher accuracy, we use the result of edge background subtraction as the initial contour of SMS. The procedure to extract the initial contour on the basis of edge background subtraction is shown below.

1. Change regions are extracted by background subtraction.
2. Two edge images of both the background and the input frame are generated. And the change edges are extracted by edge subtraction.
3. The logical sum between change regions and change edges is calculated.
4. The blank spaces in the object region are filled.

The extraction process is shown in Fig.4. Laplacian filter is used to edge detection of both the background image and the input image.



(a) (b) (c) (d)

Fig. 4. Interpolation of the change regions based on edge background subtraction. (a) Input image, (b) Background subtraction, (c) Edge subtraction, (d) The logical sum and filling-in of blank spaces.

We use labeling processing to fill the closed region. By labeling for the image that is made as logical sum between background subtraction and edge subtraction, the areas of closed region (the number of pixels) are obtained. And some closed region, where the area is small, is filled as the object region.

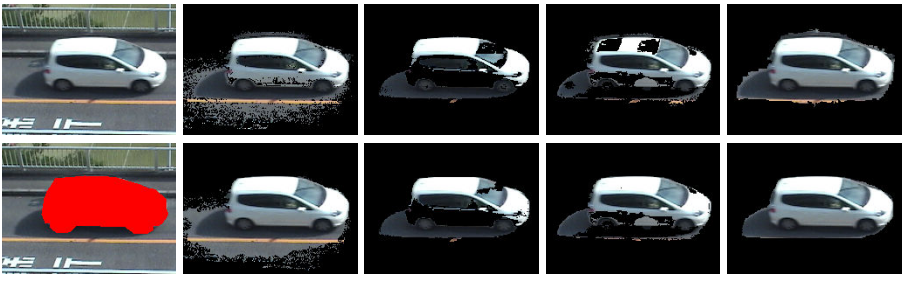
Even if brightness values of both the background and the object are similar, our method becomes the good result, while the background subtraction or the frame subtraction causes some problems. It can cover the object entirely without the big lacking and expanding in comparison with the actual object region by interpolation of the change region based on edge subtraction. We use the interpolation result based on edge subtraction as initial contour of SMS algorithm. As shown in Fig.4(b) and (c), many noises are remained around the contour. But by using this extraction result as the initial contour of SMS, the object region can be extracted with good precision.

4 Experiments

Experiments of object region extraction are done for images of a car, a hand and an origami by using the proposed method and other conditions. The input image and correct answer (object region) of the car image are shown in Fig.5(a).

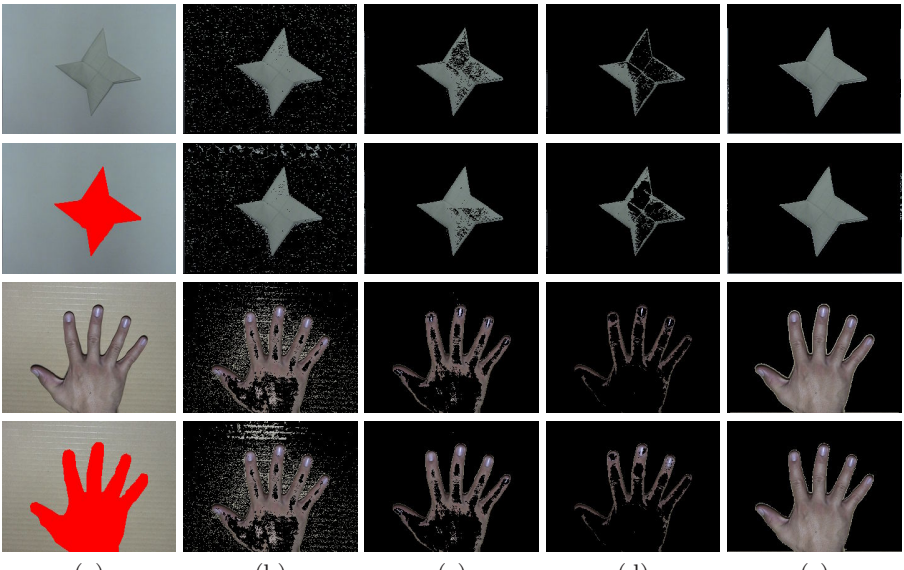
Each initial contour and the result due to SMS are shown in Fig.5(b)–(e). The upper row in Fig.5 shows the initial contour, while the lower row shows the result of SMS. It is shown that the initial contours by the background subtraction in Fig.5(b)(c), by the frame subtraction in Fig.5(d) and by proposed method in Fig.5(e). Child blocks kept making small gradually from large ones, and processing was repeated. The threshold value of the background subtraction for Fig.5(b) is set by the discriminant analysis[7]. This method can automatically determine a threshold with statistical approach of gray level histogram. The object region could be guaranteed, but peripheral noise has blended in large quantities. While the threshold for Fig.5(c) is high. It could not guarantee the object region, but peripheral noise has not blended.

As shown in Fig.5(c)(d), when the initial contour with the background subtraction or the frame subtraction is used, the large gaps and holes (blanks) have occurred in comparison with the actual object region. It can not converge to the actual contour when a gap between the initial contour and the actual one is large



(a) (b) (c) (d) (e)

Fig. 5. Experiment results (upper: initial contour, lower: extracted result by SMS). (a) Initial contour and correct answer of the object region. (b) Background subtraction (low threshold), (c) Background subtraction (high threshold), (d) Frame subtraction, (e) Proposed method.



(a) (b) (c) (d) (e)

Fig. 6. Experiment results (upper: initial contour, lower: extracted result by SMS). (a) Initial contour and correct answer of the object region, (b) Background subtraction (low threshold), (c) Background subtraction (middle threshold), (d) Background subtraction (high threshold), (e) Proposed method.

for the size of the child block. Therefore, it is not possible to closely resemble self-similarity of the object and to adjust the contour for good precision. Since lack and protrusion are little and the suitable initial contour of the object region is given with proposed method, it is possible to adjust the contour precision well as shown in Fig.5(e). Similarly as the experiment for the image of a car, only the proposed

method has accurately been able to extract for other images as shown Fig.6. It is also noted that the part of the shadow of the object is extracted because proposed method uses the result of edge background subtraction absolutely.

Next, numerical precision is calculated for the extraction result under each condition. For each method to set an initial contour (background subtraction, frame subtraction or interpolation based on edge subtraction), detection ratio by SMS is calculated. Detection efficiency is defined by the ratio that it could not be detected in spite of the object region (false negative error rate) and by the ratio that it is detected in spite of the background region (false positive error rate).

$$FalseNegativeErrorRate = \frac{LeakageArea}{ActualArea}. \tag{5}$$

$$FalsePositiveErrorRate = \frac{ExcessiveArea}{ActualArea}. \tag{6}$$

On the basis of these criteria, a mean value of the extraction result with each extraction method for any 50 images is shown in Table 1. The theoretical value of the object region to be extracted is made by hand. Proposed method has improved in comparison with background subtraction and frame subtraction at both false negative error rate and false positive error rate. We confirmed that our proposed method can work effectively for any kinds of image.

Table 1. Extraction rate. (a) Background subtraction (low threshold), (b) Background subtraction (high threshold), (c) Frame subtraction, (d) Proposed method

Method	(a)	(b)	(c)	(d)
False Negative Error Rate (%)	3.16	19.31	6.57	4.42
False Positive Error Rate (%)	37.37	4.58	18.47	6.06

Finally, to evaluate the effectiveness of proposed method, ROC (Receiver Operator Characteristics) curved line is shown in Fig.7. Fig.7 plots the relation of

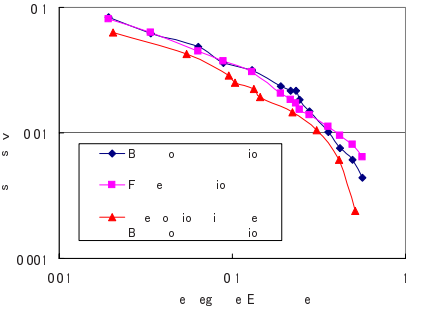


Fig. 7. ROC curved line of each method

the false negative error rate vs the false positive error rate with the log scale with changing the threshold value of edge background subtraction variously for an image. This graph shows the fact that the extraction efficiency becomes high under the condition where the curved line locates under the left. In Fig.7, the curved line of proposed method has been located under the left in comparison with the other extraction method. This means that our proposed method can extract the object region without influences of threshold value.

5 Conclusion

In this paper, we proposed a new method that interpolates of change regions based on edge subtraction for an initial contour of SMS. We ensured that it is possible to extract the object region with higher accuracy in comparison with other initial contour setting method. The evaluation was done through computer experiments. In addition, it was confirmed from ROC curved line that the proposed method has pliability against threshold value. Future works include a correspondence to the shadow areas of an object, or an interpolation method for non-object regions inside closed region at the first stage.

Acknowledgment. Kawanaka's research is supported by the Nitto Foundation. Iwahori's research is supported in part by Chubu University Grant, and the JSPS Grant No. 16500108. Woodham's research is supported by the Natural Science and Engineering Research Council (NSERC).

References

1. Boykov, Y.Y., Jolly, M.: Interactive graph cuts for optimal boundary & region segmentation of objects in N-D images. In: Proc. of ICCV, pp. 105–112 (2001)
2. Blake, A., Rother, C., Brown, M., Perez, P., Torr, P.: Interactive image segmentation using an adaptive GMMRF model. In: Pajdla, T., Matas, J.(G.) (eds.) ECCV 2004. LNCS, vol. 3024, pp. 428–441. Springer, Heidelberg (2004)
3. Kass, M., Witkin, A., Terzopoulos, D.: Snakes: active contour models. IJCV 1(2), 321–331 (1988)
4. Amini, A.A., Weymouth, T.E.: Using dynamic programming for solving variational problems in vision. IEEE Trans. PAMI 12(9), 855–867 (1990)
5. Paragios, N., Deriche, R.: Geodesic active contours and level sets for the detection and tracking of moving objects. IEEE Trans. PAMI 22(3), 266–280 (2000)
6. Ida, T., Sambonsugi, Y.: Self-affine mapping system and its application to object contour extraction. IEEE Trans. TIP 9(11), 1926–1936 (2000)
7. Otsu, N.: A threshold selection method from gray-level histograms. IEEE Trans. SMC 9(1), 62–66 (1979)

Improvement of Accuracy for Gaussian Curvature Using Modification Neural Network

Yuji Iwahori[1], Takashi Nakagawa[1], Shinji Fukui[2], Haruki Kawanaka[3], Robert J. Woodham[4], and Yoshinori Adachi[1]

[1] Faculty of Engineering, Chubu University
Matsumoto-cho 1200, Kasugai 487-8501, Japan
iwahori@cs.chubu.ac.jp, nakagawa@cvl.cs.chubu.ac.jp,
adachiy@isc.chubu.ac.jp
http://www.cvl.cs.chubu.ac.jp/
[2] Faculty of Education, Aichi University of Education
Hirosawa, Igaya-cho, Kariya 448-8542, Japan
sfukui@auecc.aichi-edu.ac.jp
[3] Faculty of Information Science and Technology, Aichi Prefectural University
Nagakute-cho, Aichi-gun 480-1198, Japan
kawanaka@ist.aichi-pu.ac.jp
[4] Department of Computer Science, University of British Columbia
Vancouver, B.C. Canada V6T 1Z4
woodham@cs.ubc.ca

Abstract. This paper proposes a new approach to recover the relative magnitude of Gaussian curvature of the test object from four shading images using modified neural network. The method is expanded to an object with color texture using four shading images taken under the different light source directions. Neural network mapps four image irradiances on the test object onto a point on a sphere. The area value surrounded by four mapped points onto a sphere gives an approximate value of Gaussian curvature. To get more accurate Gaussian curvature, the modification neural network is introduced and learned for the synthesized 2-D basis function consisting of 2-D cosine function. It is shown that learnt NN gives better accuracy for the relative magnitude of Gaussian curvature of the test object.

Keywords: Neural Network, Gaussian Curvature, Shape Recovery, Photometric Stereo.

1 Introduction

Surface gradient and curvature are the essential information for the shape representation. Especially, the surface curvature is the invariant and effective feature for the viewing direction, and curvature feature can be used to many applications such as the shape recovery, shape modeling, segmentation, the object recognition and pose determination in the field of computer vision.

Based on the physics-based vision approach, Woodham [1] developed a method to get surface curvature using the values of the surface gradients. Using the LUT

B. Apolloni et al. (Eds.): KES 2007/ WIRN 2007, Part II, LNAI 4693, pp. 1013–1020, 2007.
© Springer-Verlag Berlin Heidelberg 2007

(Look Up Table), the method obtains the local surface gradients by the empirical photometric stereo using a calibration sphere.

Iwahori has pursued neural network implementations of photometric stereo.

In [2] [3], neural network implementation with the PCA (principal component analysis) was proposed.

Angelopoulou and Wolff [4], Okatani and Deguchi [5] proposed a method to recover the local sign of the Gaussian curvature from three images taken under different light source directions without using the values of the surface gradient or the correct light source directions. These methods are applicable to the diffuse reflectance.

While Iwahori *et al.* [6] proposed the method to classify local surface from multiple shading images including the sign of Gaussian curvature using neural network. As with both previous non-parametric, empirical implementations [1,2] [3], no explicit assumptions need to be made either about light source directions or about the functional model of surface reflectance.

Iwahori *et al.* [7] further extended the approach not only to recover the sign but also the relative magnitude of local Gaussian curvature of the object. The key idea is that the area value surrounded by the four mapped points onto a sphere corresponds to the relative magnitude of Gaussian curvature.

This area value is an approximate value of the relative magnitude. This paper extends the approach with higher confidence of relative magnitude of Gaussian curvature. A new idea is developed by introducing the modification neural network learned for the basis function with different curvatures.

The modification neural network improves the accuracy of local Gaussian curvature for a test object. The advantage of the method is Gaussian curvature can be directly extracted from shading images and the proposed approach can be applied for non-Lambertian case with different albedo. The real phenomena of the surface reflectance is directly treated and the effect is evaluated through experiments.

2 Principle

2.1 Empirical Constraint

The original purpose of photometric stereo uses three light source to determine the surface normal vector (n_1, n_2, n_3) from the observed image irradiances (E_1, E_2, E_3) locally. While, this approach tries to get the local curvature information directly from (E_1, E_2, E_3, E_4) at the local four points on the test object.

Let the image irradiances at (x_{obj}, y_{obj}) on the test object be $(E_{1obj}, E_{2obj}, E_{3obj}, E_{4obj})$ and that at (x_{sph}, y_{sph}) on the sphere be $(E_{1sph}, E_{2sph}, E_{3sph}, E_{4sph})$. If the surface material is the same for both of a test object and a sphere, the following constraint

$$\begin{cases} E_{1obj}(x_{obj}, y_{obj}) = E_{1sph}(x_{sph}, y_{sph}) \\ E_{2obj}(x_{obj}, y_{obj}) = E_{2sph}(x_{sph}, y_{sph}) \\ E_{3obj}(x_{obj}, y_{obj}) = E_{3sph}(x_{sph}, y_{sph}) \\ E_{4obj}(x_{obj}, y_{obj}) = E_{4sph}(x_{sph}, y_{sph}) \end{cases} \tag{1}$$

is satisfied. This constraint for E_1 to E_4 leads to the condition that the corresponding surface normal vector should be the same between a test object and a sphere. This constraint is used to determine the relative magnitude of a test object using neural network.

2.2 RBF Networks and OLS Learning

Neural networks are attractive for non-parametric functional approximation. As an algorithm to learn the mapping of the nonlinear functions using neural network, back propagation algorithm is popular and it can often used for this kind of problem. As another candidate and choice except the back propagation, a radial basis function (RBF) neural network [8] is one choice suitable for many applications. In particular, it has been widely used for strict interpolation in multidimensional spaces. In the case that we have many sampled points, RBF neural network has an advantage that the learning is much faster and efficiently done than the standard feed forward neural network with the back propagation algorithm. This is based on the fact that he weights of network can be partially determined for each of outputs and nonlinear mapping problem can be effectively solved using the orthogonal least squared algorithm [8] proposed by Chen *et al.*

3 Improvement of Accuracy for Relative Magnitude of Gaussian Curvature

3.1 Case of Non-uniform Albedo

Let j be the j-th light source, the image irradiance E_j is observed under the illumination from each light source j. For four light source photometric stereo under the orthographic projection, the following image irradiance equations hold.

$$\begin{cases} E_1(x,y) = \rho(x,y)R_1(n_x,n_y,n_z) \\ E_2(x,y) = \rho(x,y)R_2(n_x,n_y,n_z) \\ E_3(x,y) = \rho(x,y)R_3(n_x,n_y,n_z) \\ E_4(x,y) = \rho(x,y)R_4(n_x,n_y,n_z) \end{cases} \tag{2}$$

where R is the reflectance map (image irradiance map) at the surface normal vector (n_x, n_y, n_z). ρ represents the albedo (i.e., reflectance factor) at the point (x,y). Under this condition,

$$E_1 + E_2 + E_3 + E_4 = \rho(R_1 + R_2 + R_3 + R_4) \tag{3}$$

holds. The effect of ρ is eliminated as follows.

$$e'_j = \frac{E_j}{E_1 + E_2 + E_3 + E_4} \quad (for\ j = 1, 2, 3, 4) \tag{4}$$

The objective now is to determine the mapping of (e'_1, e'_2, e'_3, e'_4) to the corresponding coordinate (x, y) on a sphere.

3.2 Mapping onto Sphere by Neural Network

A RBF neural network (NN) is used to do the mapping of the observed image irradiances $(e'_1(x_{sph}, y_{sph}),\ e'_2(x_{sph}, y_{sph}),\ e'_3(x_{sph}, y_{sph}),\ e'_4(x_{sph}, y_{sph}))$ to the corresponding coordinate (x_{sph}, y_{sph}) on a sphere.

With this learning procedure, NN is trained using input/output data from a sphere. Many training vectors are available since data from the calibration sphere are dense and include all possible visible surface normal.

The learning procedure builds an RBF neural network one neuron at a time. Neurons are added to the network until the sum-squared error falls beneath an error goal or a maximum number of neurons has been used. The resulting network generalizes in that it predicts a position (x_{sph}, y_{sph}) on the sphere, given any input values, $[e'_1, e'_2, e'_3, e'_4]$ of the test object.

3.3 Relative Magnitude of Gaussian Curvature

The kind of local surface curvature can be classified from multiple shading images [6]. However, even if the local surface is classified into the same class, the magnitude of Gaussian curvature is the additional information.

The magnitude of Gaussian curvature consists of the absolute and the relative one. Absolute value of Gaussian curvature is very small value in general, while the relative magnitude can be used to recognize the feature points. The principle to recover the relative magnitude is as follows. The coordinates (x, y) of four mapped points onto a sphere are used to recover the relative value of Gaussian curvature. Calculating the area value that is surrounded by four mapped points onto a sphere results in obtaining the relative magnitude of Gaussian curvature. This area value corresponds to the relative magnitude of Gaussian curvature. To recover the relative magnitude of Gaussian curvature, [7] shows the results for test objects with uniform albedo. The key idea of [7] is that the area value surrounded by the four mapped points onto a sphere corresponds to the relative magnitude of Gaussian curvature.

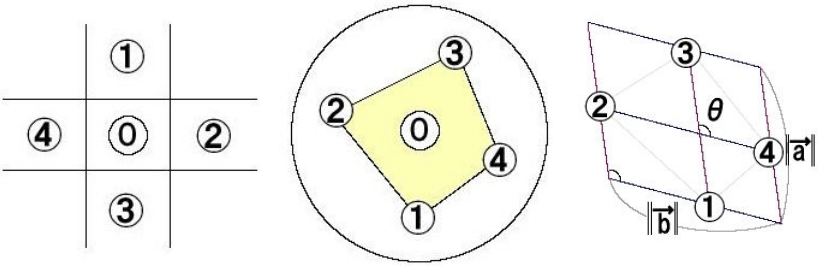
Let $z = f(x, y)$ be the height distribution of the object, then Gaussian curvature is defined as

$$G = \frac{1}{1 + f_x^2 + f_y^2} \begin{vmatrix} f_{xx} & f_{xy} \\ f_{yx} & f_{yy} \end{vmatrix} \tag{5}$$

Here, $f_x = \partial z/\partial x$ and $f_y = \partial z/\partial y$. For the smooth surface, the relation of $f_{xy} = f_{yx}$ holds. Suppose the four mapped points are located as shown in Fig.1-(a), the surrounded area value corresponds to the relative magnitude of Gaussian curvature. Consider the area value S calculated from the square with four mapped points as shown in Fig.1-(b), then the square area S is given as

$$S = \frac{1}{2}|\overrightarrow{a}||\overrightarrow{b}|\sin\theta = \frac{1}{2}|\overrightarrow{a} \times \overrightarrow{b}| = \frac{1}{2}(4\Delta x \Delta y)(f_{xy}^2 - f_{xx}f_{yy}) \tag{6}$$

This area value corresponds to the relative magnitude of Gaussian curvature, that is, the approximation of the relative value is obtained according to Eq.(6).



(a) Area Value (b) Calculation

Fig. 1. Area Value and Calculation

3.4 Modification Neural Network

Here, we introduce a modification neural network to improve the accuracy for the relative magnitude of Gaussian curvature. The area value calculated from four points is the approximate value of the magnitude. To improve the accuracy, we introduce the mapping of area value S to the true value of Gaussian curvature using the synthesized images of the following basis function $f(x, y)$.

$$f(x, y) = \alpha(\cos x \times \cos y) \qquad (-\frac{\pi}{2} < x, y < \frac{\pi}{2}) \tag{7}$$

This basis function can represent various curvatures by changing the parameter α. The key idea here is to use this basis function and to learn the neural network. The mapping of input to output of the neural network is shown in Fig.2. This neural network is learned with many data of Eq.(7).

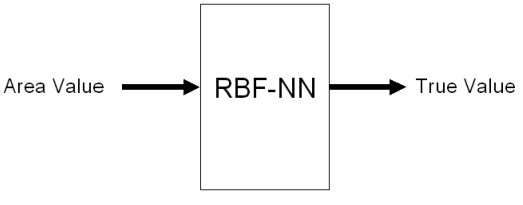


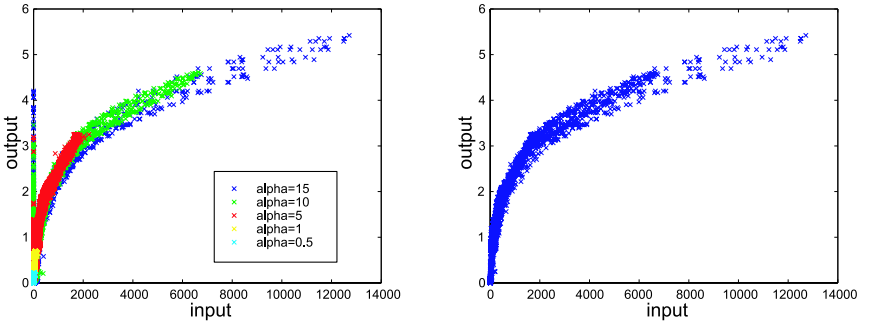
Fig. 2. Modification Neural Network

After the learning, the output of this modification neural network gives more reliable result for the magnitude of Gaussian curvature.

4 Experiments

Four input images are synthesized under the same illuminating condition for two different functions with the same reflectance property.

The learning is done for the shape of $f(x, y)$ given in Eq.(7). The learning data are prepared by combining the case of various α in the basis function Eq.(7).



(a) Candidate of Learning Data (b) Selected Learning Data

Fig. 3. Learning Data Extracted from Basis Function Eq.(7)

In the experiment, five cases (where $\alpha = 0.5$ to 15) are prepared for the learning data. The set of learning data consist of 2,225 pairs of input and output. This data set was selected and obtained by sorting the candidate of the learning data using the true value of Gaussian curvature among the total of 327,680 pairs. Fig.3-(a) shows the original data consisting of 327,680 pairs of the candidate, while Fig.3-(b) shows the used learning data consisting of 2,225 pairs.

Number of learning epochs was 100 with the spread constant `sc`, where `sc` was set to be 6000 since the area value used as input was in the range of 0 to 12612.

The performance of the generalization was investigated using this modification neural network after the learning. The test object used here is the synthesized images with the Phong model for the following 2-D sinc function $g(x, y)$.

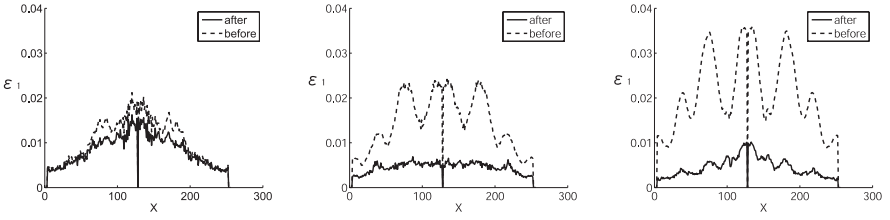
$$g(x, y) = \alpha\left(\frac{\sin x}{x} \times \frac{\sin y}{y}\right) \qquad (-2\pi < x, y < 2\pi) \tag{8}$$

The generalization is evaluated in Table 1, where "Previous Approach" represents [7], while "Proposed Approach" uses the modification neural network. In Table 1, ρ_1 represents the mean error calculated by Eq.(9), while ρ_2 represents the mean of the relative error [%] between T_i (True Gaussian Curvature) and R_i (Recovered Gaussian Curvature) at point index i among N points (i=1,$\cdots$,N).

$$\epsilon_1 = \frac{1}{N}\sum_{i=1}^{N} |T_i - R_i| \tag{9}$$

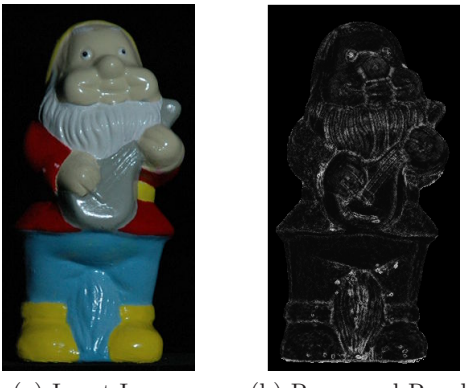
Table 1. Accuracy

		$\alpha = 0.5$	$\alpha = 1$	$\alpha = 1.5$
Previous Method [7]	ϵ_1	0.0165	0.0305	0.0356
	ϵ_2	18.2	35.3	43.8
Proposed Method	ϵ_1	0.0119	0.0112	0.0055
	ϵ_2	14.8	13.6	8.76



(a) $\alpha = 0.5$ (b) $\alpha = 1$ (c) $\alpha = 1.5$

Fig. 4. Mean Error of Gaussian curvature for 2-D Sinc Function Eq.(8)



(a) Input Image (b) Recovered Result

Fig. 5. Results for Real Object

$$\epsilon_2 = \frac{1}{N}\sum_{i=1}^{N}\frac{|T_i - R_i|}{T_i} \times 100[\%] \tag{10}$$

As the recovered results for crossing of the sinc function of α=0.5, 1.0 and 1.5, the mean error of Gaussian curvature are shown in Fig.4. Clear improvements for the accuracy can be seen from these results.

Next, the approach is examined for the real test object with color texture. Here, we use a white sphere as the known shape for the neural network learning. Both object has the same reflectance property and taken under the same four illuminating directions. Using Eq.(4), the effect of the albedo is eliminated. One input image is shown in Fig.5-(a) as an example. The recovered result of Gaussian curvature is shown in Fig.5-(b). This result is encoded as a gray scale, in which the brighter points have the larger Gaussian curvature. Reasonable results is obtained.

5 Conclusion

This paper proposed a new method to improve the accuracy of Gaussian curvature of a target object using neural network. With the neural network im-

plementation, the relative magnitude can be obtained directly from the area value surrounded by four mapped points. However, since this area value is an approximate value, we further extended the approach with modification neural network by introducing the basis function consisting of 2-D cosine function. Clear advantage can be seen in the proposed approach for the accuracy of the recovered Gaussian curvature. The recovered error is evaluated in the simulation. The robust results are obtained through the experiments. The entire approach is empirical under the condition that no explicit assumptions are not used for the surface reflectance function nor the illuminating directions. Further subjects are remained for the extension including cast shadow or interreflection problem.

Acknowledgment

This work is supported in part by Chubu University Grant and the JSPS Grant No. 16500108, Fukui's research is supported by The Hori Information Science Promotion Foundation, Kawanaka's research is spported by the Nitto foundation. Wooodham's research is supported by the Natural Sciences and Engineering Research Council (NSERC).

References

1. Woodham, R.J.: Gradient and curvature from the photometric stereo method, including local confidence estimation. Journal of the Optical Society of America, A 11, 3050–3068 (1994)
2. Iwahori, Y., Woodham, R.J., Bagheri, A.: Principal components analysis and neural network implementation of photometric stereo. In: Proc. IEEE Workshop on Physics-Based Modeling in Computer Vision, pp. 117–125. IEEE Computer Society Press, Los Alamitos (1995)
3. Iwahori, Y., Woodham, R.J., Ozaki, M., Tanaka, H., Ishii, N.: Neural Network based Photometric Stereo with a Nearby Rotational Moving Light Source. IEICE Transactions on Information and Systems E80-D(9), 948–957 (1997)
4. Angelopoulou, E., Wolff, L.B.: Sign of Gaussian Curvature From Curve Orientation in Photometric Space. IEEE Trans. on PAMI 20(10), 1056–1066 (1998)
5. Okatani, T., Deguchi, K.: Determination of Sign of Gaussian Curvature of Surface from Photometric Data. Trans. of IPSJ 39(5), 1965–1972 (1998)
6. Iwahori, Y., Fukui, S., Woodham, R.J., Iwata, A.: Classification of Surface Curvature from Shading Images Using Neural Network. IEICE Trans. on Information and Systems E81-D(8), 889–900 (1998)
7. Iwahori, Y., Fukui, S., Fujitani, C., Woodham, R.J., Iwata, A.: Relative Magnitude of Gaussian Curvature from Shading Images Using Neural Network. In: Khosla, R., Howlett, R.J., Jain, L.C. (eds.) KES 2005. LNCS (LNAI), vol. 3681, pp. 813–819. Springer, Heidelberg (2005)
8. Chen, S., Cowan, C.F.N., Grant, P.M.: Orthogonal least squares learning algorithm for radial basis function networks. IEEE Transactions on Neural Networks 2(2), 302–309 (1991)

Shadow Removal Method for Real-Time Extraction of Moving Objects

Shinji Fukui[1], Nozomi Yamamoto[2], Yuji Iwahori[2], and Robert J. Woodham[3]

[1] Faculty of Education, Aichi University of Education
Hirosawa 1, Igaya-cho, Kariya 448-8542, Japan
sfukui@auecc.aichi-edu.ac.jp
[2] Faculty of Engineering, Chubu University
Matsumoto-cho 1200, Kasugai 487-8501, Japan
yamamoto@cvl.cs.chubu.ac.jp, iwahori@cs.chubu.ac.jp
[3] Department of Conputer Science, University of British Columbia
Vancouver, B.C. Canada V6T 1Z4
woodham@cs.ubc.ca

Abstract. This paper proposes a new method for extracting moving objects in high definition video sequence. The proposed method can regard shadows as background and gives the robust result to noise and intensity changes because the method uses a* and b* components of CIELab color space and extracts moving objects by background subtraction of the estimated background image and the observed image. The computational cost is hugely increased by converting RGB color space to CIELab color space and using high definition video. The method addresses this problem by using GPU and decreasing the transfer data from GPU to CPU. Results are demonstrated with experiments on real data.

Keywords: Shadow Removal, Robustness to Illumination Changes, Estimating Background Image, Background Subtraction.

1 Introduction

Real-time extraction of moving objects from video sequences is an important topic for various applications of computer vision[1]-[3]. Applications include counting the number of cars in traffic, observing traffic patterns, automatic detection of trespassers, video data compression, and analysis of non-rigid motion.

To extract moving objects, there is a problem that shadows produced by objects are regarded as moving objects. This is one of the big problems to be solved. The methods for *shadow detection*[4] [5] or object extraction by using *chromatic information*[6] have been proposed. When the shadow detection process is added to the method for extracting moving objects, it becomes difficult to handle the whole process in real time. While, object extraction methods by using chromatic information, to which RGB color space is converted easily, can handle the whole process in real time. However, a simple background subtraction using chromatic information has the limitation to address the intensity changes.

B. Apolloni et al. (Eds.): KES 2007/ WIRN 2007, Part II, LNAI 4693, pp. 1021–1028, 2007.
© Springer-Verlag Berlin Heidelberg 2007

This paper proposes a new method for extracting moving objects. To handle shadows, the method uses CIELab color space to which RGB color space is converted. Only a* and b* components are used because both of them are chromatic information and are unchanged approximately even when a shadow is cast. We improve the method[7] to extract moving objects by using a* channel and b* channel. The proposed method is robust to noise and nonlinear intensity changes, such as the illumination changes or the automatic gain control function of the camera. Although the problem in converting RGB color space to CIELab color space takes heavy processing cost. The proposed method processes in real time and solves this problem by using GPU. Results are demonstrated on experiments with real video sequences.

2 Robust Background Subtraction for Illumination Changes[7]

Let the RGB color values at (x, y) in the original background image, BG, be denoted by $BG(x, y) = (BG^R(x, y), BG^G(x, y), BG^B(x, y))$. Let those in the observed image at time $t = 1, 2, \ldots$, be denoted by $I_t(x, y) = (I_t^R(x, y), I_t^G(x, y), I_t^B(x, y))$. The region with no moving objects at time t is defined as the background region, A_t.

When the intensity of the whole image changes, for example owing to illumination changes or the automatic gain control function of the camera, $I_t(x, y)$ in A_t differs from $BG(x, y)$. In this case, moving regions cannot be extracted by simple background subtraction. The method[7] proposed that the background image, BG_t, is estimated and then moving regions are extracted using the subtraction of BG_t and I_t.

2.1 Estimation of Present Background Image

Let (x, y) and (x', y') be different points in the image. Suppose, when the overall intensity of the whole image changes, that the value of $I_t^R(x, y)$ is equal to that of $I_t^R(x', y')$ and the value of $BG^R(x, y)$ is equal to that of $BG^R(x', y')$. From this assumption, a conversion table from $BG^R(x, y)$ to $I_t^R(x, y)$ is obtained based on the relation between $BG^R(x, y)$ and $I_t(x, y)^R$ in A_t. Conversion tables for the G and B color channels are obtained in the same way.

Let the candidate for the background region be A'_t. The procedure to get A'_t is as follows: First, the absolute value of the differences between I_t and I_{t-1} is obtained. Second, the region with the largest differences is selected by thresholding. Third, A'_t is calculated as the logical AND between the selected region and the last estimated background region A_{t-1}. Here, A_0 is the entire image, which is necessary to estimate BG_1.

The conversion table of each channel from $BG(x, y)$ to $I_t(x, y)$ is made by the RGB values in the obtained region A'_t. These tables convert the RGB color values of BG to the estimated background image BG_t. A histogram shown in Figure 1 is produced from each set $(I_t^R(x, y),\ BG^R(x, y))$ in the region A'_t to make the conversion table for R channel.

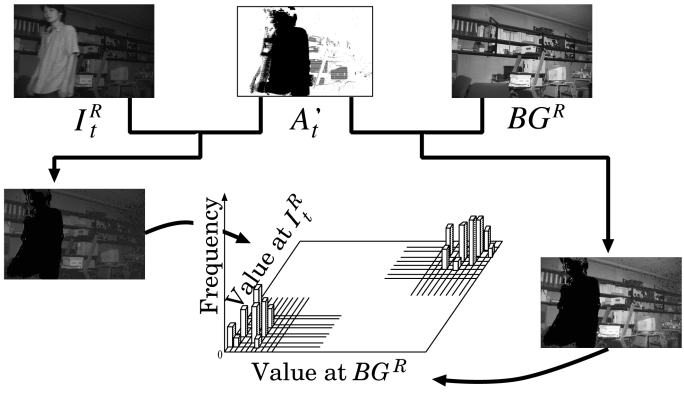


Fig. 1. How to Make Histogram

The conversion table for R channel is obtained from the histogram as shown in Figure 2. As shown in the histogram, at the 'value at $BG^R = 26$' in Figure 2, some pixels with the same value of BG^R may not have the same value ofI_t under the effect of noise. Therefore, the conversion table uses the median value of a set of pixels at I_t. Since the median value of I_t becomes 11 in this example, the set $(26, 11)$ is obtained and added to the conversion table. The set consists of all pixels which have the same value in $BG(x, y)$ as A_t'. Then, the linear interpolation is applied to the points that do not appear in the conversion table.

The conversion tables convert the RGB color values of BG to the RGB color values of the moving region in I_t. As a result, the background, BG_t, without moving objects is estimated.

2.2 Object Extraction by Background Subtraction

After BG_t is estimated, regions corresponding to moving objects are extracted by background subtraction.

To get a more robust result, I_t is divided into blocks. Each block is recognized as a background region or not. The total of the absolute values of differences for each block is calculated for each color channel. A block that has one (or more) total below threshold is regarded as a block of an object.

The method changes the threshold value dynamically as the intensity of the whole image changes. The histograms mentioned in **2.1** are used to determine the threshold. The procedure for determining the threshold for an R value is as follows: The range with 80% or more of frequency from the mean value is considered. That value is adopted as the threshold for that R value.

The threshold for an R value that does not appear in the conversion table can not be determined by the above procedure. In that case, the threshold for the R value is set to be one smaller value than the R value. Thresholds for G and B channels are determined in the same way.

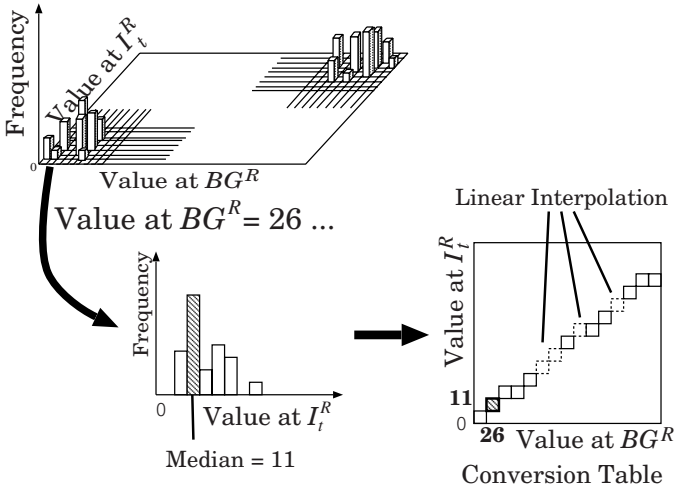


Fig. 2. Generating Conversion Table from Histogram

The threshold for each pixel is determined for each color channel. Next, the sum of the absolute value of the differences between $BG_t(x, y)$ and $I_t(x, y)$ and the sum of threshold values for each channel are calculated at each block. Finally, blocks where the sums of the absolute value of the differences become larger than those of the thresholds of all channels are regarded as blocks in the background. Otherwise, they are regarded as blocks in the moving object.

Each block is analyzed by the above process. Subsequently, a more detailed outline of the moving object is extracted. For blocks that overlap the boundary of a moving object, each pixel is classified into either the moving region or the background region, as was the case for the whole block itself.

3 Handling of Shadow by Using CIELab Color Space

To extract the moving objects, one of the big problems to be solved is how to handle shadows. In the case of using RGB color space, shadows are regarded as moving objects by many methods of object extraction because RGB values of a pixel in a shadow area differ from those of the same pixel in original background image.

Extracting shadow area as moving objects causes the problems that two or more target objects are misrecognized as one object or that shadow is detected as moving object even though no moving object exists in the image.

To handle shadows, some methods which use the chromatic information, such as H and S components of HSV color space, U and V components of YUV color space etc., have been proposed[6]. The proposed method uses a* and b* components of CIELab color space to handle the shadows because the method using CIELab color space can give more appropriate results[8].

However, a simple background subtraction has the limitation for addressing nonlinear intensity changes even though the chromatic information is used. The proposed method improves the method mentioned in **2** to extract moving objects by using a* component and b*component. L* component does not be used because L* value of a pixel in a shadow differs from that of the same pixel in BG.

Let the image that RGB values of each pixel in BG is converted to CIELab values be denoted by BG_{Lab}, and let the image that RGB values of each pixel in I_t is converted to CIELab values be denoted by I_{Labt}.

The process of proposed method is as follows: BG_{Lab} is achieved in advance. First, conversion tables for a* component and b* component are generated by the same way mentioned in **2.1** after achieving BG_{Labt}. Next, the background, BG_{Labt}, is estimated by BG_{Lab} and conversion tables. And last, moving objects are extracted by using subtraction of BG_{Labt} and I_{Labt}.

4 Speed Up Using the GPU

The proposed method uses CIELab color space. Converting RGB color space to CIELab color space costs heavily-loaded process. The method uses the GPU to improve the processing speed and performance.

Recently, the performance of the Graphics Processing Unit (GPU) has been improved to display more polygons in shorter time. The geometry and rendering engines in a GPU are highly parallelized. Moreover, the shading process has become programmable via what is called the "Programmable Shader". Current GPUs make geometry and rendering highly programmable.

GPUs are very powerful, specialized processors that can be used to increase the speed not only of standard graphics operations but also of other processes[9]. A GPU is useful for image processing and can be programmed with a high-level language like C.

While, a texture containing a image data should be created to use GPU. The result needed by CPU should be rendered to a texture and CPU should use the texture data. Also, a processing result of a pixel can not be used in a process for other pixels. So, generating the conversion tables and determining the thresholds are processed at CPU in the proposed method.

5 Experimental Results

In the experiments, a general purpose digital video camera was used as the input device. White balance of the camera was set to indoor setting. A PC with Core Duo T2500 and GeForce Go 7900 GS was used to extract moving objects. Each target image frame was 720×480 pixels. The corresponding block size was set to 10×10.

First, an outdoor scene is used as the original background.

Figure 3 shows the observed images and experimental results. Figure 3-(a) shows the observed images and Figure 3-(b) shows the experimental results.

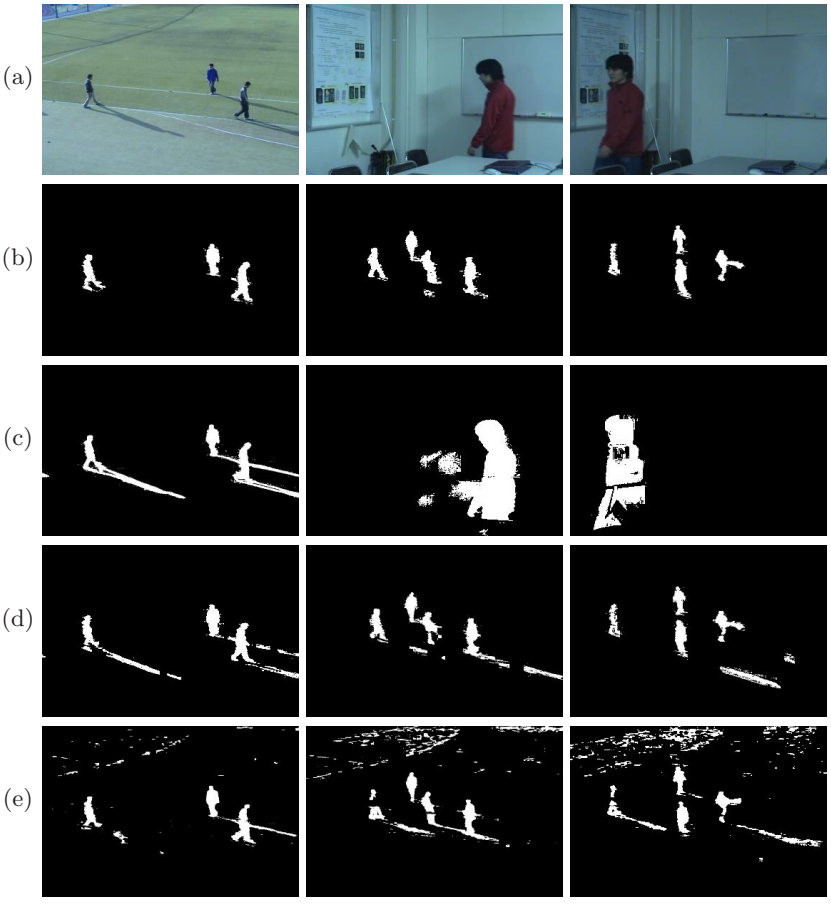


Fig. 3. Observed Images and Results

The white region shows the detected moving object, and the black region corresponds to the detected background region. Figure 3-(b) demonstrates that this approach can extract the moving object with high performance and the shadows are extracted as background. Figure 3-(c) shows the experimental results produced by the previous approach [7]. It is shown that the approach[7] do not address shadows. Figure 3-(d) and Figure 3-(e) shows the results produced by the method improved the method mentioned in **2**. The method producing Figure 3-(d) uses H component and S component of HSV and the method producing Figure 3-(e) uses U component and V component of YUV. These results show that CIELab color space is more appropriate one than HSV or YUV color space to handle shadows.

In this experiment, the processing using the GPU required about 7.8 msec/frame. In contrast, that without the GPU takes about 278.6 msec/frame. These results show that using the GPU increases processing speed dramatically so

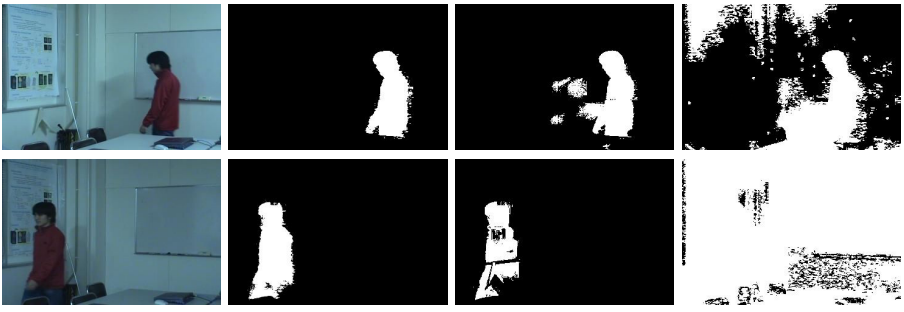


Fig. 4. Observed Images and Results

that we can extract the moving object in real time. Incidentally, the method[7] takes about 10.8 msec/frame. The proposed method can handle whole process faster than the method[7]. Even though the process for converting RGB color space to CIELab color space is heavily-loaded, it is run on the GPU and takes a very short time in the proposed method. On the other hand, the process on the CPU is decreased. The proposed approach needs two conversion tables and the method[7] needs three conversion tables in contrast. These result in a increase in processing speed.

Next, another experiment was done for indoor scene. The experiment is done in a room illuminated by fluorescent lamps. A fluorescent lamp is turned off in the middle of the image sequence, and then another lamp is turned off after a while.

Figure 4 shows the observed images and experimental results. Figure 4-(a) shows the observed images, which have different overall brightness from the original background image and Figure 4-(b) shows the experimental results. These results show that this approach can also extract the moving object with high performance even under quick illumination changes. Figure 4-(c) shows the results produced by the method [7]. It is shown that the proposed approach works better than the previous approach [7]. Figure 4-(d) shows the results by subtraction of a* and b* components of background image and those of the observed images. It is shown that simple background subtraction can not extract moving objects under illumination changes even if chromatic information is used.

6 Conclusion

A new approach is presented to extract the moving regions in a video sequence. The approach uses CIELab color space to handle shadow areas produced by objects. It estimates a* component and b* component of the present background image and makes it possible to extract a moving object without shadow areas. Moreover, the proposed method has robustness to quick intensity changes.

Using CIELab color space increases the processing cost because converting RGB color space to CIELab color space is heavily-loaded process. This problem is solved and real-time implementation is done by using the GPU. We expect the GPU to increase the speed of other image processing operations.

Cases of incorrect detection are remained. Future work includes the automatic update of background image and addressing the small movement, such as swaying trees or shake of curtain.

Acknowledgment

This work is supported by THE HORI INFORMATION SCIENCE PROMOTION FOUNDATION, the Grant-in-Aid for Scientific Research No.16500108, Japanese Ministry of Education, Science and Culture, and Chubu University grant. Woodham's research is supported by the Canadian Natural Sciences and Engineering Research Council (NSERC).

References

1. Stauffer, C., Grimson, E.: Learning patterns of activity using real-time tracking. IEEE trans. on PAMI 22(8), 747–757 (2000)
2. KaewTraKulPong, P., Bowden, R.: An improved adaptive background mixture model for real-time tracking with shadow detection. In: Proc. of the 2nd European Workshop on Advanced Video-Based Surveillance Systems (2001)
3. Sato, Y., Kaneko, S., Igarashi, S.: Robust Object Detection and Segmentation by Peripheral Increment Sign Correlation Image. Trans. of the IEICE (in Japanese) J84-D-II(12), 2585–2594 (2001)
4. Prati, A., Mikic, I., Trivedi, M.M., Cucchiara, R.: Detecting Moving Shadows: Algorithms and Evaluation. IEEE Trans. on PAMI 25(7), 918–923 (2007)
5. Salvador, E., Cavallaro, A., Ebrahimi, T.: Cast shadow segmentation using invariant colour features. Computer Vision and Image Understanding 95(2), 238–259 (2004)
6. Gonzalez, R., Woods, R.: Digital Image Processing. Addison-Wesley Longman Publishing Co. Inc., Redwood City,CA, USA (2001)
7. Fukui, S., Iwahori, Y., Itoh, H.: Robust Method for Extracting Moving Object to Intensity Change Using Color Image Sequence (in Japanese). In: Proc. of Meeting on Image Recognition and Understanding 2005, pp. 814–821 (2005)
8. Khan, E.A., Reinhard, E.: A Survey of Color Spaces for Shadow Identification. In: ACM Symposium on Applied Perception in Computer Graphics and Visualization, pp. 160–160. ACM Press, New York (2004) (abstract)
9. `http://www.gpgpu.org/`

Delayed Learning and the Organized States

Toshinori Deguchi[1] and Naohiro Ishii[2]

[1] Gifu National College of Technology, Motosu, Gifu, 501–0495, Japan
deguchi@gifu-nct.ac.jp
[2] Aichi Institute of Technology, Toyota, Aichi, 407-0392, Japan
ishii@aitech.ac.jp

Abstract. Elman presented a network with a context layer for time-series processing. The context layer keeps the output of the hidden layer, and the outputs are inputted to the hidden layer for the next calculation of the time-series. In this paper, the context layer is reformed into the internal memory layer, which is connected from the hidden layer with the connection weights to make the internal memory. The new learning algorithm, called the time-delayed back-propagation learning, is developed for the internal memory. The ability of the network with the internal memory layer and the organized states of the internal memory are demonstrated by applying it to the simple sinusoidal time-series.

Keywords: Internal memory, delayed learning, back-propagation.

1 Introduction

For time-series learning, Elman's network[1,2] has been used[3,4,5]. Elman's network is a simple recurrent network which has a context layer as a copy of the hidden layer, and is able to use the information of the context layer for the next calculation.

Since the context layer is a simple copy of the hidden layer, the outputs of the hidden layer are sent directly to the context layer.

We introduced the connection weights to the connections from the hidden layer to the context layer.[6] With these connections, the context layer is no longer a simple copy of the hidden layer but the internal memory of the network. Therefore, this layer is named the internal memory layer. It is expected that the network with the internal memory layer be more powerful for learning in time-series.

For learning of this network, we developed a new learning algorithm, called here, the time-delayed back-propagation learning.[6,7] Through our simulations, it turned out that the network does not need many neurons to learn simple time-series.

In this paper, the internal memory network and the time-delayed back-propagation learning are summarized, and the advantage of this network and the organized state of the internal memory is presented with a simple example of time-series.

B. Apolloni et al. (Eds.): KES 2007/ WIRN 2007, Part II, LNAI 4693, pp. 1029–1036, 2007.
© Springer-Verlag Berlin Heidelberg 2007

2 Network and Learning

2.1 Internal Memory Network

Elman introduced the context layer to the 3-layer back-propagation network. In this paper, the connection weights from the hidden layer to the context layer are introduced to Elman's network as shown in Fig.1. While the context layer in Elman's holds the outputs of the hidden layer, this layer keeps the internal state calculated from the outputs of the hidden layer. Therefore, this layer is named the internal memory layer, and the 3-layer network with the internal memory layer is called the internal memory network.

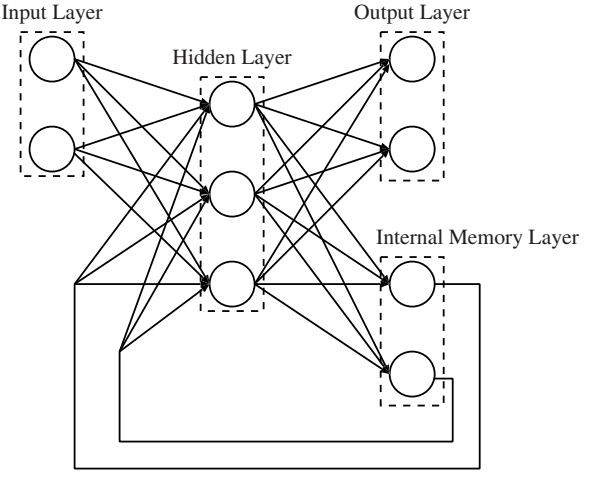


Fig. 1. Internal memory network

Because the network has the internal memory calculated from the output of the hidden layer, the learning ability of this network is expected to be higher than Elman's.

The equations of the internal memory network are defined as follows.

$$x_j^h(t) = f\left(\sum_{k=0}^{N_i} w_{kj}^{ih} x_k^i(t) + \sum_{k=0}^{N_m} w_{kj}^{mh} x_k^m(t-1) - \theta_j^h\right), \quad j = 1, \cdots, N_h \tag{1}$$

$$x_j^m(t) = f\left(\sum_{k=0}^{N_h} w_{kj}^{hm} x_k^h(t) - \theta_j^m\right), \quad j = 1, \cdots, N_m \tag{2}$$

$$x_j^o(t) = f\left(\sum_{k=0}^{N_h} w_{kj}^{ho} x_k^h(t) - \theta_j^o\right), \quad j = 1, \cdots, N_o \tag{3}$$

where $x_j^i(t)$, $x_j^h(t)$, $x_j^m(t)$, and $x_j^o(t)$ are the output of the j-th neuron at step t, in the input layer, in the hidden layer, in the internal memory layer, and in the output layer, respectively. θ_j^h, θ_j^m, and θ_j^o are the threshold value of the j-th neuron in the hidden layer, in the internal memory layer, and in the output layer, respectively. w_{kj}^{ih}, w_{kj}^{mh}, w_{kj}^{ho}, and w_{kj}^{hm} are the connection weight from the k-th neuron in the input layer to the j-th neuron in the hidden layer, from the k-th neuron in the internal memory layer to the j-th neuron in the hidden layer, from the k-th neuron in the hidden layer to the j-th neuron in the internal memory layer, and from the k-th neuron in the hidden layer to the j-th neuron in the output layer, respectively. N_i, N_h, N_m, and N_o are the number of neurons in the input layer, in the hidden layer, in the internal memory layer, and in the output layer, respectively, and f is the output sigmoid function described below in Equation (4),

$$f(x) = \frac{1}{1 + \exp(-x)} \tag{4}$$

2.2 Delayed Learning Algorithm

Back-propagation learning[8] is not applicable directly to the internal memory network, because the internal memory layer does not have the desired output, while the output layer does. For the internal memory networks, time-delayed back-propagation learning — or simply 'delayed learning' — is developed.

The basic idea of the delayed learning is the expansion of the network in time. The outputs of the internal memory layer at step $t-1$ are the inputs of the hidden layer at step t. The errors are to be back-propagated from the hidden layer at step t to the outputs of the internal memory layer at step $t-1$. It is not enough when there are few neurons in the output layer, because the errors from the output layer are sometimes so small that the hidden layer is not able to learn.

In the delayed learning, the hidden layer at step $t-1$ receives the errors from both the output layer at step $t-1$ and the internal memory layer at step $t-1$.

Let $y_i(t)$ be the desired value of i-th neuron in the output layer at step t. The error signal of i-th neuron in the output layer is shown as follows:

$$\delta_i^o = x_i^o(t)(1 - x_i^o(t))(y_i(t) - x_i^o(t)) \tag{5}$$

As same as the usual back-propagation, the error signal of i-th neuron in the hidden layer is

$$\delta_i^h = x_i^h(t)(1 - x_i^h(t)) \sum_{j=0}^{N_o} \delta_j^o \, w_{ij}^{ho} \tag{6}$$

Then, calculate the error signal of the internal memory layer at step $t-1$.

$$\delta'^m_i = x_i^m(t-1)(1 - x_i^m(t-1)) \sum_{j=0}^{N_h} \delta_j^h \, w_{ij}^{mh} \tag{7}$$

Again, the errors of the output layer and of hidden layer at step $t-1$ are calculated by the next equations.

$$\delta'^{o}_{i} = x^{o}_{i}(t-1)(1-x^{o}_{i}(t-1))(y_i(t-1)-x^{o}_{i}(t-1)) \tag{8}$$

$$\delta'^{h}_{i} = x^{h}_{i}(t-1)(1-x^{h}_{i}(t-1))\sum_{j=0}^{N_o}\delta'^{o}_{j}\,w^{ho}_{ij} + \sum_{j=0}^{N_m}\delta'^{m}_{j}\,w^{hm}_{ij} \tag{9}$$

Then, at step $t-1$, the connection weights are changed as follows:

$$\begin{aligned} w^{pq}_{ij}(t+1) &= w^{pq}_{ij}(t) + \eta\,\delta'^{q}_{j}x^{p}_{i}(t-1), \\ (p,q) &= (i,h),(m,h),(h,m),(h,o) \end{aligned} \tag{10}$$

When you adopt a momentum term,

$$\begin{aligned} w^{pq}_{ij}(t+1) &= w^{pq}_{ij}(t) + \eta\,\delta'^{q}_{j}x^{p}_{i}(t-1) \\ &\quad +\alpha(w^{pq}_{ij}(t) - w^{pq}_{ij}(t-1)), \\ (p,q) &= (i,h),(m,h),(h,m),(h,o) \end{aligned} \tag{11}$$

If you need another back-step, you can calculate the error signals at $t-2$. If you need more, you can get another back-step in the same way. Therefore, the delayed learning provides 1-delay, 2-delay, ..., n-delay learning.

3 Simulations on Simple Time-Series

3.1 Learning

In the internal memory network, the outputs are calculated not only from the inputs but from the internal states. This means that outputs with the same inputs may be different because of the internal states.

To demonstrate this feature, the sinusoidal values of $\sin(x)$ are used by sampling at every $\pi/6$. This values include the value 0 twice in a cycle, and the following values are different. Therefore, the network has to make the internal state which can tell the difference.

Each value is transformed by $\frac{\sin(x)+1}{2}$ to fit to the range of the neuron from 0 to 1 as in Equation (12).

$$a(t) = \frac{\sin\left(\frac{\pi t}{6}\right)+1}{2} \tag{12}$$

In our former works[7], we used quite a large number of neurons to learn this time-series. The following simulations revealed that we need fewer neurons in the network with this simple time-series, although the delay time in learning increases. Through these simulations, it turned out that a network with 2 neurons for the hidden layer and 1 neuron for internal memory layer was able to handle it using 8-delay learning.

In this network, the input of the network at step t is $a(t)$ and the desired output is $a(t+1)$. The parameter of the learning is set to $\alpha = 0.03$, $\eta = 0.003$ in this learning. One cycle of sampled values of $\sin(t)$ by $\pi/6$ is counted as one learning time. The network learned for 100,000,000 learning times with 7-delay and 8-delay learning. Each maximum error during a cycle at every 200,000 learning times is shown in Fig.2. The results are shown until 20,000,000 learning times, because there aren't any remarkable movements after 20,000,000.

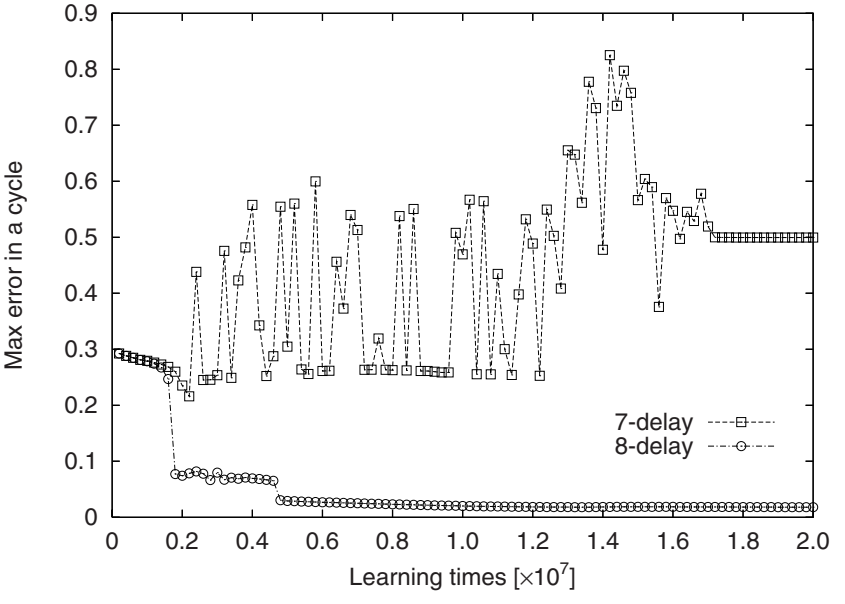


Fig. 2. Maximum errors during one cycle at every 200,000 learning times

In Fig.2, the error of 8-delay learning descended to 0.074 through 2 rapid downward-movements around 1,700,000 and 4,700,000 learning time, whereas the error of 7-delay learning didn't descend. In our former work[7], it required 4-delay learning to make the network with 30 neurons in hidden layer and with 30 in internal memory layer learn a sinusoidal series. This suggests the variation of the suitable delay time in networks.

After 10,000,000 learning of 8-delay delayed learning, the outputs of the network are shown in Fig.3. In this simulation, the network output was sent to its input without any external signals.

The outputs are lined in sinusoidal curve. This means that the network has learned the time-series and that the network has made proper internal states in the internal memory layer.

For viewing the internal memory, the outputs of the neuron in the internal memory layer are also shown in Fig.3. The output of the internal memory is a similar shape of $-\cos(t)$ function.

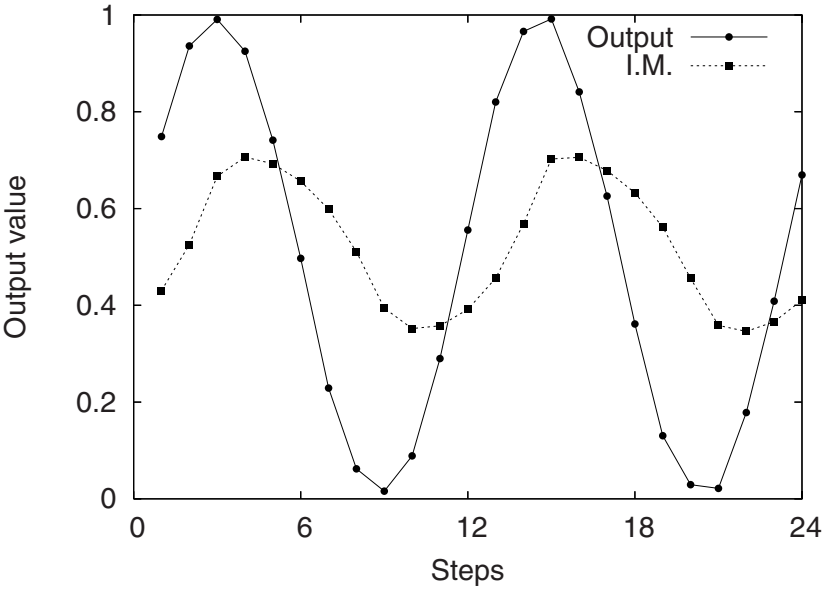


Fig. 3. Output and internal memory in 8-delay learning

3.2 Internal Memory

At first, the internal memory was connected with random-valued weights. After 10,000,000 learning, its output was organized into the shape of $-\cos(t)$. Therefore, the delayed learning organized the internal memory layer.

To show how the internal memory was formed, we focused on the maximum value, the minimum value, and the cycle length of the internal memory outputs. In learning, the whole network condition was saved at every 100,000 learning times. By using each condition, we can investigate the behavior of the internal memory at that time. The maximum value and the minimum value are collected during 24 steps which are 2 cycles in this time-series.

The maximum and minimum values are shown in Fig.4.

Under 1,700,000 learning times, the maximum value and the minimum value are almost the same. This means that there are no obvious cyclic movements in the internal memory. During the learning after 1,700,000 steps, the difference between the maximum and the minimum widened along the learning. Around 4,700,000 steps, the difference grew wide rapidly. Then, in this simulation, the internal state was organized with 2 breakthroughs. In Fig.4, the first downward-movement point of errors is around 1,700,000 steps, and the second is around 4,700,000 steps. These are the times when the error dropped in Fig.2. Therefore, the organization of the internal state affects the decrease of the error.

For cyclic length, the peaks are counted while the network works for 1,000,000 cycles. When the current value is greater than both the previous value and the

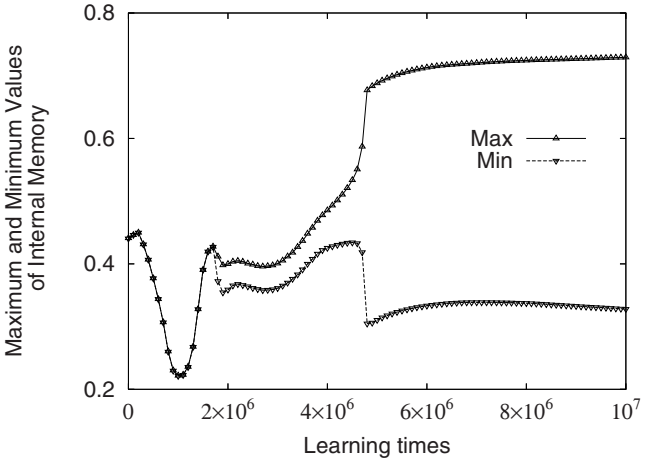


Fig. 4. The maximum value and the minimum value of the internal memory

following value, we called it a peak. The number of the peaks is shown as the ratio to 1,000,000. Because there is one peak in a cycle in this series of sinusoidal curve, the ratio would be 1.0 if the network had learned the series perfectly. This ratio is inversely proportional to the length of the cycle that the internal memory organized.

The ratios are shown in Fig.5. As same as Fig.4, the ratios are almost 0 under 1,700,000 learning steps. This is because the internal memory has no cyclic movements.

After that, it grew slowly to 1.0, and it converged to a value little high above 1.0 around 4,700,000 steps. As the difference between the maximum and the

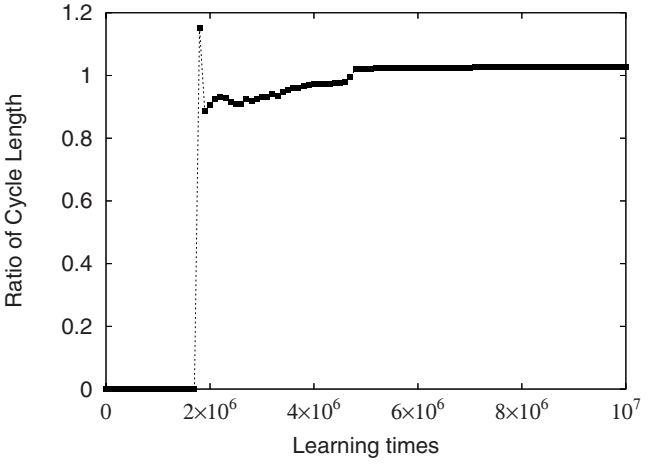


Fig. 5. The ratio of the cycle length in the internal memory outputs

minimum grew, the length of the cycle that the internal memory kept converged to a little less value than the one of the original series.

4 Conclusion

In this paper, the context layer was reformed into the internal memory layer and the time-delayed back-propagation learning was developed for the internal memory. The ability of the network with the internal memory layer and the organized states of the internal memory are demonstrated by applying it to the simple sinusoidal time-series. With 8-delay learning, the network was able to learn the series and organize the internal memory whose cycle was a little less than the one of the series.

References

1. Elman, J.L.: Finding Structure in Time. Cognitive science 14, 179–211 (1990)
2. Elman, J.L.: Learning and Development in Neural Networks: The Importance of Starting Small. Cognition 48, 71–99 (1993)
3. Koskela, T., Lehtokangas, M., Saarinen, J., Kaski, K.: Time Series Prediction with Multilayer Perceptron, FIR and Elman Neural Networks. In: Proc. of the World Congress on Neural Networks, pp. 491–496. INNS Press, San Diego, USA (1996)
4. Cholewo, T.J., Zurada, J.M.: Sequential Network Construction for Time Series Prediction. In: Proc. of the IEEE Intl. Joint Conf. on Neural Networks, Houston, Texas, USA, pp. 2034–2039. IEEE Computer Society Press, Los Alamitos (1997)
5. Giles, C.L., Lawrence, S., Tsoi, A.C.: Noisy Time Series Prediction Using a Recurrent Neural Network and Grammatical Inference. Machine learning 44(1/2), 161–183 (2001)
6. Iwasa, K., Deguchi, T., Ishii, N.: Acquisition of the Time-Series Information in the Network with Internal Memory (in Japanese). IEICE Technical Report NC 2001–71, pp. 7–12 (2001)
7. Deguchi, T., Ishii, N.: Delayed Learning on Internal Memory Network and Organizing Internal States. In: Wang, J., Yi, Z., Zurada, J.M., Lu, B.-L., Yin, H. (eds.) ISNN 2006. LNCS, vol. 3971, pp. 502–508. Springer, Heidelberg (2006)
8. Rumelhart, D.E., Hinton, G.E., Williams, R.J.: Learning internal representations by error propagation. In: Rumelhart, D.E., McClelland, J.L. (eds.) Parallel Distributed Processing: Explorations in the Microstructure of Cognition, vol. 1, pp. 318–362. MIT Press, Cambridge, MA (1986)

Educational System Using Self-monitor Study and Streaming

Masahiro Ozaki[1], Yoshinori Adachi[1], Saori Takeoka[2], Ai Sugimura[2], and Naohiro Ishii[3]

[1] Chubu University, 1200 Matsumoto-Cho, Kasugai, Aichi, Japan 487-8501
ozaki@isc.chubu.ac.jp, adachiy@isc.chubu.ac.jp
[2] Nagoya Women's University, 3-40 Shioji-Cho, Mizuho, Nagoya, Japan 467-8610
saori@nagoya-wu.ac.jp, sugimura@nagoya-wu.ac.jp
[3] Aichi Institute of Technology, Yakusa-Cho, Toyota, Aichi, Japan 470-0392
ishii@aitech.ac.jp

Abstract. It was understood that the learning behavior has been decided not only by studying situation of him or her but also by the monitored environment of the entire class such as people surrounded. We designed the self-monitor study system that integrates the class at the university and the studying at home, and have been examining the education effect. As a result, it is understood that it is effective to do the streaming when the self-monitor study is executed. Then in this study, the learning system that combined the self-monitor study with the streaming was constructed, and the education effect was examined. As the result, fairly good effect was obtained.

1 Introduction

The course content and problems executed in an ordinary face-to-face class are almost all the same and uniform. Moreover, individual learner's abilities are not considered. Actually, it is difficult for the professor who guides a lot of learners in a class to evaluate one by one properly in the limited class time. To execute the test in the class within a limited time, the large encumbrance is forced on the learner. And, the overtime studying at home is often necessary. However, it is impossible that the professor supports individual learner's studying at home directly.

To attempt the improvement of learner's greediness for learning, we have repeatedly executed the experiment by the self-monitor up to now in the large class and the semi class [1-4]. The experiment that we had executed aimed to examine whether the greediness for learning and the learning outcome improved by the self-monitor study by advancing study while understanding an own study situation. The self-monitor study was directed to an individual learner, and the relation between the learned behavior and the study result was analyzed from the record. As a result, the difference was obviously seen in the study result between a group that executed the self-monitor study and a group without it [1, 3].

Moreover, the learner classing was organized according to the ability. And, learner's proficiency and the relation of the self-monitor were investigated. As a

B. Apolloni et al. (Eds.): KES 2007/ WIRN 2007, Part II, LNAI 4693, pp. 1037–1044, 2007.
© Springer-Verlag Berlin Heidelberg 2007

result, the tendency to do stabilizing the self-monitor study was seen in the learner with high proficiency level [4]. In addition, it was examined about four learners who had moved to the subordinate position class due to the decrease at the study level. As a result, study results of all four people had decreased after they had moved to the subordinate position class. Even the content of study was reduced in that class.

It was shown to study while understanding an own study situation from those results and was shown an effective possibility to the improvement of the study result. Moreover, it was shown that the belonging level of the class and the greediness for learning of the class influenced the study result.

In this study, the influence that the self-monitor study method that we developed and the streaming exerted on the study efficiency was observed. Further, their problems were extracted there, and the strategy to solve those problems was shown.

2 Self-monitor Study and Streaming

Most of the class executed at the university is a group class that one professor teaches with a textbook. And, quizzes are executed and the understanding level of the content of study is sometimes confirmed. And, when it is understanding shortage, the review is directed to promote greater understanding. However, studies other than the class time are independent, and it is usual not to compel study. Therefore, the understanding level of the content of study will be evaluated by the final test overall. Then, to attempt the improvement of the learning effect, various didactics of the development of a comprehensible teaching material and the development of the teaching method that adopts the computer and Web, etc. are researched. Because effective didactics are not developed, it has left enormous difficulties in each technique.

This study aimed to develop didactics that was more effective to adopt the self-monitor study in English Language Education. Figure 1 shows the flow of the self-monitor study. After studying, quizzes are done. Their self-grades and the learning effect are confirmed. The problem answered wrongly in quizzes is reviewed after the class. Further, reports are taken on the study situation based on the result (confirmation vote). The learner self-manages all own study situations. It was examined whether the improvement of the greediness for learning was able to be attempted by this method. To confirm whether the self-monitor study was executed, it was directed that the confirmation vote that recorded the study situation for each learner was submitted. Those a series of learned behavior is not compulsion but learner's independent self management. Therefore, whether it had been reflected in the study result decided to be analyzed.

The experimental result of the self-monitor study that was executed in 2003 and 2004 [8-10] did not achieve a clear result for improvement of the greediness for learning. The learning effect was a problem to the learner who did not have the desire to work on study positively, often said, "Study dislike". In a word, it was thought that the learning effect was not able to be expected so much for the learner who did not have the greediness for learning from the beginning. It was decided to think the self-monitor study for the overall self management including monitoring the interpersonal relationship and learner learning environments between learners. Then, the class

composition that divided the learner into the class according to the ability was executed, and what influence the learning environment to which the class was organized according to the ability gave to the self-monitor study was analyzed.

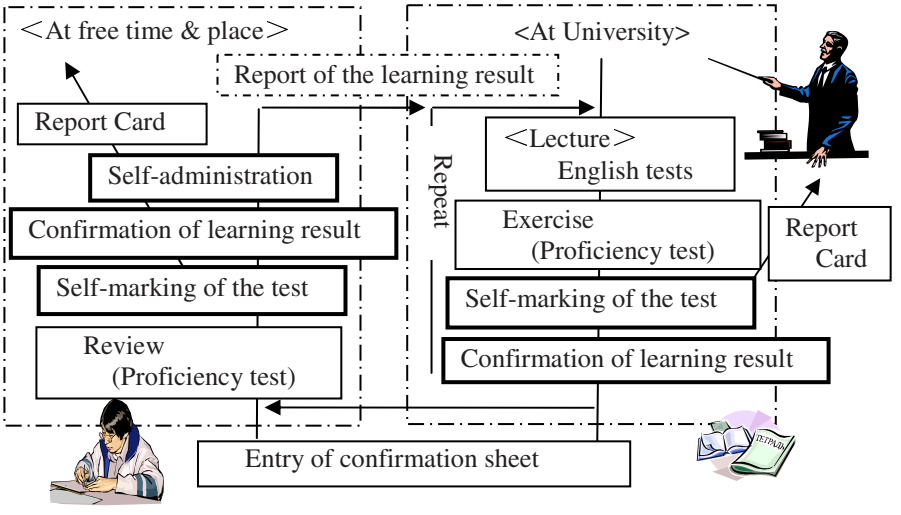


Fig. 1. Flow diagram of Self-monitor study

3 Experiment Using STEP Teaching Material

(1) STEP teaching material based on ability

It is very serious that the learner studies English whose mother tongue is not English. Moreover, there is a big difference on learner's English ability, and guidance is very difficult. We execute the English ability test beforehand, and execute the streaming based on the result. As a result, guidance that considers learner's ability is done.

Three classes (classes A, B, and C) were organized from the high rank with learner's English ability as shown in Table 1. The exercise that evaluated the learning effect was assumed to be a content of study along the streaming. A basic problem is a problem of studying in the text used by the class. The applied problem is along the content of the studying text. However, it is a problem that is more difficult than the content of the text. A basic problem and the applied problem were combined additionally at each ability level as shown in the table.

(2) Method of experiment

The experimental period was executed for one year separately for the first half and the latter half. The examination was executed before it began to teach it and after the streaming was done. In the latter half, the replacement of the class was executed based on the study result in the first half.

The experiment adopted quizzes, the final exam, the self-monitor methods and techniques of instruction executed in each class. The self-monitor methods and techniques of instruction had the advantage of uniformly appreciable of learner's

learning ability by adopting the methods and techniques of instruction method fixed beforehand. Then, the professor became possible to rate the ability of the learner.

Moreover, the class compositions in the latter half became possible by evaluating learner's understanding level in the first half. As a result, it was thought that evaluating the effect of the self-monitor study by the review of the class composition in the latter half became possible.

Table 1. Problem level in classes streamed according to degree of advancement

Class	Understanding level	Number of Quizzes	
		Basic	Application
A	The 2nd STEP class	10	10
B	The semi-2nd STEP class	15	5
C	The 3rd STEP class	20	0

Details of the experiment are described as follows;

① Experimental period: One year (First half: April - July, latter half: September - January).

Learner: In the first half, class A =21 people, class B =41 people, and class C =12 people. 74 people in total. In the latter half, class A =20 people, class B =37 people, and class C =15 people. 72 people in total

Subject: "English maneuver"

Self-monitor study method: It is executed along the following procedure.

Do the following explanations at the first class about the self-monitor study.

a) Execute quizzes after the each class. The quizzes are self-graded and are filled in on "Confirmation vote", and it is submitted by the next class.

b) The problem of miss-answering by quizzes must be reviewed by next week, and report on it.

c) The understanding level of the quizzes result and the content of study is recorded voluntarily and self-managed.

Streaming: The class composition was executed in the first half and the latter half.

a) The class composition according to the test before the first class in the first half and the learner's interview.

b) It reorganizes the class by the study result (result of the final exam, quizzes execution, and condition of the self-monitor study) in the first half and the learner’s interview.

⑥ Analysis of the experimental result: The experimental result was analyzed by each test and the execution condition of the self-monitor study. Moreover, the effect of the streaming was analyzed.

4 Evaluation of Experimental Results

Figure 2 shows the average score of each test in three classes. Because the examination contents were the same in three classes, as the result of the pretest at the

first term, average point (69.7) of class A was highest, 53.6 points of class B was next, and 49.8 points of class C was lowest which reflects the streaming. In the result of the test at the end of the first term, though the average points of three classes rise, average point of class A is 77.1, class B is 69.1, and class C is 55.3, and class B's growth rate is the highest.

The latter term pretest was a review problem of the test at the end of the first term, and how much the test was understood at the end of the first term were examined. As a result, the growth rate of the average point of class C was the highest, and the average point of class B was next, and only the average point of class A lowered than that of the tests at the end of the first term. In the tests at the end of latter term, the average point of class A was 72.2, class B was 67.5, and class C was 50.4. The average point of class A was highest.

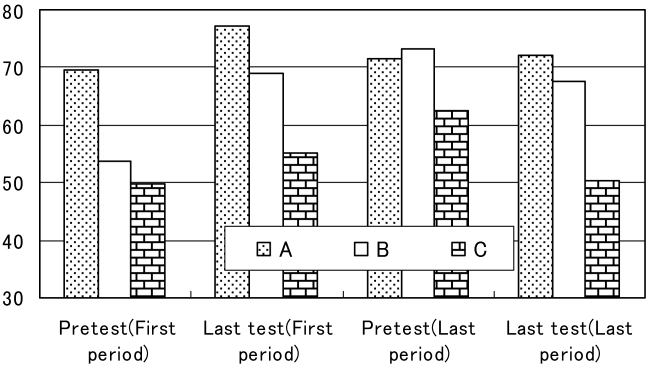


Fig. 2. Class average points

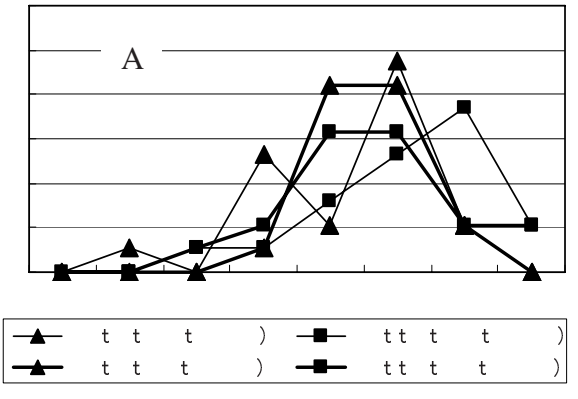


Fig. 3. Point distribution in class A

Figure 3 is score distribution of each test in class A. This class has a lot of experienced people of an overseas studying abroad for short time. There are lots of learners of high score (90.0) in the test at the end of the first term. However, the result of the term exam was not better than that of the pretest because it was guessed that the contents of study were easy. On the other hand, the greediness for learning to English of this class was high, and the studying at home was actively done. It seems that such enthusiastic study led the result of showing the high score in all the tests.

Figure 4 is score distribution of each test in class B. In the first term pretest, there are two groups whose mark centered 60 points and 80 points, respectively. In the test at the end of the first term, the score distribution around 60 points is compactly

scattered. In the test at the end of latter term, the score is distributed widely within the range from 40 to 80 points. In class B, there were included a lot of learners, who should be appropriate to belong to a higher level class or a lower level class. It is thought that those factors influenced the result of the test at the end of latter term.

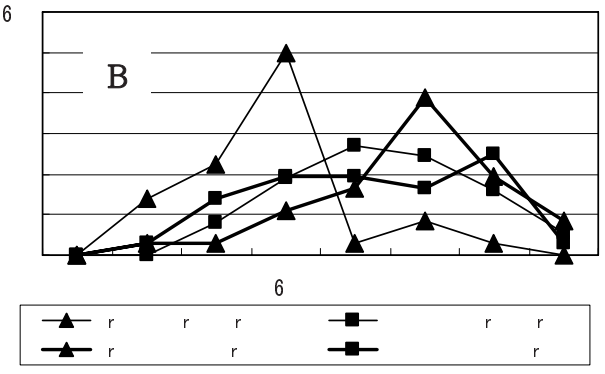


Fig. 4. Point distribution of class B

Figure 5 shows the score distribution of each test in class C. In this class, there are a lot of learners attending a class who have various problems in study. It seems reflect in the test result. The result of the first term pretest is divided into three small groups that are centers on 40, 60, and 80 point level. In the test at the end of the first term, it divided into two groups that are 70 point level group and the low score group of 40 point level. The low score group had a problem in the greediness for learning and also showed a low score at latter term. In the test at the end of latter term, though there was somewhat small number of 80 point level, low score distribution that centered on 50 point level was observed on the whole.

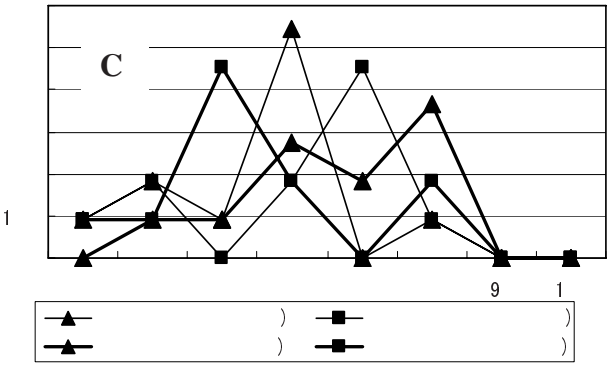


Fig. 5. Point distribution of class C

Class C consists of about 20 percent of the entire student. And most of them have a problem in the study of English or problems in basic knowledge. There were roughly two types of participants. The attitude of attending a lecture is serious though English language skill is low is one type. They review at home, and submit the confirmation vote almost every time. In the other type, there are a lot late and absence besides low English skill. From those results, it was shown that the educational environment of the class that belonged greatly influenced learner's greediness for learning and study result.

Figure 6 shows a transition of the average point of quizzes (basic problem). The number of problems set in class A was 10, that in class B was 15, and that in class C was 20. To compare those results, the score of each class was converted into 100 points. Because the

number of problems is different, it is not simply comparable but gives inspiration of the tendency. The average points of classes A, B, and C are 89.6, 81.3, and 79.9 respectively. The average point of class A is high and steady. The average point of class B is not too low, but has instability in the study. The average point of class C varies widely, and the score is low on the whole.

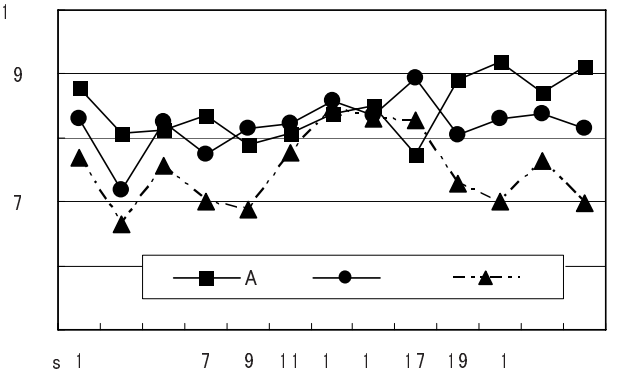


Fig. 6. Average point change

Figure 7 is an average rate of the confirmation of the wrong answer of quizzes described in the confirmation vote that the learner submitted. The average confirmation rate at latter term of class A was very high, 93.0%, though it decreased somewhat at the end of the first term. And that of class B was 80.0%, and obviously there are a difference between class A and class B. The average wrong answer confirmation rate of class C varies widely very much compared with class A and class B. Class A surely executed the process that had been directed by the self-monitor study which is confirmed from the rate of the confirmation of the wrong answer of quizzes. As the result, though a problem that was more difficult than other classes was given, the result was high. However, because there were a lot of learners with low greediness for learning in class C, the learning environment of the class was negatively affected.

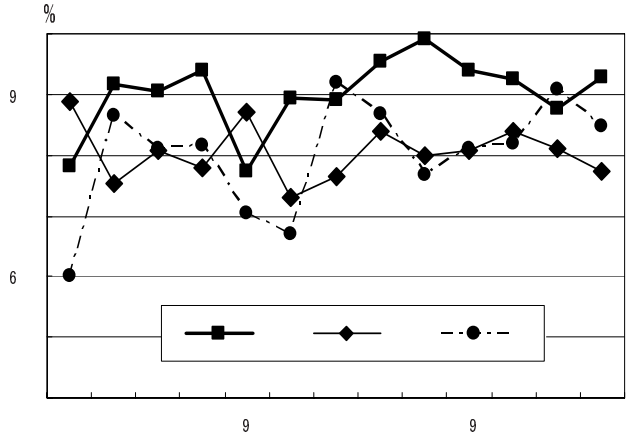


Fig. 7. A change of each class average erroneous answer confirmation rate

5 Conclusion

In this study, the learning system that combined the self-monitor study with the streaming was constructed, and the education effect was examined. As the result, fairly good effect was obtained.

However, those experimental results are the results of the analysis from the record that the learner self-declared to the confirmation vote and the confirmation seat.

To adopt the self-monitor study for an actual class in the future, the analysis of more detailed data is necessary.

References

[1] Ozaki, M., Sugimura, A., Adachi, Y.: Influence that learner's self management exerts on study. Bulletin of Chubu University 19(1.2), 67–82 (2005)

[2] Hashimoto, S., Ozaki, M., Koyama, K., Takeoka, S., Adachi, Y.: As for the influence that learner's knowledge and ability give to "Information education". In: Proc. 21 annual meeting of Japanese educational informatics association, pp. 180–183 (2005)

[3] Sugimura, A., Takeoka, S., Ozaki, M.: Effect that self-monitor causes for English study. In: Humanities and social chapter, Bulletin of Nagoya Women's College, vol. 52, pp. 133–144 (2006)

[4] Sugimura, A., Takeoka, S., Ozaki, M.: Effect of the self-monitor causes for English study(the second report). In: Humanities and social chapter, Bulletin of Nagoya Women's College, vol. 53 (in printing)

Influence of Presence of Frame on Writer Recognition

Yoshinori Adachi[1], Masahiro Ozaki[1], Yuji Iwahori[1], and Naohiro Ishii[2]

[1] Chubu University, 1200 Matsumoto-Cho, Kasugai, Aichi, Japan 487-8501
ozaki@isc.chubu.ac.jp, adachiy@isc.chubu.ac.jp
[2] Aichi Institute of Technology, Yakusa-Cho, Toyota, Aichi, Japan 470-0392
ishii@aitech.ac.jp

Abstract. There is a custom that a lot of people put out the New Year greeting postcard at the New Year's Day in Japan. The sentences of "　賀　年" or "　　して　めで　う　　す" are well used as objections of the New Year greeting postcard. It is general that those characters are written without frames usually. In this work, those characters were used for writer recognition. Especially, the influence that the presence of the frame exerted on the recognition accuracy was examined. The new local arc method was used as a feature extraction method. As a result of the examination, it has been understood that the presence of the frame has great influence on the character shape and that a dictionary made from the framed character has to be used.

1 Introduction

We have been studied writer recognition using hand-written characters up to now by several methods [1-7]. In the research process, it has been confirmed that the new local arc method (using curvature distribution of the strokes in a character) is excellent in the feature extraction, and moreover accuracy of writer recognition improves remarkably by comparing three characters at the same time, which are different kinds of characters [4-7]. However, the character data used in these studies was made from characters written in squares, and the influence on the shape of the character caused by the square frame has been disregarded. Externals of the character of Japan, China, and South Korea are near the square. Therefore, it is not easy to think that the existence of the frame exerts a big influence on the writer recognition. However, it can be guessed that the existence of the frame causes the writer a psychological effect and makes an original writing habit disappear from the character written down.

In this work, 14 characters "あ　　め　と　" in hiragana and "　　年" in Chinese character popularly used for the New Year greeting postcard were collected by two types of sheets (framed sheet and non-framed sheet), and the influence of the presence of the frame exerted on the writer recognition was examined.

2 Collection of Characters

The character collection form with frames and the character collection form without frames were used for the collection of characters. In the framed form, the four kinds

B. Apolloni et al. (Eds.): KES 2007/ WIRN 2007, Part II, LNAI 4693, pp. 1045–1050, 2007.
© Springer-Verlag Berlin Heidelberg 2007

of square frames, 9mm, 14mm, 18mm, and 23mm side, were prepared to be able to collect different size of characters. In the non-framed form, the filling in column in which the phrase was written down was framed to specify the rough size of a character. The examples of the form used are shown in Figures 1 to 2.

Subjects were nine seniors, and they wrote down the characters six times for each character. As shown in the figures, two sizes of character, 14mm and 18mm that is a little bigger than the usual writing, were written down according as the column width in the non-framed form. And in the case of framed form, four sizes of character, 9mm, 14mm, 18mm, and 23mm, were written down in the specified squares. Then, the influence of the size of character on the extraction of characteristic features was also made to be able to examine.

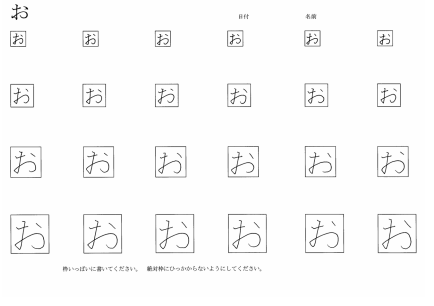


Fig. 1. Example of framed form

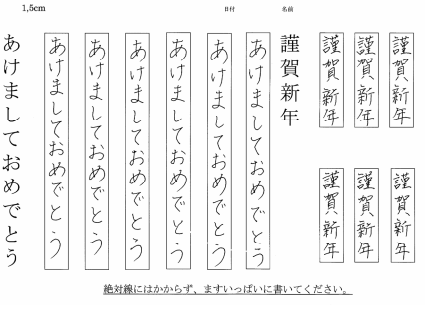


Fig. 2. Example of non-framed form (with 14mm framed column)

3 Extraction of Characteristic Feature

The new local arc method that we had developed before was used for the extraction of the characteristic feature of hand written characters. The six kinds of chord lengths that formed the circular arc were used, from 5 dots to 25 dots by every 4 dots. Then, the proper chord length was examined for extraction of characteristic features. The detail of the local arc method was shown in the previous study [6].

The degree of similarity was calculated from the inner product of curvature distribution vectors of the dictionary and of the test character. If the degree of similarity was less than 0.5, the writers of the dictionary and of the test character were considered to be unconditionally different. And if it was larger than 0.5, the writer who gave the largest degree of similarity was considered to be the writer in the case of one character judgment. Furthermore, final writer recognition was done by the new similarity evaluation function using three different types of character as shown by Equation (1).

$$\eta = 1-(1-\eta_1)(1-\eta_2)(1-\eta_3) \tag{1}$$

4 Experiment Result

4.1 Recognition Result of the Framed Character

The recognition rates of four character size, 9mm, 14mm, 18mm, and 23mm, and 6 chord lengths, 5 dots, 9 dots, 13 dots, 17 dots, 21 dots, and 25 dots, are depicted in Figures 3 and 4. From Figure 3, it is understood that the recognition ratio has decreased with an increase in the chord length, and it has increased with an increase in the character size. From Figure 4, the recognition ratio of the Chinese character is better than that of the hiragana. And the recognition ratio of the hiragana has decreased with an increase of the chord length but that of the Chinese character has not changed. From these results, the recognition ratio increases with the character size increases for the framed character (Especially, in the Chinese character). Shortening the chord length increases the recognition ratio for the hiragana but not for the Chinese character.

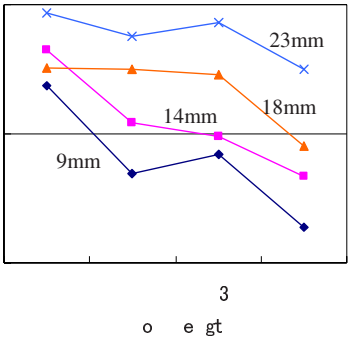


Fig. 3. Recognition ratio change according as chord length change

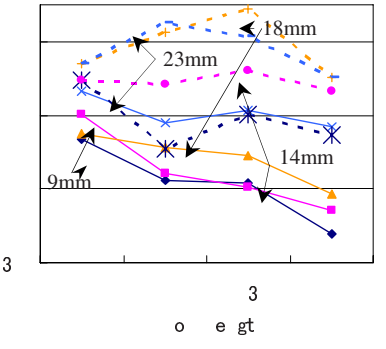


Fig. 4. Recognition ratio changes of hiragana and Chinese character according as chord length change. Solid line is hiragana and broken line is Chinese character.

4.2 Recognition Result of Non-framed Character

There are two kinds of character sizes, 14mm and 18mm. The results are shown in Figures 5 and 6. Comparing with the framed character, the characteristic feature of non-framed character is not steady, and the recognition ratio has lowered. Moreover, in the case of the hiragana, it is hard to write big characters steady and the recognition ratio has decreased. But in the case of the Chinese character, because of a lot of strokes existing, the recognition ratio has increased with an increase of the character size.

4.3 Examination of Non-framed Character by Framed Dictionary

Two sizes of non-framed character were examined by the four sizes of framed dictionaries. The result is shown in Figures 7 to 8. In the figures, D indicates

dictionary size and C indicates character size. The recognition ratio of the Chinese character increases with increases of the size of the dictionary as shown in Figure 7. Especially, it is remarkable for 23mm size of character. A smaller chord length gives better recognition ratio for the Chinese character, but there is small change for the hiragana. The chord length does not influence recognition ratio as shown in Figure 8. Moreover, the recognition ratio of the non-framed character is roughly half of the one of the framed character with using the framed dictionary.

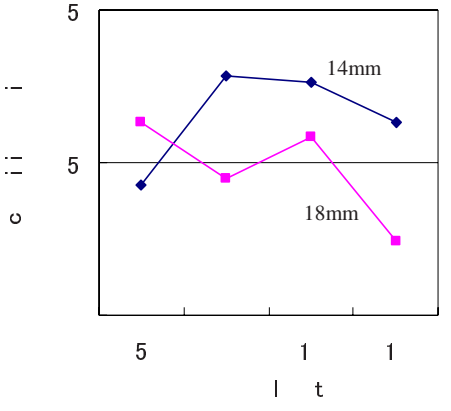


Fig. 5. Recognition ratio change according as chord length change

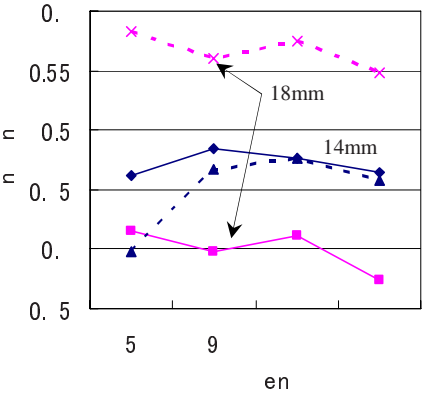


Fig. 6. Recognition ratio change of hiragana and Chinese character according as chord length change. Solid line is hiragana and broken line is Chinese character.

4.4 Examination of Framed Character by Non-framed Dictionary

A part of the result of examining framed character by the non-framed dictionary is shown in Figures 9 to 10. In Figure 9, the recognition ratio of the Chinese character is overall higher. The recognition ratio of the hiragana by 18mm dictionary is lower than that of 14mm dictionary. In Figure 10, high recognition ratio is obtained at small size of character, and no substantial change is obtained at big size of the dictionary. Anyway, the recognition ratio is very small about 0.25.

4.5 Result by New Similarity Evaluation Function

The accuracy improvement was attempted by the new similarity evaluation function, Equation (1), with three framed characters. The result is shown in Figures 11 and 12. From Figure 11, the recognition ratio of small character size is roughly same at the chord length 13 dots or less but it decreases at the chord length above 17 dots. From Figure 12, it is understood that the recognition ratio of the Chinese character is higher than that of the hiragana. In this case, it has been understood that the accuracy of the 18mm Chinese character of 13 dots chord length or that of the 14mm character of 5 dots chord length is high.

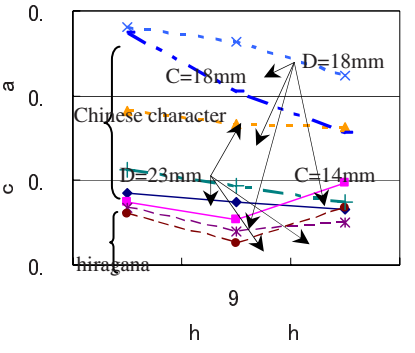


Fig. 7. Result of framed dictionaries (18mm and 23mm)

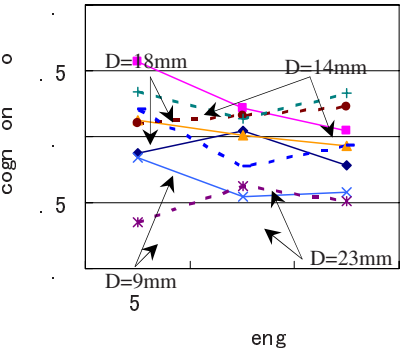


Fig. 8. Result of framed dictionaries for 2 character sizes (14mm and 18mm). Solid line is 14mm and broken line is 18mm.

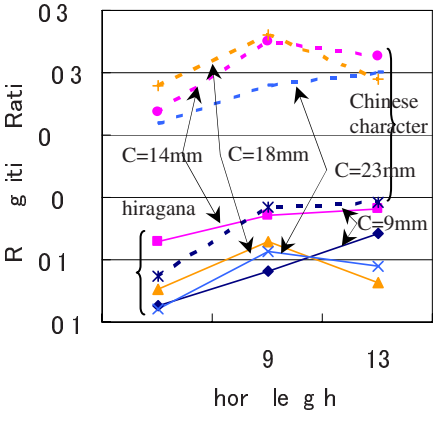


Fig. 9. Result of 18mm non-framed dictionary

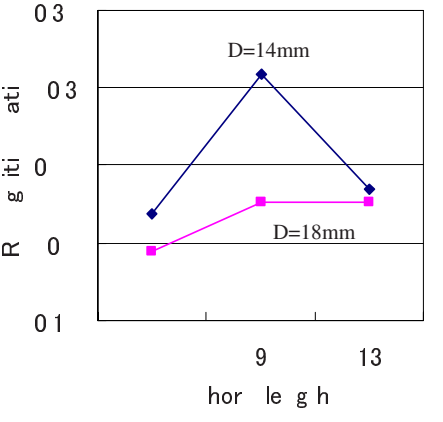


Fig. 10. Result of 14mm character

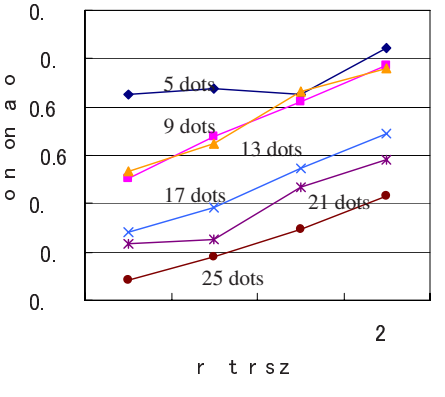


Fig. 11. Accuracy change by character size

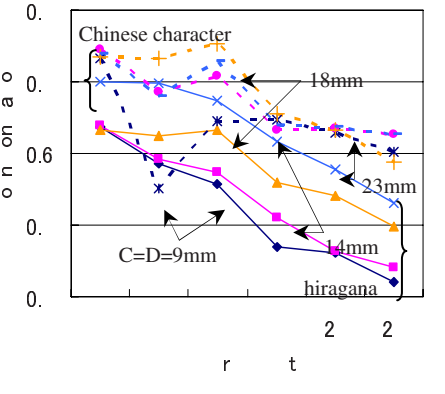


Fig. 12. Accuracy change by chord length

Similarly, the recognition ratio of three characters overlapping was examined about the no framed character. The influence of the character size change to the recognition ratio is small. And the result of 9 dots chord length is the best, and that of 25 dots is the worst. Moreover, it is understood that the recognition ratio of the Chinese character is higher than that of the hiragana at big character and the recognition ratio of chord length 9 dots is highest.

5 Conclusion

From the above-mentioned results, we obtained the followings:

(1) The presence of the frame influences even the shape of each character. And it has been understood that framed character is suitable for the writer recognition.
(2) In the range of character size studied, it has been understood that Chinese characters are more suitable for the writer recognition than hiragana.
(3) It is understood that the 13 dots chord length and the 18mm character size is most suitable for the writer recognition.

This time, each character was written for only 6 times, and then only 5 characters were used to make a dictionary. Because of this, the writer recognition ratio is only about 75%.

In the future, it will be necessary to increase the number of the characters and of writers, and to examine the influence of the character type, Chinese character or hiragana, of the chord length, and of the character size, on the writer recognition further.

References

[1] Ozaki, M., Adachi, Y., Ishii, N., Koyazu, T.: Fuzzy CAI System to Improve Hand Writing Skills by Using Sensuous Information. Trans. of IEICE J79-D-II(9), 1554–1561 (1996)
[2] Ozaki, M., Adachi, Y., Ishii, N.: Writer Recognition by means of Fuzzy Similarity Evaluation Function. In: Proc. KES 2000, pp.287–291 (2000)
[3] Ozaki, M., Adachi, Y., Ishii, N.: Study of Accuracy Dependence of Writer Recognition on Number of Character. In: Proc. KES 2000, pp.292–296 (2000)
[4] Ozaki, M., Adachi, Y., Ishii, N., Yoshimura, M.: Writer Recognition by means of Fuzzy Membership Function and Local Arcs. In: Proc. KES 2001, pp.414–418 (2001)
[5] Ozaki, M., Adachi, Y., Ishii, N.: Development of Hybrid Type Writer Recognition System. In: Proc. KES 2002, pp.765–769 (2002)
[6] Adachi, Y., Liu, M., Ozaki, M.: A New Similarity Evaluation Function for Writer Recognition of Chinese Character. In: Negoita, M.G., Howlett, R.J., Jain, L.C. (eds.) KES 2004. LNCS (LNAI), vol. 3215, pp. 71–76. Springer, Heidelberg (2004)
[7] Ozaki, M., Adachi, Y., Ishii, N.: Writer Recognition by Using New Searching Algorithm in New Local Arc Method. In: Khosla, R., Howlett, R.J., Jain, L.C. (eds.) KES 2005. LNCS (LNAI), vol. 3684, pp. 775–780. Springer, Heidelberg (2005)

Information Extraction by XLM

Masashi Okada, Naohiro Ishii, and Nariaki Kato

Aichi Institute of Technology
Yachigusa, Yakusacho, Toyota, Japan 470-0392, Japan
j03031@aitech.ac.jp, ishii@aitech.ac.jp

Abstract. XML(eXtensible Markup Language) is used as the description form of the documents and the data exchanged on the Web. Now, the usage of the XML is extending to not only the exchange of the XML document but also that of the XML database for classification and retrieval. This paper develops how to retrieve the related objective data from the tree structure in the XML document for the classification. In this paper, an ordered preserving relation is defined as the cue of the retrieval of the objective pattern. Then, the problem is to find the ordered preserving relations in the document. In the document, the ordered and the unordered subtree structure are important in the retrieval and classification. Then, the retrieval value is calculated in the tree structure. A method developed here was applied to the practical XML document.

Keywords: xml, information extraction, tree structure, embedded subtree.

1 Introduction

XML(eXtensible Markup Language) is a standard that W3C(World Wide Web Consortium) enacted to provide for the description form of documents and the data exchange on the Web. The usage of the XML is extending to that of the database like storage, the retrieval of the XML documents and it's data transfer, which were used basically for the exchange of the XML documents. The advantage of XML in the database can retrieve two or more similar databases and reuse, reconstruct parts of the data base as the integrated document. Then, it becomes the integrated database to store and to retrieve the database and contents more flexibly. This type of XML documents processing are extensively studied[1,2,3,4].

This paper develops how to retrieve the related objective data from the tree structure in the XML documents for the classification. In the document, the ordered and the unordered subtree structure are important in the retrieval. The retrieval of the both subtrees are developed. Then, the retrieval value is calculated in the tree structure. A method developed here was applied to the practical XML document. In this paper, an ordered preserving relation is defined as a cue of the retrieval of the objective pattern. Then, the problem is to find the ordered preserving relations in the document.

B. Apolloni et al. (Eds.): KES 2007/ WIRN 2007, Part II, LNAI 4693, pp. 1051–1058, 2007.
© Springer-Verlag Berlin Heidelberg 2007

2 Tree Structure of XML Document

The XML document and data have the layered structure that consists of the nest of the element (inclusion relation). Then, the tree structure model is regarded as a logical model of the XML document and the XML data. The element can be shown by labeled tree with only one route node. Figure 1 shows an example of showing the XML document by the tree structure. When a left XML document is shown by the tree structure, it becomes like the right. Here, a circle node is an element node with it's name and a square node is a text node with the text.

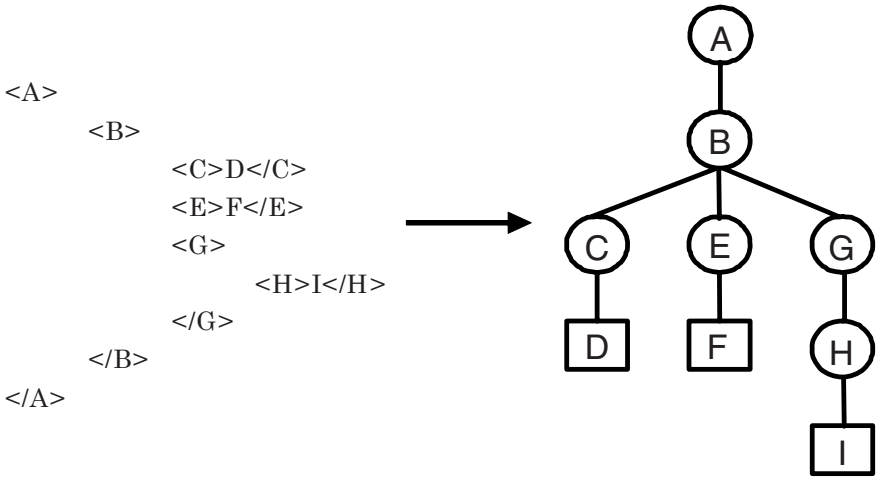


Fig. 1. Tree structure of XML document

The paper aims to retrieve the related objective subtree in the document (data) tree by giving typical subtree, which is called here a pattern tree. The related objective subtree is searched by matching algorithm of the pattern tree developed here in the document (data) tree. Then, the number of the searched subtrees in the data tree, is counted, which is called the number of supports. There are some trees in the document. The total number of the searched subtrees in the document, is called the umberof the weighted supports.

2.1 Embedded Subtree Pattern

A retrieval method is developed here by giving the objective subtree pattern. For the retrieval of the subtrees, a relation between the document (data) tree T and the objective subtree pattern S, is defined here. An embedded subtree is defined as follows. All nodes of the tree S are included also in the tree T. Assume that the node Nx is a parent node of the node Ny in the tree S. Then, a branch between Nx and Ny, is defined in the tree S as (Nx, Ny). We introduce here an ordered preserving relation (mapping) from the tree S to the tree T. The corresponding (mapping) nodes and branch in the tree T from Nx, Ny and (Nx, , Ny) in the tree S, are defined to be N'x

,N'y and (N'x, N'y), respectively. Note here that N'x becomes an ancestor of N'y in the order preserving relation. Thus, the problem is to find order preserving relations from S to T. An example of the order preserving relation is shown in Fig. 2.

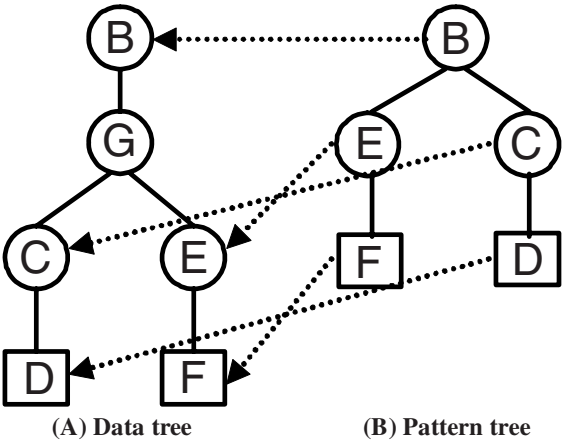


Fig. 2. Example of order preserving relation between trees

The nodes B and C in the pattern tree(B) in Fig. 2, show a parent–child relation, which connects B and C, directory. Similarly, the nodes B and E shows another parent-child relation in the pattern tree (B) in Fig. 2. The nodes B and C in the data tree, shows a parent- grandchild relation, since the node G exists between B and C in the data tree. These parent–child, parent-grandchild, parent- great grandchild relations et al are called here ancestor relations. Similarly, the nodes B and E in the data tree, shows another ancestor relation.

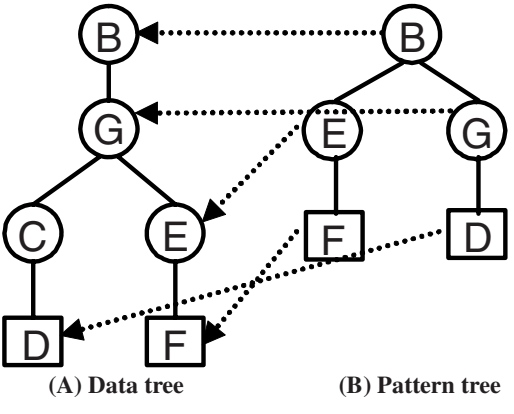


Fig. 3. Example which does not include embedded subtree

The nodes B and C or B and E, which have parent-child relation in the pattern tree, correspond to the nodes Band C or B and E in the data tree, respectively, by dotted arrows as shown in Fig. 2. Thus, an embedded subtree of the pattern (B) is found in the data tree (A). Since the pattern tree (B) is assumed here to be unordered between nodes C and E, then the relation in the data tree, is permitted as the correspondence of the nodes between E and C in the pattern tree.

Fig. 3 shows an example, which is not an embedded subtree in the data tree. Though the nodes E and G have a brother relation(on the save level in the tree) in the pattern tree(B), the corresponding nodes G and E in the data tree (A), do not hold order preserving relations as a parent relation ,or an ancestor relation.

2.2 Support and Weighted Support

To characterize documents, it is important to count the number of occurrences of the embedded subtree in the document data trees. The value of the support becomes to be 1, when the pattern tree appears as an embedded subtree in one objective data tree.

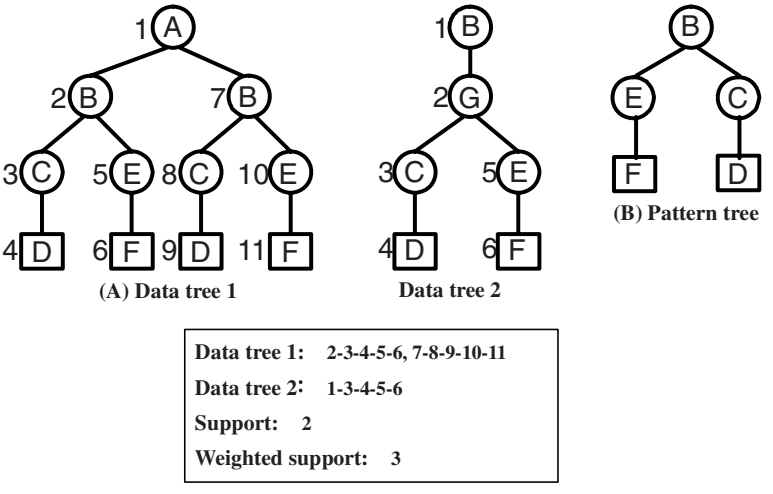


Fig. 4. Example of support and weighted support

In one data tree, the pattern tree might appear two or more times as the embedded subtree. Then, the total number of those occurrences is defined as a weighted support among data trees. An example of the support and the weighted support are shown in Fig. 4. The embedded subtrees in the data tree 1 in Fig. 4, are found as two sequences of the numerical value of node: 2-3-4-5-6 and 7-8-9-10-11, which correspond to the pattern tree of right side, while only one embedded subtree is found in the data tree 2 as that of 1-3-4-5-6. Thus the value of the support becomes 2 by the data tree 1 and the data tree 2, while the value of the weighted support shows 3.

3 Search Algorithm for Embedded Subtree

A search algorithm whether the pattern tree appears in the data tree, is developed here. First, the array of MatchNode with the node CorrespondNode in the data tree, is made, which corresponds to the node SearchNode in the pattern tree. The number of elements of nodes in the array is the number of nodes in the pattern tree. The pattern tree is sequentially substituted from the 0th element of the array in SearchNode of each MatchNode. The foolowing operations are repeated from the root node of the pattern tree.

1) Corresponding label to the node in the pattern tree is searched in the data tree by ordered scanning of the pattern tree.
2) When a corresponding node is discovered, a corresponding node to CorrespondNode of MatchNode is substituted.
3) CorrespondNode of MatchNode corresponding to the last index is set from the following index of the index in the array of the MatchNode to null.
4) If the first child exists in SearchNode, the search procedure is carried out from 1) to 6) in the hierarchy that is deeper than the position of CorrespondNode.
5) If the brother node exists in the next, the search procedure is carried out from 1) to 6) in the hierarchy that is deeper than the position of CorrespondNode corresponding to the parents node of SearchNode.
6) Afterwards, the search procedure is carried ouit in the data tree according to SearchNode.

The embedded subtree is judged to be true, if the number of MatchNode with CorrespondNode is equal to the number of nodes of pattern trees. Moreover, when the node that the fraternity of the pattern tree has becomes an ancestor descendant relation in the data tree, the subtree is not counted. Moreover, it is assumed that the subtree is not counted, when the node of the data tree corresponding to the brother node exists in the hierarchy less than the node with corresponding label to that in the pattern tree. Iteration of searching algorithms is carried out for the classification of documents. The key words are computed[6,7], which are represented as nodes in the tree of documents. Then, the pattern tree based on the key word is made as the parent node, child node, et al. Then, several pattern trees are made as the cues for the characterization of the document. Basic processing here is the search of the embedded subtrees, which satisfy the relations in the pattern tree.

4 Practical Application of XML

XML retrieval developed here, was applied to practical data on the Web, which is Reed University XML data. This data contains the educational campus data as course name, credit unit, room, time, instructor, building et al. Under some conditions, the objective search was carried out. A part of the XML document is shown in Fig. 5.

```
<?xml version='1.0' ?>
        <root>
              <course>
                   <reg_num>10577</reg_num>
                     <subj>ANTH</subj>
                     <crse>211</crse>
                     <sect>F01</sect>
                                 <title>Introduction to Anthropology</title>
                           <units>1.0</units>
                                <instructor>Brightman</instructor>
                            <days>M-W</days>
                              <time>
                                                <start_time>03:10PM</start_time>
                                                <end_time>04:30</end_time>
                              </time>
                              <place>
                                                <building>ELIOT</building>
                                      <room>414</room>
                              </place>
                    </course>
                     <course>

                     </course>
                     <course>

          </root>
```

Fig. 5. Example of campus data on the Web

```
<?xml version="1.0"?>
    <course>
        <units>1.0</units>
        <room>414</room>
        <building>ELIOT</building>
        <days>M-W</days>
    </course>
```

Fig. 6. Searched document on the campus data

The XML searched document for the retrieval is shown in Fig. 6. The objective XML document is searched on the campus data. Then, the support and the weighted support are calculated. The tree structure of the searched pattern in Fig. 6 is shown in Fig. 7.

From the result in Fig. 8, the discovered embedded subtree, ‘building’ and ‘room’ has the node parent of ‘place’ in the document data tree on the campus, while the pattern tree in Fig. 6, has not it’s word. The searched algorithm shows a discovery of the ancestor relation in the data tree. Further, the support and weighted support were calculated on the campus data as shown in Fig. 5. Then, the value of the support was 3, while the vale of the weighted support was 6 on the document database. This shows that the document data has 3 trees and the total of 6 embedded subtrees exists in these 3 trees.

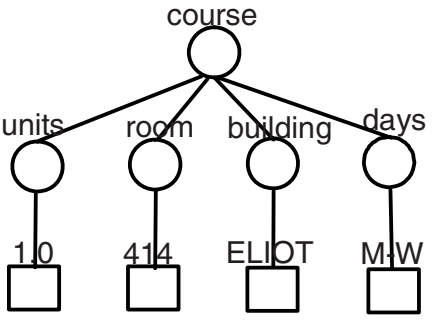


Fig. 7. Pattern tree for the retrieval

```
<course>
                    <reg_num>10577</reg_num>
                    <subj>ANTH</subj>
                    <crse>211</crse>
                    <sect>F01</sect>
                <title>Introduction to Anthropology</title>
                <units>1.0</units>
                <instructor>Brightman</instructor>
                <days>M-W</days>
                <time>
                    <start_time>03:10PM</start_time>
                    <end_time>04:30</end_time>
                </time>
                <place>
                                    <building>ELIOT</building>
                                    <room>414</room>
                </place>
    </course>
   <course>
                <reg_num>20572</reg_num>
                <subj>ANTH</subj>
                <crse>369</crse>
                <sect>S</sect>
                  <title>Media and Culture Post-MaoChina</title>
                 <units>1.0</units>
                 <instructor>Makley</instructor>
                 <days>M-W</days>
                 <time>
                          <start_time>03:10PM</start_time>
                          <end_time>04:30</end_time>
                 </time>
                 <place>
                          <building>ELIOT</building>
                          <room>414</room>
                 </place>
   </course>
```

Fig. 8. Example of searched results from document on the campus

5 Conclusion

In this paper, the tree structure model was adopted for the retrieval of the embedded subtree by giving the pattern tree, which is typical pattern to be searched in the document. The searched embedded tree is permitted to have the ancestor relation between nodes, while the given pattern has parent-child relation. The searched algorithm was developed here and was applied to the practical problem as the campus XML document to discover the given information. The support and the weighted support were computed as the number of the embedded subtrees in the document. The discovery of the embedded subtrees, will be useful to characterize the documents and to classify it.

References

1. Zaki, M.J.: Efficiently Mining Frequent Trees in a Forest: Algorithms and Applications. IEEE Trans. on Knowledge and Data Engineering 17(8), 1021–1035 (2005)
2. Zaki, M.J.: Efficiently Mining Frequent Embedded Unordered Trees. Fundamenta Informatica 65, 1–20 (2005)
3. Zaki, M.J., Aggarwal, C.C.: XRules, An Effective Structural Classifier for XML Data. In: Proc. of the 2003 Int. Conf. Knowledge Discovery and Data Mining, pp. 316–325 (2003)
4. Asai, T., Arimura, T., Uno, T., Nakano, S.: Discovering Frequent Substructures in Large Unordered Trees. In: Proc. Sixth Int. Conf. Discovery Science, pp. 47–61 (October 2003)
5. Chi, Y., Yang, Y., Munz, R.R.: Indexing and Mining Free Trees. In: Proc. Third IEEE Int. Conf. Data Mining, pp. 509–512. IEEE Computer Society Press, Los Alamitos (2003)
6. Bao, Y., Tsuchiya, E., Ishii, N., Du, X.: Classification by Instance-Based Learning Algorithm. In: Gallagher, M., Hogan, J.P., Maire, F. (eds.) IDEAL 2005. LNCS, vol. 3578, pp. 133–140. Springer, Heidelberg (2005)
7. Yamada, T., Yamashita, K., Ishii, N., Iwata, K.: Text Classification by Combining Different Distance Functions with Weights. In: SNPD2006, pp. 85–90. IEEE Computer Soc. Publication, Los Alamitos (2006)

The Characteristics Evaluation of Stacked-Type Electrostatic Actuators by the Neural Network

Masashi Kawaguchi[1], Katsuhisa Yamano[1], Kazuo Okuda[1], Takashi Jimbo[2], and Naohiro Ishii[3]

[1] Department of Electrical & Electronic Engineering, Suzuka National College of Technology, Shiroko, Suzuka Mie 510-0294, Japan
{masashi, okuda}@elec.suzuka-ct.ac.jp
[2] Department of Environmental Technology and Urban Planning Graduate School of Engineering, Nagoya Institute of Technology, Gokiso-cho, Showa-ku, Nagoya, 466-8555 Japan
jimbo.takashi@nitech.ac.jp
[3] Department of Information Network Engineering, Aichi Institute of Technology, Yachigusa, Yagusa-cho, Toyota, 470-0356 Japan
ishii@aitech.ac.jp

Abstract. We developed the electrostatic and layered actuator with the structure which alternately tucked 2 ribbons film electrode. The aim is for the application to the artificial muscle. We proposed the prediction system of spring characteristic which is beforehand estimated in order to develop the actuator efficiently. This system was almost completed as a system of the prediction. However, the prediction result is sometimes not suitable the practicable accuracy.

Keywords: Electrostatic Actuator, Neural Network, Spring Characteristic, Stacked-type.

1 Introduction

Recently, many kinds of electrostatic actuators have been developed for application in Micro-Electro-Mechanical Systems (MEMS) [1-3]. An electrostatic actuator built from many small-sized actuators can generate a sufficiently large force, because the electrostatic force is intensified by the narrowing of the electrode interval. Electrostatic actuators are of two types, parallel and perpendicular, according to the difference of the direction in which the movable electrodes operate. In the parallel-type actuator, the movable electrodes operate in parallel, with the electrode interval kept constant.

The actuator is structured in such a way that two insulated thin plastic films are alternately folded around each other; the electrode part is thick, and the hinge part is thin. The spring characteristics of the actuator obtained in the simulation

B. Apolloni et al. (Eds.): KES 2007/ WIRN 2007, Part II, LNAI 4693, pp. 1059–1065, 2007.
© Springer-Verlag Berlin Heidelberg 2007

corresponded almost entirely to the measurement results, and it was confirmed that this spring characteristic is crucial for the stacked-type electrostatic actuator to work.

It is required that it is pliant in the operating region and it is inexpensive over the necessity. Therefore, it is necessary to carry out measurement and improvement of the spring characteristic. In the reason, the trial manufacture which changes various data of the actuator is required many times. In order to develop the actuator efficiently with the better spring characteristic, it is desired to develop the system which can estimate the actuator with the good characteristic in a basement by various data of existing actuator.

In this research, we predicted the spring characteristic of producing actuator using the measurement data of stacked-type electrostatic actuator by multi-layered neural network.

2 Electrostatic Actuators and Neural Network

The ribbon films used for the actuator are each composed of a metallic thin film lined on both sides by thin insulating plastic films. Because the stiffness of the actuator constructed out of only these thin ribbon films is very low, the length of the actuator can be varied greatly by applying a very small external force.

2.1 Structure of Stacked-Type Electrostatic Actuators

Fig.1 shows the manufacture method of the actuator. The details are as follows:

1. A part of the ribbon is fixed and is alternately folded.
2. Plus and negative voltage is applied.
3. The contraction power by the electrostatic is generated.

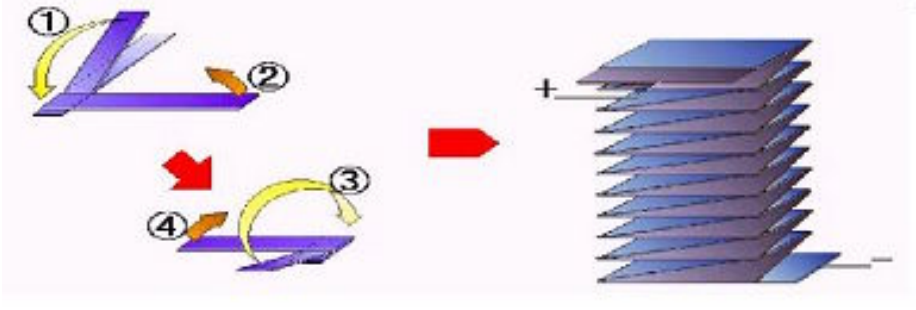


Fig. 1. The manufacture method of the actuator

Fig.2 shows the outline of the ideal spring characteristic of the actuator. There are three regions for operation, before operating region, Operating region and Overload region.

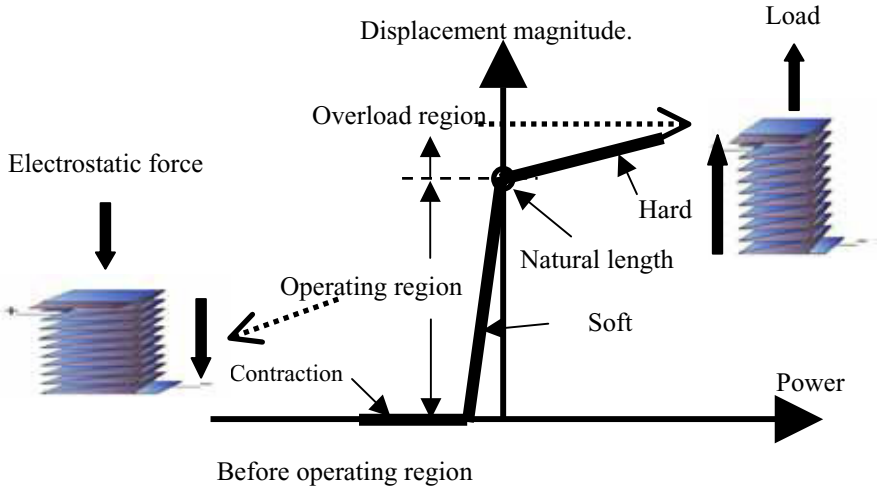


Fig. 2. The outline of the ideal spring characteristic

2.2 Neural Network

We constructed neural network system as shown in Fig. 3. This neural network consists of three layers. There are 8 units in the first layer. 8 units mean the 8 measurement factors. There are another three units in the second and third layer. Three units in the third layer mean the spring characteristics of each factor, Load in before operating region, Load in the operating region and load in the overload region. We have predicted the spring characteristics at each region of this stacked-type electrostatic actuator.

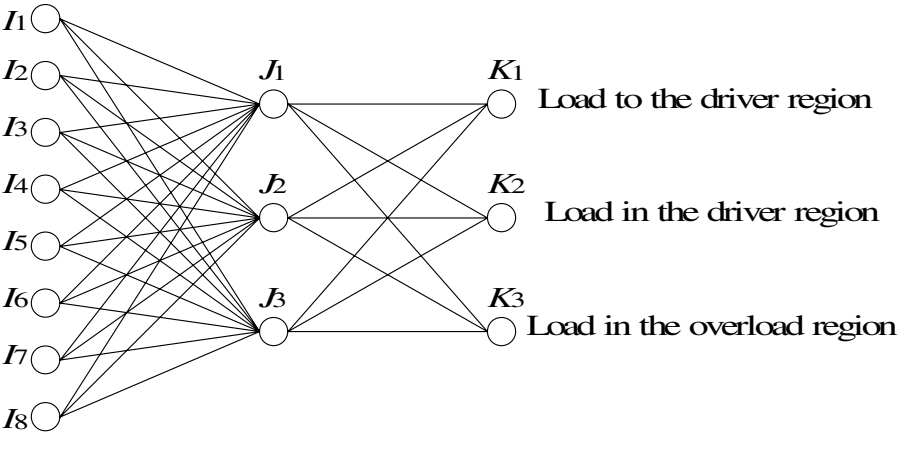


Fig. 3. Structure of Neural Network of Spring Characteristics Prediction System

2.3 Learning and Prediction Using Neural Network

We constructed neural network system to predict spring characteristics. We used here back-propagation model for data learning [4]. Input signal is used to measure each factor. We predicted the spring characteristics of each factor, Load in before operating region, load in the operating region and load in the overload region. The example of learning data is shown in Table 1. Table 2 means the teaching signal for the each region of measured data. These spring characteristic data are calculated from Displacement magnitude and Load. Equation (1) and equation (2) are used for learning of this network.

$$w_{ij} = w_{ij} - 0.3\sum(e_i j_j (1 - j_j) i_i) \tag{1}$$

w is connecting weight between input layer and middle layer. *j* is output value of middle layer. *i* is input value of input layer. *e* is learning error.

$$v_{ij} = v_{ij} - 0.3\sum(e_i k_j (1 - k_j) j_i) \tag{2}$$

v is connecting weight between middle layer and output layer. *k* is output value. *j* is output value of middle layer. *e* is learning error. These equations are typical learning rules of each learning time.

Table 1. The example of learning data

	Data1	Data2	Data3	Data4	Data5
Width of a film(mm)	5	5	7.5	7	7
Temperature of the heat treatment(C^o)	80	120	140	100	160
Displacement magnitude in before operating region(mm)	0.521	0.112	0.143	0.170	0.082
Displacement magnitude in operating region(mm)	1.758	1.823	1.944	2.187	1.878
Displacement magnitude in overload region(mm)	3.311	3.755	3.161	3.583	3.052
Load in before operating region(mN)	9.457	10.780	16.333	8.428	8.428
Load in the operating region(mN)	16.072	24.010	27.440	22.148	27.440
Load in overload region(mN)	39.298	39.288	34.055	32.732	34.055

3 Experiment

We proposed the prediction system of spring characteristic which is beforehand estimated in order to develop the actuator efficiently using neural network. The improvement on the stacked-type electrostatic actuator is carried out with the aim of the application to the artificial muscle. We developed the electrostatic and layered actuator. The structure of stacked-type electrostatic actuator is alternately tucked 2 ribbons film electrode [5].

1. The measurement data is collected from previous actuator. We measured the width of a film, the existence of the heat treatment, the temperature of the heat treatment, the load in before operating region, the load in the operating region, the load in the overload region, the displacement magnitude before operating region, the displacement magnitude in operating region and the displacement magnitude in the overload region.
2. Making the neural network, learning signal is measured data from the previous actuator.
3. In general, neural network can accept the input value in between 0 and 1. So, we converted the measurement data to the input data, which is between 0 and 1. These values are reported in Table. 3

Table 2. Spring characteristic data as the teaching signal of neural network

	Data1	Data2	Data3	Data4	Data5
Load in before operating region	19.025	97.402	117.085	54.781	103.989
Load in the operating region	9.312	16.404	16.362	10.479	16.742
Load in the overload region	11.846	10.447	10.775	9.126	11.158

Table 3. The conversion style for changing the input singal from 0 within 1's

	Conversion style
Width of a film	$I_1=0.05l_1$ (mm)
Temperature of the heat treatment	$I_2=0.005t$ (C^o)
Displacement magnitude in before operating region	$I_3=l_2$(mm)
Displacement magnitude in operating region	$I_4=0.3l_3$ (mm)
Displacement magnitude in the overload region	$I_5=0.1l_4$ (mm)
Load in before operating region	$I_6=0.05F_1$ (N)
Load in the operating region	$I_7=0.01F_2$ (N)
Load in the overload region	$I_8=0.01F_3$ (N)

We show the prediction results for the experiment. The data of Table 1 and Table 2 was used for the experiment. Spring characteristic (teaching signal) uses 3 points, operating region, in the overload region and to operating region. The data used for the learning of various electrode sizes, 5mm, 7mm and 7.5mm. Thickness the electrode is 38μm. We used electrode as the Polyethylene terephthalate film, PET. The spring characteristic is calculated using following equation (3). K is spring Characteristic. P is road and δ is displacement magnitude.

$$k = \frac{P}{\delta} \tag{3}$$

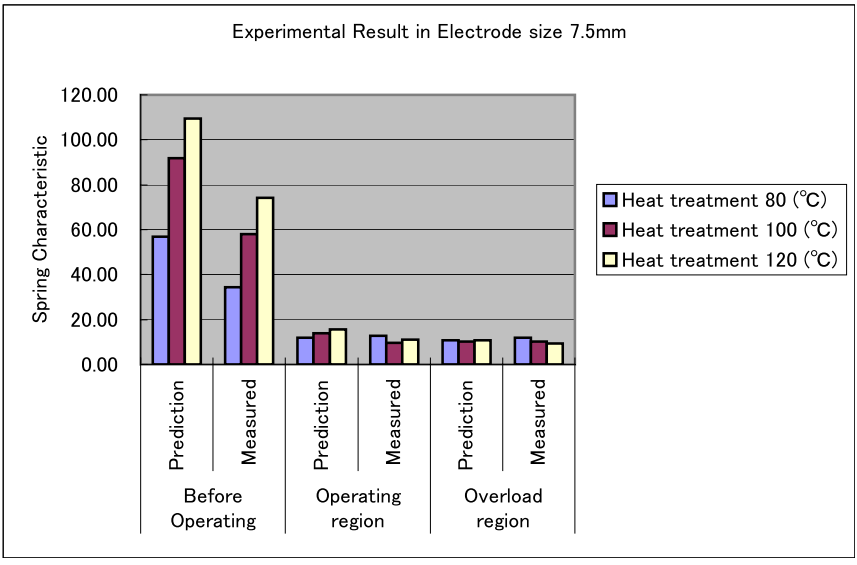


Fig. 4. The outline of the ideal spring characteristic

Fig. 4 shows the one of the example of experimental result to predict the spring characteristic. The electrode size is 7.5mm. There are three regions, before operating region, operating region and overload region. There are three different temperatures for heat treatment, 80, 100 and 120 Centigrade. In the input signal of this neural network, we used the learning signal data as the Displacement magnitude and Load. Thus, this neural network is stabilized.

4 Discussion

The average error is 9.77% in load region, and overload region with 10 test data. It is not the practicable accuracy. However the result in the fixed condition with small error is also observed. We used only five learning data, so we have to use much more learning data.

There is sometimes large error in to the operating region. The slight difference of the prediction becomes large error, because the measured value is small in before operating region. The element of the actuator gives large effect in the slight change of the characteristic. The future problem is as follows. It is necessary to obtain more much accurate and detailed data. Furthermore, the program is improved in order to handle much data.

Acknowledgments. This research was financially supported by Grant-in-Aid from the Japanese Ministry of Education, Science, Sports and Culture.

References

1. Shutov, M.V., Howard, D.L., Sandoz, E.E., Sirota, J.M., Smith, R.L., Collins, S.D.: Electrostatic inchworm microsystem with long range translation. Sensors and Actuators A 114, 379–386 (2004)
2. Matsunaga, T., Kondoh, K., Kumagae, M., Kawata, H., Yasuda, M., Murata, K., Yoshitake, M.: Fabrication of a New Electrostatic Linear Actuator. Jpn. J. Appl. Phys. 39, 7115–7119 (2000)
3. Chu, P.B., Nelson, P.R., Tachiki, M.L., Pister, K.S.J.: Dynamics of polysilicon parallel-plate electrostatic actuators. Sensors and Actuators A 52, 216–220 (1996)
4. Rumellhart, D.E., Hinton, G.E., Williams, R.J.: Learning internal representations by error propagation. In: Parallel distributed processing, pp. 318–362. MIT Press, Cambridge (1986)
5. Yamano, K., Okuda, K., Kawaguchi, M.: The Characteristics Evaluation of Stacked-type Electrostatic Actuators by the Neural Network. Record of Tokai-Section Joint Conference of the Eight Institutes of Electrical and Relater Engineers, P-058 (2006)

A Piecewise Linear Representation Method of Time Series Based on Feature Points

Yuelong Zhu, De Wu, and Shijin Li

Hohai University, Nanjing, Jiangsu 210098 China
{ylzhu, Wdcc12, lishijin}@hhu.edu.cn

Abstract. In recent years, there has been an explosion of interest in mining time series databases. Representation of the data is the key to efficient and effective solutions. One of the most commonly used representation is piecewise linear approximation, which has been used to support clustering, classification, indexing and association rule mining of time series data. In this paper, we propose a method of piecewise linear representation (PLR) based on feature points. Experiment shows that the method has less fit error to the original time series and has a better ability of adaptation, which can be applied to diverse data environments.

Keywords: time series, piecewise linear representation, feature points.

1 Introduction

Time series is a kind of important data type and is ubiquitous in many fields such as economy, finance, stocks and hydrology etc. Analysis on these data can reveal their intrinsic laws and help us make scientific decisions. Due to such features as high dimension, large quantity and severe noises of time series, mining directly on them may not only pay high costs on their storage and computation, but also affect the accuracy and reliability of the algorithm. Representation of the data is the key to efficient and effective solutions. Many researchers proposed pattern representation of time series to describe their main shapes and ignore the local details.

Several high level representation of time series have been proposed, including Fourier Transforms [1], Wavelets [2] Symbolic Mapping [3,4,5] and Piecewise Linear Representation (PLR). In this paper, we confine our attention to PLR, perhaps the most frequently used representation. Intuitively PLR refers to the approximation of a time series T, of length n with K straight lines. Because K is typically much smaller than n, this representation makes the storage, transmission and computation more efficient.

In this paper, we propose a method of PLR based on Feature Points (PLR_FP) to deal with the time series with obvious periodicity and violent pattern fluctuation during a specific period. The result of experiment shows that our method has a good ability of adaptation.

The rest of the paper is organized as follows. In section 2, we provide the related work briefly; in section3, we propose the method of time series PLR based on Feature

B. Apolloni et al. (Eds.): KES 2007/ WIRN 2007, Part II, LNAI 4693, pp. 1066–1072, 2007.
© Springer-Verlag Berlin Heidelberg 2007

Points (PLR_FP); section 4 carries out the experiment based on the section 3; section 5 offers summary and conclusions.

2 Related Work

For the method of PLR, the number of line segments determines the granularity of approximation. Keogh [6] calls the algorithms, which input a time series and return a PLR, as Linear Segmentation Algorithms.

[7] and [8] independently suggested the method of Piecewise Aggregate Approximation (PAA). Keogh's experiments show that PAA is simple, intuitive and is surprisingly competitive with the more sophisticated transforms, such as DFT, DWT and SVD. Park et al. suggested modifying the Sliding Window algorithm to create "monotonically changing" segments [9]. They also introduced a modification where they first perform a scan over the entire dataset marking every peak and valley [10]. These extreme points are used to create an initial segmentation, and the Top-Down algorithm is applied to each of the segments. These modifications worked well on the smooth synthetic dataset it was demonstrated on. But on real world datasets with any amount of noises, the approximation is greatly over-fragmented. In data mining, the Bottom-Up algorithm has been used extensively by two of the current authors to support a variety of time series data mining tasks [11,12,13]. Some other researchers select some data points that have obvious characteristics to segment time series data through heuristic rules. In General, this kind of algorithm has no restrictions on the number / error of segments. Peng et al. used landmark model to achieve the segmentation of time series [14]. Pratt and Frink took advantage of important points to segment time series [15]. Important points are defined as the extremal points in local scales where the ratios of them to end points are greater than parameter R . Time series can be approximated at different levels according to the values of R .

Focusing on the time series with features: a) obvious periodicity; b) violent pattern fluctuation during a specific period, we introduce a method of time series PLR based on feature points (PLR_FP), which is simplicity, intuitive and can keep the trends of changing patterns effectively.

3 PLR Based on Feature Points

As mentioned above, we focus on time series with the following features (as is showed in Fig.1.):

a) obvious periodicity; b) violent pattern fluctuation during a specific period.

In Fig.1, the amplitudes of the data between the periods are stable and smooth relatively, while during each specific period they change violently and frequently. The former is susceptible to noises and for which the main task is to filter them. And the later is slightly influenced by noises because the changing scale of their amplitudes is large enough to cover the effects of them and for which the main task is to retain the changing patterns. This section introduces a method of PLR_FP for time series with obvious periodicity and violent pattern fluctuation during a specific period.

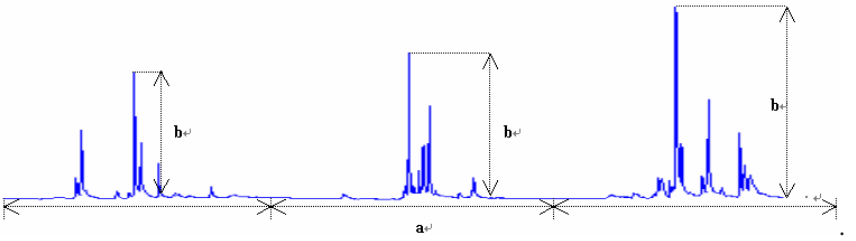


Fig. 1. An example of time series with features: a) obvious periodicity; b) violent pattern fluctuation during a specific period

Def. 1. Feature Points

Given constant $R \geq 0$ and time series $X = \langle x_1, x_2, \cdots, x_n \rangle$, feature point x_m $(1 \leq m \leq n)$ is the one that satisfies one of the following three conditions:

$$\begin{aligned} &1)\ m = 1 \text{ or } m = n; \\ &2)\ x_m - x_{m+1} \geq R \quad (1 \leq m < m+1 \leq n); \\ &3)\ x_m - x_{m-1} \geq R \quad (1 \leq m-1 < m \leq n). \end{aligned} \tag{1}$$

R is the parameter that we can control. The greater the value of R, the less feature points we can get, and the line segments description of time series is more rough. Otherwise, if the value of R is smaller, the number of fragments is larger. Particularly, if R=0, the result of segmentation is the original time series, and if $R \to \infty$, we will eventually get a straight line. Therefore, the value of R determines the granularity that we mine on time series.

The following is the pseudo code for the algorithm based on Def.1.

Table 1. The algorithm frame of PLR_FP

```
Algorithm: Select_Feature_Points(S, R, S^C)
Input: time series S (length= N) and parameter R
Output: series S^C, the set of feature points
Steps:
      1. i=1;
          Add_Feature_Points(v_i, S^C);
          j=i+1;
      2. while i<N do{
          if Abs(v_i - v_j) < N then
            j++;
          else{
            Add_Feature_Points(v_j, S^C);
            i=j;
      }
      }
      3. Add_Feature_Points(v_N, S^C);
```

From Table 1, in the process of selecting feature points, the algorithm scans the entire time series only once and involves only additive and minus operation. Obviously, its time complexity is $O(N)$. Particularly, each computation involves just two adjacent points so the algorithm also supports online extraction of feature points.

4 Experiment

As mentioned in section 2, PLR of PAA is surprisingly competitive with the more sophisticated transforms, such as DFT, DWT and SVD. Meanwhile, our method is also unique. This section carries out an experiment to make a comparison with PAA in two aspects: 1) the main trend held on the whole; 2) the fit error to the original time series. Because we focus on time series with obvious periodicity and violent pattern fluctuation during a specific period, here, we select hydrological time series from the daily water flow data of Ziluo Hill station of the year 1997 in Shaying River, one of the water systems of Huaihe watershed. The original time series is figured in Fig. 2, and we set parameter R = 1, 2, 3 and 4 respectively and the corresponding data compression ratio is given in Table2.

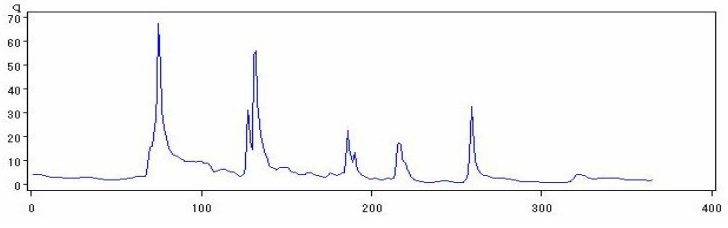


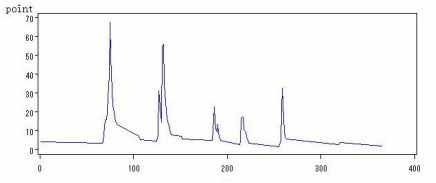
Fig. 2. The time series of daily water flow from Ziluo Hill station (length=365)

Table 2. The data compression ratio according to the values of R (length=365)

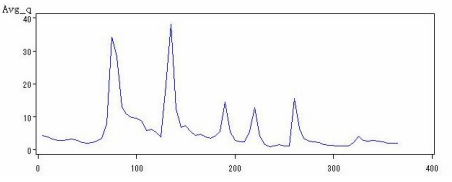
R	The number of segments	Compression Ratio
1	70	81%
2	54	86%
3	42	89%
4	38	90%

Fig.3 a-g show the comparison of the patterns retained between the methods of PLR_FP and PAA at different levels of compression ratio listed in Table 2. In Fig.3, it is clear that the method of PLR_FP can extrude the feature of patterns more efficiently than PAA. Especially to patterns that change violently at local scale during a specific period, PLR_FP can well retain them, while to those which are stable and smooth this method can filter them with an ideal effect.

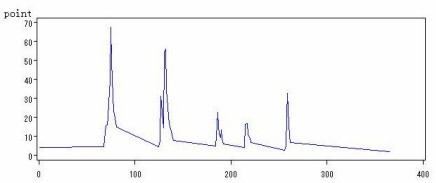
All segmentation algorithms also need some methods to evaluate the quality of fit for a potential segment. Here, the measure used in conjunction with linear regression is the sum of squares, or the residual error. The result is shown in Table 3.



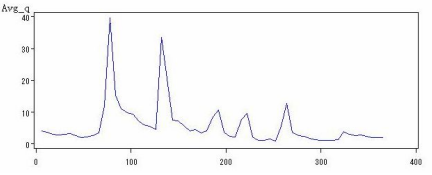
a. PLR_FP Compression ratio: 81%



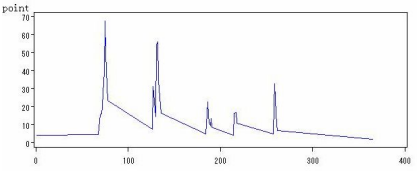
b. PAA Compression ratio: 81%



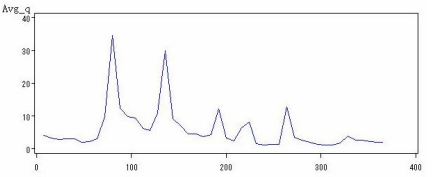
c. PLR_FP Compression ratio: 86%



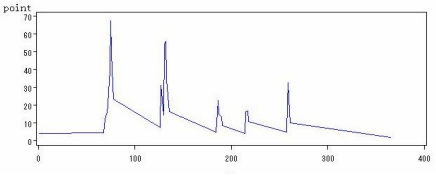
d. PAA Compression ratio: 86%



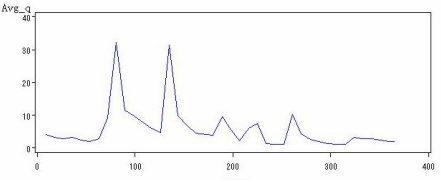
e. PLR_FP Compression ratio: 89%



f. PAA Compression ratio: 89%



f. PLR_FP Compression ratio: 90%



g. PAA Compression ratio: 90%

Fig. 3. The comparison of the patterns retained between PLR_FP and PAA at different levels of compression ratio listed in Table 2

Table 3. Fit error of the two methods to the original time series at different compression ratio levels (length of original time series: 365)

Compression Ratio	PAA	PLR_FP
81%	70.3	**25.6**
86%	402	**39.1**
89%	353.1	**78.8**
90%	**80.8**	89.9

From Table.3 we can see that the fit error of PLR_FP is ascending monotonously along with the increase of compression ratio and is less than that of PAA on the whole

except at the level of 90% (as the boldfaced number shows). This means that at some points PAA may result in good effect, but under most circumstances it is hard for us to determine how many segments we should have in advance. If the number of segments selected is not appropriate, the result may be even worse.

The experiment implies that the advantage of PLR_FP is unique in the aspects of pattern extraction and fit error of time series with obvious periodicity and violent pattern fluctuation during a specific period.

5 Summary and Conclusion

Pattern representation can achieve the compression of data and keep the main features while filtering the detail noises so as to describe the intrinsic character of time series, which benefits the improvement of efficiency and accuracy of data mining. Piecewise linear representation (PLR) is one of the important and perhaps the most frequently used means to accomplish pattern representation of time series. The method of PLR_FP not only has less time complexity,, but also can avoid the influence of noises effectively. Experiment shows that this method can well hold the trend on the whole, produces high quality approximations of the data and works well on time series with obvious periodicity and violent pattern fluctuation during a specific period.

Acknowledgements

This research is partially supported by NSFC 60673141.

References

1. Agrawal, R., Faloutsos, C., Swami, A.: Efficient similarity search in sequence databases. In: Proceedings of the 4th Conference on Foundation of Data Organization and Algorithms (1993)
2. Chan, K., Fu, W.: Efficient time series matching by wavelets. In: Proceedings of the 15th IEEE International Conference on Data Engineering, IEEE Computer Society Press, Los Alamitos (1999)
3. Agrawal, R., Lin, K.I., Sawhney, H.S., Shim, K.: Fast similarity search in the presence of noise, scaling, and translation in time series databases. In: Proceedings of the 21st International Conference on Very large Data Bases, pp. 492–500 (1995)
4. Das, G., Lin, K., Mannila, H., Renganathan, G., Smyth, P.: Rule discovery from time series. In: Proceedings of the 3rd International Conference on Data Engineering (1998)
5. Peng, C., Wang, H., Zhang, S., Parker, S.: Landmarks: a new model for similarity based pattern querying in time series databases. In: Proceedings of the 16th International Conference on Data Engineering (2000)
6. Keogh, E., Chu, S., Hart, D., et al.: Segmenting Time Series: A Survey and Novel Approch.[M]. Data Mining in Time Series Databases. World Scientific Publishing Company, Singapore (2003)

7. Keogh, E., Chakrabarti, K., Pazzani, M.J., et al.: Dimensionality Reduction for Fast Similarity Search in Large Time Series Databses[J]. Knowledge and Information Systems 3(3), 263–286 (2001)
8. Yi, B.K., Faloustsos, C.: Fast Time Sequence Indexing for Arbitrary Lp Norms[A]. In: Proceeding of the 26th International Conference on Very Large Databases[C], pp. 385–394. Morgan Kaufmann Publishers Inc., San Francisco (2000)
9. Park, S., Kim, S.W., Cho, J.S., et al.: Prefix-Querying: An Approach for Effective Subsequence Matching Under Time Warping in Sequence Databases[A]. In: Proceedings of the 10th International Conference on Information and Knowledge Management[C], pp. 255–262. ACM Press, New York (2001)
10. Park, S., Lee, D., Chu, W.W.: Fast Retrieval of Similar Subsequences in Long Sequence Databases. In: Proceedings of the 3rd IEEE Knowledge and Data Engineering Exchange Workshop, IEEE Computer Society Press, Los Alamitos (1999)
11. Keogh, E., Smyth, P.A: probabilistic approach to fast pattern matching in time series databases. In: Proceedings of the 3rd International Conference of Knowledge Discovery and Data Mining, pp. 20–24 (1997)
12. Keogh, E., Pazzani, M.: An enhanced representation of time series which allows fast and accurate classification, clustering, and relevance feedback. In: Proceedings of the 4th International Conference on Knowledge Discovery and Data Mining, pp. 239–241. AAAI Press, Stanford, California, USA (1998)
13. Keogh, E., Pazzani, M.: Relevance feedback retrieval of time series data. In: Proceedings of the 22nd Annual International ACM-SIGIR Conference on Research and Development in Information Retrieve, ACM Press, New York (1999)
14. Perng, C.S., Wang, H., Zhang, S.R., et al.: A New Model for Similarity-based Pattern Querying in Time Series Databases[A]. In: Proceedings of the 16th International Conference on Data Engineering[C], pp. 33–42. IEEE Computer Society, Washington (2000)
15. Prat, K.B., Fink, E.: Search for Patterns in Compressed Time Series[J]. International Journal of Image and Graphics 2(1), 89–106 (2002)

A Regional Safety Information Sharing System Based on CMS and Online Map System

Akira Hattori[1], Masato Goto[2], Takami Yasuda[2], and Shigeki Yokoi[2]

[1] Faculty of Information Technology, Kanagawa Institute of Technology, 1030, Shimo-ogino, Atsugi, Kanagawa, 243-0292, Japan
ahattori@ic.kanagawa-it.ac.jp
[2] Graduate School of Information Science, Nagoya University, Furo-cho, Chikusa-ku, Nagoya, Aichi, 464-8601, Japan
masato@nagoya-u.jp, {yasuda, yokoi}@is.nagoya-u.ac.jp

Abstract. We describe a regional safety information sharing system based on CMS and online map system. The system is characterized by extending the functional capability of XOOPS, which is a type of CMS, with online map system. It accumulates local knowledge through mapping of safety information by local residents. To evaluate it by a field experiment, we introduced it into Higashi ward in City of Nagoya. As the result of the experiment, we found that our proposed system was effective for sharing regional safety information among the community.

Keywords: Regional Safety Information Sharing, CMS, Online Map System.

1 Introduction

Recently, there are a variety of mapping application system on the web to communicate geographical knowledge of crime prevention to the public for their safety, such as crime map and traffic accident map [1]. Being informed about crime in one's community is the first step in preventing future occurrences. Although creating, maintaining, and handling map data require an amount of effort, publication of online map system, such as Google Maps, gives us the environment in which we can easily introduce a mapping application system.

On the other hand, local knowledge based on community participation is vital to making daily living safe. That is to say, it is also useful to share information which local residents exchange with neighborhood friends, such as "there was a suspicious person hanging, so you had better watch out the person", among the community. However, in a system for sharing regional information in the community, it is important to manage the contents easily, to customize the system functions based on the situation in each community flexibly, and to manage user accounts precisely. A CMS (Contents Management System), which manages contents and functions of a web site efficiently, offers the promise of meeting these requirements. In recent years, open source software with the fundamental functions of CMS has been developed. XOOPS, a type of such CMS, has been used to manage regional information [2].

B. Apolloni et al. (Eds.): KES 2007/ WIRN 2007, Part II, LNAI 4693, pp. 1073–1080, 2007.
© Springer-Verlag Berlin Heidelberg 2007

Therefore, the purpose of this study is to develop a system based on CMS and online map system, in which users (local residents) can manage safety information linked to a map easily and share it in an easy-to-understand manner.

2 Related Work

In this section, we describe related works and the feature of our study. There have been a number of research and development of a mapping application system on the web to provide safety information. Some are based on GIS (Geographic Information System) [3][4]. However, these have problems with maintenance of the map data and user-friendliness, so it is hard to use them in a local community. There have also been systems based on Blog, a type of CMS, in which the information is managed by just a small staff [5][6]. But, since Blog is a personal tool, it is not appropriate for community participation. In addition, as far as we know, no mapping application system using Blog has been developed. When CMS is not used, to meet requirements as mentioned above, we have to build all functions from scratch.

We proposed a concept of a system for sharing information in local crime prevention last year [7]. Our study's unique contribution is using both CMS and online map system. In this paper, we present system architecture to extend the functional capability of XOOPS with Google Maps [8] and external web service called geocoding service which converts an address into the corresponding geographical position data [9] by mechanism using potential web technologies, such as XML (eXtensiblie Markup Language), web service, and Ajax (Asynchronous Javascript + XML). This makes our system easy-to-use, reduces an amount of effort for handling a map, makes it easy to manage contents, and makes it flexible for various changes. Therefore, in our system, it is easy to manage contents and users can share safety information in an easy-to-understand manner. In addition, we confirmed effectiveness of developed system by a field experiment.

3 Development of a Regional Safety Information Sharing System

In this section, we describe a regional safety information sharing system based on CMS and online map system.

3.1 Features of the System

Figure 1 indicates the outline of our system. The system is characterized by extending the functional capability of XOOPS with online map system, which is Google Maps and a geocoding service. Consequently, since it provides an easy environment for information manage and search, users can register safety information linked to a map easily and to share it in an easy-to-understand manner. In addition, our system arranges an information input screen according to the type of incident. This helps them to input information on the incident.

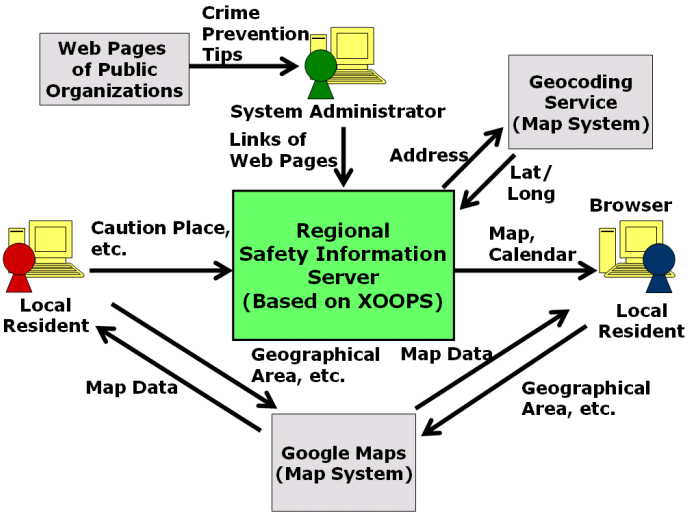


Fig. 1. Structure of our system

Moreover, because it displays the result of search on a map and a calendar, they can visually understand what happened in surroundings.

3.2 Functions of the System

Fundamental functions of our system are registration of and searching for safety information. Figure 2 and 3 indicate the flows of them respectively. The following sections describe the functions focusing around the features.

(1) Linking of safety information to a map
Geographical position data, such as latitude and longitude, is necessary to link safety information to a map. To make it easy for users to input such data, our system displays the area around the selected address by them on the resulting

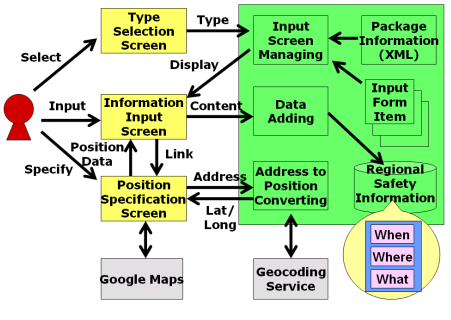


Fig. 2. The flow of information registration

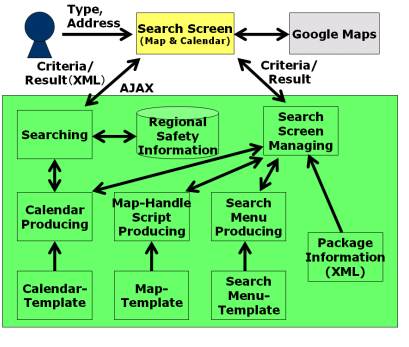


Fig. 3. The flow of information search

map (see Fig. 4). They can input it by just clicking the position on the map. To display a map containing appropriate area, it is necessary to convert a user-selected address into the corresponding position data. But, that requires a large amount of data and a complex program. Therefore, our system makes use of an external web service with such conversion functionality. As shown Fig. 2, we added an address to position converting function to XOOPS, which receives a user-selected address, invokes the service and returns the response from it in the XML format. A web browser calls the function to display position specification screen by using Ajax technology. Users can input geographical position data on the screen. For this reason, they can easily link safety information to a map.

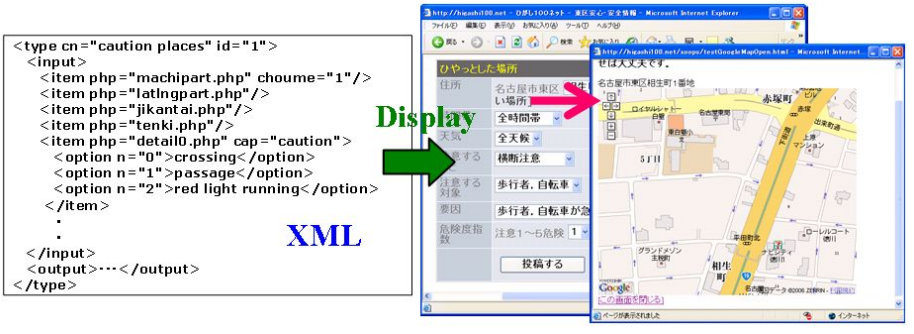


Fig. 4. Package information and information input screen

(2) Displaying safety information on a map
To display search results on a map, it is necessary to associate XOOPS with Google Maps. We realized this association by making a search function available as a web service. As shown Fig. 3, the service returns search results in the XML format. When a search screen management function receives a request from a browser, it sets search criteria, such as address and type of incident, to a map-template using Smarty Template Framework and returns it. The map-template contains a Javascript to handle Google Maps. And the script sends the search criteria to the search function and plots the results on a map. Since the map contains the area of user-selected address, users can understand what happened in surroundings easily. On the other hand, there is information which should not be displayed on a map, for example, when it contains a privacy issue. Such information is displayed on a calendar. Therefore, they can understand what and when happened. Figure 5 and 6 are examples of search results.

(3) Packaging of safety information
Formalized information is useful for users, because it is easy to input and use it. And we noted that an event becomes clear when based on 5W1H (when, where, who, what, why, and how) information. Therefore, in our system, users-supplied information is managed in the form of when, where, and what (detailed information about incident).

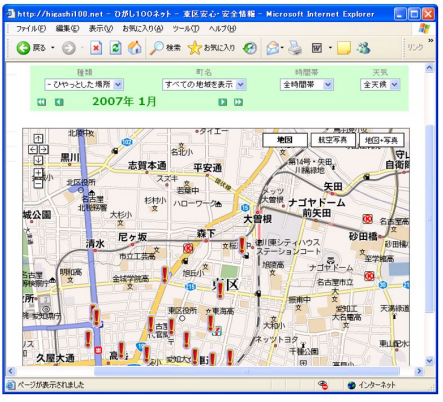


Fig. 5. Search results on a map

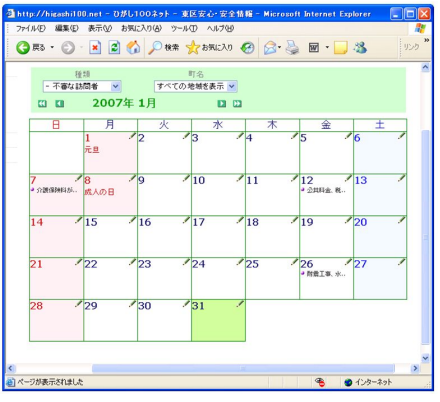


Fig. 6. Search results on a calendar

However, there are a variety of safety information, including those for being useful for preventing traffic accidents, onefs life as a consumer, and the safety of the children. The natures of information, such as a privacy issue, differ from each other, and we must gather safety information based on their natures. Therefore, we classified each of when, where, and what information based on their details. For example, where information has a town name only or includes a position data, and so on. And we prepared an input form item for every class, which is a PHP program. Our system arranges an information input screen by packaging each item. The package of items is managed by using XML (see Fig. 4). In the XML, a type element is corresponding to each type of incident, and it contains item elements for input form items. Through it, since information input screen according to the type of incident is displayed, users can easily input information and our system can gather formalized it

On the other hand, in our system, users can search for safety information by address and type of incident, and search results are displayed on a map and a calendar. However, it is important to display them based on the nature of information as information gathering. Therefore, it also packages how to display them by XML. Figure 5 and 6 are displaying them through the XML.

4 Field Experiment

We have been building practical system in collaboration with citizen volunteers working in Higashi ward of Nagoya. They are interested in an online portal in the area. In this study, we carried out a field experiment using our system. In this section, we describe the result of the experiment.

4.1 Overview of the Experiment

We evaluated the following characteristics of our system through the experiment:

(1) It is easy to manage the contents
(2) Users can share safety information in an easy-to-understand manner

Twenty-one people who live in Higashi ward of Nagoya participated in the experiment. They used the system for one month from mid-January. We sent a questionnaire to them by e-mail and received answers from fourteen people. In the questionnaire, we asked about ease of operation and effectiveness of using a map and a calendar, as well as usefulness of the system. We used a 5-point scale with 5 being the highest score for each question.

4.2 Introduction of the System into the Field

To evaluate our system in the field, target information in the experiment are determined from the discussion with the volunteers. They are about two types of incident, which are "caution places" and "suspicious visitors". It packages input form items for them as shown Table 1. The information on "caution places" is linked to a map. In addition, we excluded input form including free description, because improper purposes can not be input and administrator no need for checking users-supplied information at all times. Moreover, when users supply information on "suspicious visitors", the system invites them to proper web pages that they can get crime prevention tips provided by public organizations. We could easily customize our system for the field.

Also, the options in each input form item and web pages that users can get crime prevention tips were discussed with them and selected. And then, even those who are unfamiliar with a program could incorporate voices in the discussion, because all they need to do is just modifying XML file to manage package. This feature enables our system to flexibly correspond to changes of options and web pages selection.

Table 1. Types and input items

caution places	address (town name and street), position data, time zone, weather, three concrete content, level of risk
suspicious visitors	address (town name), two concrete content, level of risk

4.3 Results and Discussion

We discuss the result of the experiment in accordance with the purpose of it.

(1) It is easy to manage the contents
Six people registered safety information. All of them registered it on "caution places", on the other hand, one person of them registered it on "suspicious visitors". The result shows that local residents who can exchange e-mail without difficulty can link safety information to a map.

However, they divided on whether they were able to register safety information easily. Participants who answered "difficult" chose more unnecessary input items

than participants who answered "easy". With regard of excluding input form with free description, participants who answered "easy" commented that they want to register more specific information. With regard of arranging information input screen according to the type of incident, average of evaluation points was 3.8. But, we received comment that it is inevitable. Therefore, we found that it was important to arrange information input screen according to user's position as well as the type of incident.

(2) Users can share safety information in an easy-to-understand manner
One of the features of our system is displaying search results of safety information by a combination of a map and a calendar. Table 2 shows the evaluation results about this function.

Table 2. Evaluation results about display method of search results (averaged)

displaying them on a map	4.6
displaying them on a calendar	3.7
using both as the type of incident	3.4

(answeres=14, higest score=5)

With respect to displaying them on a map, we got a good result. But, the others were not so good. This may be explained by the facts that almost all of safety information registered by participants are about "caution places" and that map and calendar with all registered information are displayed at the first screen. In fact, we received the following suggestions:

- It is better to display a larger map to show the position of the information more clearly.
- It is better to display the map above the calendar at least.

However, the person who gave above suggestions evaluated using both a map and a calendar as the type of incident highly, the point was 4. In addition, with respect to displaying the map and calendar, we received the following favorable comments:

- It is possible to catch what happened.
- It may be easy to come to understand what day of the week an incident occurs a lot in.

From these results, we can see that each of map and calendar was effective for providing users-supplied regional safety information. In addition, displaying information according to a school district and listing it were also required.

Let us summarize the main points that have been made clear in the experiment as follows:

- Our system has the flexibility important to use it in a local community.
- Local residents who can exchange e-mail without difficulty can link safety information to a map to register it.

- Displaying of search results on a map and a calendar is effective for providing safety information in an easy-to-understand manner.
- How to input safety information and how to display it besides them need further discussion.

Therefore, we confirmed that our system based on CMS and online map system was effective for sharing regional safety information among the community.

5 Conclusions

In this paper, we developed a regional safety information system based on CMS and online map system. The system is characterized by extending the functional capability of XOOPS with Google Maps and geocoding service. We introduced it into Higashi ward in City of Nagoya and conducted a field experiment in the area. As the results, we found that our system had the flexibility and the effectiveness for sharing regional safety information among the community. The future work is to extend the function of it based on the results of the experiment. And we will continue a field experiment to use practically it at a portal site.

References

1. Tohru, O., Naoyuki, N., Keitarou, H.: Relationship between Prevention Actions and Cognitive Appraisals of a Snatch Occurrence Map on the Internet. Theory and Application of GIS 13(1), 1–10 (2005)
2. GIJOE & matchan: Customizing XOOPS, Mainichi Communications (2005)
3. LAPD crimemaps, http://www.lapdcrimemaps.org/
4. San Diego Regional Crime MAPS, http://mapping.arjis.org/
5. Souraku Safety Network, http://www.souraku-anzen.net/
6. Nichinan Area Crime Prevention Group, http://www.scramble-box.net/bohan/
7. Masato, G., Akira, H., Takami, Y., Shigeki, Y.: A Web-Based System for Gathering and Sharing Experience and Knowledge Information in Local Crime Prevention. In: Gabrys, B., Howlett, R.J., Jain, L.C. (eds.) KES 2006. LNCS (LNAI), vol. 4252, pp. 1079–1086. Springer, Heidelberg (2006)
8. Google Maps, http://www.google.com/api/maps/
9. CSIS Geocoding Tools & Utilities, http://pc035.tkl.iis.u-tokyo.ac.jp/~sagara/geocode/

Avoidance of Traffic Delay for Panicking Crowds Subject to Information Propagation

Naoto Mukai[1], Toyohide Watanabe[1], and Jun Feng[2]

[1] Department of Systems and Social Informatics,
Graduate School of Information Science, Nagoya University
Furo-cho, Chikusa-ku, Nagoya, 464-8603, Japan
naoto@watanabe.ss.is.nagoya-u.ac.jp,
watanabe@is.nagoya-u.ac.jp
[2] Hohai University, Nanjing, Jiangsu 210098, China
fengjun-cn@vip.sina.com

Abstract. In disaster situations, a panicking people tend to follow other leading people when they must select the next route from alternative routes. In some cases, such tendency may cause traffic congestion in escape routes, and increases the total amount of escape time. Therefore, in this paper, we propose a traffic congestion model and an information propagation model in emergency situations, and consider how to propagate escape route information through panicking crowds in order to avoid traffic congestion. The selection of pairs of "voice alarms" and "escape gates" is an important factor to change the propagation of escape route information. Thus, we evaluate the effect of the selection of pairs by computer simulation using the traffic congestion and information propagation models. Our simulation results indicate that effective pairs can keep the standard deviation of escape time low.

1 Introduction

Rescue systems play an important role in disaster situations, and have been developed as a big project in Japan. For example, information providing/sharing systems using ad-hock network or RFID tags make it possible that rescue teams can obtain disaster information rapidly without network infrastructures [1,2]. Moreover, disaster simulation systems enable evacuees to find useful escape strategies in disaster situations [3,4]. In this paper, we consider a problem "how to propagate route information among evacuees?". In disaster situations, people are in a panic and tend to follow other leading people when they select the next route from alternative routes. Such tendency causes traffic congestion in narrow ways in some cases, and increases the total amount of escape time. Therefore, we try to find the optimum pairs of "voice alarms" and "escape gates" to minimize the amount of escape time by avoiding traffic congestion. The voice alarms are information sources, and evacuees can get route information from the warning sounds of the voice alarms. The escape gates are destination points, and

B. Apolloni et al. (Eds.): KES 2007/ WIRN 2007, Part II, LNAI 4693, pp. 1081–1088, 2007.
© Springer-Verlag Berlin Heidelberg 2007

evacuees are headed to one of the escape gates. Our simulation model is based on Adriana's model [5]. In Adriana's model, a moving rule of evacuees is followed by a force filed. In particular, if people are under the influence of a disaster (e.g., fire gas), moving speeds of evacuees slow according to the danger level of the disaster. In addition to the influence of disaster, our simulation model considers the influence of traffic congestion. The traffic congestion in escape routes is modeled by the concept of "selfish routing" [6,7,8]. The selfish routing represents a delay problem in a system whose resources are shared among selfish people (e.g., communication network).

The remainder of this paper is as follows: Section 2 introduces the basic concept of traffic congestion: "selfish routing". Section 3 formalizes the simulation model which includes environments, evacuees, and alarms. Section 4 reports simulation results. Finally, Section 5 concludes and offers our future works.

2 Selfish Routing

In this section, we introduce "selfish routing" which is the basic concept of traffic congestion in our simulation model. Figure 1 shows an example of a traffic congestion route network caused by selfish routing. There are start node S and goal node D, which represent traffic intersections. Moreover, there are two routes from S to D. Let's assume that there are N evacuees at start node S, and they have two route options, i.e., upper route or lower route, to go to goal node D. The selection probabilities of upper route and lower route are given by p and $1 - p$, respectively. The delay functions caused by congestion, which represent travel time from S to D, of upper route and lower route are given by p and 1. For example, if half of the people $1/2 \cdot N$ selects upper route, delay time of upper route and lower route are $1/2 \cdot N$ and N, respectively. Our objective for this problem is to minimize the expected travel time of all evacuees calculated by Equation (1). Here, let's consider two types of evacuees: altruistic and selfish evacuees. Altruistic evacuees represent that there is a difference of travel time among evacuees, and selfish evacuees represent that there is no difference of travel time among evacuees. If evacuees are altruistic, the expected travel time T is $3/4 \cdot N$, and the selection probability p is $1/2$. On the other hand, if evacuees are selfish, the expected travel time T is N, and the selection probability p is 1. This result indicates that selfish route selections of evacuees may increase the expected travel time T according to traffic conditions at escape routes. Therefore, in this paper, we attempt to find effective navigation rules (i.e., a pair of "voice alarms" and "escape gates") to minimize the expected travel time T for selfish evacuees (i.e., panicking evacuees) by simulation analysis.

$$T = \sum_{i=1}^{N} \frac{d_i}{N} \tag{1}$$

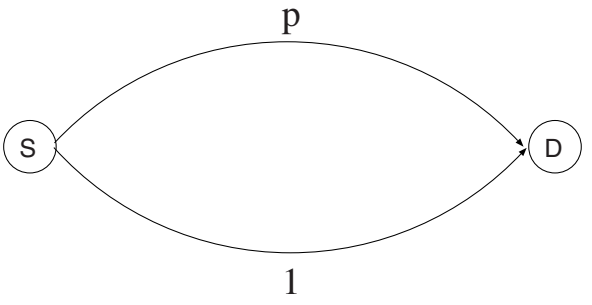


Fig. 1. Selfish routing

3 Simulation Model

In this section, we formalize a simulation model of disaster, which includes environment, evacuees, and alarms.

3.1 Environment

A disaster environment Env is given by a network structure as Equation (2). N is a set of nodes (rooms), L is a set of links (routes) between nodes, and G is a set of escape gates. A node n is represented by union of rectangles, and a link l is represented by a linear rectangles (whose width is always 1), as shown in Figure 3. At initial simulation time, each evacuee is located at any node n, randomly. If an evacuee senses an occurrence of a disaster, the evacuee starts to go to any escape gate g according to the route information from voice alarms or information propagation from other evacuees. Our model does not consider the spread of disaster (e.g., gas leaking) in order to simply evaluate the delay influence caused by traffic congestion in disaster environments.

$$\begin{cases} Env = (N, L, E) \\ N \quad = \{n_1, n_2, \cdots, n_h\} \\ L \quad = \{l(n, n') | n, n' \in N\} \\ G \quad = \{g_1, g_2, \cdots, g_k\} \end{cases} \tag{2}$$

3.2 Evacuees

A set of evacuees E is given by Equation (3). If there is no traffic congestion, moving rules of evacuees are followed by uniform accelerated motion: α is an accelerated velocity, $e(t)$ is a position of evacuee e at time t, and $v(t)$ is a speed at time t.

$$\begin{cases} E \qquad = \{e_1, e_2, \cdots, e_m\} \\ e_j(t+1) = e_j(t) + v_j(t) \\ v_j(t+1) = v_j(t) + \alpha_j \end{cases} \tag{3}$$

Figure 2 illustrates the traffic delay model caused by congestion. There are two evacuees e_1 and e_2. The accelerated velocities of e_1 and e_2 are α_1 and α_2, respectively. Evacuee e_1 is affected by evacuee e_2 according to the distance between these two evacuees $dist(e_1, e_2)$, and the angular difference between these two evacuees $angle(e_1, e_2)$ $(= \theta)$. The accelerated velocity α_2 can be divided into two directions $\alpha_2 \cdot cos(\theta)$, and $\alpha_2 \cdot sin(\theta)$. The opposite accelerated velocity $beta_1$ calculated by Equation (4), where w is a weight value, is added to evacuee e_1. Consequently, if a lot of evacuees pass a narrow route at the same time, they cannot move smoothly by an influence of traffic congestion.

$$\beta_1 = \frac{w_2 \times \alpha_2 \cdot cos(\theta)}{dist(e_1, e_2)^2} \tag{4}$$

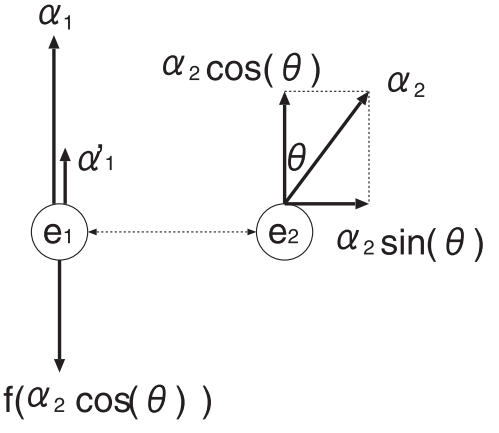


Fig. 2. Delay caused by traffic congestion

3.3 Voice Alarms

A set of voice alarms A is given by Equation (5), where g is a destination escape gate, and r is an information propagation speed. Each voice alarm is statically located at any node n. If a disaster occurs, the route information to destination escape gate g of alarm a is spread in a circle with speed r (i.e., the radius of circle is $r \times t$). Evacuees start to go to destination g as soon as the circle contains the positions of evacuees (i.e., evacuees can get the route information to destination g at that moment). Obviously, the decision of pairs of "voice alarms" and "escape gates" is an important factor to avoid traffic congestion. Therefore, in the next section, we evaluate the effect of the pairs on the expected escape time by a simulation experiment.

$$\begin{cases} A = \{a_1, a_2, \cdots, a_o\} \\ a \ = (x, r) \end{cases} \tag{5}$$

4 Experiments

In this section, we report our simulation results. Adriana et al. had been also reported the effect of "voice alarms" by crowd simulation in [5]. Their simulation considered the types of alarms (local, communication, or global), but did not consider the pairs of "voice alarms" and "escape gates". On the other hand, our aim of this simulation is to investigate the impacts of the pairs of "voice alarms" and "escape gates" on the traffic delay of evacuees caused by congestion at escape routes.

4.1 Experimental Setting

Our simulation compares two types of pairs: nearest pairs and optimum pairs. Nearest pairs represent that evacuees are head to the nearest escape gates from voice alarms. Optimum pairs represent that evacuees are head to the escape gates which minimize the total amount of escape time (the optimum pairs can be found by full search of all possible pairs).

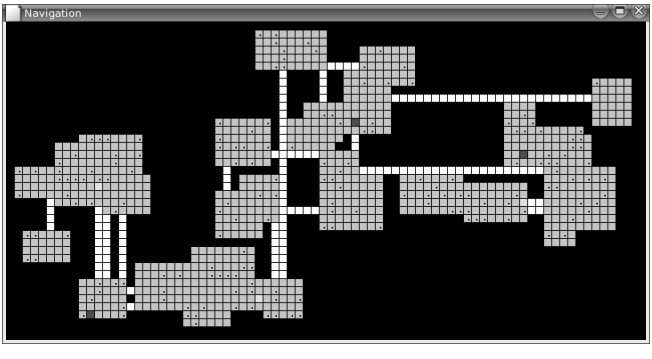


Fig. 3. Environment

An environment of disaster situation is given by a 2-dimensional space as shown in Figure 3. The size of the space is 800×400 pixels. In the space, there are 100 - 300 evacuees, 3 voice alarms, and 3 escape gates. The moving speed of evacuees is set to 2 pix/t as maximum value and 1 pix/t as minimum value. The accelerated speed of evacuees is set to 0.5 pix/t. The information propagation

Table 1. Parameter setting

Parameters	Value
Width	800
Height	400
Evacuees $\|E\|$	100 - 300
Gates $\|G\|$	3

Parameters	Value
Alarms $\|A\|$	3
Acceleration α	0.5
Maximum Speed	2
Minimum Speed	1

speed r is set to 3 or 7 pix/t. Initial positions of evacuees are changed randomly in each simulation process. There are three evaluation items: total of escape time, average of escape time, and standard deviation of escape time. Simulation parameters are summarized in Table 1.

4.2 Experimental Results

Figure 4 shows the sum of escape time. This result indicates that the sum of escape time increases in proportion to the number of evacuees. The difference among four patterns also increases with the number of evacuees. Figure 5 shows the average of escape time. This result denotes similar tendency of the sum of escape time. However, the difference among four patterns is kept although

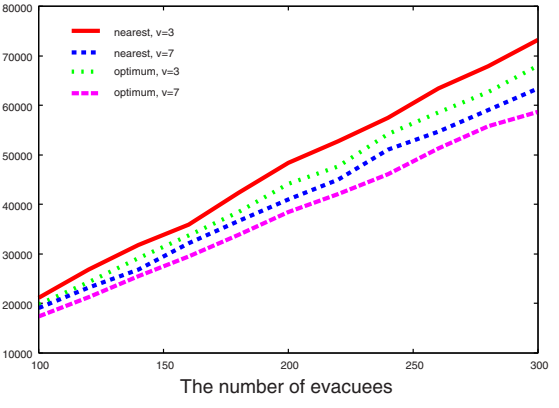


Fig. 4. Sum of escape time

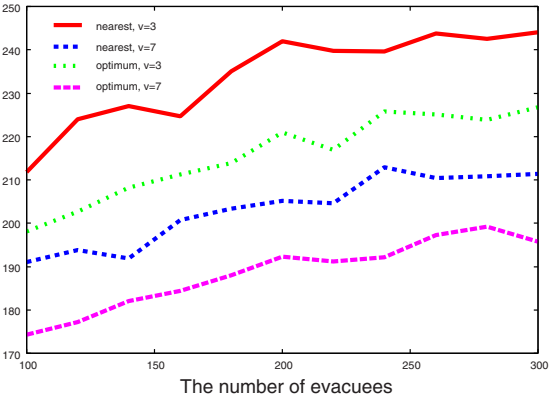


Fig. 5. Average of escape time

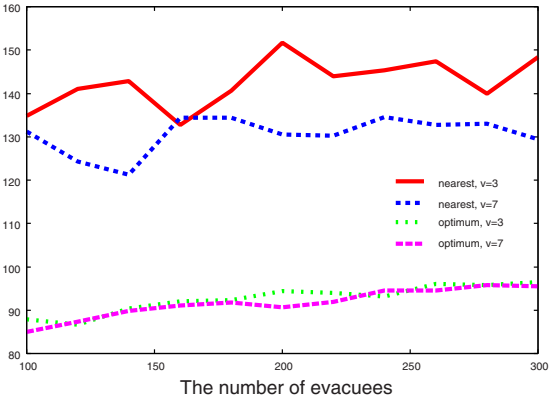


Fig. 6. Standard deviation of escape time

the number of evacuees increases. From these results, we figure out that the optimum pairs of "voice alarms" and "escape gate" can reduce the expected escape time, and rapid information propagation among evacuees helps the effect of the escape performance. Figure 6 shows the standard deviation of escape time. This result indicates that the optimum pairs can reduce the standard deviation among evacuees than nearest pairs. Furthermore, the standard deviation of nearest pairs strongly depends on initial positions of evacuees. On the other hand, the standard deviation of optimum pairs is independent of the initial positions of evacuees. The reason of this trend is that optimum pairs can decreases the number of traffic congestion routes by changing the flow direction of evacuees.

5 Conclusions

In this paper, we focused on navigation for panicking people in disaster situations. Such people tend to follow other leading people by information propagation. Therefore, we proposed a simulation model of traffic congestion and information propagation. The traffic congestion is based on the concept of selfish routing, and its effect is represented by an opposite accelerated velocity. Moreover, the information propagation is represented by information spreading in a circle from voice alarms. We performed simulation experiments to evaluate the effect of the selection of pairs of "voice alarms" and "escape gates". Our results indicate that effective pairs reduce the standard deviation of escape time among evacuees. Our future task is to introduce the reliability of the disaster information which is propagated through the grapevine communication.

Acknowledgment

This research is supported by Japan Society for the Promotion of Science (JSPS).

References

1. Asama, H., Hada, Y., Kawabata, K., Noda, I., Takizawa, O., Meguro, J., Ishikawa, K., Hashizume, T., Ohga, T., Hatayama, M., Matsuno, F., Tadokoro, S.: Rescue infrastructure for global information collection. In: Proc. of SICE-ICCAS International Joint Conference 2006, pp. FP4–11 (2006)
2. Shinoda, K., Noda, I., Kunifuji, S.: Emergency communication support system used by human relationship (in Japanese). In: Proc. of DICOMO 2003, pp. 45–48 (2003)
3. Muraki, Y., Kanoh, H.: The effectiveness of communication on wide area simulation of disaster evacuation using multiagent model (in Japanese). Technical Report 130, IPSJ-MPS04052018 (2004)
4. Kyogoku, M., Takeuchi, I.: Incorporating global decision in robocuprescue simulation (in Japanese). Technical Report 23, IPSJ-GI06015004 (2006)
5. Braun, A., Bodmann, B.E.J., Musse, S.R.: Simulating virtual crowds in emergency situations. In: Proc. of Virtual Reality Software and Technology (VRST), pp. 244–252 (2005)
6. Koutsoupias, E., Papadimitriou, C.H.: Worst-case equilibria. In: Proceedings of the 16th Annual Symposium on Theoretical Aspects of Computer Science, pp. 387–396 (1999)
7. Roughgarden, T., Tardos, E.: How bad is selfish routing? In: IEEE Symposium on Foundations of Computer Science, pp. 93–102. IEEE Computer Society Press, Los Alamitos (2000)
8. Roughgarden, T.: The price of anarchy is independent of the network topology. In: Proceedings of the 34th ACM Symposium on the Theory of Computing, pp. 428–437. ACM Press, New York (2002)

Building the Virtual Community to Support Interregional Exchange Between Rural and Urban

Yoshida Chiho, Yasuda Takami, and Yokoi Shigeki

Graduate School of Information Science, Nagoya University, Fro-cho, Chikusa-ku, Nagoya-city, Aichi, 464-8601 Japan
po_chiho@nagoya-u.jp, {yasuda, yokoi}@is.nagoya-u.ac.jp
http://www.mdg.human.nagoya-u.ac.jp/

Abstract. The rural area faces to collapse the social infrastructure, such as a medical treatment and education, due to depopulation, which is as serious a problem as ever. We can see many interregional exchange programs between rural and urban as one of the popular solutions for depopulation all over the country. Thus our goal is to seek for the best method by ICT to support the exchange programs. We built a virtual community by XOOPS and tried to store a variety of opinions and information in virtual community from rural and urban so as to find the best solutions for depopulation. XOOPS is very useful and is low cost to build a website. We discussed the effect of virtual community. As a result, we concluded that the virtual community on the Internet would be an effective tool for exchanges.

Keywords: Sharing knowledge, Depopulated area, Virtual community, Supporting interregional exchange program, XOOPS.

1 Introduction

1.1 Present Situation of Rural Areas

Depopulation is growing in rural areas like mountains and coastal sides as opposed to the metropolis. It has encouraged the destruction of the local community by collapsing the basis of the social infrastructure, such as medical treatment and education. In spite of many policies such as a revival plan in rural areas and the improvement project of IT infrastructure by the government, we have not seen their effect on limiting depopulation yet. Instead, the regional gaps have been widening.

1.2 The Content of Study

The purpose of study was to look for the method to limit depopulation, which is unsolved problem. To this end, we tried to get a variety of opinions and information by sharing depopulation problems with urban and we proposed a support tool by ICT for finding the best solution through these opinions.

B. Apolloni et al. (Eds.): KES 2007/ WIRN 2007, Part II, LNAI 4693, pp. 1089–1096, 2007.
© Springer-Verlag Berlin Heidelberg 2007

The method of this study was to collaborate with the virtual community on the Internet and interregional exchange in Toei-cho. Toei-cho, kita-shitara-gun, Aichi Prefecture has serious problems of depopulation with the 4,400 inhabitants and 44.3 percent of its population consisting of people with more than 65 years of age. On the other hand, you can come across a lot of fascinating things, such as it being close to nature, the strong bonds between the local people and a specialized festival called 'Hanamatsuri', which started 800 years ago. Toei-cho has positively adopted the exchange program between rural and urban for limiting depopulation. Their objective is to make opportunities for urban people to stay and live in Toei-cho through the exchange program. They hope that Toei-cho will be a favorite town for urban people.

In this study, we built the virtual community on the Internet to support the exchange activities in the real world. A large number of urban people would take an interest in the fascination of the rural area through exchange activities by the virtual community. The virtual community simultaneously would deliver the problem of depopulation to urban areas and provide a place to share and discuss the problem of depopulation.

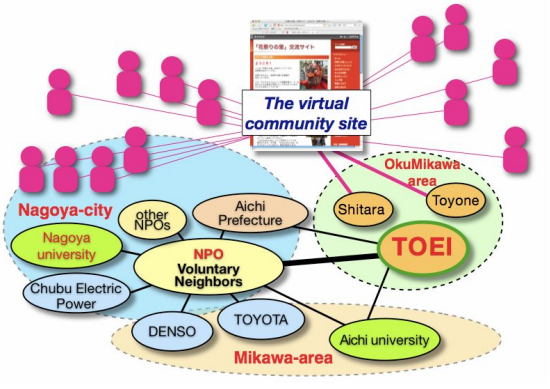


Fig. 1. Combination Image between exchange network and virtual community

The originality of study is to be based on the exchange program in the real world. It is important to collaborate with the virtual community on the Internet and interregional exchange in the real world. The virtual community on the web might be so interactive as to collect a variety of opinions and information anytime and anywhere.

The significance to focus on Toei-cho in this study is in the followings:

The Populated area is situated within 2 hours by car from Toei-cho. The people from Toei-cho have easily moved between rural and urban areas and kept sustainable family bonds without living in Toei-cho. However, if the family bond were lost, they wouldn't have a reason to visit. As a result, depopulation would be advancing and they would lose the social welfare function in Toei-cho.

As the economic growth is more rising, these phenomena would be expanded not only in rural area but also all over the Japan. Therefore, this study intends to develop

advanced research in other areas by the original method different from depopulated and closed area.

2 Activity of Exchange Network in Real World

2.1 Contents of Activity

The activities of exchange program were based on the exchange network and the exchange network committee was established by industry-academia-government collaboration, such as Toei-cho, Aichi Prefecture autonomy, NPO, university and company's volunteer centers etc. They planned and performed two events in Toei-cho and they conducted a promotion campaign in the urban area to gather participants for the events.

The first event was an experience workshop. The workshop's theme was nature and traditional culture in Toei-cho. The inhabitants in Toei-cho participated as teachers, and urban people were students. The purpose was to let urban people discover fascination of Toei-cho by learning directly from Toei-inhabitants about valuable regional resources and build reliance and friendship with Toei inhabitants.

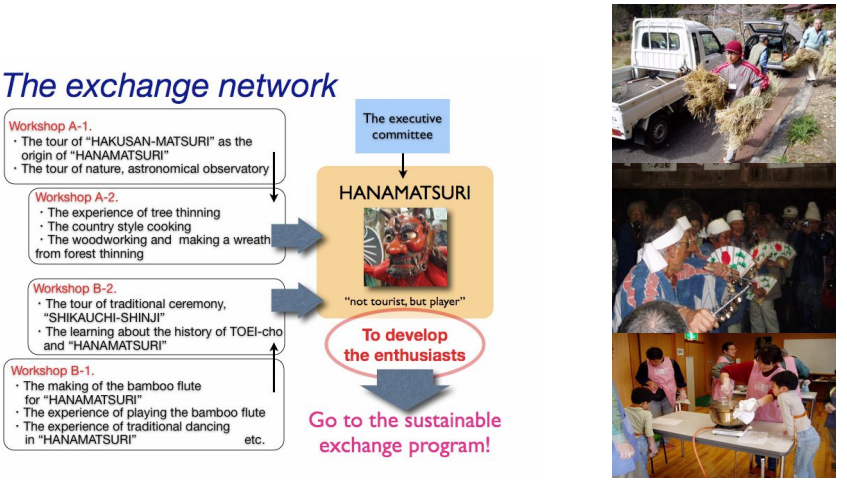


Fig. 2. Contents and purpose about lecture of experience work shop

The second event was the meeting for interregional exchange opinion. Participants of the regional-side and the urban-side shared their own problems with each other and discussed general methods to solve each problem by collaboration.

2.2 Actualized Problems

According to the result of the questionnaire survey on the exchange network (66 responders) and the opinion from meeting, we could find several problems about real activities as follows:

- Advertising and enlightenment.
- There is no distinction between applicants for sightseeing and for the exchange program.
- The number of participants was less compared with the amount of activity for advertising.
- Exchange activities were not acknowledged among the Toei habitants.
- The number of participants was less from near local area, which was situated a half to an hour away by car.
- Self-Participating and sharing of roles.
- The number of participants in Toei-cho was less especially for youngsters and women.
- Each of the participants lacks human communication together.
- It is necessary to overcome inconvenience of access and transportation.
- It has to take a lot of time to build reliable friendship, especially with the mountain area paying so much importance to the bonding of the region.
- Some participants in the urban area let inhabitants in Toei-cho feel belittled because of a lack of understanding of traditional culture 'Hanamatsuri'.
- Sustainability of activity.
- Funds for activity are limited.
- Sustainability and development is not expected in the future. We have to accept and respond faithfully about some comments of inhabitants in Toei-cho, that they said, "If tentatively, don't cooperate with it".

The main point mentioned above is listed in three respects as follows:

- A way of ensuring the applicants have a positive sense of exchange activity.
- The importance of putting out and sharing information in a wide area (Toei-cho/neighboring town/urban).
- The importance of long-term and sustainable organization of exchange activities (building reliance by understanding mutual culture and customs).

3 Building the Virtual Community

3.1 Case Study

There are several cases about the method of software by specializing in ICT in depopulated area. For instance, the website for depopulation networking is called "e-kasochi.net"[1], which introduces the case for a revival project as a portal site for the regional. "Kouryu-kyoju.net"[2] is the promotion website for the exchange living and guidance to live in a rural area. "Muramachi-net"[3] is a portal site maintained by non-profit organization and it delivers information about interregional exchange between rural and urban in Nara Prefecture. The activity called 'Muramachi-sales-lep' on this site exists and forms a virtual organization for sales promotion and promotes regional

[1] http://www.e-kasochi.net/index.html
[2] http://kouryu-kyoju.net/index.php
[3] http://www.channel-e.tv/muramati/

souvenir and products on blog. We found that the virtual community could be an effective tool for promoting exchanges between rural and urban but a definitive method has not been found yet.

3.2 Design

Based on the problem mentioned in 2.2 and case study 1.4, we estimated that virtual community had to install functions as follows:

- Their needs to be a distinction between a function for putting out information for unlimited applicants and a function for sharing information for limited member.
- A function to provide a feedback to inhabitants as a regional-knowledge database in Toei-cho.
- A function to access information of similar activities in neighborhood and of events in the real world.
- A function to achieve interactive communication.
- Facilitating building and operation by limited funds.
- The performance of adding on a flexibly in function according to the request of the users.
- A lot of operation results and stabilization performances.

This study adopted XOOPS as a based system through the result of previous work.

XOOP has many samples of introductionary cases by the government and the municipal body[4]. We didn't need to purchase software because XOOPS and the runtime environment, such as PHP, APACHE and MySQL were Open Source License. Furthermore, if you just know their basic operation, you easily could build and maintain a website. The predominant point of XOOPS is to provide various kinds of functional module free of charge. The website for the same purpose could be built in a flexible fashion by combining module.

3.3 Building

We began to build a virtual community on the base of the earlier study.

Firstly, the function to administer as a user is to be installed to distinguish users between Member and Guests. We tried to keep a high level of information and to attain a sense of purpose by each user through the function that Guests were available to read and member both to read and write.

Secondly, information about regional resources and regional knowledge, which is wisdom of life and traditional culture, were stored on virtual community by function of album, BBS and blog. The detail of regional resources and location information were illustrated on map by WebGIS.

Thirdly, RSS reader function was installed to keep RSS information of registered websites and to display this headline information anytime. The administrator could manually register the RSS of a website describing similar activity and events. Now we have a plan to give member rights to register RSS in the future.

[4] For instance, the Tsukuba City Board of Education has introduced XOOPS to elementary schools in the city (http://www.tsukuba-ibk.ed.jp/online/), and "NetCommons" system by National institute of informatics has based on XOOPS (http://www.netcommons.org).



Fig. 3. Image of system framework and display capture

4 Evaluation

4.1 Variability and Inquiry Survey

According to access logs in the virtual community, we found about 4,000 hits per month (about 950 in number for unique access). Guest hits were almost 90% of them and we supposed that there was something of significance to build as a portal site for users who are interested in "Hanamatsuri". Moreover, the access logs did not decline after the experience workshop. Thus the virtual community would be able to support the sustainable exchange program continuously.

According to the results of the questionnaire survey on usability, 31% of responders said “Very good” or “Good”. 35% said “Not too bad”. 4% said “Bad”.

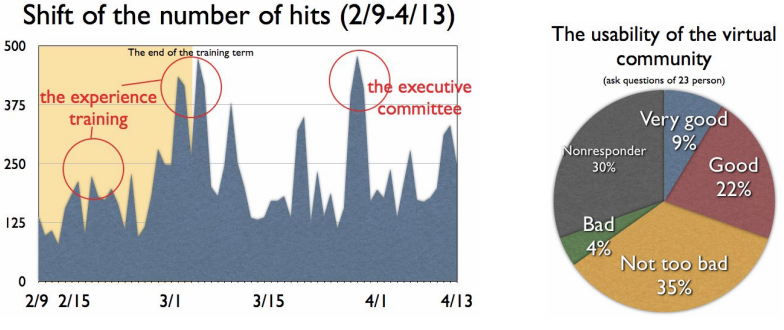


Fig. 4. The numbers of access (07/02/09-07/04/13) and inquiry survey

4.2 Evaluation for Study

The virtual community could store several opinions and information about other exchange activities. Thus the virtual community could make opportunities to find the

solutions but we could not see the active posting by members. In particular, the posting on the virtual community by Toei-cho residents are less than others. We supposed that the reason would have no motivation to post for users and no useful interface of the virtual community.

On the other hand, the virtual community connected to four organizations with the same activities as the exchange network. (i.e. the activity of Shitara-cho, Toyone-mura and Nagoya University of Arts etc.)

Furthermore, the exchange network would be re-founded as an official organization in the near future. Then the committee decided to develop exchange activities based on the virtual community. They said that they would like to combine their activities with the virtual community.

As a result, we supposed that the virtual community was evaluated as an important tool in the exchange.

5 Future Issues

To promote more posting on the virtual community, we suppose it is necessary to post by rural residents. We can see a lot of people in rural areas who use a cellular phone without distinction of age or sex. Therefore the virtual community should be available from many kinds of device such as cellular phone and PDA.

We also have to consider about the usability of the virtual community. According to the result of our research about WebGIS, the effectiveness of it was indicated by many official organizations. For example, “Study group about ICT tool available for the regional community” produced by LADC (Local Authorities systems Development Center) introduces BBS by map on the Internet. Furthermore, MLIT (the Ministry of Land, Infrastructure and Transport) introduces effectiveness and impact of it. Consequently we hope that WebGIS would be considered as the effective factor in the virtual community. Now we make a prototype.

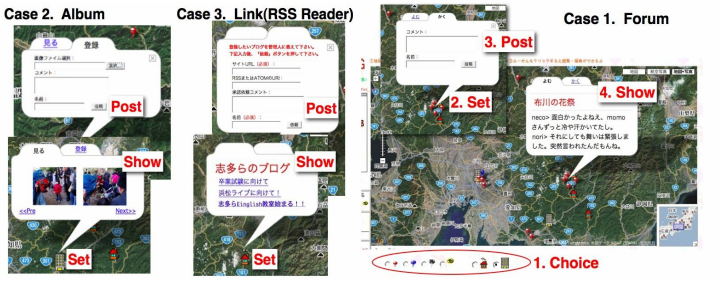


Fig. 5. Web Interface used by map (Draft)

6 Conclusion

The purpose of the study was to propose a support tool to find a method to limit depopulation through an interregional exchange between rural and urban. The method

of study was to build the virtual community that is available to share the information of exchange and the depopulation problems with the urban area and to discuss about them. We tried to find the best solution for depopulation in discussing it with a lot of people in any place.

The present result suggested that the virtual community could make opportunities to find the solutions. At first, the virtual community could store opinions and information of personal activities in other area. Secondly, the virtual community was evaluated as a good tool for exchange program and was adopted as a base of exchange in the real world.

However the less comments issue was left behind. To find more active posters, it is necessary that we build up the system and redesign user interface of the virtual community. Further studies and experiments are needed in order to achieve our purpose. We continue our efforts to study about new values created by the virtual community.

References

[1] Shoji, M.: Nikkei BP Government Technology14 (in Japanese) (December 2006)
[2] Securing the permanent resident outside of the region in Kitashitara-Area and proposal of both administrative service and foundation of living: white paper in Aichi prefecture (in Japanese) (2006)
[3] Maki, S.: Activation for community by ICT -promotion taking part in 'Public' by inhabitants (in Japanese), NTT Data Co., ltd. (2006)
[4] O'reilly, T.: What Is Web2.0 (2005), http://www.oreillynet.com/pub/a/oreilly/tim/news/2005/09/30/what-is-web-20.html
[5] Manabe, R.: 'KAKIKO-Map' as a BBS by map left a message on, The 5th study group. Local Authorities systems Development Center (in Japanese) (July 24, 2006), http://upmoon.t.u-tokyo.ac.jp/kakikodocs/select.html
[6] Inauguration of Study group for community: MIC's Press release (in Japanese) (February 1, 2007), http://www.soumu.go.jp/menu_03/shingi_kenkyu/kenkyu/community/070207_1.html
[7] Maruta, H.: New land created by Web -Invitation for regional information. Koudansya (in Japanese) (2007)
[8] The numbers of households available for broadband service counted on the Provinces: MIC's white paper (in Japanese) (December 2007), http://www.soumu.go.jp/joho_tsusin/broadband/info/info_03/index.html

Indexing of Moving Objects on Road Network Using Composite Structure

Jun Feng[1], Jiamin Lu[1], Yuelong Zhu[1],
Naoto Mukai[2], and Toyohide Watanabe[2]

[1] Hohai University, Nanjing, Jiangsu 210098 China
{fengjun,welmanwenzi}@hhu.edu.cn
[2] Nagoya University, Nagoya, Aichi 464-8603 Japan

Abstract. Composite structures are proposed to index moving objects in road network. However, there are many problems for indexing moving objects for current and forecasting queries on road network, efficiently. This paper proposes methods for improving such kinds of structure, and gives a new structure R-TPR$^{\pm}$ Tree. Evaluation shows the new structure outperforms that of R-TPR-tree in query validity and disk access.

1 Introduction

With the improvements of geographic positioning technology and the popularity of wireless communication, location-based service (LBS) is proposed to provide the user with a service that is dependent on positional information associated with the user [1]. Prominent examples of LBS concern vehicle navigation, tracking and monitoring, where the positions of air, sea, or land-based equipment such as airplanes, boats and cars are of interest.

This paper is concerned with the indexing of moving objects on road network for real time and forecasting purposes, for example, transportation control and real time transportation guidance. It is quiet different to the indexing methods for objects moving in 2D free space, because not only the moving directions but also the distance among moving objects are constrained by the road network. Therefore the neighboring relations among moving objects should take the underlying road network into consideration. Past works on indexing of spatiotemporal data concerns either past data or present and future data. The former direction includes works in 2D space [2,3,4] or in spatial networks [5]. The latter direction includes works on indexing the present positions of moving points [3,6,7,8,9,10], all these approaches deal with spatial data in 2D space and cannot be applied to the moving objects in spatial network.

A straightforward method for indexing the static road network and the moving objects at the same time is the Composite Structure, shown in Fig. 1. The Composite Structure is made up of two parts: one is a Static Tree managing the information of road network, its leaf node such like the R1 in Fig. 1 denotes a region of the road network, points to a dynamic tree indexing moving objects move in this region, and all these dynamic trees compose the Dynamic Index

B. Apolloni et al. (Eds.): KES 2007/ WIRN 2007, Part II, LNAI 4693, pp. 1097–1104, 2007.
© Springer-Verlag Berlin Heidelberg 2007

under the Static Tree. Using Composite Structure, we can insert and query spatial locations efficiently with the Static Tree, and update the positions of moving objects quicker as we divide them into small subsets and index them separately.

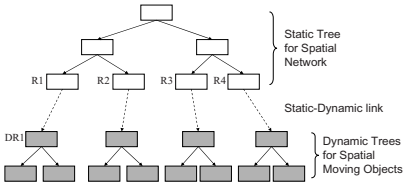


Fig. 1. Image of a Composite Tree

However, to assure the effective of the composite structure for indexing current positions of moving objects on road network, there are two issues need to be solved:

- **Update cost.** The update cost includes the cost for one operation and the times of operations. In current composite structures, index element of the Static Tree is road segment (in other words, the forecasting region is limited by the segment), and results in frequent updates when the objects move into other segments. Therefore, the issue of making the index element longer, enlarge the forecasting region and decrease the update times is the key to a "really" efficient composite tree.
- **Query cost.** The queries on moving objects includes the region query, nearest neighbor query, aggregate query and so on. In applications, aggregate queries are important as they can answer such kinds of questions: a specific road segment is fluent or not, or the average speed on a road. Therefore, how to make an index structure support kinds of queries should be considered.

In this paper, we propose methods to getting an efficient composite structure for indexing moving objects on road network. We bring in the idea of road connection algorithms [11] for dividing the approximate regions of Static Tree, which use the road network within which the objects are assumed to move for predicting their future positions, and can increase the forecasting period of moving object and decrease the update times of the database. Furthermore, we propose new spatiotemporal index structure for moving objects for every road region, which arranges the objects by their moving directions as well as their locations. This structure supports the aggregate and spatiotemporal queries more efficiently than other structures.

In the followings of this paper, Section 2 gives the road connection methods and the analysis based on experiments; Section 3 depicts our new spatiotemporal structure and Section 4 is experiments and analysis; conclusions are given in Section 5.

2 Road Connection Algorithms and Analysis

There is reality meaning in connecting several road segment as an index element, as on one hand, regarding the connected segments as a group will decrease the element number of the Static Tree, and increase the query efficiency; on another hand, the relativity of road segment provides the important cause for segment selection for a moving object when it faces several segment at a cross node. The better the connection algorithm is, the more precision the forecasting is. However, how to decide the size of forecasting region or the length of forecasting path is a problem in monitoring moving objects on road network. When the length of forecasting path is too long, the forecasting precision cannot be approved; while it is too short, the gathering of sample locations will be a heavy burden.

To find out a better way for road connection, we apply LSC algorithm (Length and Side-Based Segment Connection Algorithm) and ASC algorithm (Angle-based Segment Connection Algorithm) [11] to the real road networks (California Road Network from Geographic Data Technology Inc, 1999). Experiments shows that the numbers of road segments are decreased more than 59 percent, the length of segment and the nodes number are increased. And the road networks processed by ASC is better for representing the real road network for car tracking, because the number of main roads is less than those non-main roads, and the nodes included in the main roads are more, the length is longer than those of the non-main roads. These properties are just similar to the driving properties of cars in road networks.

In our composite tree for the moving objects in road networks, the problem of creating a "good" Static Tree for road network is how to decrease the update times occurred at every changing segment time of moving objects, the segment connection algorithms is just bring a practical way for solving this problem. In the followings, we use the road segment processed by ASC algorithm as the basic index element in our Static Tree.

3 R-TPR$^{\pm}$ Tree

R-TPR$^{\pm}$ Tree is a composite index structure used for managing the real-time data of moving objects on the road network. It is combined by two parts: a static 2D R^*-tree indexing the road network on the top of a forest of dynamic 1D TPR$^{\pm}$-trees indexing the moving objects. Objects on every road segment are supposed to be moving at a specific speed just the same as that at the sampling time. Though the directions of them may change all the time along the road segments, if they are still in one segment, it is thought to be keeping the same direction. Therefore, in a road segment, the moving directions of moving objects are just two: from start node s to end node e(denoted as + direction), or from e to s(denoted as - direction). Furthermore, in the real road networks, the cost or the weight of a segment in two directions are not always as the same, we propose a TPR$^{\pm}$-tree for indexing the spatio-temporal information in one road segment.

TPR$^{\pm}$-tree is a tree with a root node and two subtrees. In every subtree, there are only objects in the same direction are managed, in other words, the

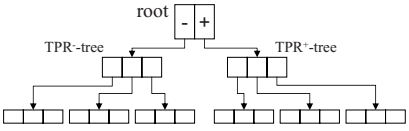


Fig. 2. Image of TPR$^{\pm}$ Tree

two subtrees are managing the objects in different moving directions. In the root node, there are two pointers point to these two subtrees, one is +, another is -. Fig. 2 depicts this situation. In a more formal speaking, there are two elements of R-TPR$^{\pm}$ Tree's data set, R and O, R is the road network and O is the moving objects set. After dividing R into segments by ASC algorithm, there will be a set of independent segments, R=$\{S_1, S_2, , S_n\}$, where $S_i (1 \leq i \leq n)$ means a segment of the road network, and n is the number of the segments in R. While R is divided, the set of moving objects O is also divided into n subsets, represented as O = $\{O_1, O_2, \cdots, O_n\}$, and $O_i = \{mo_p, mo_{p+1}, \cdots, mo_q\} (1 \leq i \leq n\ )$, where mo_j $(p \leq j \leq q\ )$denotes a moving object that moves in the segment S_i, and we will get: NUM(O) = $\sum_{i=1}^{n} NUM(O_i)$, Where NUM(O_i) means the amount of moving objects in the subset O_i. As we said before, in each segment of the road network, a moving object moves from the start point to the end, or opposite. We use TPR^{+}-tree(TPR^{-}-tree) to index the data of the set O_i in +(-) direction. The leaf node of the TPR^{+}-tree(TPR^{-}-tree) is denoted as: (j-Identifier, P_j, V_j), where j-Identifier refers to the object O_j. $P_j = \frac{Length(mo_j,A)}{Length(A,B)}$ is the relative position of the segment its on, where A is the start point of the segment, and B is the end point, Length(A, B) means the length between A and B. P_j approaches to 0 means O_j is as close to the start point A as much, and if O_j is more close to the end point B, P_j will approaches to 1 as much. At the same time, V_j means the velocity of O_j, when O_j moves from A to B, then $V_j > 0$, and $V_j < 0$ when it moves opposite. The non-leaf node of this TPR^{+}-tree(TPR^{-}-tree) is denoted as: (P_s, P_e, V_s, V_e, child-point). Where child-point is the address of a lower node, $|P_s, P_e|$ covers the position interval of its lower node's entries, and $|V_s, V_e|$ covers the velocity interval. As there are n subsets of O, so there will be n TPR$^{\pm}$-trees, and we can use a forest to indexing the moving objects in the road network.

Segments R are indexed by a 2D R^*-tree. The leaf node of the index is denoted as: (i-Identifier, MBR_i, TPR$^{\pm}$-point, P_i). Where i-Identifier refers to the segment S_i, MBR_i is a minimal bounding rectangle of the segment, TPR$^{\pm}$-point the pointer to the TPR$^{\pm}$-tree indexing the objects of O_i, and P_i is something other properties of the segment, such like length. The non-leaf node of this R^*-tree is denoted as: (child-MBR, child-pointer), where child-pointer points to the a lower node, and child-MBR covers all MBRs of its lower nodes.

4 Forecasting Query Method

In application, it is more important to get aggregate information than calculating exact positions of cars moving on the road network. A forecasting aggregate

query Q is denoted as: (R, t), where R describes the forecasting range, and t refers to the forecasting period. Using the 2D R^*-tree, we can find out the segments which we want to calculate inside R, and those TPR$^{\pm}$ Trees indexing moving objects in these segments. And we will get how many cars will be still in this segment after t seconds later according to their current speeds and positions with this forecasting method.

The non leaf node of the TPR$^{\pm}$ Tree is denoted as: (P_s, P_e, V_s, V_e, child-point), as the moving directions of those objects indexed by the same node are the same, and we assume their speeds keep constant, so we can get:

$$NP_s = P_s + V_s * t$$

$$NP_e = P_e + V_e * t$$

Here, $|NP_s, NP_e|$covers the position interval which those objects will not overstep after t seconds. So if NP_e is less than 1, which means all those objects indexed by this node will not exceed this segment t seconds later, and if NP_s is larger than 1, which means all those objects will leave this segment t seconds later. When NP_s is less than 1 and NP_e is larger than 1, we need recurs to its sub nodes and repeat this algorithm until arriving to the leaf nodes.

Those objects leave the segment after t seconds will select proper segments, insert into those other TPR$^{\pm}$ Trees , and deleted after the algorithm end. Those objects still in will be added into the traffic flow of current segment.

Algorithm Forecasting Traffic Flow FTF(R, t)
/*The number of segments of this road network is n. There is a global queue *TransArray* to save nodes will be inserted into other trees and a global array *FLOW[]* for recording the result. **Input**: R is the predict Range, and t is the forecasting period. **Output**: FLOW[] contains the traffic flows every segment will have after t seconds. */

```
1.   int FLOW[n]
2.   initialize a node array TransArray
3.   TPR±-tree trees[]    //find those TPR-trees inside R
4.   trees = 2D R*-tree.RangeQuery(R)
5.   for each tree in trees
6.      FLOW[tree.ID] = Single Segment Forcast(tree.+node, t)
7.      FLOW[tree.ID] = Single Segment Forcast(tree.−node, t)
8.   while TransArray is not empty
9.      tree = transArray.headNode.selectRoute
10.     tree.Insert(transArray.headNode)
11.     FLOW[tree.ID] = Single Segment Forcast(tree.+node, t)
12.     FLOW[tree.ID] = Single Segment Forcast(tree.−node, t)
13.     tree.delete(transArray.headNode)
14.     transArray.removeHead
```

End Algorithm Forecasting Traffic Flow FTF

Algorithm Single Segment Forecast SSF(N, t)
/∗ **Input**: N is a node of TPR tree, at the beginning of this algorithm, it means the root node of TPR$^+$- tree(TPR$^-$-tree); t is the forecasting period. **Output**: FLOW is the number of cars which will be still in the segment t seconds later indexed by N ∗/

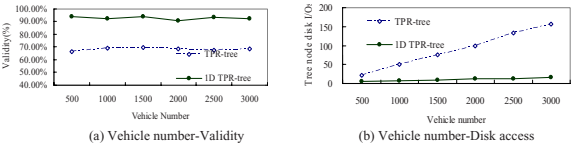
```
1.   int FLOW;
2.   if ( NPe ≤1 )
3.      /∗all cars indexed by N will not exceed the segment t seconds later ∗/
4.      FLOW += N.carNum;
5.    else if (NPs > 1)
6.      /∗all cars indexed by N will exceed the segment t seconds later ∗/
7.      For each subNode Ni in N
8.         TransArray.add(Ni);
9.    else if (N.level == 0)
10.     /∗part of cars will be in the segment and N is a leaf node ∗/
11.     For each car Moi in N
12.        If (Moi.position + Moi.velocity*t ≤1)
13.           FLOW++;
14.        Else
15.           transArray.Add(Moi);
16.   else
17.     /∗part of cars will be in the segment and N is an non leaf node ∗/
18.     For each subNode Ni in N
19.        FLOW += flowCal(Ni);
20.   return FLOW;
```

End Single Segment Forecast SSF

5 Experiments and Analysis

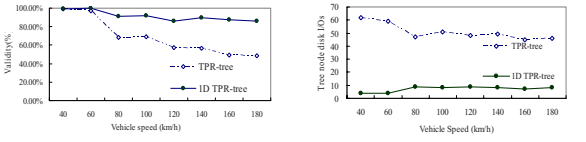
In this section, we evaluate the method proposed by this paper compared with that using TPR-tree [8] in the composite structure in terms of forecasting query validity and disk accesses (here refers to tree node access times). For our experiments, we used a PC of Pentium IV CPU with 3.19 GHz and 512M bytes of main memory. The Trees and the algorithms were implemented in C++ and compiled using Visual C++ version 2003.

Our simulating environment for experiments are: the moving objects in a road network are indexed by R-TPR$^\pm$ Tree and R-TPR-tree (the Static Tree is R*-tree and the Dynamic Tree is TPR-tree). The validity tests and tree node disk access tests are carried out for the vehicles moving on a 1 km road segment. For the queries, we varied vehicle numbers or forecasting period or vehicle speed. For every query, we performed it 3 times and the values shown in this section are the average per query. Fig. 3 depicts the results varied with different vehicle numbers when vehicle speed is 100 km/h and forecasting period is 5 seconds.



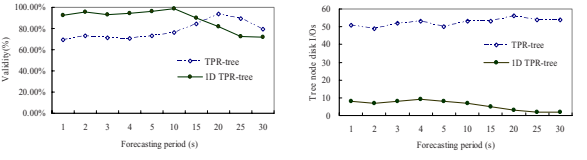
(a) Vehicle number-Validity (b) Vehicle number-Disk access

Fig. 3. Situations of vehicle speed is 100km/h and forecasting period is 5s



(a) Vehicle speed-Validity (b) Vehicle speed-Disk access

Fig. 4. Situations of vehicle number is 1000 and forecasting period is 5s



(a) Forecasting period-Validity (b) Forecasting period-Disk access

Fig. 5. Situations of vehicle number is 1000 and speed is 100km/h

Fig. 4 depicts the results varied with different vehicle speed when there are 1000 vehicles moving on that segment in two directions, and the forecasting period is 5 seconds. Fig. 5 depicts the results varied with different forecasting period when there are 1000 vehicles moving on that segment in two directions, and vehicle speed is 100 km/h.

By observing the results, there are: first, as expected, the number of disk accesses of our method is quite fewer than that of TPR-tree, and keeps a stable line. Second, it is clear that the validity of our method is better than that of TPR-tree while the forecasting period is shorter.

6 Conclusion

In this paper, we proposed R-TPR$^{\pm}$ Tree for indexing moving objects in road network for current and forecasting query. It is a composite structure consists of a R*-tree and a forest of TPR$^{\pm}$ Tree proposed by this paper. Evaluation shows that our method outperforms that of R-TPR-Tree in query validity and disk access. In our future work, the more forecasting queries under more complex situations will be tested, such as forecasting aggregate queries for several connected road segments and so on.

Acknowledgements

This research is supported by NSFC(granted number 60673141, Research on the Index Structure of Spatial Network-based Moving Objects).

References

1. Chung, J.D., Peak, O.H., Lee, J.W., Ryu, K.H.: Temporal pattern mining of moving objects for location-based services. In: Proc. Int'l Conf. Database and Expert Systems Applications (2002)
2. Hadjieleftheriou, M., Kollios, G., Tsotras, V., Gunopulos, D.: Efficient indexing of spatiotemporal objects. In: Proc. of the 8th Int'l Conference on Extending Database Technology, pp. 251–268 (2002)
3. Kollios, G., Gunopulos, D., Tsotras, V., Delis, A., Hadjieleftheriou, M.: Indexing animated objects using spatiotemporal access methods. IEEE TKDE 13(5), 758–777 (2001)
4. Lazaridis, I., Porkaew, K., Mehrotra, S.: Dynamic queries over mobile objects. In: Proc. of ACM-SIGMOD Conference on the Management of Data, pp. 269–286. ACM Press, New York (2002)
5. Pfoser, D., Jensen, C.S.: Indexing of network constrained moving objects. In: Proc. of GIS 2003, pp. 25–32 (2003)
6. Agarwal, P.K., Arge, L., Erickson, J.: Indexing moving points. In: Proc. of the 19th ACM Symposium on Principals of Database Systems, pp. 175–186. ACM Press, New York (2000)
7. Pfoser, D.: Indexing the trajectories of moving objects. IEEE Data Engineering Bulletin (2), 3–9 (2002)
8. Saltenis, S., Jensen, C.S., Leutenegger, S., Lopez, M.: Indexing the positions of continuously moving objects. In: Proc. of ACM-SIGMOD Conference on Management of Data, pp. 331–342. ACM Press, New York (2000)
9. Saltenis, S., Jensen, C.S.: Indexing of moving objects for location-based services. In: Proc. of the 18th Int'l Conference on Data Engineering, pp. 463–472 (2002)
10. Sistla, A.P., Wolfson, O., Chamberlain, S., Dao, S.: Modeling and querying moving objects. In: IEEE Int'l Conf. on Data Engineering, pp. 422–432. IEEE Computer Society Press, Los Alamitos (1997)
11. Civilis, A., Jensen, C.S., Pakalnis, S.: Techniques for efficient road-network-based tracking of moving objects. IEEE Trans. on Knowledeg and Data Engineering 17(5), 698–712 (2005)

Optimization of Vehicle Assignment for Car Sharing System

Kentaro Uesugi, Naoto Mukai, and Toyohide Watanabe

Graduate School of Information Science, Nagoya University,
Furo-cho, Chikusa-ku, Nagoya, 464-0803, Japan
Faculty of Engineering Division 1, Tokyo University of Science,
Kudan-kita, Chiyoda-ku, Tokyo, 102-0073, Japan
{uesugi,watanabe}@watanabe.ss.is.nagoya-u.ac.jp,
mukai@ee.kagu.tus.ac.jp

Abstract. Recent advances in information technology enable realization of new on-demand transportation systems. A car sharing system is one of the new on-demand transportation systems and its concept is that people share vehicles to save maintenance cost. A car sharing system is expected as a solution for traffic jams and lack of parkings. There are two types of car sharing systems: the one-way type and the round-trip type. In this paper, we focus on the one-way type. In the one-way type, users can return vehicles to any stations at any time. Thus, it is hard to keep distribution balance of parked vehicles among stations. We propose a method for optimizing vehicle assignment according to distribution balance of parked vehicles. Finally, we report experimental results of our simulation, and evaluate the effect of our method.

Keywords: car sharing, vehicle assignment.

1 Introduction

In these years, developments of information and communication technologies become the basis of Intelligent Transport System (ITS)[1]. The advances of global positioning systems and communication techniques enables realization of new on-demand transportation system such as Demand Bus System[2]. We focus on the car sharing system [3,4,5] which is one of the new on-demand systems. The concept of the system is to share a small number of vehicles among a large number of users. There are three contributions by the system: cost sharing, earth sharing, and road sharing.

The car sharing system is classified into the one-way type and the round-trip type according to where users return a vehicle. In the round-trip type, a user has to return a vehicle at the station where the user rent it. Therefore, the round-trip type applies to a user who shall return again to the original position. On the other hand, in the one-way, a user can return a vehicle at any station. Thus, users can use vehicles like a public bus or taxi. Consequently, the one-way type is more flexible and useful than the round-trip type. However, it is hard

B. Apolloni et al. (Eds.): KES 2007/ WIRN 2007, Part II, LNAI 4693, pp. 1105–1111, 2007.
© Springer-Verlag Berlin Heidelberg 2007

to keep distribution balance of parked vehicles among stations. If distribution balance is disrupted, a lack of vehicles is occurred at any stations. As a result, it is impossible to assign vehicles to new users immediately. Therefore, the one-way type needs a mechanism for balancing parked vehicles among stations. In this paper, we propose a method for optimizing distribution balance of parked vehicles by assigning the optimum number of used vehicles to users. Thereby, it is possible to keep distribution balance of parked vehicles and keep a convenience of car sharing system. Finally, we report experimental results of our simulation, and evaluate the effect of our method.

The remainder of this paper is organized as follows: Section 2 formalizes one-way car sharing system. In Section 3, we mention assignment of the number of vehicles. In Section 4, we report simulation results and consider effectiveness and possibility of our method. Section 5 concludes and offers future works.

2 Formalization

2.1 Service Area

A road network in a service area is given by a graph G which consists of nodes N and edges E as Equation 1. The nodes represent intersections and the edges represent road segments between intersections. For simplicity, we do not consider one-way traffic and traffic jam. We denote the distance between n and n' as $d(n, n')$.

$$G = (N, E), N = \{n_1, n_2, ...\}, E = \{e(n, n') | n, n' \in N\} \tag{1}$$

There are h stations where users get on/off cars in a service area. The stations are corresponded to a subset of nodes (Equation (2)).

$$S = \{s_1, s_2, ..., s_h\} \in N \tag{2}$$

All users aim to move from a node to another node. And all users select n_i as a start node or a goal node with a choice probability: p_i. A probability p_i is given by Equation(3) which contains a weight of choice probability: w_i. A weight w_i of choice probability is set from 0 to 1 randamly.

$$p_i = \frac{w_i}{\sum w_k} \tag{3}$$

2.2 Station

A station s_i is given by Equation (4), where V_i is a set of parked vehicles, R_i is a list of reservations, and α_i is a optimum number of parked vehicles. The optimum number α_i of parking vehicles is obtained by the number of reservations at s_i: $|R_i|$, the number of reservations at all the stations: $|R_{all}|$, and the number of parked vehicles at all the stations which is given by Equation (5). For simplicity

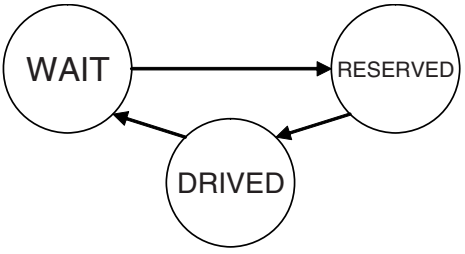


Fig. 1. State transition of a vehicle

of problems, we do not consider a limit of the number of parked vehicles i.e, it is possible that infinite vehicles park at a station.

$$s_i = (V_i, R_i, \alpha_i) \tag{4}$$

$$|R_{all}| = \sum_{j=0}^{K-1} |R_j|, \alpha_i = |V| \times \frac{|R_i|}{|R_{all}|} \tag{5}$$

2.3 Vehicle

Vehicles are given by V in Equation (6). The traveling speed of vehicles is fixed as q_v. The state transition of a vehicle is shown in Figure 1. In WAIT, a vehicle waits for reservation at a station. A request is given as a pair of ride-on and drop-off stations. If a vehicle received a reservation from a user, the vehicle is reserved for the user and waits for his/her arrival (RESERVED). In RESERVED, the vehicle does not receive other reservations. The user drives the vehicle from the station to his/her drop-off station (DRIVEN). After the driving, the state of the vehicle transits again to WAIT. We assume that β persons can ride on a car at the same time.

$$V = \{v_1, v_2, ..., v_m\} \tag{6}$$

2.4 User

Users are given by U in Equation (7). The walking speed of users is fixed as q_u. A user u_i is given by Equation (8) which contains a reservation: r_i, the number of persons: γ_i, a start node: n_i^s, and a goal node: n_i^g. A user u_i is a group of γ_i persons, but we call just a user in this paper. A reservation r_i is given by Equation (9) which contains a start time: t_i^s, a end time: t_i^e, a station where u_i will ride on: s_i^r, and a station where u_i will drop off: s_i^d. The state transition of a user is shown in Figure 2. In SEARCH, a user u_i decides the ride-on station s_i^r and the drop-off station s_i^d. Here, let S_{wait} be the set of stations where some vehicles are available (in WAIT). The station, which is closest to the start node n_i^s, is selected from S_{wait} as ride-on station s_i^r. And, the station which is closest to the goal node n_i^g is selected from S as the drop-off station s_i^d. Next, the user u_i decides whether they use vehicles or not by the condition $T_i^t \leq T_i^f$. The

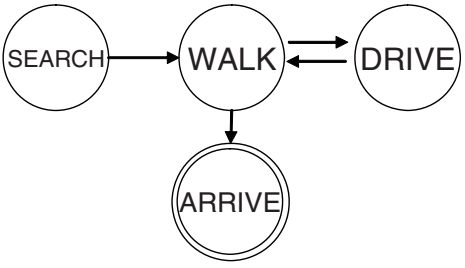


Fig. 2. State transition of a user

travel time for the user u_i by using the system is given by T_i^t which is defined in Equation (10). The travel time for the user u_i by walking is given by T_i^f which is defined in Equation (11). If the condition is satisfied, firstly the user u_i sends the reservation r_i and walks to the ride-on station s_i^r (WALK). Then, the user u_i drives the car to the drop-off station s_i^d (DRIVE). Finally, the user u_i walks to the goal node n_i^g (WALK), and finally arrives at his/her destination (ARRIVE). Otherwise, the user u_i walks to the goal node n_i^g directly by walking (WALK)

$$U = \{u_i, u_2, ...\} \tag{7}$$

$$u_i = (r_i, \gamma_i, n_i^s, n_i^g) \tag{8}$$

$$r_i = \left(t_i^s, t_i^e, s_i^r, s_i^d, v_j\right) \tag{9}$$

$$T_i^t = \frac{d(n_i^s, s_i^r) + d\left(s_i^d, n_i^g\right)}{q_u} + \frac{d\left(s_i^r, s_i^d\right)}{q_c} \tag{10}$$

$$T_i^f = \frac{d(n_i^s, n_i^g)}{q_u} \tag{11}$$

3 Optimization of Vehicle Assignment

The probability with which a user selects a node as a start node or a goal node is not uniform among stations. Moreover, a user can return a vehicle to any stations at any time. It is hard to keep distribution balance of parked vehicles among stations. Disruption of distribution balance causes decreasing convenience of a car sharing system. Therefore, it is necessary to optimize the distribution of idle vehicles. In this paper, we propose a method for assigning the optimum number of using vehicles to users, according to distribution of parked vehicles, in order to assign the number of in/out vehicles between ride-on/drop-off stations. There are three ways of assignment: normal assignment, divided assignment, and combined assignment.

3.1 Way of Assignment

Normal assignment. In normal assignment, a user ($n\,(n \leq \beta)$ persons) rides on one vehicle. As a result, the user subtracts one vehicle from the ride-on station, and adds the vehicle to the drop-off station.

Divided assignment. In divided assignment, a user ($n\,(2 \leq n \leq \beta)$ persons) rides on $m\,(2 \leq m \leq n)$ vehicles with the divided group. As a result, the user subtracts m vehicles from the ride-on station, and adds m vehicles to the drop-off station. Compared with the normal assignment, the divided assignment increases the number of moved vehicles from 1 to m, so that it has advantages to assist "decreasing parked vehicles at a ride-on station" and "increasing parked vehicles at a drop-off station".

Combined assignment. In combined assignment, k user groups (total of $n\,(k \leq n \leq \beta)$ persons) which have the same drop-off station, ride on one vehicle with the combined group. As a result, the users subtract one vehicle from the ride-on station, and add one vehicle to the drop-off station. Compared with the normal assignment, the combined assignment decreases the number of moved vehicles from k to 1, so that it has advantages to prevent "decreasing parked vehicles at a ride-on station" and "increasing parked vehicles at a drop-off station".

3.2 Decision of the Number of Assigning Vehicles

In a station which has excessive vehicles, it is possible to reduce the number of vehicles by assigning a divided ride. On the other hand, in a station which has less vehicles, it is possible to increase the number of the vehicles by assigning a division ride. Thus, it is necessary to decide the number of assigned vehicles depend on distribution balance of parked vehicles. In this paper, decision of the number of assigned vehicles is based on a square residual error sum for the number of parked vehicles and the optimum number of parked vehicles. Residual sum of squares T between the optimum number of parking vehicles α and the number of parking vehicles $|V|$ is shown by Equation (12).

$$T = \sum_{i}^{K} \left(\alpha_i - |V_i|\right)^2 \tag{12}$$

All users select the number of using vehicles which minimizes residual sum of squares T. Thus, all users select the number of using vehicles which minimizes residual sum of squares T_{ij} between a ride-on station and drop-off station which is shown by Equation (13).

$$T_{ij} = \left(\alpha_i - |C_i|\right)^2 + \left(\alpha_j - |C_j|\right)^2 \tag{13}$$

4 Experiments

There are four experimental patterns shown in Table 1. A service area is set to a square (600×600 pixels). A road network in the area is set to a grid network (30×30), i.e., the length of all road segments is 20 pixels. The service process is repeated until the max time $5.0 \times 10^6 t$. Other parameter setting is shown in Table 2.

First, we report experimental results related to occurrence rate of reservations. The occurrence rate of reservations is set from 1% to 20%. The number of vehicles

Table 1. Experimental pattern

Pattern	Normal assignment	Divided assignment	combined assignment
PT1		—	—
PT2			—
PT3		—	
PT4			

Table 2. Experimental parameter

Parameter	Value
$\|S\|$	20
$\|C\|$	35
$\|q_v\|$	20
$\|q_u\|$	2
γ	2
β	4

is set to a fixed value 100. Figure 3(a) indicates that PT4 could decrease the residual sum of squares compared with PT1. However, PT3 could not decrease the residual sum of squares compared with PT4. This is due to rare selection of combined assignments. It is necessary to assign combined assignment that users have the same drop-off station. As a result, it is thought that combined assignment was not assigned so much. Moreover, Figure 3(b) indicates that PT2, PT3, and PT4 can improve the utilization rate compared with PT1. Altogether to control the number of assigning vehicles, in order to control the number of in/out vehicles at ride-on/drop-off stations improve the utilization rate.

Second, we report experimental results related to the number of vehicles $|V|$. The number of vehicles is set from 40 to 200. The occurrence rate of driving

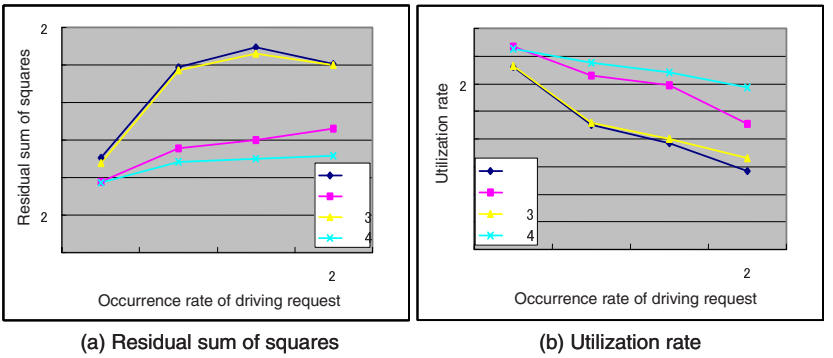


Fig. 3. Experimental results related to occurrence rate

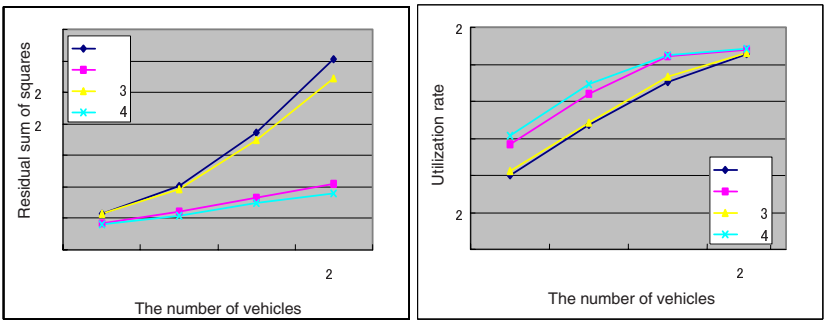


Fig. 4. Experimental results related to the number of vehicles

requests is set to a fixed value 5%. Figure 4(a) indicates that increasing the number of vehicles makes residual sum of squares increase, especially as PT1, PT3: not assigned combined assignments.

5 Conclusion

In this paper, we focus our attention on the one-way type car sharing system. It is hard to keep a distribution balance of parked vehicles among stations in the one-way type, because the one-way type user can return a vehicle at any station. The distribution balance of vehicles is a key problem to be solved for improving convenience of the system. Therefore, we propose a method for optimizing distribution balance of parked vehicles by assigning the optimum number of vehicles to users. The effectiveness of our method is estimated by our computer simulation. The results of the simulation show that our method is effective for one-way car sharing system. In our future work, we must consider incentives for users to behave according to proposed model. In addition, we should perform the simulation by using real city maps.

References

1. Nakashima, H., Kurumatani, K., Itoh, H.: Supporting a society with ubiquitous computing. IPSJ-MGN450904t (2004)
2. Ohta, M., Shinoda, K., Noda, I., Kurumatani, K., Nakashima, H.: Usability of demand-bus in town area. IPSJ-MBL02023033 2002 (2002)
3. Abraham, J.E.: A survey of preferences in carsharing. The Journal of World Transport policy & Practice (2000)
4. Britton, E.: Executive summary in carsharing 2000: Sustainable transport's missing link. The Journal of World Transport Policy & Practice (2000)
5. Mukai, N., Watanabe, T.: Dynamic Location Management for On-Demand Car Sharing System. In: Khosla, R., Howlett, R.J., Jain, L.C. (eds.) KES 2005. LNCS (LNAI), vol. 3681, Springer, Heidelberg (2005)

An Application of Decision Trees for Rule Extraction Towards Telecommunications Fraud Detection

Constantinos S. Hilas[1] and John N. Sahalos[2]

[1] Dept. of Informatics and Communications, Technological Educational Institute of Serres, Serres, GR-621 24, Greece
chilas@teiser.gr
[2] Radiocommunications Laboratory, Aristotle University of Thessaloniki, Thessaloniki, GR-541 24, Greece
sahalos@auth.gr

Abstract. Telecommunications fraud has drawn the attention in research due to the huge economic burden on companies and to the interesting aspect of users' behavior modeling. In the present paper, an application of decision trees to fraud detection is presented. The appropriate user behavior modeling is, also, discussed. The trees are used for rule extraction in order to distinguish between normal and fraudulent activities in a telecommunications network. Several real cases of defrauded user accounts are modeled by means of selected usage features. Decision trees are applied in order to identify the critical values that separate fraud from legal use. These thresholds are expressed in the form of rules that will be used in a rule based expert system.

Keywords: Fraud detection, user profiles, rule extraction, feature selection, telecommunications.

1 Introduction

Telecommunications fraud can be simply described as any activity by which telecommunications service is obtained without intention of paying [1]. This kind of fraud has certain characteristics that make it particularly attractive to fraudsters. The main one is that the danger of localization is small. This is because all actions are performed from a distance which in conjunction with the mess topology and the size of networks makes the process of localization time-consuming and expensive. Additionally, no particularly sophisticated equipment is needed, if one is needed at all. The simple knowledge of an access code, which can be acquired even with methods of social engineering, makes the implementation of fraud feasible. Finally, the product of telecommunications fraud, a phone call, is directly convertible to money [2].

Several categories of telecommunications fraud have been reported. The main are the technical fraud, the contractual, the hacking, and the procedural fraud [1]. All fraud cases can actually be view as fraud scenarios which are related to the way the access to the network was acquired. Detection techniques tailored to one case may fail to detect other types of fraud. For example, velocity traps which can identify the use

B. Apolloni et al. (Eds.): KES 2007/ WIRN 2007, Part II, LNAI 4693, pp. 1112–1121, 2007.
© Springer-Verlag Berlin Heidelberg 2007

of a cloned cell phone will fail to detect a case of contractual fraud. So, fraud detection focuses on the analysis of users' activity. The related approaches are divided into two main subcategories. The absolute analysis searches for limits between legal and fraudulent behavior, and the differential approach tries to detect extreme changes in the user's behavior. In both cases, analysis is achieved by means of statistical and probabilistic methods, neural networks and rule based systems. The authors in reference [3] criticize the use of indicators of excessive usage as they may not only imply fraud but may also point to the best customers. A comparison of probabilistic methods with those that use rules is given in [4]. In 1999, Fawcett and Provost, [5], proposed a combination of rules and profile extraction, in order to detect fraud. The outputs of their system are combined via a trained linear model in order to produce alarms. Rosset et al report encouraging results from the use of rules that are exported with a variant of the C4.5 algorithm, [6]. Alves et al. propose two anomaly detection methods based on the concept of signatures for the detection of superimposed fraud [7]. The appropriate feature extraction procedure is dealed with in [8]. In a previous work, [9], the authors of the present paper concluded to a user behavior characterization model that gives good results towards superimposed fraud detection.

In the present paper, our model is combined with decision trees in order to identify thresholds that separate normal from fraudulent usage. The thresholds obtained are represented in the form of rules and will be used in a rule-based expert system for fraud detection. The expert system shall operate in the telecommunications system of a large organization. The paper proceeds as follows. In the next chapter a brief introduction to decision trees is made. In chapter 3, the data, the user modeling and the feature selection process are presented. Experimental results are given in chapter 4. In the last chapter conclusions are drawn.

2 Decision Trees

Learning algorithms usually use a divide-and-conquer approach. The input space is incrementally divided using splits that maximize information gain or some other expression of the change in knowledge. This approach leads to tree-like data structures. Decision trees are realizations of such structures. Nodes in a decision tree involve testing of a particular attribute. The test compares an attribute value with a constant or another attribute, or uses some appropriate function. Objects are assigned to each node according to their attribute value. If all the objects in a node belong to the same class then the division of this node is stopped. However, this is not always feasible in real world problems. The aim is to have leaves that are as pure as possible. Appropriate purity measures should be used and the procedure will ideally lead to pure leafs [10]. The entropy, $H(N)$, of each node is usually used as a measure of its purity.

Suppose a tree that is partially developed up to some node N. We want to know which the best parameter s for the next split is. This should be the parameter that minimizes $i(N)$ [13]:

$$\Delta i(s) = i(N) - \sum_{k=1}^{B} p_k i(N_k) \cdot \tag{1}$$

B is the number of branches in which node N splits, p_k is a fraction of training patterns sent down the link to node N_k, $\sum_{k=1}^{B} p_k = 1$, and $i(N_k) \equiv H(N_k)$ is the entropy of branch k. Equation 1 gives the difference between the information needed to recognize an element of N and the information needed to recognize an element of N with the prior knowledge of the value of a parameter s. $\Delta i(s)$ is called *information gain*. This is an expression of the information of the parent node minus the information in any possible division. The second part in the right hand side of equation 1 is called average specific conditional entropy. Information gain is also known as *Kullback - Leibler distance*, or *relative entropy*.

The drawback of information gain is that it inherently favors decisions with a large B. To avoid the drawback, the change in impurity is scaled [13], according to:

$$\Delta i_B(s) = \frac{\Delta i(s)}{-\sum_{k=1}^{B} p_k \log p_k}. \tag{2}$$

The optimal split is the one that maximizes $\Delta i_B(s)$, which is called *gain ratio impurity*. The algorithm that constructs decision trees without using gain ratio impurity is known as ID3. The use of gain ratio impurity led to one of the most popular machine learning algorithms, the C4.5 [11]. Further improvements in C4.5 allow the use of numeric variables, missing values, rule extraction, etc. Its J48 implementation, [10], is used in the present paper.

Decision trees can fit very well to training data. However, they may become large and complex. As a consequence they become over-fit and unable to generalize to a similar input. One of the most common techniques to recognize the generalization ability of an approach is to apply the technique of cross-validation. During cross-validation the training set is divided into N stratified subsets. The training procedure is repeated N times and each one of the N subsets is hold out as a validation set while the N-1 others are combined into one training set. The model is judged by the mean error over the N repeats. N is usually 10, and if this is the case we refer to *stratified tenfold cross-validation* [14].

Another common technique against over-fitting is pruning. During pruning some parts of the tree are substituted by a leaf. One of the pruning techniques uses the *minimum description length principle* [12]. Another approach retards branch evolution if a predefined criterion is met. This could be the absolute number of objects in a node or a percentage of the objects in the original data set [13].

Tree performance cannot be evaluated by a single numeric measure like mean squared error. The number of correct and false classifications are used instead, and they are expressed with measures like the true positive (TP) or the false positive (FP) rate. Two parameters that are commonly used are the recall and the precision. Recall is actually the same as TP. Precision is the number of objects correctly classified over the total number of objects that were classified in the same class, *precision*=TP/(TP+FP). The *F-measure* statistic is equally often used:

$$F - measure = \frac{2 \times recall \times precision}{recall + precision} = \frac{2\mathrm{TP}}{2\mathrm{TP}+\mathrm{FP}+\mathrm{FN}}. \tag{3}$$

Tree structures can be easily transformed into decision rules of the IF…THEN… form. Another feature that makes them popular in pattern recognition is that they can cope with numeric or non-numeric variables equally well.

3 User Modeling and Feature Selection

In telecommunications systems user transactions and hence user behavior is contained in the Call Detail Record (CDR) of any Private Branch Exchange (PBX). The CDR contains data such as: the caller ID, the chargeable duration of the call, the called party ID, the date and the time of the call, etc, [15].

User behavior profiles are constructed from the raw data contained in the CDRs. In [9], five different user profiles were proposed and studied. Tests were performed to evaluate the fraud detection ability of each different profile structure. It was concluded that accumulated characteristics of a user give better discrimination results. The most promising user profile is shown in Fig. 1. It consists of seven fields which are the mean and the standard deviation of the number of calls per week (calls), the mean and the standard deviation of the duration (dur) of calls per week, the maximum number of calls, the maximum duration of one call and the maximum cost of one call. All maxima are computed within a week's period.

The daily data used for the construction of the aforementioned profile are the number of calls per day (Calls), the duration of these calls (Dur), the corresponding charging units (Units), the maximum duration of one call (MaxDur), and the maximum units for one call in this day (MaxUnits). These features were also combined in one profile, (Fig. 2), to test the daily characteristics of fraud.

Our experiments are based on real data extracted from a database that holds the CDR records from an organization's PBX. The data span in a period of eight years. Several defrauded user accounts have been identified. All contain both examples of legitimate and fraudulent activity. The detailed accounts were examined by a field expert and each phone call was marked as either normal or defrauded. If during a day no fraudulent activity was present then the whole day was marked as normal. If at least one fraudulent call was present then the whole day was marked as fraud.

22,000 phone calls were examined in total. The calls correspond to 5,541 days (2,702 days of legal behavior and 2,839 of fraudulent one). The weekly aggregation of these calls for each user is represented by means of the user profile (Fig. 1). This yielded 2,014 vectors (weeks) with 7 variables each. Typical or legal use constitutes of 51.1% of the whole data set (class 1) while fraud is the remaining 48.9% (class 2). It is pointed out that the rate of weeks and days is not 1:7, because the weekly aggregation was performed per user and days with zero activity were omitted.

C4.5, [11], was used to classify calls into two classes. The task was to identify the values of the appropriate variables that better separate the classes. Stratified tenfold cross-validation was applied. Data were normalized in [0, 1] but in the figures all variables are presented with their original values.

Theoretically, the number of variables that are used in the description of a problem can improve the discriminating power of machine learning algorithms. However, some variables may be correlated or irrelevant to the problem. The use of such variables "confuses" machine learning algorithms [16]. So, it is common to precede

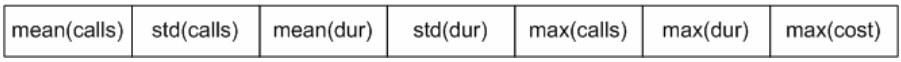


Fig. 1. The basic vector for the weekly user behavior representation

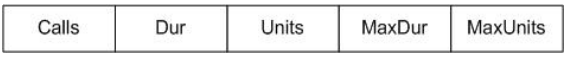


Fig. 2. The basic vector for the daily user behavior representation

learning with an *attribute* or *feature selection* process that tries to eliminate all but the most relevant attributes.

The best feature selection process is done manually and is based on specialized knowledge of the problem. The selection of the user profile, (Fig. 1), is actually a realization of such a process. Several automated methods exist to assist in the feature selection process. The reduction of the problem's dimensionality yields a more easily interpretable representation of the problem.

There are two main approaches in feature selection, namely the filter methods and the wrapper methods [17]. Most methods search the attribute space in order to find the subset that best describes the problem. The task is to select the attributes that are strongly correlated with the problem classes and less correlated with each other. Feature selection works either with forward selection or with backward elimination [18].

In the present paper four different feature selection methods were applied. These are the Best First, the Exhaustive Search, the Greedy Backward Elimination, and the Ranker method [10]. The last one was applied with both the Information Gain and the Gain Ratio as relevance measures. All methods gave similar results which are presented in Table 1. Attributes 1 – 4 are characterized as highly relevant by most of the methods, attribute 6 and 7 are characterized as irrelevant, while the characterization of attribute 5 depends on the method. The procedure was also applied to the daily representation of user behavior. The corresponding feature selection is shown in Table 2.

4 Experimental Results

For the weekly profiles of the users four different trees were built. In order to have a basis for comparisons, the first tree, Tree1, was built with no pruning or feature selection. A minimum of 2 objects in the leaves was set as prerequisite. For Tree2, only the number of objects in the leaves was changed to be 20. Feature selection and pruning (*subtree raising*) were applied for the third tree, Tree3. Last, reduced error pruning was additionally applied and the outcome is Tree4 (Fig. 3). Due to space limitations only the last tree is presented here. In each leaf two numbers appear that are separated by a slash (/). The first is the total number of objects in the leaf, and the second is the number of misclassified objects in the leaf. The number of the leaves for each tree and the corresponding size (no of branches) are presented in Table 3.

In the same table (Table 3) five accuracy statistics for the aforementioned trees are given, namely the TP rate, the FP rate, the Recall, the Precision, and the F-measure. The confusion matrices for the four trees are also presented, (Table 4). The percentage of objects that are correctly classified is more than 85% for all the trees. Tree

performance deteriorates slightly as we move to smaller trees after the application of pruning and feature selection (Table 3). However, there is a considerable improvement to the descriptive power of the smaller trees which are more easily interpretable. The cost of this improvement varies from 2% to 10% reduction in the accuracy. Compare, for example, the statistics and the sizes of Tree1 and Tree4.

Table 1. Feature selection for the weekly representation of the users

Attribute		Feature Selection Method				
		Best First	Exhaustive Search	Greedy Backward Elimination	Ranker (Information Gain)	Ranker (Gain Ratio)
		Number of folds (%)			Average rank (rank)	Average rank (rank)
1	MeanCalls	10(100%)	10(100%)	10(100%)	1,6 ± 0,49 (1)	1,7 ± 0,78 (2)
2	StdCalls	8(80%)	9(90%)	8(80%)	3,8 ± 0,4 (4)	2,9 ± 1,04 (3)
3	MeanDur	10(100%)	10(100%)	10(100%)	5 ± 0 (5)	3,9 ± 0,54 (4)
4	StdDur	10(100%)	10(100%)	10(100%)	1,9 ± 1,22 (2)	1,7 ± 0,64 (1)
5	MaxCalls	2(20%)	0(0%)	2(20%)	2,7 ± 0,46 (3)	4,8 ± 0,4 (5)
6	MaxDur	0(0%)	0(0%)	0(0%)	7 ± 0 (7)	6,2 ± 0,4 (6)
7	MaxUnits	0(0%)	0(0%)	0(0%)	6 ± 0 (6)	6,8 ± 0,4 (7)

Table 2. Feature selection for the daily representation of the users

Attribute		Feature Selection Method				
		Best First	Exhaustive Search	Greedy Backward Elimination	Ranker (Information Gain)	Ranker (Gain Ratio)
		Number of folds (%)			Average rank (rank)	Average rank (rank)
1	Calls	10(100%)	10(100%)	10(100%)	1 ± 0 (1)	1,1 ± 0,3 (1)
2	Dur	10(100%)	10(100%)	10(100%)	2,6 ± 0,49 (3)	2,6 ± 0,49 (3)
3	Units	10(100%)	10(100%)	10(100%)	2,4 ± 0,49 (2)	2,3 ± 0,64 (2)
4	MaxDur	0(0%)	0(0%)	0(0%)	5 ± 0 (5)	4,8 ± 0,4 (5)
5	MaxUnits	0(0%)	0(0%)	0(0%)	4 ± 0 (4)	4,2 ± 0,4 (4)

The observation of Tree4 (Fig. 3) gives clues for the construction of rules for the distinction between normal and fraudulent use. Its right branch (MeanCalls>0,857) covers 77% of the fraudulent cases. An interesting result is that the standard deviation of calls' duration is important in fraud identification. In particular, fraud cases are characterized by lower standard deviation values than normal use, which implies that fraudsters show some kind of "compact" behavior. An expected aspect of fraudsters' behavior is apparent in the lowest leaves of Fig. 3. They tend to place long calls. The most powerful rules that result are:

IF MeanCalls<0.86 THEN class=1 (confidence: 71.98%, coverage: 70.48%)

IF MeanCalls>0.86 AND StdDur<129.5 THEN class=2 (conf.: 97.5%, cov.: 41.5%).

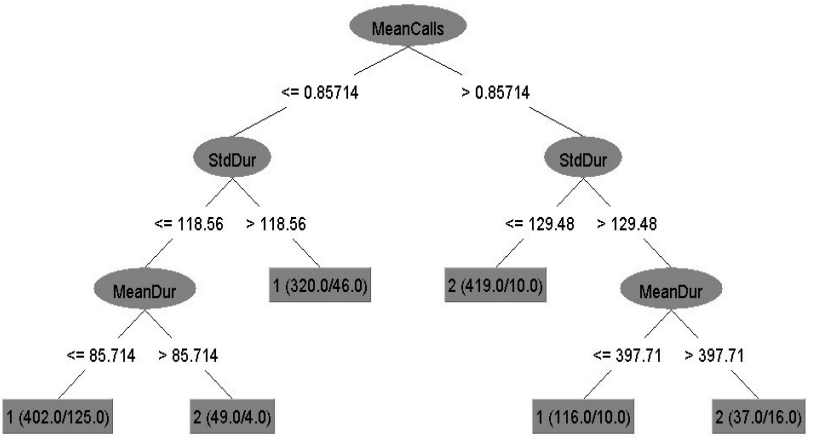


Fig. 3. Decision tree for the weekly representation of the users

Table 3. Statistics for the trees that correspond to the weekly representation of the users

Statistics for Tree1 (no. of leaves: 63, size of tree: 125)					
TP Rate	FP Rate	Precision	Recall	F-Measure	class
0.937	0.153	0.865	0.937	0.899	1
0.847	0.063	0.928	0.847	0.885	2
Statistics for Tree2, (no. of leaves: 20, size of tree: 39)					
TP Rate	FP Rate	Precision	Recall	F-Measure	class
0.918	0.168	0.851	0.918	0.884	1
0.832	0.082	0.907	0.832	0.868	2
Statistics for Tree3, (no. of leaves: 13, size of tree: 25)					
TP Rate	FP Rate	Precision	Recall	F-Measure	class
0.946	0.202	0.83	0.946	0.884	1
0.798	0.054	0.933	0.798	0.86	2
Statistics for Tree4, (no. of leaves: 6, size of tree: 11)					
TP Rate	FP Rate	Precision	Recall	F-Measure	class
0.927	0.199	0.83	0.927	0.876	1
0.801	0.073	0.913	0.801	0.853	2

The first rule says that if a user places less than 1 call per day his is a legitimate user with confidence 72%. According to the second rule, if the mean number of calls in a week is more that 1 (that is at least 7 costly calls in the week) and the standard deviation of their duration is less than 2 minutes, then the user is a fraudster with confidence 97.5%.

In the same way the daily representation of user behavior was studied. Feature selection is presented in Table 2. All five feature selection methods agree over the most relevant features. Three trees were built. For the first one, Tree5, no pruning method

or feature selection was applied. The application of feature selection yielded the second tree, Tree6, while for the third one, Tree7, both pruning and feature selection were applied. In Tree7, (Fig. 4.), two are the most prominent rules:

IF calls ≥ 3 AND Dur > 302sec THEN class = 2 (confidence: 72%, support: 52%).

IF calls = 1 AND Dur < 84sec THEN class = 1 (confidence: 68%, support: 46.7%). The first rule means that if the user places more than three chargeable calls within a day with duration more than 5 minutes, then he is a fraudster with confidence 72%. The second rule implies that if the user places at most one chargeable call in a day with duration less than 84 sec then he is the legitimate user with confidence 68%.

Table 4. Confusion matrices for the weekly user representation

Confusion matrix for Tree1			**Confusion matrix for Tree2**		
1	2	Classified as ←	1	2	Classified as ←
965	65	1	946	84	1
151	833	2	165	819	2
Confusion matrix for Tree3			**Confusion matrix for Tree4**		
1	2	Class. as ←	1	2	Class. as ←
974	56	1	955	75	1
199	785	2	196	788	2

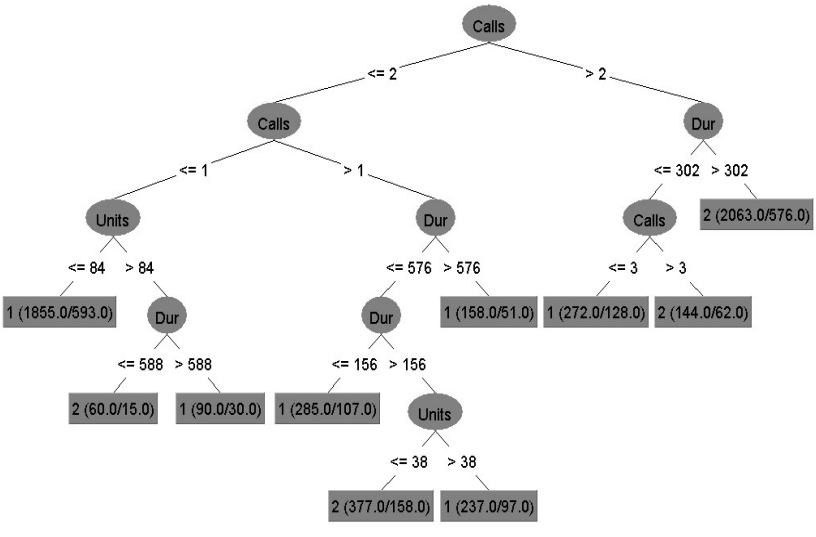


Fig. 4. Decision tree for the daily user representation

5 Conclusions

In the present paper, decision trees were applied in order to identify thresholds between normal and fraudulent user behavior in the telecommunications system of a

large organization. The data used came from real defrauded user accounts. Feature selection methods were applied before tree building.

Comparing the confusion matrices, Table 4 and Table 6, it is concluded that with the daily representation of the user behavior only the 65% of the cases is classified correctly. This is in contrast with the weekly user representation where 85% of the cases were classified correctly. The result coincides with the findings of [9] which showed that the aggregated user behavior gives better results in fraud identification. Results also coincide with those in [6] which also reports satisfactory fraud detection levels from the use of a rule discovering system such as the C4.5 algorithm.

The rules will be used in a rule based expert system which shall operate in the network from which the training data came. The specified thresholds depend on the training data so as more fraud cases come into light the thresholds may be adjusted. Exceeding the thresholds can be used as clue but no proof of fraud. More criteria can be applied in conjunction with the former. An important feature of the applied user profiles is the fact that they hide the user's personal data. This protects user's privacy.

The presented methodology may, also, be applied to similar problems in a mobile or a data network. The first step should be the appropriate user modeling, i.e. the selection of the appropriate attributes for profile building. Rules often depend on the specific nature of each problem and there is a risk when directly generalizing the findings even to similar environments.

Table 5. Statistics for the trees that correspond to the daily representation of the users

Statistics for Tree5, (no. of leaves: 17, size of tree: 33)					
TP Rate	FP Rate	Precision	Recall	F-Measure	class
0.688	0.377	0.635	0.688	0.66	1
0.623	0.312	0.677	0.623	0.649	2
Statistics for Tree6, (no. of leaves: 13, size of tree: 25)					
TP Rate	FP Rate	Precision	Recall	F-Measure	class
0.677	0.358	0.643	0.677	0.659	1
0.642	0.323	0.676	0.642	0.658	2
Statistics for Tree7, (no. of leaves: 10, size of tree: 19)					
TP Rate	FP Rate	Precision	Recall	F-Measure	class
0.677	0.36	0.642	0.677	0.659	1
0.64	0.323	0.676	0.64	0.658	2

Table 6. Confusion matrices for the daily user representation

Conf. matrix for Tree5			**Conf. matrix for Tree6**			**Conf. matrix for Tree7**		
1	2	Class. as	1	2	Class. as	1	2	Class. as
1860	842	1	1828	874	1	1830	872	1
1071	1768	2	1017	1822	2	1021	1818	2

Acknowledgements. The authors would like to thank the staff of the Telecommunications Center of the Aristotle University for their contribution of data. This work was supported in part by the Research Committee of the Technological Educational Institute of Serres, Greece.

References

1. Gosset, P., Hyland, M.: Classification, detection and prosecution of fraud in mobile networks. In: Proc. of ACTS Mobile Summit, Sorrento, Italy (1999)
2. Hoath, P.: Telecom fraud, gory details. Computer Fraud & Security, 10–14 (1998)
3. Moreau, Y., Vandewalle, J.: Detection of mobile phone fraud using supervised neural networks: A first prototype. In: Gerstner, W., Hasler, M., Germond, A., Nicoud, J.-D. (eds.) ICANN 1997. LNCS, vol. 1327, pp. 1065–1070. Springer, Heidelberg (1997)
4. Taniguchi, M., Haft, M., Hollmen, J., Tresp, V.: Fraud detection in communication networks using neural and probabilistic methods. In: ICASSP'98. Proc. of the 1998 IEEE Int. Conf. in Acoustics, Speech and Signal Processing, vol. II, pp. 1241–1244. IEEE Computer Society Press, Los Alamitos (1998)
5. Fawcett, T., Provost, F.: Adaptive fraud detection. Journal of Data Mining and Knowledge Discovery 1(3), 291–316 (1997)
6. Rosset, S., Murad, U., Neumann, E., Idan, Y., Pinkas, G.: Discovery of fraud rules in telecommunications – Challenges and solutions. In: Chaudhuri, S., Madigan, D. (eds.) Proc. 5th ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining, San Diego, CA, USA, pp. 409–413. ACM Press, New York (1999)
7. Alves, R., et al.: Discovering telecom fraud situations through mining anomalous behavior patterns. In: KDD 2006 Workshop on Data Mining for Business Applications, Philadelphia, USA (2006)
8. Dong, W., et al.: A feature extraction method for fraud detection in mobile communication networks. In: Proc. 5th World Cong. On Intelligent Control and Automation, Hangzhou, China, pp. 1853–1856 (2004)
9. Hilas, C., Sahalos, J.: Testing the fraud detection ability of different user profiles by means of FF-NN classifiers. In: Kollias, S., Stafylopatis, A., Duch, W., Oja, E. (eds.) ICANN 2006. LNCS, vol. 4132, pp. 872–883. Springer, Heidelberg (2006)
10. Witten, I.H., Frank, E.: Data mining: practical machine learning tools and techniques with Java implementations. Academic Press, London (2000)
11. Quinlan, J.R.: C4.5: Programs for machine learning. Morgan Kaufmann, San Fransisco, CA (1993)
12. Rissanen, J.: Modeling by shortest data description. Automatica 14(5), 465–471 (1978)
13. Duda, R.O., Hart, P.E., Stork, D.G.: Pattern Classification, 2nd edn. John Wiley & Sons, Inc., New York (2000)
14. Kohavi, R.: A study of cross-validation and bootstrap for accuracy estimation and model selection. In: Proc. 14th Int. Joint Conf. on Artificial Intelligence, Montreal, Canada, pp. 1137–1143 (1995)
15. Hinde, S. F.: Call Record Analysis. Making Life Easier - Network Design and Management Tools (Digest No: 1996/217), IEE Colloquium on, 8/1 – 8/4 (1996)
16. John, G.H.: Enhancements to the data mining process. PhD Dissertation, Computer Science Department, Stanford University (1997)
17. John, G.H., Kohavi, R., Pfleger, P.: Irrelevant features and the subset selection problem. In: Proc. 11th Int. Conf. on Machine Learning, New Brunswick, NJ, pp. 121–129 (1994)
18. Kira, K., Rendell, L.: A practical approach to feature selection. In: Proc. 9th Int. Workshop on Machine Learning, Aberdeen Scotland, pp. 249–258 (1992)

Assisting Dialogical Agents Modeled from Novice User's Perceptions

David Leray and Jean-Paul Sansonnet

LIMSI-CNRS BP 133
F-91403 Orsay Cedex France
{jps,leray}@limsi.fr

Abstract. This paper is based on the idea of integrating the actual cognitive features of novice users at the core of the design process of the software applications in order to significantly increase the quality of the Function of Assistance. We present the first methodological elements of the construction of such symbolic models that are user-driven assisted applications (as opposed to current code-driven assisting approaches). We present and discuss the required modeling principles for an effective and relevant assistance which are finally illustrated, as an example, on the issue of the Perceptive Aggregates.

Keywords: Contextual assistance-oriented models, Rational agent, Novice users, Perceptive aggregates.

1 Introduction

1.1 New Assisting Tools

With the development of the new technologies and the Internet at home there is an exponential growth of the class of novice users, that is ordinary people with lack of skill in computer science, hence struggling with their computer. Ordinary people are now the main target of the economical market which produces new applications very rapidly. Consequently, there is a need for new assisting tools with two ideal features:

1) The assisting content must be 'novice user'-oriented: instead of computer-oriented technical documentation which is incomprehensible for ordinary people, it must be expressed in their 'cognitive world'. This viewpoint is at the origin of most user-centered approaches and has prompted a very active research field in Contextual Help Systems (CHS) [1-3].

2) The access to the assisting content must be as natural as possible: it is a known fact in Computer Human Interaction (CHI) that when things go well, users (especially expert users) prefer to interact with the application through Graphical User Interface (GUI) actions. However, one notices that when things go wrong, users (especially novice users – but also expert ones) express their distress mainly in natural language, together with other expressive gestures, emotions etc. [4]

B. Apolloni et al. (Eds.): KES 2007/ WIRN 2007, Part II, LNAI 4693, pp. 1122–1129, 2007.
© Springer-Verlag Berlin Heidelberg 2007

This is the reason why we definitively think that there is a need for a single entity that embodies the assisting function in a natural way: an Embodied Conversational Agent (ECA) [5] will be the focal point (single input) to the urge of assistance (embodiment of assistance) when users put questions in natural language (natural interaction). In a word, the agent acts as the mediator between the user and the internal help system [6]. In support to this claim, embodied agents have been used in assistance, tutoring and learning applications, and several studies have comforted the results of Lester et al on the persona effect [7], showing better performance, better agreeability etc. when virtual characters are used.

To summarize, one can see the integration of a Contextual Help System together with an Embodied Conversational Agent into a single framework as an Assisting Conversational Agent: CHS + ECA = ACA.

1.2 The Need for Novice User Perception Handling in CHS

In this context, we developed a first ACA framework and we made experiments with novice users in front of several simple assisted applications in an attempt to characterize novices' requests. The experiments with novices resulted in the constitution of a corpus of ~11000 different requests acquired over 20 months. It will be called the "Daft corpus" in the following. The first qualitative analysis of the Daft corpus showed that novices tend to express their problems with two strong characteristics:

- First, they refer to the application's entities via their perceptual features (static like shape and color… or dynamic like movements and content changes issuing from application or user actions) rather than their actual internal features;
- Second, they express the problem in declarative (they state goals) rather than in procedural form.

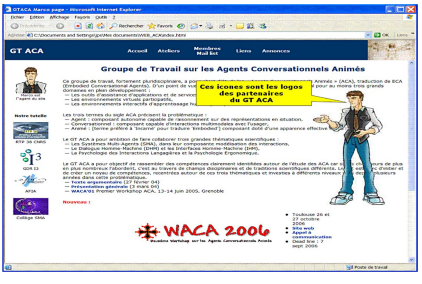
These two characteristics have the same cause: novices have no idea of the internal structure and the internal functioning of the applications. They categorize/reify entities (objects and behaviors) on the basis of their own perceptions. This can result in a dramatic 'cognitive drift' between their model and the programmer's model, thus they begin to struggle with the application. Therefore when they enter a request in the chatbox the help system is in the worst interactional situation: the request is most likely to be not coherent with the internal model because if it was, the request would not have been put in the first place!

This phenomenon opens an important scientific issue for the domain of assistance that will not be easily resolved. Therefore, in this paper we will focus on the first point that is the question of the perceptual features and we will restrict ourselves to the simplest static case. Moreover we will deal with entities representing objects as they appear in the GUI of the application i.e. the widgets. In section 2, we present the Daft corpus while focusing on the form of the referential expressions because they are mainly related to the objects. In Section 3 we present a model dedicated to the representation of the perceptual features of the users in order to be exploited by the reasoning module and in section 4 the implementation of an example illustrates the proposed approach.

2 Building and Analysis of the Daft Corpus

In order to collect the request corpus several small applications were developed in the Daft framework [8] in two main situations: 1) Java standalone applets embedded in the assisting frame of the agent, both in modal and modeless modes 2) Web sites, one with static content and the other with active content (one can edit the pages via requests to the agent). An example is given in Table 1 below.

Table 1. GTACA is the site of the French working group on ECA. The agent gives the users information about the working group. Their requests are logged on a server to build a corpus which in turn has been proposed online on the GTACA web site. On the web pages, the textual answers are produced with balloons. See http://www.limsi.fr/aca/



Like says previously, the experiments with novice subjects resulted in the constitution of a corpus of ~11000 different requests. It provides a first coverage in French of the linguistic sublanguage of dialogical assistance (for an introduction, from a computational point of view, on linguistic domains Cf. [9]). Requests can be viewed at http://www.limsi.fr/~jps/research/daft/corpus/rawcorpus.1000.txt

In this paper we focus on the referential expressions in the corpus that are mainly supported by the NP groups (Nouns Phrases). An analysis upon 5725 different requests exhibited 6680 NPs (including the named entities – people or institutions names, user or agent name etc.) that is, a ratio of 1.2 references by sentence. See http://www.limsi.fr/~jps/demos/grasp/GRASPWEB/phraselist.htm for NP analysis.

The qualitative analysis of the extracted NPs showed the main role of the Extensional Associative Reference (EAR) phenomenon: the subjects perceive atomic objects or aggregates of atomic objects (whether these entities have an internal counterpart or not) and the subjects refer to them by using their perceptual features (atomic: color, shape, label, … or aggregative: list, table, stack, map …) but also their perceived function (member, homepage, project …). Issues related to functions will not be discussed in this paper. Handling of EAR expressions extracted from the requests consists mainly in browsing the active symbolic model of the application in order to exhibit the possible internal candidates associated with them. For this approach to succeed, the model needs to incorporate special features oriented towards the users' perceptual expressions as they appear in the corpus. In the following section, we propose a symbolic model for the support of the function of assistance where perceptual features are first class citizens.

3 The Daft Model

The user assistance has already been studied several times in order to categorize the kind of requests a user is able to perform. Several taxonomies has been proposed, see for instance [10] and [11]: based on those studies one can extract three main classes of requests directly related to the application used: a) the 'what' requests, a request about what is seen on screen and not understood; b) the 'how' requests, a request about the way to achieve the goal the user has in mind; c) the 'why' requests, a request about a state or an unexpected behavior of the application.

The idea of representing the application by more abstract representation is not new: in the Human Computer Interaction literature, one may notice that the Model-View-Controller paradigm of conception shares common points with the previous class of requests: the 'what' requests would be associated with the View (classically the Graphical User Interface), as the 'how' requests would be assimilated to the Controller part and the Model part (representing the state of the application as well as the data-structures used) could possibly be seen as a kind of a context related to the 'why' requests. When looking at the UML (see [12] and [13]), even more direct links can be made with the three classes of requests. For instance the sequence diagrams and use-cases can be seen as an answer to the 'how' and the class/object diagram may be useful to answer the 'what'. On the other hand nothing is really set-up in the usual conceptual modeling languages to deal with the 'why' requests.

Based on these considerations, we propose that a model for the assistance is relevant and could be useful as soon as it stressed aspects of the application linked with the three classes of requests, that is to say defines: a) a representation of the graphical context; b) a task-model; c) a dynamic representation of the context. We already proposed a simple way to build a model in previous work. However, in order to deal with the issues exposed previously, the model has to strongly correlate the inner formal components with perceptual features so that the assisting agent would be able to map the references contained in the request with their formal counterpart. We will now briefly expose the formal part of the model and stress the perceptual features required to deal with the 'what' requests.

3.1 Perceptive Issues

We assume that an ordinary user has no knowledge of the underlying mechanisms and structures of the software or service (s)he is using. Thus the only way to build his mental representation of the application is by its own perception and interaction with it. That's why we choose to focus the way of building the model through the representation of cognitive issues. Briefly, let say that the cognitive drift is the difference between the mental representation built by the user and the formal representation expressed in the source-code of the application. This phenomenon is not described in formal ways, but the Ozkan experiment exemplifies the issue: a user is supposed to command a drawer application. (S)He first asks the system to draw a brown rectangle, then a green circle, then to put the green circle on the top of the brown rectangle, then to move the tree on the left-side. The reference to the "tree" is clearly an artifact of the mental representation of the task the user is doing and has no formal counterpart in the formal representation of the world led by the system. Some similar effects could

appear in assisting tasks as the view of the user may significantly differ from the view of the system on the software [14]. To deal with this issue, we claim that, if the assisting agent relies on a model expressing the inner mechanism by their perceptual features, it would be easy for it to understand the requests of the user as it will mainly contains perceptive cues (at least for the 'what' requests). To support this claim, we will study how a model only built with an aggregating algorithm can better expresses the graphical organization of a web page than a model only built from its inner structure.

3.2 Building the Perceptual Layer of the Daft Model

The building process of the perceptive model is based on the perceptual features that one can extracts for a given application. To simplify, we restrained ourselves to the assistance of html-based application. For an application, there are two ways for building a model. Either the software already exists so we just have to extract the desired information from the source-code, or the software is about to be built so we can use the conception step to build the assisting model as the designer build the software. In order to build the layer of the model that integrates perceptual features, we use an algorithm based on the Gestalt theory from psychology ([15] and [16]). In HTML, all the tags have an equivalent DOM object, which all the properties can be accessed through JavaScript instructions. In particular they have a set of sub-properties attached to the 'style' property (the same as the CSS property).

The perceptive algorithm tries to assign each widget to the most probable group. That is to say, it tries to aggregates widgets that may share a strong semantic relationship because of their perceptive features [17]. The perceptive features implied are: proximity, color, similarity, redundancy of elements. Because a webpage mainly consists on a set of hyperlink and texts organized, the algorithm particularly takes in account the fonts properties.

Here is the principle of the algorithm:

- The first step is to fuse very salient elements together based on some heuristics: i.e. if a 'radiobutton' is followed by a text or a picture or whatever, it is likely that the 'radiobutton' element applies to the following elements;
- Then it computes a hierarchical clustering of the elements based on their proximity: for each element involved, the set of its neighbors is computed. Based on the distance of each neighbor is computed a proximity threshold: each neighbor whose distance is lesser than the threshold is discarded;
- The algorithm goes on again, assuming that each group of elements computed at the previous iteration is taken as a whole, until there is just one group left;
- Finally perceptive heuristics are applied at each level of the hierarchy in order to extract the most salient groups. Three operations are available: correction, merging and weighting. A correction is applied when two heuristics give overlapping groups; the correction is given by the stronger heuristic (explicit > color > hierarchical). The weighting is a way to compute similarity between groups and to assign a saliency to the aggregates and possibly to merge groups together: if two groups close to each other share two many similarities, they are merged together.

4 Experiment

In order to evaluate the relevance of our perceptive model, we propose a simple experiment. Given a sample website (a page in eBay), we will compare the model outputted by a naïve algorithm and the one produced by the perceptive model. We will assume that the relevant clusters are clarified by the inner structure of the website so that an assisting agent may assume that widgets nested inside particular HTML tags are most likely to be regrouped. The naïve algorithm proceeds as follows: the structure is explored and each time one of the {DIV, TABLE, TD, and TR} tags is used to nest one or more component, they are likely to be a cluster. The elements found are then displayed on a screen and an operator decides whether or not they correspond to a relevant aggregate of components. In order to build the perceptive model, we need to obtain all the elements involved in the visual scene presented on screen. The process relies once again on the DOM tree representing the web page. For each node, we check whether it is a widget (direct representation on screen, like for instance, the <input> tag) or whether it is a container (, <table>…). In this latter case, the DOM elements are kept only if they implement a perceptive attribute (especially the CSS attributes 'border' or 'background-color'). Then, for all the nodes extracted by this process, a JavaScript function computes their absolute position and size inside the document. After this step, the algorithm (also implemented in JavaScript) is applied. At each step of the algorithm, the computed aggregations are displayed on a screen and an operator annotates the validity of the aggregate.

4.1 Results

For the naïve algorithm, we extracted a total of 165 potential groups. Among these groups only 34% of them where tagged to have a relevance regarding to their semantic value. But when looking further in the results, we saw that some of the relevant groups were in fact wrapping the same set of graphical elements. Thus, when we discard them, the ratio of relevant groups falls to 25,4%. Even worse, some of them were overlapping i.e. a subset of components appeared in more than one group and the groups were not inclusive.

There are two important results to notice: first, the perceptive algorithm gives much better results. However the algorithm is perfectible: some of the aggregates found were judged irrelevant. On the contrary the operator judged that no other aggregate existed but those who were found by the algorithm. The second point is that the structure is not reliable to provide useful information for two mains reasons: 1) even if the algorithm finds relevant groups, it misses important features: for instance, if homogeneous information is split in two different structures, the information is lost; and 2) there is no way to provide a reliable heuristic to know where to cut the DOM tree to aggregate widgets as the way it is implemented really depends too much on the person who have made the design choices.

Table 2. Results of the perceptive algorithm. The first row shows a neutral partitionning using only the perceptual features explicited in the CSS properties. The next five rows shows the aggregations made during the hierachical process. The last row shows the results of the application of the perceptive heuristics. In the experiment, the model didn't evolve after the application of the first heuristic (explicit regrouping).

		#aggregates	%relevant
Explicit grouping only		15	73%
Hierarchical aggregation	Aggregating process N=1	25	64%
	Aggregating process N=2	12	33%
	Aggregating process N=3	6	66,6%
	Aggregating process N=4	2	100%
	Aggregating process N=5	1	100%
Total :		64	55,8%
Final grouping		43	81%

5 Conclusion

We have discussed the idea of integrating actual cognitive features of novice users at the core of the building process of assisting symbolic models in order to significantly increase the quality of the Function of Assistance. We presented and discussed a simple algorithm to implement the perceptive layer of the assisting model as well as an experiment designed to evaluate it. The results and other complementary studies show that the structure of websites is too heterogeneous to be reliable for an assisting agent trying to exhibit interesting visual scenes in a homogeneous way. We are now aiming at evaluate the impact of perceptive models in real assistance tasks.

References

1. Marion, C.: What is interaction design and what does it mean to information designers? http//mysite.verizon.net/reznx4g7/7/ MakeWay/ MakeWay.htm
2. Capobianco, A.: Questioning the effectiveness of contextual online help: some alternative propositions HCI. In: INTERACT '03, pp. 65–72 (2003)
3. Carenin, G., Moore, J.: Generating explanations in context. In: Int. workshop on Intelligent User Interfaces, pp. 175–182 (1993)
4. Amalberti, R., Carbonell, N., et al.: User representations of computer systems in human-computer speech interaction. International Journal of Man-Machine Studies 38 (1993)
5. Cassell, J., Sullivan, J., et al.: Embodied Conversational Agents (2000)
6. Maes, P.: Agents that reduce workload and information overload. Communications of the ACM 37(7) (1994)

7. Lester, J., et al.: The Persona Effect: Affective impact of Animated Pedagogical Agents. In: CHI'97 (1997)
8. Sansonnet, J.-P., Leray, D., Martin, J.: Architecture of a Framework for Generic Assisting Conversational Agents. In: Gratch, J., Young, M., Aylett, R., Ballin, D., Olivier, P. (eds.) IVA 2006. LNCS (LNAI), vol. 4133, Springer, Heidelberg (2006)
9. Kittredge, R.: Sublanguages and controlled languages. Ch. 23 in The Oxford Handbook of Computational Linguistic (2003)
10. Pilkington, R.M.: Question-answering for intelligent on-line help: The process of intelligent responding. Cognitive Science: A Multidisciplinary Journal 16, 455–489 (1992)
11. Roesler, A.W., McLellan, S.G.: What Help Do Users Need?: Taxonomies for On-line Information Needs & Access Methods. In: CHI (1995)
12. OMG - Object Management Group. Introduction to OMG's Unified Modeling Language, http://www.omg.org/
13. Lu, S., Paris, C., et al.: Toward the Automatic Construction of Task Models from Object-Oriented Diagrams. In: Proceedings of the IFIP Tc2/Tc13 Wg2.7/Wg13.4 Seventh Working Conference on Engineering For Human-Computer interaction (1998)
14. Sansonnet, J., Leray, D.: KIWI: An environment for caapturing the Perceptual Cues of an application for an Assisting Conversational Agent. In: AISB'07 (2007)
15. Wertheimer, M.: Untersuchungen zur Lehre von der Gestalt. Psychologische Forschung. A Source Book of Gestalt Psychology, vol. 4 (1938)
16. Thorisson, K.: Simulated Perceptual Grouping: An Application to Human-Computer Interaction. In: Proceedings of the Sixteenth Annual Conference of the Cognitive Science Society, Atlanta, Georgia, August 13-16, pp. 876–881 (1994)
17. Gibson, J.J.: The Theory of Affordances. In: Shaw, R.E., Knowing, J. (eds.) Lawrence Erlbaum Associates, Hillsdale, NJ (1977)

Combining Empirical Studies of Audio-Lingual and Visual-Facial Modalities for Emotion Recognition*

M. Virvou, G.A. Tsihrintzis, E. Alepis, I.-O. Stathopoulou, and K. Kabassi

Department of Informatics
University of Piraeus
Piraeus 185 34
Greece
{mvirvou, geoatsi, talepis, iostath, kkabassi}@unipi.gr

Abstract. In this paper, we present and discuss two empirical studies that we have conducted involving human subjects and human observers concerning the recognition of emotions from audio-lingual and visual-facial modalities. Many researchers agree that these modalities are complementary to each other and that the combination of the two can improve the accuracy in affective user models. However, there is a shortage of research in empirical work concerning the strengths and weaknesses of each modality so that more accurate recognizers can be built. In our research, we have investigated the recognition of emotions from the above mentioned modalities with respect to 6 basic emotional states, namely *happiness, sadness, surprise, anger* and *disgust* as well as the emotionless state which we refer to as *neutral*. We have found that certain states such as neutral happiness and surprise are more clearly recognized from the visual-facial modality whereas sadness and disgust are more clearly recognized from the audio-lingual modality.

Keywords: Empirical studies, affective user modeling, facial expression analysis, multi-modal interfaces, audio-lingual affect recognition, visual-facial affect recognition.

1 Introduction

Recently, the recognition of emotions of users while they interact with software applications has been acknowledged as an important research topic. How people feel may play an important role on their cognitive processes as well [11]. Thus the whole issue of human-computer interaction has to take into account users' feelings. Picard points out that one of the major challenges in affective computers is to try to improve the accuracy of recognising people's emotions [14]. Improving the accuracy on emotion recognition may imply the combination of many modalities in user interfaces. Indeed, human emotions are usually expressed in many ways. For example, as we articulate

* Support for this work was provided by the General Secretariat of Research and Technology, Greece, under the auspices of the PENED-2003 program.

B. Apolloni et al. (Eds.): KES 2007/ WIRN 2007, Part II, LNAI 4693, pp. 1130–1137, 2007.
© Springer-Verlag Berlin Heidelberg 2007

speech we usually move the head and exhibit various facial emotions [12]. Ideally evidence from many modes of interaction should be combined by a computer system so that it can generate as valid hypotheses as possible about users' emotions. This view has been supported by many researchers in the field of human computer interaction [13], [3], [4]. However, progress in emotion recognition based on multiple modalities has been quite slow. Although several approaches have been proposed to recognize human emotions based on facial expressions or speech, relatively limited work has been done to fuse these two and other modalities to improve the accuracy and robustness of the emotion recognition system [2].

In view of the above, it is our aim to combine modalities in order to improve the accuracy of emotion recognition. A system that we have already constructed combining two modalities, namely the keyboard and the voice, is described briefly in [1]. In this paper, we focus on the audio-lingual and the visual-facial modalities. Towards combining the audio-lingual and the visual-facial modalities, we had to determine the extent to which these two different modalities can provide emotion recognition from the perspective of a human observer. Moreover, we had to specify the strengths and weaknesses of each modality. Hence, we conducted empirical studies concerning emotion recognition based on two modalities from the perspective of a human observer: the audio-lingual and the visual-facial.

These empirical studies provide the basis towards the combination of modalities into the affective user modelling component of our bi-modal system and also give evidence for other researchers to use since there are not many results from such empirical studies in the literature. Indeed, after an extensive search of the literature, we found that there is a shortage of empirical evidence concerning the strengths and weaknesses of these modalities. The most relevant research work is that of De Silva et al. [4] who performed an empirical study and reported results on human subjects' ability to recognize emotions. They showed video clips of facial expressions and corresponding, synchronised emotional speech clips to human subjects who were not familiar with the languages used in the video clips (Spanish and Sinhala). Then, they compared human recognition results in three tests: video only, audio only, combined audio and video. However, De Dilva et al. [4] focus on the audio signals of voice concentrating on the pitch and volume of voice rather than lingual keywords that convey affective information. On the other hand, in our research we have included the lingual aspect of users' spoken words on top of the pitch and volume of voice and have compared the audio-lingual results with the visual -facial results so that we can see which modality conveys more information for human observers for six basic emotions, namely happiness, sadness, surprise, anger and disgust as well as the emotionless state which we refer to as neutral.

In this paper, we present each of the empirical studies and show and discuss the comparison results. This paves the way to combination methods that take into account the human observers' performances in the recognition of emotions.

2 Empirical Study for Audio-Lingual Emotion Recognition

Our first empirical study concerns audio-lingual emotion recognition with the microphone as an input device. The empirical study aimed at finding out common user

reactions that express user feelings while they interact with computers and at associating these reactions with particular feelings.

The empirical study involved 100 users (male and female) of varying educational backgrounds, ages and levels of computer experience. The participants were asked to use an application which incorporated a user monitoring component. The basic function of this application was to capture all the data inserted by the user orally. The data was recorded to a database and video clips. This component recorded the actions of users from the microphone. Then the transcripts collected were given to 10 human expert-observers who were asked to perform emotion recognition with regard to the six emotion states, namely happiness, sadness, surprise, anger, disgust and neutral. Human expert-observers analyzed the data corresponding to the audio-lingual input only. Thus, they were asked to listen to the video tape and were also given a print out of what the user had said just in case the user's words were better captured by the computer's audio recorder rather than the video clip. The human observers were asked to justify the recognition of an emotion by indicating the weights of the criteria that s/he had used in terms of words, pitch of voice and volume.

The user monitoring component can be incorporated in any kind of application. However if the application is not sophisticated enough it may not generate sufficient user feelings for the purposes of our research. Therefore, for the purposes of the empirical study, we incorporated it into an educational application. Part of the interaction included knowledge tests, while the users were requested to use the oral interaction via the microphone way. In order to record the users' oral behaviour, the usual educational application incorporated the monitoring module that was running "quietly" in the background and which the user could not notice. Moreover, users were also video taped while they worked.

The audio-lingual empirical study gave important results about the strengths and weaknesses of emotion recognition that is based on the audio-lingual modality. The human experts' recognition rates of the six emotions showed that some emotions are easily recognised by the audio-lingual modality. Such an emotion is sadness. On the other hand, other emotions are more difficult to recognise by the audio-lingual modality. Such an emotion is surprise. That made the application more accurate in recognizing feelings thus the database that had been resulted from the empirical study could be made more accurate. In particular the results from the empirical study could be either confirmed be the recognition of the users' emotions in real-time (by using the user modelling component), or in some cases take specific results into reconsideration. Combining the two phases of the empirical study we come up with a database constructed from the users' oral input and contained words, phrases and exclamations. At the same time changes in voice volume and voice pitch were recorded in relation with the oral input.

Table 1 illustrates the results of the empirical study in terms of the audio-lingual input via microphone and the six basic recognized emotions (neutral, happiness, sadness, surprise, anger, disgust) by human expert-observers. For each emotion we can remark the percentages of the users' oral reaction or the absence of audio input. Furthermore we can consider the changes in the users' voice (volume and pitch) while saying a word or phrase or saying an exclamation in each emotional situation. For example, a user in surprise may have said an exclamation (58 %) rather than having spoken a word (24 %). This action is recognised to a degree of 66 % accompanied by

an increase in the user's volume. In addition, the audio-lingual empirical study supplied us with a significant number of words, phrases and exclamations, which are used in the creation of an "active grammar", with specific words and phrases for emotion recognition.

Table 1. Basic Emotional States through Microphone - Empirical Study Results

	Say an exclamation		Say a certain word or phrase		Say nothing	
Emotions	Change in volume	Change in pitch	Change in volume	Change in pitch	Change in volume	Change in pitch
Neutral	6%		22%		72%	
	45%	18%	37%	12%		
Happiness	31%		45%		24%	
	40%	37%	55%	25%		
Sadness	8%		28%		64%	
	52%	34%	44%	14%		
Surprise	58%		24%		18%	
	66%	23%	60%	21%		
Anger	39%		41%		20%	
	62%	58%	70%	62%		
Disgust	50%		39%		11%	
	64%	43%	58%	38%		

We may remark that the neutral emotion and sadness could not be recognized easily in this modality because users did not express themselves orally when they were in these moods. Users expressing the emotions of surprise and disgust usually said an exclamation (58 % and 50 % respectively) while happy users and users in anger would likely say a word or phrase contained in our 'emotional database' of words and phrases (45 % and 41 % respectively). Particularly in regard to the emotions of surprise and anger, users would increase the volume of their voice while saying something, while in anger and disgust we could remark the greatest changes in the users' voice pitch.

3 Empirical Studies for Visual-Facial Emotion Recognition

The question of how to best characterize perception of facial expressions has clearly become an important concern for many researchers in affective computing. Ironically, this growing interest is coming at a time when the established knowledge on human facial affect is being strongly challenged in the basic research literature. In particular, recent studies have thrown suspicion on a large body of long-accepted data, even on studies previously presented by the same people.

In the past, two main studies regarding facial expression perception have appeared in the literature. The first study is the classic research by psychologist Paul Ekman

and colleagues [5], [6], [7], [8] in the early 1960s, which resulted in the identification of a small number of so-called "basic" emotions: *anger, disgust, fear, happiness, sadness and surprise (contempt was added only recently).* In Ekman's theory, the basic emotions were considered to be the building blocks of more complex feeling states [5], although in newer studies he is sceptical about the possibility of two basic emotions occurring simultaneously [6]. Ekman and Friesen [9], [10] developed the ‘‘facial action coding system (FACS),’’ a method for quantifying facial movement in terms of component muscle actions. Recently automated, the FACS remains the one of the most comprehensive and commonly accepted method for measuring emotion from the visual observation of faces.

Secondly, in the past few years, psychologist James Russell and colleagues [15], [16] have strongly challenged the classic data, largely on methodological grounds. Russell argues that emotion in general (and facial expression of emotion in particular) can be best characterized in terms of a multidimensional affect space, rather than discrete emotion categories. More specifically, Russell claims that two dimensions, namely “pleasure” and “arousal,” are sufficient to characterize facial affect space.

Despite the fact that divergent studies have appeared in the literature, most scientists agree that:

- Human experience emotions in subjective ways.
- The “basic emotions” deal with fundamental life tasks.
- The “basic emotions” mostly occur during interpersonal relationships, but this does not exclude the possibility of their occurring in the absence of others.
- Facial expressions are important in revealing emotions and informing other people about a person’s emotional state. Indeed, studies have shown that people with congenital (Mobius Syndrome) or other (e.g. from a stroke) facial paralysis report great difficulty in maintaining and developing interpersonal relationships.
- Each time an emotion occurs, a signal will not necessarily be present. Emotions may occur without any evident signal, because humans are, to a very large extent, capable of suppressing such signals. Also, a threshold may need to be exceeded to bring about an expressive signal and this threshold may vary across individuals.
- Usually, emotions are influenced by two factors, namely *social learning* and *evolution*. Thus, similarities across different cultures arise in the way emotions are expressed because of past evolution of the human species, but differences also arise which are due to culture and social learning.
- Facial expressions are emotional signals that result into movements of facial skin and connective tissue caused by the contraction of one or more of the forty four bilaterally symmetrical facial muscles. These striated muscles fall into two groups:
  - four of these muscles, innervated by the trigeminal (5th cranial) nerve, are attached to and move skeletal structures (e.g., the jaw) in mastication
  - forty of these muscles, innervated by the facial (7th cranial) nerve, are attached to bone, facial skin, or fascia and do not operate directly by moving skeletal structures but rather arrange facial features in meaningful configurations.

Visual-facial empirical study on subjects: In our visual-facial empirical study, first we created our own database of facial expressions, by photographing individuals aged 19-35 years while they were forming various expressions. To ensure spontaneity, each subject was presented with pictures on a screen behind the camera. These pictures were expected to generate such emotional states that mapped on the subject's face as the desired facial expression. For example, to have a subject assume a "smile" expression, we showed him/her a picture of funny content. We photographed the resulting facial expression and then asked him/her to classify this expression. If the image shown to him/her had resulted in the desired facial expression, the corresponding photographs were saved and labeled; otherwise, the procedure was repeated with other pictures. The final dataset consists of the images of 250 different individuals, each forming the seven expressions: "neutral", "smile", "sad", "surprise", "angry", "disgust" and "bored-sleepy".

Studying this dataset, we identified differences between the "neutral" expression of a model and its deformation into other expressions. We quantified these differences into measurements of the face (such as size ratio, distance ratio, texture, or orientation), so as convert pixel data into a higher-level representation of shape, motion, color, texture and spatial configuration of the face and its components. Specifically, we locate and extract the corner points of specific regions of the face, such as the eyes, the mouth and the brows, and compute their variations in size, orientation or texture between the neutral and some other expression. This constitutes the *feature extraction process* and reduces the dimensionality of the input space significantly, while retaining essential information of high discrimination power and stability.

Visual-facial empirical study by observers: In order to understand how a human classifies someone else's facial expression and set a target error rate for automated systems, we developed a questionnaire in which each we asked 300 participants to classify the facial expressions that appear in 14 images. Each participant could choose from six common facial emotions, namely: "neutral", "anger", "surprise", "happiness", "disgust" and "sadness," or specify any other emotion that he/she thought appropriate.

From the results of the questionaire, we observed that the "surprise'' emotion was the one recognized with the lowest error rate of 10%. Next, follow the "anger'' and "happiness'' emotions, with error rates of 23% and 30%, respectively. The highest error rate corresponded to the "sadness'' and rose to 66%.

Clearly, the facial image classification task is quite challenging. The reasons why humans achieved big classification error rates for specific emotions, which can be considered as "negative" emotions, "sad" and "disgust", may be found in the fact that these expressions seem to differ significantly from person to person or some people may be too shy to form these expressions clearly. The combination of the results in the two empirical studies leads to Table 2.

4 Combination of Results from the Two Empirical Studies

In Figure 1, we illustrate the percentages of successful emotional recognition through audio and visual means. Analyzing Figure 1, we may come up with considerable conclusions. There are cases where both the audio-lingual and the visual-facial

emotion analysis recognize an emotion significantly (anger). In other cases like the neutral emotion and the emotion of surpise the percentages differ in the two modes. When a user is bored (state contained in the neutral emotion), this emotion is more likely to be recognized by the visual mode than the audio mode. Opposite is the situation of disgust, where the audio mode can recognize the emotion well (57 %), while the visual mode provides us with low percentages of success (37%). The two modes are complementary in a high degree while in other cases where both have high or low percentages of successful emotion recognition we still have gaining since we improve the probability of emotion recognition by adding emotion recognition data from the two modalities.

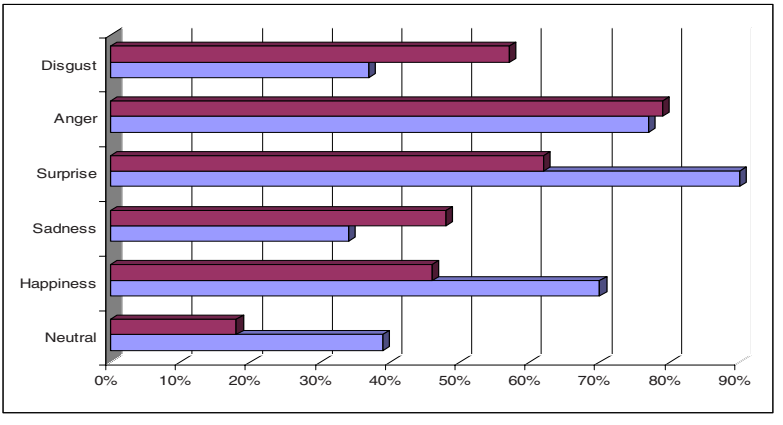


Fig. 1. Recognition of emotions through facial expression and voice

5 Conclusions and Future Work

In this paper we have described and discussed two empirical studies that concern the audio-lingual and the visual-facial recognition of human users' emotions from the perspective of human observers. These two modalities are complementary to each other and, thus, can be used in a bi-modal affective computer system that can perform affect recognition taking into account the strengths and weaknesses of each modality.

From the empirical studies we found that certain emotion states such neutral and surprise are more clearly recognized from the visual-facial modality rather than the audio-lingual. Other emotion states, such as anger and disgust are more clearly recognized from the audio-lingual modality. These results are only partially in accordance with another empirical study that did not take into account the lingual aspect in the audio modality. This is not surprising, since certain words convey emotions and this is evidence that can be added to the pitch and volume of the voice. Moreover, we have constructed a basic affective vocabulary for the Greek language that can be used in the audio-lingual modality.

There is ongoing research of construction of an affective user interface that will use the different modalities as criteria for recognition of emotions and will use the results

of performances for each modality as the basis for the specification of weights. This and other related work is going to be presented in a future publication.

References

1. Alepis, E., Virvou, M., Kabassi, K.: Affective student modeling based on microphone and keyboard user actions. In: ICALT. Proceedings of the Sixth IEEE International Conference on Advanced Learning Technologies, Kerkrade, The Netherlands, pp. 139–141. IEEE Computer Society Press, Los Alamitos (2006)
2. Busso, C., Deng, Z., Yildirim, S., Bulut, M., Lee, C.M., Kazemzadeh, A., Lee, S., Neumann, U., Narayanan, S.: Analysis of emotion recognition using facial expressions, speech and multimodal information. In: Proceedings of the 6th international conference on Multimodal interfaces, State College, PA, USA, pp. 205–211 (2004)
3. Chen, L.S., Huang, T.S., Miyasato, T., Nakatsu, R.: Multimodal human emotion/expression recognition. In: Proc. of Int. Conf. on Automatic Face and Gesture Recognition, IEEE Computer Soc., Nara, Japan (1998)
4. De Silva, L., Miyasato, T., Nakatsu, R.: Facial emotion recognition using multimodal information. In: ICICS'97. Proc. IEEE Int. Conf. on Information, Communications and Signal Processing, pp. 397–401. IEEE Computer Society Press, Los Alamitos (1997)
5. Ekman, P.: Emotion in the human face. Cambridfe University Press, New York (1982)
6. Ekman, P.: Basic Emotions. In: Dalgleish, T., Power, T. (eds.) The Handbook of Cognition and Emotion, pp. 45–60. John Wiley & Sons, Ltd, Sussex, U.K (1999)
7. Ekman, P.: An argument for basic emotions. Cognition and Emotion 6, 169–200 (1992)
8. Ekman, P., Davidson, R.J.: The Nature of Emotion. Fundamental Questions. Oxford University Press Inc., Oxford (1994)
9. Ekman, P., Friesen, W.V.: Facial action coding system: A technique for the measurement of facial movement. Consulting Psychologists Press, Palo Alto, Calif (1978)
10. Ekman, P., Friesen, W.V.: Unmasking the face. A guide to recognizing emotions from facial clues. Prentice-Hall, Englewood Cliffs, New Jersey (1975)
11. Goleman, D.: Emotional Intelligence. Bantam Books, New York (1995)
12. Graf, H.P., Cosatto, E., Strom, V., Huang, F.J.: Visual Prosody: Facial Movements Accompanying Speech. In: Proceedings of the fifth IEEE International Conference on Automatic Face and Gesture Recognition, pp. 381–386. IEEE Computer Society Press, Los Alamitos (2002)
13. Pantic, M., Rothkrantz, L.J.M.: Toward an affect-sensitive multimodal human-computer interaction. Proceedings of the IEEE 91, 1370–1390 (2003)
14. Picard, R.W.: Affective Computing: Challenges. Int. Journal of Human-Computer Studies 59(1-2), 55–64 (2003)
15. Russell, J.A.: Is there universal recognition of emotion from facial expression? Psychological Bulletin 95, 102–141 (1994)
16. Russell, J.A., Fernandez-Dols, J.M.: The Psychology of Facial Expression. Cambridge University Press, New York (1997)

Learning Objects Repository for Training of Power Systems Operators

G. Arroyo-Figueroa, L. Argotte-Ramos, and G. Rodríguez-Ortíz

Instituto de Investigaciones Electricas
Reforma # 113, Col. Palmira, 62490 Cuernavaca, Morelos, Mexico
{garroyo, largotte, gro}@iie.org.mx

Abstract. An important problem to be faced by power systems is the training of operators and the updating of their knowledge and skills. An operator must comprehend the physical operation of the process and must be skilled in handling a number of normal and abnormal operating problems and emergencies. We are developing an intelligent environment for training of power system operators. This paper presents the architecture of the intelligent environment and describes one of its components: the Domain Knowledge Module. The general aim of our work is to provide operators of complex industrial environments with a suitable training from a pedagogical and affective viewpoint to certify operators in knowledge, skills, expertise, abilities and attitudes for operation of power systems.

Keywords: Intelligent tutorial systems, learning objects, power systems.

1 Introduction

Derived from governmental policies related to training management and considering its strategic planning objectives, CFE (Comisión Federal de Electricidad –the National Electric Utility in Mexico) has establish an strategic program to improve its human capital and to align it with its mission and future vision. CFE is responsible for generating, transmitting and distributing electricity through out the Mexican nation.

As part of this program, CFE has generated an ambitious program for learning and training of personnel. The training impact is direct on the production efficiency increase and in the decrease of non-planned production equipment outages and failures. The constant technology changes make it necessary a permanent training of the CFE personnel. In general, the traditional training systems are used: classroom instruction, laboratory practices, workshops, training in specialized simulators, organized seminars, and short courses lectured by specialists, and annual congresses like IEEE, CIGRE, etc.

These training traditional systems have shown effectiveness and consistency for decades. However, CFE has a clear view of its training needs according with its expansion plans and modernization; some actions are mentioned as follows [1]:

B. Apolloni et al. (Eds.): KES 2007/ WIRN 2007, Part II, LNAI 4693, pp. 1138–1145, 2007.
© Springer-Verlag Berlin Heidelberg 2007

- Construction a collection of 150 Job Skills Technical Standards (JSTS).
- To have the distance training technology consolidated, including the technological infrastructure (Internet and Intranet) and the design of courses (self-instruction and traditional) to satisfy the requirements of the JSTS.
- To have an environment with a suitable training from a pedagogical and affective viewpoint to certify personnel in knowledge, skill, expertise, abilities and attitudes for operation of power systems.

The training requirements for utilities ask for powerful interfaces, a more efficient and better adaptive training, by means of incorporating artificial intelligent (AI) techniques, adaptive interfaces, simulation tools, learning objects based on multimedia and virtual reality components.

Intelligent Tutoring Systems [2] are interactive learning environments that have the ability to adapt to a specific student during the teaching process. In general, the adaptation process can be described by three phases: (i) getting the information about the student, (ii) processing the information to initialize and update a student model, and (iii) using the student model to provide the adaptation.

After a summary of the architecture of an intelligent environment (IE) for the training of operators presented in [3], this paper describes in detail one of the components of the IE, the Domain Knowledge Module. In contrast with a traditional training system, the main goal of the intelligent environment is to certify operators in knowledge, skills, expertise, abilities and attitudes for operation of power systems.

2 Architecture of the Intelligent Environment

The architecture of the intelligent environment is based on dynamic course generating systems proposed by Brusilovsky [4]. The intelligent environment is composed of four main components (see Figure 1): the domain knowledge module, the tutor, the operator model, and the learning management system (LMS).

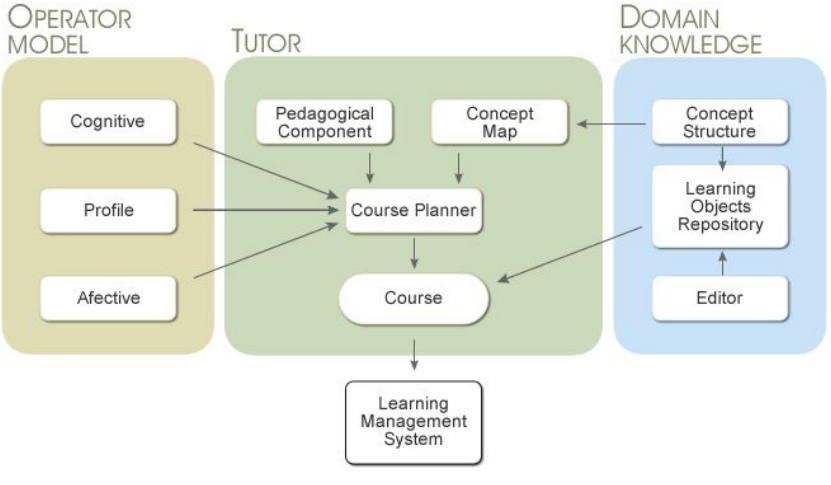


Fig. 1. Architecture of the intelligent environment

The first module contains the knowledge of the domain in form of learning objects and concept structure maps. The tutor module is the component that generates the sequence of learning objects to be presented to operator in form of course. Taking as a basis the concept map, the pedagogical component and the operator model, the training course planner generates a specific course for each operator. The operator model is used to adapt the intelligent environment (IE) to each specific operator. The operator model is divided into three subcomponents: cognitive component, operator profile and affective component [5]. Finally, the LMS controls the interactions with the operator, including the dialogue and the screen layout. The main purpose of an LMS is to present to the operator the learning materials in the most effective way. In this paper a detailed description of the domain knowledge module is presented and the three other modules will be described in a future paper.

3 Domain Knowledge Module

The domain knowledge module has three main components: the concept structure map, the editor and the repository of learning objects. The domain knowledge contains the expert's operation abilities in procedure and malfunction operations and its representation considers both theoretical and practical concepts.

3.1 The Concept Structure Map

The concept structure contains the concept/topic structure of the subject knowledge to be taught (see figure 2). It is possible to organize the domain concepts/topics into a set of smaller, possibly interrelated AND/OR graphs, representing relatively independent sub-areas of the knowledge, different views, or different levels of granularity.It is represented as an AND/OR graph, where nodes represent the concepts domain or elements of knowledge, such as electrical topics, components of control board, rules, procedures and so on; and arcs represent relationships between concepts, such as a prerequisite for learning a concept or a sequence. Every node is associated with a set of teaching and testing materials labeled as Reusable Learning Object (RLO), which instantiate different ways to teach the concept/topic (e.g. introduce, explain, give an example, and give a simulation, exercise, or test).

For the training of power plant operators, the concept structure map is made based on the structural decomposition of the generation process of electricity. The generation process can be divided in unit, structure, systems, subsystems, equipment and component. The concept structure has a higher expressive power because it allows representing not only prerequisites, but also many different types of relationships between concepts; and it enables the use of AI planning techniques for the generation of alternative courses. Therefore, it guarantees a wide variety of different teaching goals and several courses for achieving these goals.

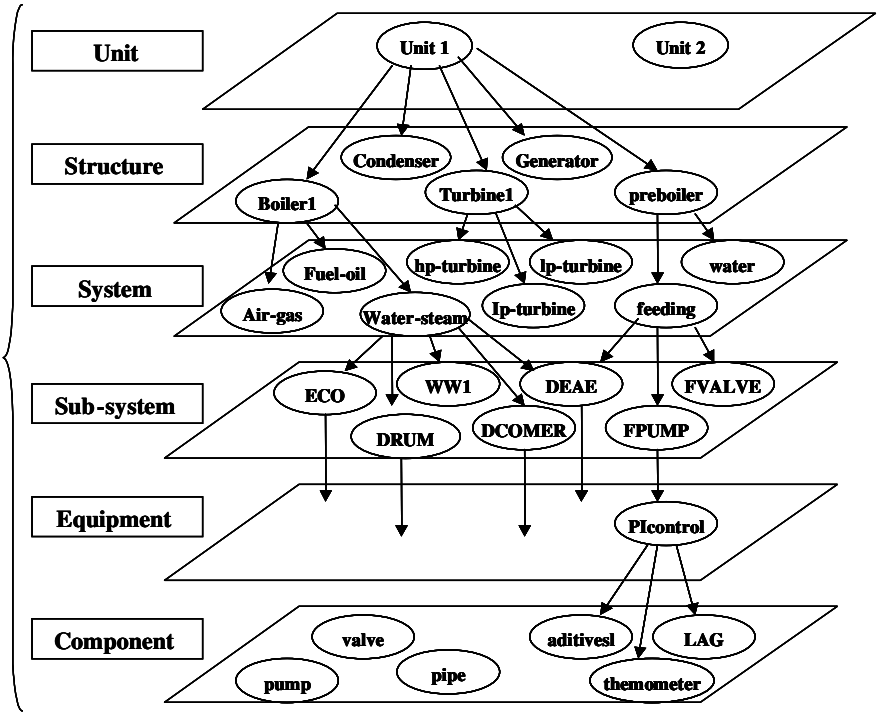


Fig. 2. Concept structure map for training of power plant

3.2 Editor

The editor contains tools for edition of teaching and testing materials based on learning objects. These learning materials consider both theoretical and practical concepts contained in: electronic books, simulation tools, multimedia and virtual reality tools, to present to the operator pedagogical actions such as explanations, exercises, tests, and so on. Each material is labeled as Reusable Learning Object (RLO) or Shared Content Object (SCO) according with the SCORM (Sharable Content Object Reference Model) terminology [6]. Learning objects are self-contained learning components that are stored and accessed independently. RLO are any digital resource that can be reused to support Web-based learning using learning management systems (LMS). The authors or learning developers can create, store, reuse, manage and deliver digital learning content. The same object may be used as many times and for as many purposes as is appropriate.

The learning objects are either dispensed to users individually or used as components to assemble larger learning modules of full courses, depending on individual learning needs. The instructional output may be delivered via the Web, CD-ROM or printed materials. Lately, some work has been done at CFE to create a valve maintenance course and an energized power-line maintenance course using RLO based on virtual reality [7]. Also, and only for the generation process, CFE has a

collection of more than 400 instructional courses developed in house during the last 10 years, these RLO will be integrated to the Domain Knowledge Module.

The editor allows the repository manager to take learning content, generate a SCORM 2004 compliant label o manifest and to compress all together in a zip file (see Figure 3). The result is a SCORM compliant Learning Object (SLO) that can be managed by any SCORM compliant LMS.

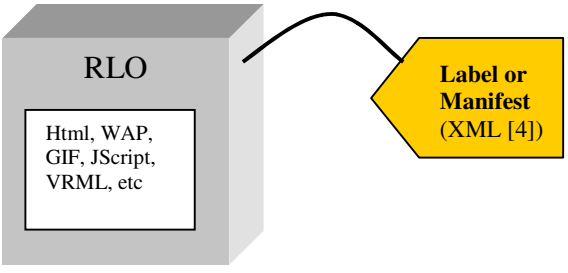


Fig. 3. SCORM complaint learning objects (SLO)

The SCORM compliant manifest contains metadata for the classification and recovery of the SLO. The classification categories have been extended to satisfy the needs of CFE, some were inspired from Figure 2. The nine categories of SCORM meta-data elements are: *General, Life Cycle, Meta-metadata, Technical, Educational, Rights, Relation, Annotation, and Classification.* This extended classification categories are as follows.

Activity branch, it identifies established criteria according to the labor contract.
Specialty, the department, area or worker's function with acquired specialty along their labor life.
Process, generation, transmission, control, distribution, or support processes (management).
Equipment, the equipments are classified as main, auxiliary and miscellaneous according to their importance in the electric generation process.
Career Plan, it is the functional position in the activity branch and specialty in which the worker is currently assigned and that he can ascend in terms of his knowledge, abilities and attitudes or demonstrated competence.
Functional position, it identifies the role of the worker's formation in his current position and immediate higher position according to his career plan.
Course type, it is institutional, foundation, development, or others. A course type is also self training or traditional.
Activity type, it is traditional or distance learning.
Department, it is the workspace according to the conceptual and functional organic structure.
Functional level, Directives, controls means, supervisors and operative.

The SLO is a set of items to structure a course, workshop, or other aggregation of learning resources, the organization of the SLO is as shown in Figure 4.

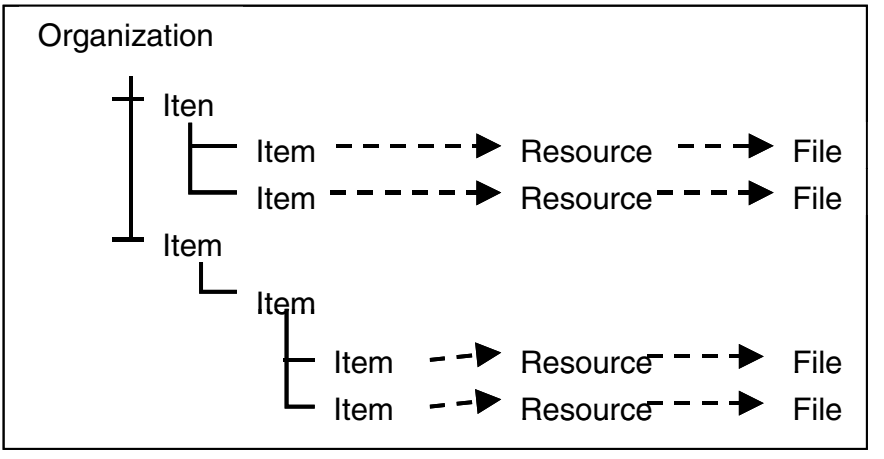


Fig. 4. The organization of a SLO

The editor has an interface with the instructor an a presentation manager that allows the user to search the repository for a SLO that fits his needs, and once the user selects a SLO, the system presents the SLO index and the user can display the learning content objects using hyperlinks to these resources as show in Figure 5.



Fig. 5. An example of actual screen showing a SLO organization

3.3 Learning Object Repository

The Domain Knowledge module has a Learning Object Repository (LOR). The LOR is a central database in which learning content (SLO) is stored and managed (see Figure 6). The Repository main component is the database.

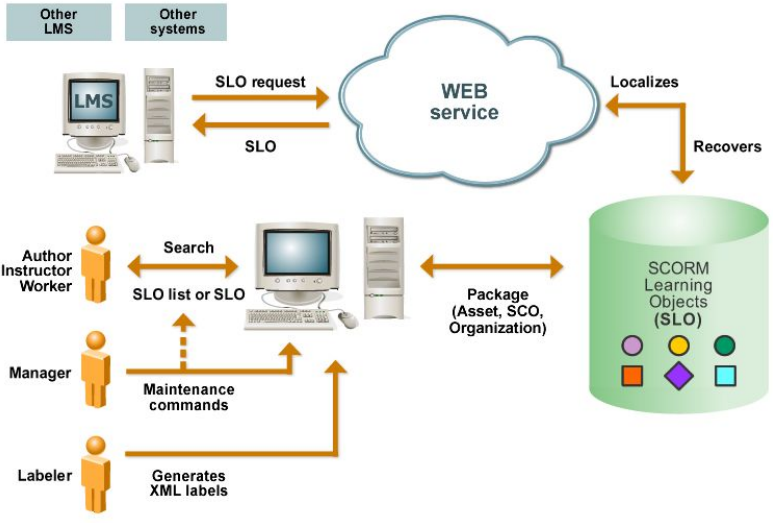


Fig. 6. The environment for the Learning Object Repository

The repository is modeled as a semantic network (see Figure 7) and implemented using a relational database management system. The learning objects are packed and an SCORM compliant LMS can access them through the Web using either an HTML page or a Web service [8]. The organization component of the manifest is a sequencing map that the LMS uses to present to the trainee the different learning items. Each item has objectives to be satisfied and rules to evaluate if a tree of learning items in the organization is completed by the student or not. Also, rules are used to allow the trainee to follow a different order than the linear top-down left-to-right order and skip items if he shows that he already has the appropriate competence to satisfy the objectives of the items to be skipped.

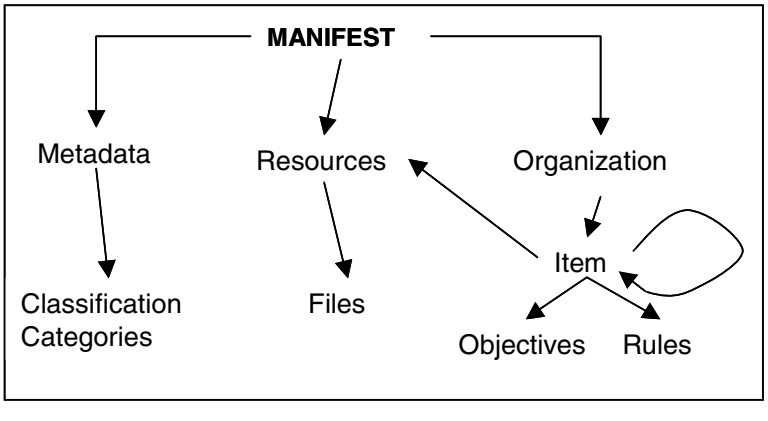


Fig. 7. The semantic network for the database structure

By the end of the year 2006, the database is loaded with 115 power plant courses wrapped as SCORM complaint Learning Objects. The repository can be accessed nation wide through Internet.

4 Conclusions and Future Work

In this paper we have presented the architecture of an Intelligent Environment for training of Power Systems Operators. One of the components was described: the Domain Knowledge Module that contains the Learning Object Repository. The concept structure map of a power plant domain, where the nodes represent concepts of the domain and the links relationships between concepts was used as the basis to build the semantic network that supports the storage and retrieval of SCORM compliant Learning Objects (SLO).

Currently, the repository is loaded with 115 power plant courses wrapped as SLO and it is fully operating at CFE.

The other two modules of the Intelligent Environment: the intelligent tutor module based on course planner, and the operator models based on cognitive and affective components are under development. In contrast with a classical ITS, the IE builds a course based on cognitive, affective and pedagogical aspects. The intelligent tutor module will be a dynamic course generating system interacting with the operator model to produce training material specific to each operator.

References

1. Rodríguez-Ortiz, G., Paredes-Rivera, J., Argotte-Ramos, L., Arroyo-Figueroa, G.: Learning Objects Planning for the Training of the Power Generation Operation and Maintenance Personnel. In: Procc. IEEE Electronics, Robotics and Automotive Mechanics Conference, vol. II, pp. 349–354. IEEE Computer Society Press, Los Alamitos (2006)
2. Peachey, D.R., McCalla, G.I.: Using planning techniques in intelligent tutoring systems. Int. Journal Man-Machines Studies 24, 77–98 (1986)
3. Arroyo-Figueroa, G., Hernandez, Y., Sucar, E.: Intelligent Environment for Training of Power Systems Operators. In: Gabrys, B., Howlett, R.J., Jain, L.C. (eds.) KES 2006. LNCS (LNAI), vol. 4251, pp. 943–950. Springer, Heidelberg (2006)
4. Brusilovsky, P., Vassileva, J.: Course sequencing techniques for large-scale web based education. Int. Journal Cont. Engineering Education and Lifelong Learning 13(1/2), 75–94 (2003)
5. Hernández, Y., Noguez, J., Sucar, E., Arroyo-Figueroa, G.: A probabilistic model of affective behavior for Intelligent Tutoring Systems. In: Gelbukh, A., de Albornoz, Á., Terashima-Marín, H. (eds.) MICAI 2005. LNCS (LNAI), vol. 3789, pp. 1175–1184. Springer, Heidelberg (2005)
6. ADL. Sharable Content Object Reference Model Version 1.2:The SCORM Overview. Advanced Distributed Learning (2001), URL http://www.adlnet.org
7. Perez, M.: Virtual Reality: introduction, prototypes, applications and tendencies (In Spanish). Boletin IIE 28(2) (2004), http://www.iie.org.mx
8. W3C: Extensible Markup Language (XML) 1.0. World Wide Web Consortium 2 ed. W3C Recommendation (2000), http://www.w3c.org/TR/2000/REC-xml-2001006

Mining Meaningful Student Groups Based on Communication History Records

Yang-Sae Moon, Hun-Young Choi, Hea-Suk Kim, and Jinho Kim

Department of Computer Science, Kangwon National University
192-1, Hyoja2-Dong, Chunchon, Kangwon 200-701, Korea
{ysmoon, hychoi, hskim, jhkim}@kangwon.ac.kr

Abstract. In this paper we propose a novel approach that identifies meaningful student groups more objectively. As the data for objective analysis, we use *communication history records* that are collected from various communication tools such as telephones, e-mails, and messengers. We use the simple intuition that communication history records implicitly contain peer relationship information. We first formally define the notion of *degree of familiarity* between students and present mathematical formulas that compute the degree based on the history records. We then adopt a clustering technique to mine meaningful groups. To use the clustering technique, we define the measure of *similarity* between friends based on the degree of familiarity, and perform clustering using the measure. To show the practicality of the proposed method, we have implemented it and interpreted the meaning of experimental results.

1 Introduction

In recent years, delinquent groups[9] in teenagers cause many serious social problems. The number of delinquent groups such as "Iljinhoe," which is a notorious delinquent student group in Korea [7], is increasing due to infelicitous or bad relationships among students [9]. These delinquent groups cause a lot of serious deviations such as using violence and committing blackmail. We think these delinquent groups are caused by infelicitous or bad relationships among students.

In this paper we propose a novel approach that identifies meaning student groups more objectively. The traditional analysis methods studied in sociology can be divided into two categories. The first one is an observer-oriented approach that an advisor consults students about their personal affairs or manages history records of students' life. The second one is a student-oriented approach that students ask for an interview with an advisor and give their information to an advisor. However, these traditional approaches may find out incorrect or distorted student groups since they strongly depend on observers' or students' subjective decisions or opinions. To solve this problem, we suggest two requirements: 1) the data used for identifying the groups should be objective ones that can be obtained from students' daily life, and 2) the analysis method using the objective data should be the more systematic one. To satisfy these requirements,

B. Apolloni et al. (Eds.): KES 2007/ WIRN 2007, Part II, LNAI 4693, pp. 1146–1153, 2007.
© Springer-Verlag Berlin Heidelberg 2007

we use communication history records as the data for objective analysis, and we adopt a clustering technique [3,5] as the method for systematic analysis.

The analysis method to be proposed in the paper is based on *communication history records*. In general, we use various communication tools such as telephones, e-mails, and messengers to communicate with each other. These communication tools usually store communication history records generated in communication processes. Here, we focus on that the communication history records may contain human relationship information implicitly. That is, we think that the more communications try, the closer relationship has. Also, the history records are generated automatically in communication processes, so we can use the records as the objective data to mine the meaningful student groups.

In this paper we propose a systematic approach that identifies the meaningful student groups by using the communication history records. For this, we first define the notion of *degree of familiarity* between students and present mathematical formulas that compute the degree based on communication history records. Here, we use the intuition that the degree of familiarity from student x to student y becomes higher as the number of communications from x to y increases. Thus, we can say that the degree of familiarity from x to y is a quantitative criterion that represents how much x likes y as a friend. We then adopt a clustering technique[2,3,5] to identify meaningful student groups. To use the clustering technique, we define the measure of *similarity* between friends based on the degree of familiarity. Intuitively speaking, the similarity between two students becomes higher as the number of their common friends increases. As the clustering technique, we use the ROCK algorithm[2,3].

We have implemented the proposed method and interpreted the meaning of experimental results. The communication history records used in analysis are collected through a survey. By using the survey data, we first compute the degrees of familiarity for every pair of students. We then mine meaningful groups by using the clustering technique, and interpret the meaning of the results.

2 Related Work

In sociology, researchers have focused on major causes of forming complex social relationships[8] and tried to reveal mutual relationships in the causes. Many investigation methodologies such as conflict theory, symbolic interactionism, and structural-functionalism have been studied for human relationship analysis as one of social relationship analyses [4]. These methods can be classified into two categories:

- *Interaction analysis*: By analyzing interaction patterns happened among members, the method identifies relationships between individuals. Bales's "system of categories used in observation and their major relations" and Comel's class observation method are widely used as the representative ones.
- *Sociometry*: By evaluating degrees of attraction, exclusion, and indifference between individuals, the method finds each individual's position in a community and analyzes relationships between individuals.

However, since the methods in sociology make a conclusion basically through group members' reports or supervisors' observations, they may cause a lot of misjudgment or misunderstanding in analyzing relationships between individuals[4]. First, wrong conclusions would occur due to inaccurate observations or abnormal subjective views since the methods derive conclusions based on personal observations or subjective views. Second, the methods may commit a serious mistake that derives an over-generalized conclusion based on a small number of observation cases. Third, since the methods are based on observers' or supervisors' selective data or experience, they may derive an exaggerated conclusion for whole members rather than a small number of members.

Likewise, the traditional methods in sociology have a serious drawback of being hard to ensure objectivity since they have the inherited problem in data collection processes. That is, the input data itself may lack objectivity since the collection of data depends on observers' subjective decisions, individuals' direct reports, and unrelated persons' opinions. Therefore, we need to collect the data from the more objective information sources that can be obtained from individuals' daily life. To accomplish this purpose, we try to find the degree of familiarity between students from communication history records which are automatically accumulated in students' daily life.

3 Degree of Familiarity

Recent advances in science and computers enable people to use various communication tools to have conversation with each other. Most of these communication tools store communication history records that are automatically generated in communication processes. Here, we note that these communication history records may contain human relationship information implicitly, i.e., we may assume that the more communications between two people, the closer relationship has. Thus, by applying this intuition to students' communication history records, we are able to analyze peer relationships among students. In general, the communication history records are automatically stored by computers. Therefore, we may guarantee objectivity of the stored history records and use the records as the objective data to analyze peer relationships.

Each communication history record contains sender, receiver, and communication date & time[1]. Table 1 shows the examples of communication history records for representative communication tools. As shown in Table 1, most of history records contain the information of a sender (or an originator) and a receiver (or a terminator). In this paper, we use this limited information of senders and receivers to mine meaningful student groups. We summarize in Table 2 the notation to be used throughout the paper.

We now define the degree of familiarity based on the intuition that "the more communications tried from student x to student y, the closer friendship from x to y." According to this intuition, we define the degree of familiarity as follows.

Definition 1. Given communication history records, the *degree of familiarity* from student x to student y is defined as the ratio of "the number of all

Table 1. Examples of communication history records

Communication tools	Configuration of each history record
(cellular) phone	date, outgoing number, incoming number, interval
messenger	date, time, sender, receiver, message
e-mail	sender, receiver, data, time, subject, content

Table 2. Summary of notation

Symbols	Definitions
x, y, z	Students to be analyzed
S	A set of students to be analyzed (i.e., $x \in S, y \in S, z \in S$)
m	The number of communication tools
$f_i(x, y)$	The number of communications from student x to student y using the i-th communication tool f_i $(1 \leq i \leq m)$

communications that student x tried" to "the number of communications that student x tried to student y." That is, we mathematically define the degree of familiarity $dof(x, y)$ from x to y as the following Eq. (1):

$$dof(x, y) = \frac{\sum_{i=1}^{m} f_i(x, y)}{\sum_{\text{for all } z \in S} (\sum_{i=1}^{m} f_i(x, z))} \tag{1}$$

According to Definition 1, the degree of familiarity from x to y becomes higher as the number of communications from x to y increases. In general, a student who formed a close friendship with another student may attempt frequent communications to that student. Thus, we think the degree in Definition 1 is a reasonable and intuitional measure that reflects students' daily life.

Figure 1 shows the algorithm *ComputeDof*() that computes the degrees of familiarity between students. The input to the algorithm is a set of communication history records $f_i(x, y)$ automatically accumulated in communication processes. As shown in Figure 1, we can easily compute each degree of familiarity from x to y for every pair of students x and y by using Eq. (1) in Definition 1.

procedure *ComputeDof* (A set of $f_i(x,y)$'s)
begin
(1) **for each** x **in** S
(2) **for each** y **in** S $(y \neq x)$
(3) $$dof(x,y) = \frac{\sum_{i=1}^{m} f_i(x,y)}{\sum_{\text{for all } z \in S} \left(\sum_{i=1}^{m} f_i(x,z) \right)};$$
end

Fig. 1. The algorithm *ComputeDof*() of computing the degrees of familiarity

4 Mining Meaning Student Groups

We now use the degree of familiarity to mine meaningful student groups from communication history records. However, we cannot directly use the degree of familiarity for identifying the groups since the degree is not adequate to use in existing clustering algorithms. Hence, we introduce the similarity between students as a new measure. Based on the measure, we then adopt a clustering technique[2,3,5] to distinguish clusters, i.e., groups whose members are similar with each other. As a clustering algorithm, we use ROCK[2,3], which is widely used in the case where the similarity between two objects can be obtained in advance.

To use the algorithm ROCK, we have to define the similarity between two objects, i.e., two students. It is because most of clustering techniques including ROCK perform clustering based on the similarity(or the distance) between two objects. Hence, to identify meaningful student groups using ROCK, we first compute the similarity between two students based on communication history records. By the way, in Section 3 we have already defined the degree of familiarity between students and also proposed an algorithm of computing the degrees from communication history records. Thus, by using the degree of familiarity, we formally define the measure of similarity between two students and propose an algorithm of computing the similarity.

Intuitively speaking, the similarity between two students becomes higher as the number of their common friends increases. That is, if two students have a large number of common friends, we can say that they are in close friendship and are similar with each other. By this intuition, we conceptually define the similarity between two students x and y as "how many friends of x are similar with those of y," i.e., "how many common friends x and y have." To compute the similarity between students based on such an intuitional concept, we need to obtain the list of friends for each student first. We include student y into the list $\mathit{flist}(x)$ of friends for student x if the degree $\mathit{dof}(x, y)$ of familiarity from x to y is greater than another user-specified threshold δ. Here, the threshold δ will also be given by an analyzer, and its value can be more than 0. That is, the threshold δ is a quantitative criterion of how often student x tries to communicate with student y in order to regard y as x's friend (To simplify the problem, we use 0 as δ, i.e., we assume the students with whom x tries to communicate at least one time as x's friends.). According to this assumption, if $\mathit{dof}(x, y)$ is greater than δ, then we include student y into $\mathit{flist}(x)$, the list of friends for student x.

Next, by using the list of friends for each student, we define the similarity $\mathit{sim}(x, y)$ between two students x and y as follows.

Definition 2. Given $\mathit{flist}(x)$ and $\mathit{flist}(y)$ that are the lists of friends for students x and y, respectively, the *similarity* $\mathit{sim}(x, y)$ between x and y is defined as the ratio of the union of two lists to the intersection of two lists. That is, the similarity between x and y is defined as the following Eq. (2):

$$\mathit{sim}(x, y) = \frac{|\ \mathit{flist}(x) \cap \mathit{flist}(y)\ |}{|\ \mathit{flist}(x) \cup \mathit{flist}(y)\ |} \tag{2}$$

In brief, the similarity in Definition 2 means the ratio of common friends of two students. That is, the divisor in Eq. (2), the union of two lists, represents all the friends of two students; the dividend, the intersection of two lists, the common friends of them. Thus, we may regard Eq. (2) as the degree of how many common friends two students have, and we are able to use the similarity in Definition 2 as the similarity measure in clustering algorithms.

Figure 2 shows the algorithm *ComputeSim*() that computes the similarity between each pair of students based on the degree of familiarity. The inputs to the algorithm are a set of $dof(x, y)$'s and the threshold δ. In Steps (1)∼(3), we first compute the list $flist(x)$ of friends for each student x by using the degree of familiarity. In Steps (4)∼(6), we then compute the similarity between two students x and y by using the lists of friends for x and y. After computing the similarity between every pair of students using the algorithm *ComputeSim*(), we now perform clustering by using the obtained similarity measures. That is, we can identify clusters, i.e., meaningful student groups, by applying the similarity between students to the algorithm ROCK[3].

procedure *ComputeSim*(A set of $dof(x,y)$'s, Threshold δ)
begin
(1) **for each** x **in** S
(2) **for each** y **in** S $(y \neq x)$
(3) **if** $dof(x,y) > \delta$ **then** Include y into $flist(x)$;
(4) **for each** x **in** S
(5) **for each** y **in** S $(y \neq x)$
(6) $$sim(x,y) = \frac{|flist(x) \cap flist(y)|}{|flist(x) \cup flist(y)|};$$
end

Fig. 2. The algorithm *ComputeSim*() of computing the similarity between students

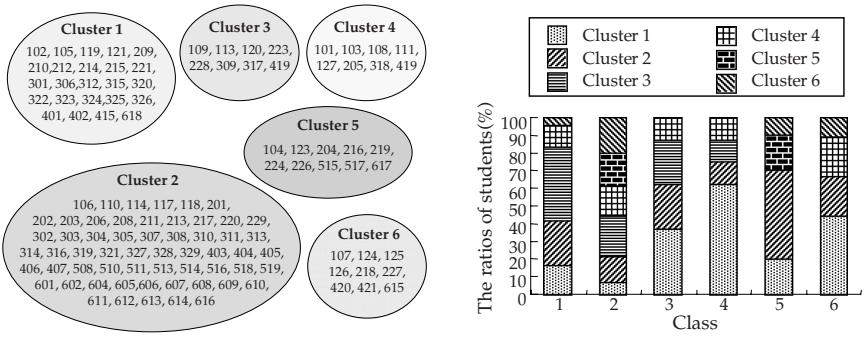
5 Experiments and Analysis

We implemented all the algorithms proposed in Sections 3 and 4. We conducted all the experiments on a Windows 2000 server with Intel Pentium 4 3.0GHz CPU, 392MB RAM, and 30GB hard disk. And, we used Borland Delphi language to implement the algorithms.

We have collected the communication history records through a survey. Actually, we can obtain the real history records from communication companies only after we get legal agrees of individual students due to "communication secrete guard law[1]." In our situation, however, it is difficult to get all the necessary agrees from every student, and thus we collected the history records through a survey. The survey was performed on all the second year high school students of six classes located in a city of Korea. The items surveyed for each student (i.e., for each sender) consist of receivers(or terminators) and the numbers of

communications for representative communication tools including cellular phones, SMS messages, messengers, and e-mails.

In the experiment, we fix the number of groups(clusters) to six to distinguish student groups from all students. In other words, we perform the experiments by setting the number of clusters to be extracted to six which is the same as the number of classes surveyed for the experiments. We intuitively expected that the pattern of extracted clusters would have some meaningful relationship with the structure of classes since the number of clusters is identical to the number of classes. That is, we expected the students in the same cluster might be contained in the same class. However, the expectation is not correct. Experimental results show that the students in the same cluster are widely distributed over all the classes. Figure 3 (a) shows six clusters and their student identifiers, and Figure 3 (b) a chart of students' distributions over classes for each cluster. From Figure 3, we can see that the clusters are formed over several classes rather than over one or two specific classes. In particular, we can observe that Cluster 2 in the figure contains about 41% of all students, and this confirms the fact that student groups are formed over all the classes rather than within an individual class. This conclusion indicates that it is required for teachers or parents to guide individual students by considering inter-class student relationships as well as intra-class student relationships.



(a) Clusters and their students (b) Distributions over classes

Fig. 3. Student's distribution over clusters extracted by the clustering algorithm

In Figure 3, Clusters 3, 4, 5, and 6 are small clusters that have a very small number of members. We think these small clusters will be different from the normal clusters, Clusters 1 and 2, that contain most of students. That is, these small groups may be identified as abnormal groups that are isolated from other students or as the potentially delinquent groups that may use violence or commit blackmail. Therefore, we recommend that teachers or parents may give much more attention and special guidance to these small student groups.

6 Conclusions

In this paper we have proposed an objective and systematic approach to mining meaningful student groups by using the communication history records. Contributions of the paper can be summarized as follows. First, we have formally defined the degree of familiarity between students and presented mathematical formulas that compute the degree based on communication history records. Second, we have adopted the clustering technique to identify meaningful student groups using the degree of familiarity. To use the clustering technique, we have defined the measure of similarity between students and proposed an algorithm of computing the similarity from the degrees of familiarity. Third, we have shown the practicalty of our analysis method by implementing the method and by interpreting the meaning of experimental results. These results indicate that our research result provides a useful framework that analyzes peer relationships among students more objectively and more systematically.

Acknowledgements

This work was supported by the Ministry of Science and Technology (MOST)/Korea Science and Engineering Foundation (KOSEF) through the Advanced Information Technology Research Center (AITrc).

References

1. Communication Secret Guard Law: Initial Legislation in 1993 (Law No. 4650), The 8th Revision in 2004 (Law No. 7138) (in Korean)
2. Dutta, M., Mahanta, A.K., Pujari, A.K.: QROCK: A Quick Version of the ROCK Algorithm for Clustering of Categorical Data. Journal of Pattern Recognition Letters 26(15), 2364–2373 (2005)
3. Guha, S., Rastogi, R., Shim, K.: 'ROCK: A Robust Clustering Algorithm for Categorical Attributes. In: Proc. Int'l Conf. on Data Engineering(ICDE), pp. 512–521. IEEE Press, Sydney, Australia (1999)
4. Herbert, B.: Symbolic Interactionism: Perspective and Method. Prentice-Hall, Englewood Cliffs (1969)
5. Hwang, J.-J., Whang, K.-Y., Moon, Y.-S., Lee, B.-S.: A Top-down Approach for Density-Based Clustering Using Multidimensional Indexes. Journal of Systems and Software 73(1), 169–180 (2004)
6. Lee, C.: Preventing Bullying in Schools. Sage, Thousand Oaks (2004)
7. Media News Daily, AsiaMedia (March 23, 2005), `http://www.asiamedia.ucla.edu/article.asp?parentid=22103`
8. Saltz, J.S., Hiltz, S.R., Turoff, M.: Student Social Graphs: Visualizing a Student's Online Social Network. In: Proc. of Conference on Computer Supported Cooperative Work, ACM CSCW, Chicago, Illinois, pp. 596–599 (2004)
9. Warr, M.: Organization and Instigation in Deliquent Groups. Criminology 34(1), 11–37 (1996)

Ontological Engineering for Practical Knowledge Work

Tatiana Gavrilova

Saint-Petersburg State Polytechnic University
Intelligent Computer Technologies Dpt.
195251, Politechnicheskaya 29, St. Petersburg, Russia
tgavrilova@gmail.com

Abstract. The paper describes one approach to teaching and training of analysts for practical knowledge processing. The training framework is targeted on the development of skills that will allow facilitating the process of knowledge elicitation, structuring and ontology development. The structuring procedure is the kernel of ontological engineering. The 5-steps ontology designing process is proposed. Special stress is put on "beautification" principles of ontology creating. The curriculum includes interactive game-format training of lateral thinking, interpersonal cognitive intellect and visual mind mapping techniques. Latter works as a powerful mind tool.

Keywords: knowledge engineering, ontologies, analyst training.

1 Introduction

Knowledge work deals with structuring. New information age with huge information overload makes knowledge processing more and more sophisticated. Sophistication needs professionals.

How many professional knowledge engineers need the company targeted at information processing? Where to find them?

Knowledge Engineering traditionally emphasized and rapidly developed a range of techniques and tools including knowledge acquisition, conceptual structuring and representation models [1], [24]. But for practitioners as enterprise analysts it is still a rather new, eclectic domain that draws upon areas like cognitive science. Accordingly, knowledge engineering has been, and still is, in danger from fragmentation, incoherence and superficiality. Still few universities deliver courses in practical knowledge engineering.

Many companies take decision to teach and train their IT-staff and developers. This paper describes recent experience in such training for some Russian subsidiaries of international companies (British-American Tobacco, Siemens Business Services, etc.). The total number of trainees that received certificates of knowledge analysts is more than 60.

Training for Knowledge Engineering (TKE) is based on university courses in intelligent-systems development, cognitive sciences, user modeling and human-computer interaction delivered by author in 1992-2005. TKE proposes information

B. Apolloni et al. (Eds.): KES 2007/ WIRN 2007, Part II, LNAI 4693, pp. 1154–1161, 2007.
© Springer-Verlag Berlin Heidelberg 2007

structuring multi-disciplinary methodology, including the principles, practices, issues, methods, techniques involved with the knowledge elicitation, structuring and formalizing. Emphasis is put not on technologies and tools, but in training of analytical skills. Ontological Engineering is further development of knowledge engineering towards ontology design and creating.

2 Teaching Outline

Trainees are introduced to major issues in the field and to the role of the knowledge analyst in strategic information system development. Attention is given both to developing inter-personal information communication skills and analytical cognitive creative abilities. One group is not more than 8 persons.

Participants gain an understanding the role of knowledge engineering and knowledge management in companies and organizations; in decision-making by members of an organization; in developing information framework. They study and are trained in practical methods mainly by doing.

The class feature short lectures, discussions, tests, quizzes and exercises. Lectures are important but the emphasis is put on learning through discussions, simulation, special games, training and case studies. A good deal of the course focuses on auto-reflection and auto-formalizing of knowledge, training of analytical and communicative abilities, discovery, creativity, cognitive styles features, and gaining new insights.

TKE consists of 4 inter-connected modules:

- Getting Started in KE (12 hours),
- Practical KE in depth (12 hours),
- Ontological Engineering (12 hours),
- Business Processes Modelling (12 hours).

Different combination of sub-topics is possible. Fig.1 illustrates the structure of one variant chosen by Business Engineering Group Company.

The main difference of TKE to existing methodologies is cognitive (not technological) bias. The topics of exercises cover categorization, observation, laddering, lateral thinking and other problem solving cognitive methods. IT-managers often under-value the significance of psychological background of categorization, laddering and lateral thinking. But during training some of them feel “insight” and become very enthusiastic.

But only knowledge structuring exercises show the importance of obtained analytical skills in practice. Even simple tests from their domains are rather difficult at the first workshops.

The training is aimed on semantics not syntax of knowledge engineering. We suppose that systems and languages may be self-studied while general scope and knowledge-stressed approach should be trained thoroughly. Practical specialists often under-estimate the role of cognitive styles, verbal skills and logics in information processing. It is supposed to be common sense while it needs to be taught.

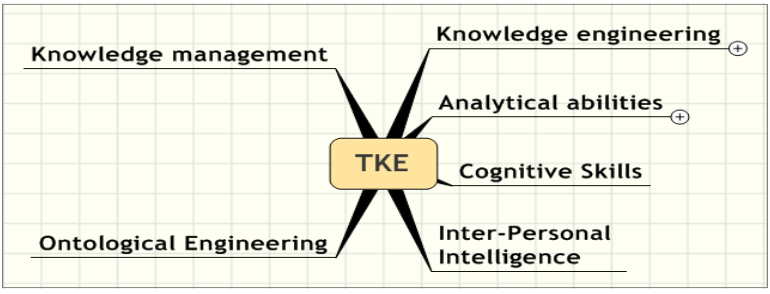


Fig. 1. Outline of training on knowledge engineering

3 Ontological Engineering Framework

Ontologies are fashionable now. But our experience in training show that nobody can deal with ontologies without knowledge engineering practice. How to teach ontology design? The theory differs from practical needs…

There are numerous well-known definitions of this milestone term (Gruber , 1993; Guarino and Giaretta, 1998; Jasper and Uschold, 1999; Mizogushi and Bourdeau, 2000; Neches, 1991) but they may be generalized as "Ontology is a hierarchically structured set of terms for describing an arbitrary domain"[9]. In other words "ontologies are nothing but making knowledge explicit" [12].

Since 2000 a major interest of researchers focuses on building customized tools that aid in the process of knowledge capture and structuring. This new generation of tools – such as Protégé, OntoEdit, and OilEd - is concerned with visual knowledge mapping that facilitates knowledge sharing and reuse [18], [19], [22]. The problem of iconic representation has been partially solved by developing knowledge repositories and ontology servers where reusable static domain knowledge is stored. Ontolingua, and Ontobroker are examples of such projects [20], [21].

But practitioners from companies need simple and constructive algorithms for their activity.

Ontology creating also faces the knowledge acquisition bottleneck problem. The ontology developer encounters the additional problem of not having sufficiently tested practical methodologies, which would recommend what activities to perform. An example of this can be seen when each development team usually follows their own set of principles, design criteria, and steps in the ontology development process. The lack of structured guidelines hinders the development of shared and consensual ontologies within and between the teams. Moreover, it makes the extension of a given ontology by others, its reuse in other ontologies, and final applications difficult [11].

Several effective methodological approaches have been reported for building ontologies [4]; [25], [16]. What they have in common is that they start from the identification of the purpose of the ontology and the needs for the domain knowledge acquisition. However, having acquired a significant amount of knowledge, major researchers propose a formal language expressing the idea as a set of intermediate representations and then generating the ontology using translators. These represent-tations bridge the gap between how people see a domain and the languages in which

ontologies are formalized. The conceptual models are implicit in the implementation codes. A re-engineering process is usually required to make the conceptual models explicit.

Fig. 2 presents our vision of the mainstream state-of-the-art categorization in ontological engineering [12], [13], [28] and may help the knowledge analyst to figure out what type of ontology he/she really needs. We use Mindmanager™ and Cmap as they proved to be powerful visual tools.

We try to simplify a bunch of different approaches, terms and notations for practical use and dare to propose a 5-steps recipe for practical ontology design.

3.1 Ontology Design Process

The existing methodologies describing ontology life cycle [28], [16], [9] deal with general phases and sometimes don't discover the design process in details. Five simple practical steps were proposed.

Step 1. Glossary development: The first step should be devoted to gathering all the information relevant to the described domain. The main goal of this step is selecting and verbalizing all the essential objects and concepts in the domain.

Step 2. Laddering: Having all the essential objects and concepts of the domain in hand, the next step is to define the main levels of abstraction. It is also important to elucidate the type of ontology according to Fig. 1 classification, such as taxonomy, partonomy, and genealogy. This is being done at this step since it affects the next stages of the design. Consequently, the high level hierarchies among the concepts should be revealed and the hierarchy should be represented visually on the defined levels.

Step 3. Disintegration: The main goal of this step is breaking high level concepts, built in the previous step, into a set of detailed ones where it is needed. This could be done via a top-down strategy trying to break the high level concept from the root of previously built hierarchy.

Step 4. Categorization: At this stage, detailed concepts are revealed in a structured hierarchy and the main goal at this stage is generalization via bottom-up structuring strategy. This could be done by associating similar concepts to create meta-concepts from leaves of the aforementioned hierarchy.

Step 5. Refinement: The final step is devoted to updating the visual structure by excluding the excessiveness, synonymy, and contradictions. As mentioned before, the main goal of the final step is try to create a beautiful ontology. We believe what makes ontology beautiful is harmony.

Using these tips the trainees develop several huge company ontologies.

3.2 From Good Gestalt to Ontology Harmony

The idea of the good shape strength is rather common in science and modelling. One of substantial impulse to it was given by German psychological school of M. Wertheimer. His idea of good Gestalt (image or pattern) may be transferred into ontological engineering. Some essential Gestalt principles of this school [30]:

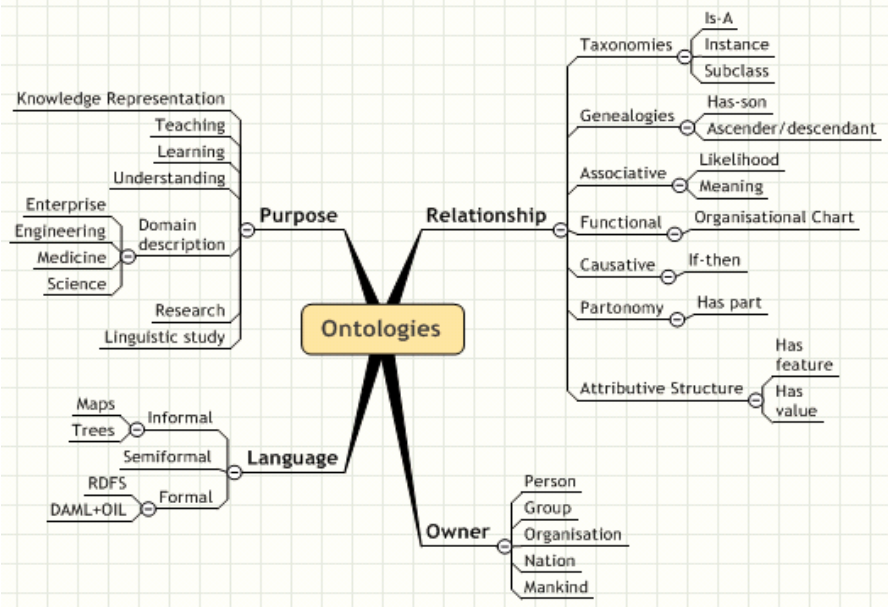


Fig. 2. Ontology mind map

- Law of Pragnanz (M. Wertheimer) - organization of any structure in nature or cognition will be as good (regular, complete, balanced, or symmetrical) as the prevailing conditions allow (law of good shape).
- Law of Proximity – objects or stimuli that are viewed being close together will tend to be perceived as a unit.
- Law of Similarity – things that appear to have the same attributes are usually perceived as being a whole.
- Law of Inclusiveness (W.Kohler) - there is a tending to perceive only the larger figure and not the smaller when it is embedded in a larger.
- Law of Parsimony – the simplest example is the best or known as Ockham's razor principle (14-th century): ``entities should not be multiplied unnecessarily".

We suggest to use these laws for pursuing conceptual balance and clarity.

3.2.1 Conceptual Balance

A well-balanced ontological hierarchy equals a strong and comprehensible representation of the domain knowledge. However, it is a challenge to formulate the idea of a well-balanced tree. Here we offer some tips to help formulate the "harmony":

- Concepts of one level should be linked with the parent concept by one type of relationship such as is-a, or has part.
- The depth of the branches should be more or less equal (±2 nodes).

- The general outlay should be symmetrical.
- Cross-links should be avoided as much as possible.

Fig.3 illustrates the idea of well-balanced (A) and ill-balanced (B) ontology design.

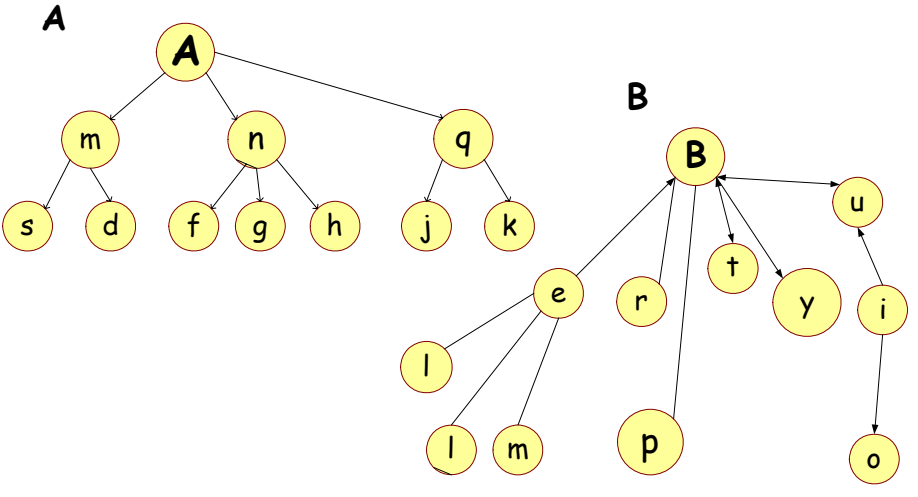


Fig. 3. "Well-balanced" and "ill-balanced" ontologies

3.2.2 Clarity

Moreover, when building a comprehensible ontology it is important to pay attention to clarity. Clarity may be provided through number of concepts and type of the relationships among the concepts. Minimizing the number of concepts is the best tip according to Law of Parsimony.

The maximal number of branches and the number of levels should follow Miller's magical number (7±2) [15]. Furthermore, the type of relationship should be clear and obvious if the name of the link is missed.

Some practical tips to refine and illuminate the ontology's design layout stage can be proposed:

- Use different font sizes for different levels.
- Use different colours to distinguish particular subsets or branches.
- Use a vertical layout of the tree structure/diagram.
- If needed, use different shapes for different types of nodes.

At the first stages it is possible to use any of the available graphical editors to design an ontology, e.g. PaintBrush, Visio, Inspiration. A nice layout can be reached by using mindmapping tools as MindManager™ or Visual Mind ™.

The trainees really enjoy the process of "beatification" of their ontologies during test exercises.

4 Discussion

IT-specialists enter "the world of ontologies" with interest and begin to use it in their practical work. They are knowledge workers. Our experience in training of knowledge analysts in the period of 1999-2006 confirm the unique role of knowledge structuring for developing ontologies quickly, efficiently and effectively. We follow David Jonassen's idea of "using concept maps as a mind tool" [14]. The use of visual paradigm to represent and support the training process not only helps a professional trainer to concentrate on the problem rather than on details, but also enables students to process and understand greater volume of information.

The described approach is twice ontological as the development of educational knowledge structures in the form of ontologies provides training and learning support. Teaching ontologies used in the course scaffold and improve trainees' understanding of the courseware and later help to realize substantive and syntactic company knowledge. As such, they can play a part in the overall pattern of learning, facilitating for example analysis, comparison, generalization and transferability of understanding to analogous problems.

Acknowledgements

This work is partly supported by the grant of Russian Foundation for Basic Research No. 06-01-81005.

References

1. Adeli, H.: Knowledge Engineering. McGraw-Hill, New-York (1994)
2. Boose, J.H.: Knowledge Acquisition Tools, Methods and Mediating Representations. In: Motoda, H., et al. (eds.) Knowledge Acquisition for Knowledge-Based Systems, pp. 123–168. IOS Press, Ohinsha Ltd, Tokyo (1990)
3. Eisenstadt, M., Domingue, J., Rajan, T., Motta, E.: Visual Knowledge Engineering. IEEE Transactions on Software Engineering 16(10), 1164–1177 (1990)
4. Fensel, D.: Ontologies: A Silver Bullet foe Knowledge Management and Electronic Commerce. Springer, Heidelberg (2001)
5. Gavrilova, T., Voinov, A.: Work in Progress: Visual Specification of Knowledge Bases. In: Mira, J.M., Moonis, A., de Pobil, A.P. (eds.) Tasks and Methods in Applied Artificial Intelligence. LNCS, vol. 1416, pp. 717–726. Springer, Heidelberg (1998)
6. Gavrilova, T.A., Voinov, A., Vasilyeva, E.: Visual Knowledge Engineering as a Cognitive Tool. In: Mira, J.M. (ed.) IWANN 1999. LNCS, vol. 1606, pp. 123–128. Springer, Heidelberg (1999)
7. Gavrilova, T.: Teaching via Using Ontological Engineering. In: Proceedings of XI Int. Conf. Powerful ICT for Teaching and Learning PEG-2003, St.Petersburg, pp. 23–26 (2003)
8. Gavrilova, T., Kurochkin, M., Veremiev, V.: Teaching Strategies and Ontologies for E-learning. Int. J. Information Theories and Applications 11(1), 35–42 (2004)

9. Gómez-Pérez, A., Fernández-López, M., Corcho, O.: Ontological Engineering with examples from the areas of Knowledge Management, e-Commerce and the Semantic Web. Springer, Heidelberg (2004)
10. Gruber, T.: A translation approach to portable ontology specifications. Knowledge Acquisition 5, 199–220 (1993)
11. Guarino, N., Giaretta, P.: Ontologies and Knowledge Bases: Towards a Terminological Clarification. Towards Very Large Knowledge Bases: Knowledge Building & Knowledge Sharing, pp. 25–32. IOS Press, Amsterdam (1998)
12. Guarino, N., Welty, C.: A Formal Ontology of Properties. In: Dieng, R., Corby, O. (eds.) EKAW 2000. LNCS (LNAI), vol. 1937, pp. 97–112. Springer, Heidelberg (2000)
13. Jasper, R., Uschold, M.: A Framework for Understanding and Classifying Ontology Applications. In: KAW'99. 12th Workshop on Knowledge Acquisition Modeling and Management (1999)
14. Jonassen, D.H.: Designing constructivist learning environments. In: Reigeluth, C.M. (ed.) Instructional design models and strategies, 2nd edn., Lawrence Eribaum, Mahwah, NJ (1998)
15. Miller, G.: The Magical Number Seven, Plus or Minus Two: Some Limits on Our Capacity for Processing Information. The Psychological Review 63, 81–97 (1956)
16. Mizogushi, R., Bourdeau, J.: Using Ontological Engineering to Overcome Common AI-ED Problems. International Journal of Artificial Intelligence in Education 11, 1–12 (2000)
17. Neches, et al.: Enabling Technology for Knowledge Sharing, AI Magazin, pp. 36–56 (Winter 1991)
18. Bechhofer, S., Ng, G.: OilEd (December 7, 2004), http://oiled.man.ac.uk/
19. OntoEdit, AIFB, University Karlsruhe (December 07, 2004), Accessed from http://www.ontoknowledge.org/tools/
20. OntoBroker, (December 7, 2004), Accessed from http://ontobroker.aifb.uni-karlsruhe.de/
21. Ontolingua, Stanford University (December 7, 2004) Accessed from http:// www.ksl.stanford.edu/ software/ontolingua/
22. Protégé, Stanford Medical Informatics (December 07, 2004) Accessed from http://protege.stanford.edu/ at
23. Sowa, J.F.: Conceptual Structures: Information Processing in Mind and Machine. Addison-Wesley, Reading, Massachusetts (1984)
24. Scott, A., Clayton, J.E., Gibson, E.L.: A Practical Guide to Knowledge Acquisition. Addison-Wesley, Reading (1994)
25. Swartout, B., Patil, R., Knight, K., Russ, T.: Toward Distributed Use of Large-Scale Ontologies. In: Ontological Engineering, AAAI- 97. Spring Symposium Series, pp. 138–148 (1997)
26. The CIO's Guide to Semantics, © Semantic Arts, Inc. (2004), www.semantic-conference.com
27. Tu, S., Eriksson, H., Gennari, J., Shahar, Y., Musen, M.: Ontology-Based Configuration of Problem-Solving Methods and Generation of Knowledge-Acquisition Tools. Artificial Intelligence in Medicine, N7, 257–289 (1995)
28. Uschold, M., Gruninger, M.: Ontologies: Principles Methods and Applications. Knowledge Engineering Review 1(1) (1996)
29. Wielinga, B., Schreiber, G., Breuker, J.: A Modelling Approach to Knowledge Engineering. Knowledge Acquisition 4(1) (Special Issue), 23–39 (1992)
30. Werthheimer, M.: Productive Thinking. HarperCollins, New York (1959)

Semantic Modeling of Product Manuals

Rossitza Setchi[1], Nikolaos Lagos[1], and Ammar Huneiti[2]

[1] Cardiff University, School of Engineering, Cardiff CF24 3AA, UK
[2] IT Department, University of Jordan, Amman, Jordan
{setchi, lagosN}@cf.ac.uk, a.huneiti@ju.edu.jo

Abstract. Compared to other product lifecycle phases, the development of product documentation and technical manuals in particular, is an area currently undersupported by tools. This is in contrast with the benefits associated with the production of high quality product documentation, and the effort involved to produce it. This paper addresses this problem by proposing three complementary semantic-based methods for modeling product manuals. The first approach is based on using semantically linked information objects. The second approach focuses on abstracting the intended purpose, user tasks and the functional characteristics of the product data elements. The third approach uses ontologies to capture the semantic complexity of the product support domain.

Keywords: Product documentation, technical manuals, hypermedia, semantic modeling, ontology, information object, performance support.

1 Introduction

Informative, up-to-date and well-prepared product documentation can provide time and cost savings to both users and product manufacturers/supplier. However, authoring product documentation is not easy. It is "a composite skilled task that requires special knowledge and experience" [1], "an active and complex process that one masters only through ongoing learning" [2]. These statements are in sharp contrast with the observation that product support is an area currently undersupported by tools. The problem of cost-effectively producing technical documentation becomes apparent especially when product support is compared to other phases of the product lifecycle, which are supported by tools for computer-aided design (CAD), product data management (PDM), computer-aided manufacture (CAM) and computer-aided planning (CAPP). Without similar tools, technical authors experience difficulties in collecting and integrating product data, and structuring and maintaining technical documentation.

The aim of this work is to demonstrate how these problems can be addressed by employing semantic modeling techniques. In particular, the paper presents three complementary approaches to semantic based authoring of hypermedia product manuals. The first approach is based on building product documentation from information objects. The second approach focuses on modeling product documentation by abstracting its intended purpose, user tasks and the functional

B. Apolloni et al. (Eds.): KES 2007/ WIRN 2007, Part II, LNAI 4693, pp. 1162–1169, 2007.
© Springer-Verlag Berlin Heidelberg 2007

characteristics of the information elements. The third approach uses ontologies to capture the semantic complexity of the product support domain. The remainder of the paper is organized as follows. Section 2 briefly reviews recent developments in modeling hypermedia documentation for performance support. Section 3 presents the three complementary approaches. Section 4 concludes the paper.

2 Background and Related Work

The importance of integrating knowledge engineering practices into the development of product support systems has been acknowledged by several researchers. Earlier studies focused on introducing rule-based and model-based reasoning in product support systems, mainly as diagnostic tools. However, most existing approaches lacked "structured representation of performance data and mechanisms for transforming that data into applicable knowledge" [3]. Gradually, the emphasis in this area of research shifted from the reasoning mechanism to the knowledge base, and a substantial body of knowledge was created on developing and integrating models of products, users, and tasks [4, 5].

In knowledge system applications, domain knowledge is encoded by a wide variety of knowledge representation techniques [6], including semantic networks [7], frames [8], and ontologies [9]. Latest research in product support indicates a trend towards semantic data modeling as it helps technical authors to categorize the information fragments used in the product documentation into abstract types and sub-types that reflect their subsequent use. The goal is to approximate the semantic content of the documentation components and represent it using semantic structures which can be further used for reasoning. The semantic-based approaches reported in the literature include the use of concept maps in knowledge elicitation [10], concept-based searches in information retrieval [11], document fragmentation and decomposition based on the relationships between document elements [12], and ontologies for modeling tasks and goals [13].

The research described in the next section extends these approaches by representing the semantic content of documentation components by capturing their significant characteristics, creating abstract descriptions and building semantic models that are then used in the provision of intelligent product support.

3 Three Approaches to Semantic Modeling of Hypermedia Documentation

3.1 Semantic Authoring Using Information Objects

The first approach advocates the building of product documentation from logically linked, semantically distinguishable and classifiable information objects (IOs). An IO is a data structure that represents an identifiable and meaningful instance of information in a specific presentation form [14]. In a virtual collaborative space, these IOs are structured using product data and product documentation models. These are

information models that describe the underlying structures and relationships of product data and documentation.

The product data model is a structured representation of information about a product. It integrates the life-cycle information necessary to completely define the product and each of its components for the purposes of design, engineering analysis, manufacture, test and product support. The structure of the product data model follows the topology of the product (product, systems, subsystems, assemblies, etc.). Similarly, the product documentation model is a structured representation of the basic document components (chapters, sections, subsections, etc.). The product documentation model contains objects are created during documentation development. It uses the product data model as a primary source of information.

To help the process of direct association between these two models, they are constructed identically as sets of manageable information elements (IEs). IEs are annotated using attributes that enable product data to be easily organized, addressed, classified and maintained. Each product or documentation element contains a definition and data. The element definition consists of attributes used to identify the element (id number, name, quantity, type, and version). Element data comprise one or more information objects (IOs). All IOs associated with a particular IE inherit its definition. The IOs are described using attributes identified through an analysis of the main documentation components and the principal sources of product data. The analysis focused primarily on the function, content and presentation form of the IOs. As Fig. 1 shows, the IOs are described using four groups of attributes which capture the semantics of these objects. Two of the groups, definition and processing, contain attributes used to address, process and maintain data. The other two categories, type and assessment, are for describing information objects in terms of their contextual meaning and presentation form.

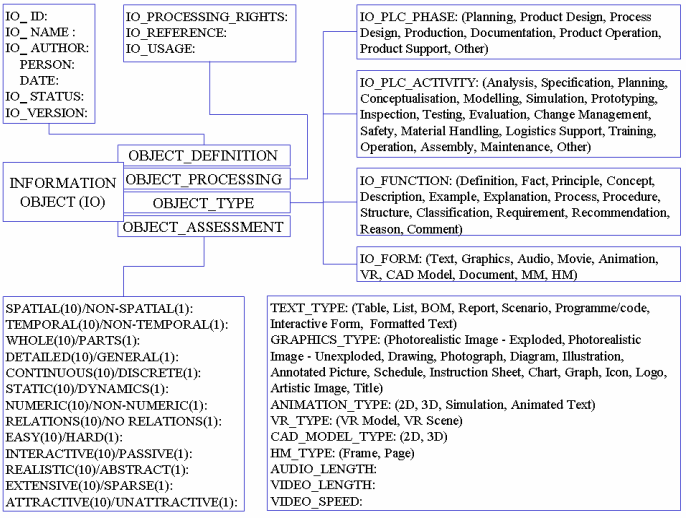


Fig. 1. IO attributes

Table 1 shows an example of existing relationships between IOs, and the attributes used to describe them. The hypermedia document in Table 1 “lubric_points_10.htm” illustrating the lubricating procedure 10 contains a video file created using the CAD models of the braking system and its components “movie_pt10.avi”. The identification number of this IE (DOC-8-1-2-4-0-0-0-0) indicates that it belongs to the product documentation model, and is included in chapter 8 (Lubrication Chart), section 1-2 (Lubrication points), and subsection 4 (point 10) of the technical documentation. The ID of the information element “outside.asm” (DV-1-1-1-2-0-0) is composed in a similar way. It shows that this IE belongs to the product data model; its coding follows the topology of the product. Table 1 also includes IOs generated using the CAD model of the subassembly "outside" such as the VR animation “outside.iv” and the video clip “outside_a_1.avi”. This table comprises data extracted from an integrated virtual environment for collaborative working containing 1558 files including 336 hypermedia documents, 656 images, 213 virtual reality (VR) models, 166 CAD models, 63 3D animations, 75 video files, 27 audio files, etc.

Table 1. Relationship report generated in a product data management system

Name	ID	Type	Phase	Activity	Form	Function
lubric_points_10.htm	DOC-8-1-2-4-0-0-0-0	Subsection	Documentation	Modeling	Hypermedia	Procedure
movie_pt10.avi	DV-1-1-1-1-0-0	Assembly	Documentation	Maintenance	Movie	Procedure
outside.asm	DV-1-1-1-2-0-0	Subassembly	Product Design	Modeling	CAD	Structure
020700_1.prt	DV-1-1-1-2-0-1	Part	Product Design	Modeling	CAD	Structure
outside1.unv	DV-1-1-1-2-0-0	Subassembly	Product Design	Modeling	Text	Definition
outside.iv	DV-1-1-1-2-0-0	Subassembly	Documentation	Modeling	VR	Structure
outside_a_1.avi	DV-1-1-1-2-0-0	Subassembly	Documentation	Training	Movie	Description
out_exp.tif	DV-1-1-1-2-0-0	Subassembly	Documentation	Prototyping	Graphics	Structure

3.2 Usage-Based Semantic Data Model for Product Documentation

This semantic data model is extracted through usage analysis of product data which abstracts a hierarchical representation of (i) the purpose of the product documentation (planning, support, action), (ii) the supported user tasks (learning or performing), and (iii) the functional characteristics of the product documentation data elements (fundamental, procedural, clarification, advice, specification, organization).

An important factor in product information retrieval is the recognition of the user’s main pursuit or activity. Within the context of product support and documentation, user activities are the user’s general motive for accessing a particular product-related concept or piece of information. Normally, when a user requests assistance from the product support system s/he has, in the back of the mind, a known purpose and an associated task to accomplish. Thus the abstract activities supported by the product documentation can be indexed using a composite attribute combining task and product data purpose.

The semantic network shown in Fig. 2 illustrates the proposed approach. The model uses high-level primitives such as documentation purpose, task, activity, function, product concept, and product data element.

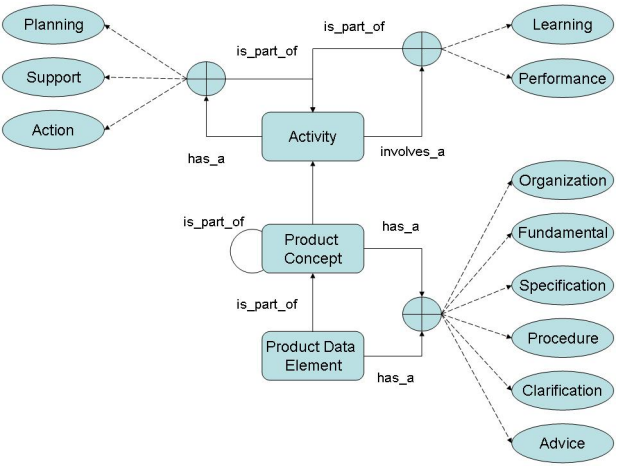


Fig. 2. Usage-based semantic network for product documentation

The semantic network is built on the following hypotheses:

- Every activity supported has a purpose and involves a practical task.
- A product concept is an abstract entity, which acts as a data structure for organizing product data elements. Every product concept is associated with a user activity, and it inherits the activity's documentation purpose and the associated task. In addition, a product concept can be aggregated from one or more concepts and/or product data elements.
- Product data elements are discrete pieces of information that describe or contribute to the description of a product concept. The aggregation of one or more product data elements creates a product concept. Within a product concept, every product data element has a distinct functional characteristic, which describes its role within the product concept or documentation (fundamental, procedure, clarification, advice, and specification). Product data elements are reusable i.e. the same element can perform different functions or roles within different concepts.

The grouping constructor $\oplus$ is used to represent entities of the same category. The activity entity performs a central role within the model. Rules and constraints have been applied to govern the validity and applicability of the relationships of the semantic data model. For example, a task of type "perform" can only be associated with documents of purpose "action" within the "perform action" abstract activity, which is described through a hierarchy of product concepts. These concepts are themselves described by product data elements of function type "procedure" and may be further illustrated by elements of function type "clarification" or "advice". Similar rules are created for other supported abstract activities. The full benefits of these rules and constraints can be appreciated when building the structure of the product documentation. A special tool has been developed to implement these rules and to guide the technical authors towards constructing a semantically valid product documentation structure.

A Web-based product support system in the form of an online technical manual has been developed using the presented semantic data model. The semantic data model is

represented in the hypermedia authoring process by a group of semantic attributes. These attributes identify the semantic properties of each product data element within the framework of the semantic data model. The values of these attributes are selected from predefined values and valid combinations identified by the domain expert who identifies the semantic properties of the current product data element by selecting a valid combination of these attribute values. The technical manual is automatically rendered in two different mark-up languages, HTML and XML. A set of CASE tools have been also developed in order to assist in the hypermedia authoring and implementation processes.

3.3 Semantic Modeling Using Ontologies

This approach aims at organizing data in a way that ensures the homogeneity, validity and effectiveness of the product support information generated. It is based on using ontological principles and takes advantage of the formality and richness of ontological knowledge modeling. This approach employs some widely accepted notions in knowledge engineering, such as concept, instance, and relation. In addition, several specialized structures, such as domain, connector, knowledge-specifier and arc, are employed to enrich the representation ability of the system.

Domain is a hierarchy of concepts (i.e. concepts linked with "is-a" relation), which describes a part of the real world and is a concept itself. For example, a domain "Product" (D) (see Fig. 3) includes the hierarchy "car is-a vehicle is-a Product". Domains identify where each abstraction hierarchy is contained, thus clarifying the mapping to different real world components. Associating concepts with the "is-a" relation and developing abstraction hierarchies (i.e. domains) is not enough in the case of a product support system. A relation that connects different domains is also needed. Take the example of "Product" (D) and "Assembly" (E) domains that are linked with the "Product has Assembly" relation (F). Relations that associate different domains are called connectors. Connectors express "part-of", "is-composed-of", "is-realized-with", and aggregation relations. Connectors can also link instances that belong to different domains, if the concepts to which these instances belong are included within these domains and the domains are also linked with connectors. Connectors cannot link information structures describing different knowledge areas (i.e. product, task, and user).

Knowledge-specifier is a property that is significant within the application domain (i.e. product support) and therefore is represented as a concept. For example, it may be useful to know whether a product is viewed as complex or not, since that determines the way in which the support provided is adapted. Therefore, a knowledge-specifier called "ProductType" (G), which defines product's complexity, is introduced. The level of the product support will be determined using mechanisms that include "ProductType". A simple example is the following predicate, which states that if the product is complex, the user is novice and the task is complex, then the product support information generated should be presented in a greater detail.

$$((ProductType = complex)\ AND\ (User = novice)\ AND\ (Task = complex)) \Rightarrow ProductSupport = detailed$$

Arc is a relation that links the knowledge-specifiers with other concepts and/or domains. For example, "ComplexProduct" (H) can refer to concept "car" and "clutch" (I). Furthermore, knowledge-specifiers and arcs are used to relate different knowledge

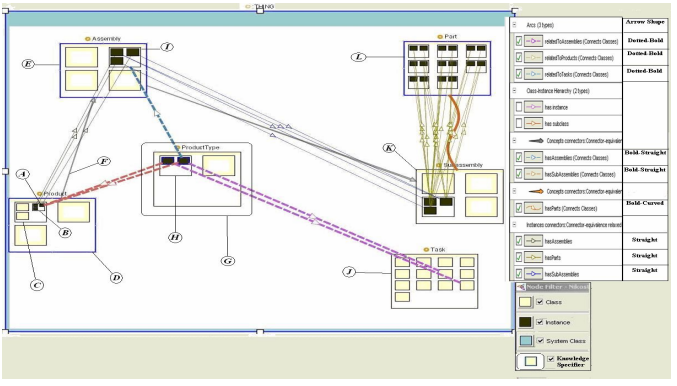


Fig. 3. A knowledge base developed using ontological principles (fragment)

areas. In Fig. 3, the task knowledge set (part of which is the task domain (J)) is related to the product knowledge set (composed by product, assembly, subassembly, and part domains) via "ProductType". More complex predicates can be developed that take in consideration the status of the user and relate it to the above arc, which determines the amount of support provided.

$$((Arc\ (ProductType, Task) = true)\ AND\ (ProductType = complex)\ AND\ (User = novice)\ AND\ (Task = complex)) \Rightarrow ProductSupport = detailed$$

The knowledge base is developed using the knowledge model described above, populated with instances and their associations. Each attribute of each concept corresponds to a small fraction of a product support document, while the value of that attribute defines the way in which the content of that fraction changes. Each concept is mapped to a part of the document that contains several of the aforementioned fragments that match the attributes that characterize each concept. The connectors and generalization relations are used to search for part of documents that relate to the required concept, while the knowledge-specifiers and arcs are utilized to establish the presentation format of the provided document.

The knowledge base is created using Protégé, and part of it is depicted in Fig. 3. The squares in the figure represent concepts. An "is-a" relation is represented by placing a square within a bigger one. The knowledge-specifier ("ProductType") is shown as a rounded square and the arcs that link it to the other concepts are represented by bold dotted lines. THING (the domain within which everything is contained) is the root concept of the ontology. Currently, the knowledge base contains 685 frames, including 162 concepts, 144 slots, 22 facets, and 357 instances.

4 Conclusions

The paper presents three semantic-based approaches to authoring of product documentation. The first approach focuses on the integration of product data and product documentation models enabling product design and documentation development and collaborative authoring of product documentation. The second approach represents product manuals in terms of their subsequent. It extends the first

approach by integrating the users' purpose into the development of the product manual from an early stage. The continuous involvement of usage considerations enables the design of a structured adaptation approach to different usage requirements. The third approach uses ontologies to represent the internal components of a document in terms of the data structures involved. It abstracts the knowledge base constructs used by the first two approaches and applies constraints on their subsequent modeling. The clarification of the data structures advances knowledge sharing and facilitates data homogeneity and validity. Promising direction for further research is the integration of these three approaches.

Acknowledgments. The authors are grateful for the support of their industrial collaborators and the European Community (ISAR project, grant LDV PL/05/B/F/PP/174007) in conducting the research described in this paper.

References

1. BS 4884 Technical manuals, Part2: Guide to content, British Standards Institution, London (1993)
2. Falkovych, K., Nack, F.: Context Aware Guidance for Multimedia Authoring: Harmonizing Domain and Discourse Knowledge. Multimedia Systems 11(3), 226–235 (2006)
3. Stary, C., Stoiber, S.: Model-based electronic performance support. In: Jorge, J.A., Jardim Nunes, N., Falcão e Cunha, J. (eds.) DSV-IS 2003. LNCS, vol. 2844, pp. 258–272. Springer, Heidelberg (2003)
4. Pham, D.T., Setchi, R.M.: Case-Based Generation of Adaptive Product Manuals. IMechE Proceedings 217(B), 313–322 (2003)
5. Brusilovsky, P., Cooper, D.W.: Domain, Task, and User Models for an Adaptive Hypermedia Performance Support System. In: Proceedings of the IUI '02, pp. 23–30. ACM, San Fransisco, California, USA (2002)
6. Yao, H., Etzkorn, L.: Automated conversion between different knowledge representation formats. Knowledge Based Systems 19(6), 404–412 (2006)
7. Khalifa, M., Liu, V.: Semantic Network Representation of Computer-Mediated Discussions: Conceptual Facilitation Form and Knowledge Acquisition, Omega, (2007) (in press) available online
8. Cornet, R., Abu Hanna, A.: Description Logic-Based Methods for Auditing Frame-Based Medical Terminological Systems. Artificial Intelligence in Medicine 34(3), 201–217 (2005)
9. Davies, J., Fensel, D., Van Harmelen, F.: Towards the Semantic Web-Ontology-Driven Knowledge Management. John Wiley & Sons, England (2003)
10. Coffey, J.W., Canas, A.J., Hill, G., Carff, R., Reichherzer, T., Suri, N.: Knowledge Modelling and the Creation of El-Tech: A Performance Support and Training System for Electronic Technicians. Expert Systems with Applications 25(4), 483–492 (2003)
11. McMahon, C., Lowe, A., Culley, S., Corderoy, M., Crossland, R., Shah, T., Stewart, D.: Waypoint: An Integrated Search and Retrieval System for Engineering Documents. Journal of Computing and Information Science in Engineering 4(4), 329–338 (2004)
12. Liu, S., McMahon, C.A., Darlington, M.J., Culley, S.J., Wild, P.J.: A Conceptual Framework for Retrieval of Document Fragments Based on Decomposition Schemes in Engineering Information Management. Advanced Engineering Informatics 20, 401–413 (2006)
13. Fox, J., Alabassi, A., Patkar, V., Rose, T., Black, E.: An Ontological Approach to Modelling Tasks and Goals. Computers in Biology and Medicine 36, 837–856 (2006)
14. Setchi, R.: Enhanced Product Support through Intelligent Product Manuals, PhD thesis, University of Wales Cardiff, UK (2000)

Appearance-Based Multi-robot Following Routes Using Incremental PCA

Luis Payá, Oscar Reinoso, Arturo Gil, Jose M. Pedrero, and Mónica Ballesta

Departamento de Ingeniería de Sistemas Industriales
Miguel Hernández University. Avda. de la Universidad s/n
03202 Elche (Alicante) Spain
lpaya@umh.es

Abstract. The appearance-based approach in visual robot navigation applied to the following of pre-recorded routes supposes several advantages, such as its application to non-structured environments and the relatively simple extraction of control laws. The classical approaches consist on two separated phases: learning and navigation, so, one robot cannot start navigation until learning has finished, and if some locations have to be added once learning has finished, the database must be created from the scratch. This work presents how an incremental PCA model can be used to overcome these limitations. With this approach, the follower robot can start while the database is still being built. Comparing the current view with those in the database, it can localize and navigate using a fuzzy controller. This approach can be applied to collaborative tasks where a team of robots must follow a guide one or in vigilance tasks where a route in a building has to be repeated continuously.

Keywords: Incremental PCA, Fuzzy Logic, Learning, Route Following, Collaborative Robotics.

1 Introduction

In the last years, the research on collaborative robotics has experienced a great development due to the advantages the use of multi-robot systems suppose comparing to only a robot performing the same task. Several works have been developed in the field of visual localization and navigation. [11] presents a probabilistic EKF algorithm where a team of robots builds a map online, while simultaneously they localize themselves. [3] takes into account the problem of the concurrent mapping and localization with extra positional information available when multiple vehicles operate simultaneously. In [4], a map is built using visual appearance. From sequences of images acquired by a team of robots, subsequences of visually similar images are detected and finally, the local maps are joined into a single map.

Some tasks imply one robot following a leader one (formations problem) or following a route that a previous robot has recorded. Different approaches have been presented to carry out this task, such as the model-based one, where some features of the image are extracted to carry out localization and navigation. An example is the

B. Apolloni et al. (Eds.): KES 2007/ WIRN 2007, Part II, LNAI 4693, pp. 1170–1178, 2007.
© Springer-Verlag Berlin Heidelberg 2007

behavior-based control [2]. Nowadays, a different approach, based on the global appearance, has attracted the interest of computer vision research. In these methods, the scenes are stored without any feature extraction, and recognition is achieved based on the matching of the whole images. This approach is useful for complicated scenes in the real world in which appropriate models for recognition are difficult to create.

Researchers have proposed several methods to carry out auto-location and navigation based on global appearance. Matsumoto et al. addressed a method in [7], consisting on the direct comparison of low-resolution images. Other approaches make use of the color histogram to perform auto-location, extracting other features such as texture, density of edges, etc., as shown in [13]. Also the complexity of the problem can be reduced working in the PCA (Principal Components Analysis) subspace as in [5] or [6], where through PCA techniques a database is created using a set of views, with a probabilistic approach for the localization. Other works have shown how this subspace reduction can be applied to other tasks in robotics, such as in loop closure detection for SLAM tasks [9].

Several works have dealt with the problem of following a pre-recorded route using an appearance-based approach [7], [10]. However, in the classical approaches, two phases need to be accomplished for this task: the learning one, where the robot stores general visual information from several points of view along the environment, and the autonomous navigation, where the robot performs its navigation comparing the current visual information with the stored one. This phase can not begin until the first one has finished. Also, a new model must be built from the scratch when we want to include information about new locations in the map. To overcome these problems, we propose in this work, using an incremental method to build the database. Then, with the system proposed in this work, a team of robots can follow the route that other robot is learning with some offset in the space and in the time.

The paper is arranged as follows. Section 2 describes how the information about the route can be stored using an incremental eigenspace approach. Section 3 shows how the follower robots use this database to carry out navigation. To finish, the results and the conclusions of the work are presented.

2 Database Construction Using Incremental PCA

In visual object recognition systems, different techniques have been used to reduce the dimensionality of the data. The goal is to discard the irrelevant or redundant data. The most popular technique for dimensionality reduction is PCA. This technique has been used both in general object recognition applied to robotic tasks [8] and in face recognition applications [12].

2.1 PCA

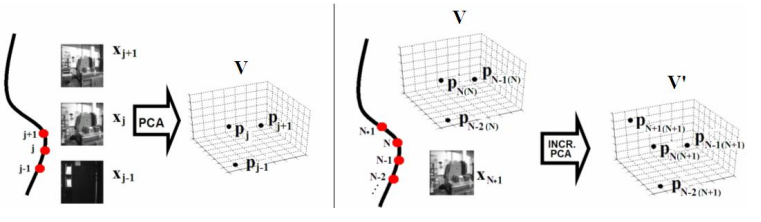
In the standard PCA procedure, the eigenspace is built using the data matrix *X*, that is composed by the training images $\vec{x}_j \in \Re^{Mx1}; j=1...N$ (arranged in 1-column vector), where *N* is the number of images and *M* the number of pixels of each image.

$$X = \left[\tilde{x}_1 \ldots | \tilde{x}_j | \ldots | \tilde{x}_N\right].$$
$$\tilde{x}_j = \vec{x}_j - \vec{m}; \quad \vec{m} = \frac{1}{N}\sum_{j=1}^{N} \vec{x}_j \,. \tag{1}$$

The PCA transformation is computed from the covariance of the data matrix using Eq. 1, e.g. using SVD and the Turk and Pentland's method [12]. If the size of the data matrix is *M x N*, (*M >> N*), a total amount of *N* eigenvectors are obtained. The dimensionality reduction is done by the projection:

$$P = V^T \cdot X \,. \tag{2}$$

where $V \in \Re^{MxK}$ is the matrix containing the *K* principal eigenvectors $v_i \in \Re^{Mx1}; i = 1 \ldots K$ (those with the higher associated eigenvalues) and $P \in \Re^{KxN}$ the reduced data matrix, where $K \leq N$. After this process, we have reduced each of the original images with *M* pixels to one vector $p_j \in \Re^{Kx1}; j = 1 \ldots N$ with the most relevant information of the image. The database will consist on these *N* projections and the matrix *V* to project the data. Fig. 1(a) shows the form of the database taking *K*=3. Each image could be reduced to a point in the 3D space. The three coordinates of each point contain the most relevant information of the learned images.



(a) (b)

Fig. 1. (a) Using PCA, the route images are dimensionally reduced. (b) Input and output data of Incremental PCA in each iteration.

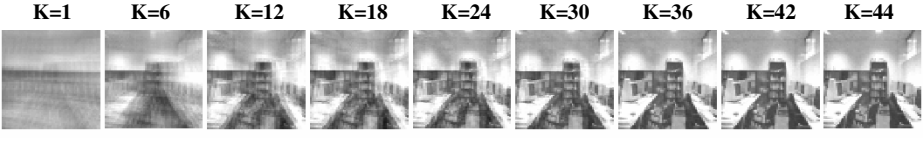


Fig. 2. Reconstruction of an image of a route of 45 points, using different quantity of eigenvectors. The reconstruction for K=44 corresponds to the original 64x64 image.

The projection matrix *P* has part of the original image information but with lower dimension vectors, depending on the number of eigenvectors used in the transformation. If *K* = *N*, the information in *P* is the same as in *X* (the reconstruction error is 0), but the dimension of the projected vectors is smaller: each projection p_j, has just *K* components. Fig. 2 illustrates this approach.

2.2 Incremental PCA

To build the database as exposed, all images must be available before carrying out the compression, so the robots that must follow the route should wait the first one to run till the end. To overcome this inconvenient, a progressively construction of the database is used. Several algorithms have been proposed to perform PCA in an incremental way [1]. As can be proved, when having a set of eigenvectors from a set of views, when a new image is added to the database, these eigenvectors and the projection of the existing images can be updated following the next four-step algorithm [1]:

1. First, the mean must be updated with the expression:

$$\vec{m}' = \frac{1}{N+1}\left(N \cdot \vec{m} + \vec{x}_{N+1}\right). \tag{3}$$

2. Now, the set of eigenvectors must be updated so that they include the information of the new image $\vec{x}_{N+1}$. To do it, we compute the residual vector, that is the difference between the reconstruction and the original N+1 image $\vec{h}_{N+1} = (V \cdot \vec{p}_{N+1} + \vec{m}) - \vec{x}_{N+1}$. This vector is orthogonal to the old eigenvectors. Then, it must be normalized (so that it becomes a unit vector), obtaining $\hat{\vec{h}}_{N+1}$.

3. The new matrix of eigenvectors V' can be obtained by appending $\hat{\vec{h}}_{N+1}$ to V and rotating them, according to the next expression:

$$V' = \left[V \cdot \hat{\vec{h}}_{n+1}\right] \cdot R. \tag{4}$$

being R the solution to the eigenproblem $D \cdot R = R \cdot \Lambda'$ and D is:

$$D = \frac{N}{N+1} \cdot \begin{bmatrix} \Lambda & \vec{0} \\ \vec{0}^T & 0 \end{bmatrix} + \frac{N}{(N+1)^2} \cdot \begin{bmatrix} \vec{p} \cdot \vec{p}^T & \delta \cdot \vec{p} \\ \delta \cdot \vec{p}^T & \delta^2 \end{bmatrix}. \tag{5}$$

where $\delta = \vec{h}_{N+1} \cdot (\vec{x}_{N+1} - \vec{m})$, $\vec{p} = V^T \cdot (\vec{x}_{N+1} - \vec{m})$ and Λ is a diagonal matrix containing the original eigenvalues. This way, if $V \in \Re^{MxK}$, then $V' \in \Re^{Mx(K+1)}$. It must be studied whether this new dimension is significant or not.

4. The image representations can be updated with the next expression:

$$\vec{p}_{i(N+1)} = (R')^T \cdot \begin{bmatrix} \vec{p}_{i(N)} \\ 0 \end{bmatrix} + \left[V \cdot \vec{h}_{N+1}\right]^T \cdot (\vec{m} - \vec{m}'). \tag{6}$$

Fig. 1(b) shows the input data and the results of this approach in each iteration, comparing to the classical PCA (Fig. 1(a)). When a new image arrives, the previous projections in the database and the eigenvectors are updated and the new projection is added.

3 Robot Navigation Using Incremental PCA

Auto-location: Once the robot has captured an image, using this information it must know which of the stored points is the closest one. A projection of the current image on the current eigenspace calculated by the leader allows determining it. This returns a *K*-components vector that contains the main information of the view. Then, this vector has to be compared with those stored in the database. The one that offers the minimum distance is the matching one, which corresponds to the current position of the robot. The criterion used here is the Euclidean distance. To normalize the values of the distances, a degree of similarity between projections is computed with Eq. 7. The image of the database that offers the maximal value of γ is the matching image. d_{max} and d_{min} depend on the number of eigenvectors chosen.

$$\gamma_j = \frac{(d_{\max}/d_j)-1}{(d_{\max}/d_{\min})-1}; \quad \gamma_j \in [0,1]; \quad j=1\ldots N. \tag{7}$$

Control: From each image stored in the database, *j*, a set of *N'* sub-windows is obtained from the whole image to form the data matrix $W_j \in \Re^{M'xN'}$ (Fig. 3):

$$W_j = \left[\tilde{w}_j^{\,1} \ldots | \, \tilde{w}_j^{\,i} \, | \ldots | \, \tilde{w}_j^{\,N'}\right].$$
$$\tilde{w}_j^{\,i} = \vec{w}_j^{\,i} - \vec{m}_j; \quad \vec{m}_j = \frac{1}{N'}\sum_{i=1}^{N'} \vec{w}_j^{\,i}. \tag{8}$$

where $\vec{w}_j^i$ is each sub-window and $\vec{m}_j$ is the mean of the set *j*. The sub-windows are obtained scanning the original scene with a step in the horizontal axis. In order to compress the sub-windows information the original data are projected:

$$F_j = U_j^T \cdot W_j. \tag{9}$$

where $U_j \in \Re^{M'xK'}$ is the matrix containing the *K'* principal eigenvectors and $F_j \in \Re^{K'xN'}$ the reduced data, where $K' \leq N'$.

During the autonomous navigation, a central sub-window is taken on the current view and tracked over the central band of the matching image. To do this, once the robot knows its location, the central sub-window of the current view is projected using U_j. This operation returns a *K'*-components vector. Then, the distance of this vector to the projections f_j^i is calculated. Knowing the most similar projections, the steering velocity can be inferred. To do it, fuzzy logic has been used, as shown on Fig. 3. The inputs to the controller are the degrees of similarity $\gamma_i^j; i=1\ldots N'$ of the central sub-window of the current image with each sub-window of the corresponding image in the database, (Eq. 7) and the global degree of similarity γ_j. The outputs are the linear and steering speeds of the robot. When starting the navigation, as the robot may be far from the route, small movements around the initial point are made so that the robot can infer its location and the optimal steering at the start to tend to the route.

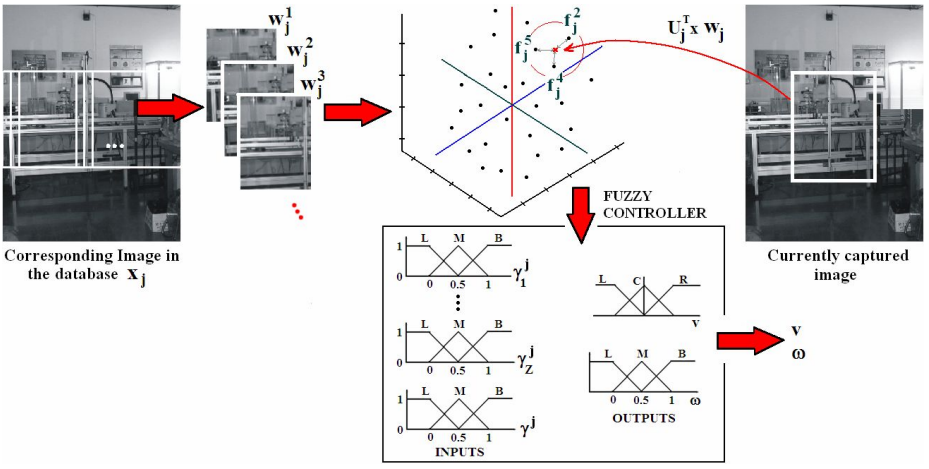


Fig. 3. Calculation of the linear and steering speeds of the robot. A central sub-window taken on the current view is compared with the sub-windows on a central horizontal band of the corresponding image, using classical PCA.

4 Results

Two Pioneer 3-AT robots with a camera on each of them, and with processors P_A and P_B on-board, have been used to test the algorithms. R_A is the leader robot, which is recording the route and R_B is the follower one. Also, an independent processor P_C has been used. All the computers are connected in a CORBA-based architecture.

To begin, the robot A is teleoperated through the desired route. After a while, R_B begins following this route. The tasks each processor must carry out are:

Processor P_A:

- Read a new image $\vec{x}_{N+1}$, compare with the previous one $\vec{x}_N$ and store it if correlation is under a threshold
- Update eigenspace V'. Calculate projection of current image $\vec{p}_{N+1}$. Delete all the previous data.
- Create sub-windows eigenspace U_j and projections f_j^i.
- Send V', $\vec{p}_{N+1}$, U_j and f_j^i to P_C.

Processor P_C:

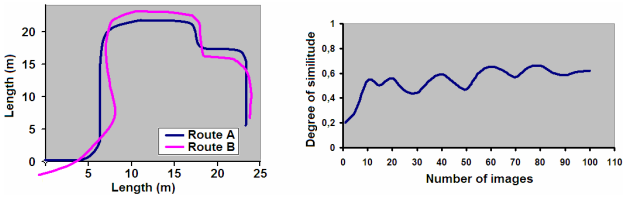
- Recalculate projections of the images $\vec{p}_{i(N+1)}$.
- Calculate the error of reconstruction using K and K+1 eigenvectors. If it is under a threshold, reduce the dimension of the eigenspace from K+1 back to K.
- Send P_B the eigenspaces V', U_j and the updated projections $\vec{p}_{i(N+1)}$ and f_j^i.

Processor P_B:

- Capture image and project in V'.

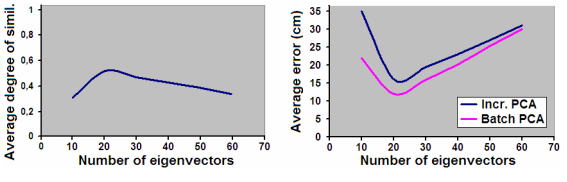
- Calculate degree of similitude with all projections in the database $\vec{p}_{i(N+1)}$. Choose the one with maximum γ.
- Study the degrees of similitude between the sub-windows f_j^i and the centered sub-window on the corresponding image.
- Use the fuzzy controller to calculate the control action to apply to R_B.

With the experiments, the performance of the incremental PCA method is tested in robot navigation and also, this method is compared with the batch PCA method. To carry out the experiments, the robot R_A has gone through a route of about 75 m in an office building including corridors and halls. R_B follows the route that R_A is recording using the incremental PCA method exposed. We have taken as a measure of the performance of the PCA method, the degree of similitude of the projected current image with the corresponding one in the database (Eq. 7). Fig. 4(a) shows the recorded route and the route R_B has followed, using K=20 eigenvectors.



(a) (b)

Fig. 4. (a) Learned route and navigated route with K=20 eigenvectors using incremental PCA. (b) Evolution of γ during the navigation.



(a) (b)

Fig. 5. (a) Average γ during navigation for different number of eigenvectors using incremental PCA. (b) Average quadratic error between routes, comparing incremental PCA and batch PCA.

In Fig. 4(b), the evolution of the degree of correlation during the navigation is shown. It reaches several maximums when R_B crosses the learned route. First, the similitude is low because R_B is initially far from the route. With this approach, R_B is able to find the route and follow it. After several experiments with different K and also with batch PCA, the results of average γ during navigation and quadratic error between the routes are shown on Fig. 5. These charts show how the best results are obtained for a medium number of eigenvectors. If K is too low, the information on the projections is too poor to have a good navigation and if K is too high, the computational cost is also

high, so the control action is refreshed with less frequency. Also, Fig. 5(b) shows that batch PCA is slightly better when calculating overall error, but using incremental PCA the task is carried out with no problem and with the advantages it supposes.

5 Conclusions

In this paper an appearance-based following route scheme is presented, using low resolution images and PCA techniques to extract the most relevant information. To allow a new robot can follow the route another robot is carrying out at the same time, an incremental PCA algorithm is employed. Based on this incremental learning stage, some robots can follow the route of the first one with a certain distance (as in space or in time) localize themselves with respect to the route made by the first robot.

A fuzzy control algorithm has been employed to determine the velocities (steering and linear) of the robots to follow the route. Some experiments have been carried out with two Pioneer 3-AT Robots. These experiments show how the process employed allows following a route. The results show that the differences between using a PCA method (with a finished route) and a PCA incremental method (with routes not finished) are not very significant, although the scene suffers small changes (illumination, position of some objects or partial occlusions in one of the cameras).

Acknowledgments. This work has been supported by MEC through project DPI2004-07433-C02-01.

References

1. Artac, M., Jogan, M., Leonardis, A.: Mobile robot localization using an incremental eigenspace model. In: Proc. of the IEEE Int. Conf. on Robotics & Automation, pp. 1025–1030. IEEE Computer Society Press, Los Alamitos (2002)
2. Balch, T., Arkin, R.: Behavior-based formation control for multi-robot teams. IEEE Transactions on Robotics and Automation 14(6), 926–938 (1998)
3. Fenwick, J., Newman, P., Leonard, J.: Cooperative Concurrent Mapping and Localization. In: Proc. IEEE Int. Conf. on Robotics and Automation, pp. 1810–1817. IEEE Computer Society Press, Los Alamitos (2002)
4. Ho, K., Newman, P.: Multiple Map Intersection Detection using Visual Appearance. In: Proc. 3rd Int. Conf. on Computational Intelligence, Robotics and Autonomous Systems (2005)
5. Kröse, B., Bunschoten, R., Hagen, S., Terwijn, B., Blassis, N.: Household robots: Look and learn. IEEE Robotics & Automation Magazine 11(4), 45–52 (2004)
6. Maeda, S., Kuno, Y., Shirai, Y.: Active navigation vision based on eigenspace analysis. In: Proc. IEEE Int. Conf. on Intelligent Robots and Systems, vol. 2, pp. 1018–1023. IEEE Computer Society Press, Los Alamitos (1997)
7. Matsumoto, Y., Ikeda, K., Inaba, M., Inoue, H.: Visual navigation using omnidirectional view sequence. In: Proc. of the IEEE Conf. on Intelligent Robots and Systems, pp. 317–322. IEEE Computer Society Press, Los Alamitos (1999)
8. Nayar, S.K., Nene, N.A., Murase, H.: Subspace methods for robot vision. IEEE Transactions on Robotics and Automation 12(5), 750–758 (1996)

9. Newman, P., Cole, D., Ho, K.: Outdoor SLAM using Visual appearance and laser ranging. In: Int. Conf. on Robotics and Automation, pp. 1180–1187 (2006)
10. Payá, L., Vicente, M.A., Navarro, L., Reinoso, O., Fernández, C., Gil, A.: Continuous navigation of a mobile robot with an appearance-based method. In: Proc. of the second Int. Conf. on Informatics in Control, Automation and Robotics, vol. 3, pp. 443–446 (2005)
11. Thrun, S.: A probabilistic online mapping algorithm for teams of mobile robots. Int. Journal of Robotics Research 20(5), 335–363 (2001)
12. Turk, M., Pentland, A.: Eigenfaces for recognition. Journal on Cognitive Neuroscience 3(1), 71–86 (1991)
13. Zhou, C., Wei, T., Tan, T.: Mobile robot self-localization based on global visual appearance features. In: Proc. of the 2003 IEEE Int. Conf. on Robotics & Automation, pp. 1271–1276. IEEE Computer Society Press, Los Alamitos (2003)

Contribution to Legged Robot Visual Servoing

Zelmar Echegoyen, Alicia d'Anjou, and Manuel Graña*

Computational Intelligence Group, Dept. CCIA
Paseo Manuel de Lardiazábal, 1 20018 San Sebastian - Spain

Abstract. This article presents a contribution to the visual tracking of objects using all of the degrees of freedom of an Aibo ERS-7 robot. Our approach follows the principled methodology of visual servoing system design. State of the art visual tracking solutions for this kind of robots inspired in the visual servoing approach either are restricted to control of the head effectors or use a learned kinematics matrix. In our work we take into account all the effectors which can affect the robot camera extrinsic parameters. We construct the matrix that specifies the robot kinematics from the robot mechanical description. Visual servoing is performed computing the pseudoinverse of this matrix.

Keywords: Legged robots, visual servoing.

1 Introduction

Visual servoing [1] is a technique for robot control which uses as a feedback signal the information extracted from the image sequences taken by one or several video cameras. In fact, it is defined as the control of the end-effector pose relative to a target object or set of features, for robotic manipulators, or the pose of the robot relative to some landmarks, in the case of mobile robotics. A major classification of visual servoing systems distinguishes position-based control from image-based control. In position- based control, the features extracted from the image are used to fit a geometric model of the target and the known camera model which gives an estimate of the robot pose relative to the target. The control rule is derived by trying to minimize the estimation error in pose space. In image-based servoing, control parameter values are computed on the basis of image features directly. We have chosen the image-based approach because it (1) reduces the computational delay, (2) avoids the need for image interpretation and (3) removes errors due to sensor modeling and camera calibration. However the image-based approach imposes a linear approximation to obtain the control parameters. This is a significant simplification of the nonlinear and highly coupled robotic system whose effects must be evaluated by physical experimentation. In the RoboCup robot soccer matches some visual servoing approaches [2,3] have been implemented in the Aibo robot to track the ball. However, these approaches are limited to the movement of the head effectors in order to keep

* The MEC partially supports this work through grant DPI2006-15346-C03-03.

B. Apolloni et al. (Eds.): KES 2007/ WIRN 2007, Part II, LNAI 4693, pp. 1179–1186, 2007.
© Springer-Verlag Berlin Heidelberg 2007

the ball inside the video image. The space in which the ball can be followed is restricted by the robot body pose.

In this paper we address the problem of maintaining the toy ball in the center of the robot camera image. The visual feature considered is the center of the ball region in the image identified by the color detection routines implemented in the robot. We have profited from the CMU's SDK [4] and the SONY's SDK [5]. The image error is the distance in the image plane to the image center of the ball blob centroid. The image features considered are very naive when compared with recent works in other domains (i.e.: [6,7]), however they are the current state of the art in the Aibo environment.

In section 2 we introduce the direct linear kinematics of the robot. In section 3 then we compute the inverse kinematics and we end up with some discussion of the physical implementation, the observed robot behavior and future work lines.

2 Direct Aibo Kinematics

We build the Aibo kinematic matrix as a transformation from the ground supporting plane to the head coordinate system, composing the transformations that correspond to the limbs and head degrees of freedom. The robot's feet support points are illustrated in figure 1. It must be noted the knees can be also support points. We need to be able to determine their 3D coordinates at any time. The position of these points is computed from the robot joint angles in a coordinate system centered in the middle of the robot body.

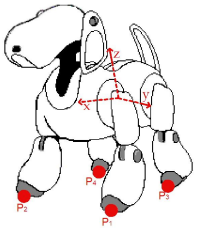


Fig. 1. Points of contact with the supporting surface

Each leg has three articulations, as shown in figure 3. The support points are the points of the robot limbs that determine the plane where it is standing on. These points must be expressed in the dog body coordinate system, because the control commands assume this frame of reference. From the point of view of this coordinate system the supporting plane apparently varies when the robot servos are affected. The first problem that we must face when building the direct kinematic transformation is the detection of the ground plane. This is done via testing all the possible planes generated as all the possible combinations of the the contact points taken in groups of three. The ground plane is characterized by all the contact points being in or above it. Assuming that the robot is standing adds a restriction illustrated by figure 2.

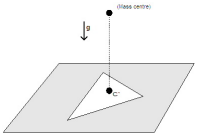


Fig. 2. Condition for standing pose based on the supporting points on the ground plane

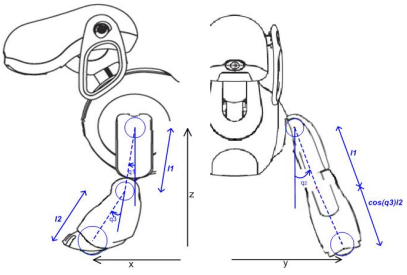


Fig. 3. Geometry of the leg articulations

Let us define the following transformations: T_1: Translation along the z-axis of length $-l_1$. R_1: Clockwise rotation about y-axis by angle q_1. R_2: Counterclockwise rotation about x-axis by angle q_2. R_1: Clockwise rotation about y-axis by angle q_3. T_2: Translation along the z-axis with length $-l_2$. T_l: Translation along the x-axis with length $\frac{1}{2}l$, being l the robot length. T_a: Translation along the y-axis with length $\frac{1}{2}a$, being a the robot width. In homogeneous coordinates the transformation from the body center to the robot front left foot coordinate system can be described as the product of transformation matrices (the transformation to the other feet follow by analogy):

$$\left(X_p, Y_p, Z_p, 1\right)^t = (R_1.R_2.T_1.R_3.T_2).\left(0, 0, 0, 1\right)^t$$

The Aibo ERS-7 has three degrees of freedom in the head. That introduces ambiguity in the control trajectories needed to track the ball trajectory. Figure 4 shows the two tilt degrees of freedom of the Aibo, denoted θ_{big} and θ_{small}. The first head tilt degree of freedom corresponds to the neck base pivoting along part of the dog chest, while the second one allows the head to move vertically using as the rotation center the joint between the neck and the head. The third degree of freedom, denoted θ_{pan}, allows a perpendicular rotation to the previous one, moving the head from side to side.

2.1 Image Features

The stated goal is to bring the ball center to the image center, so the target features are the image center coordinates and the observed features are the image coordinates of the ball region center and its diameter. These features

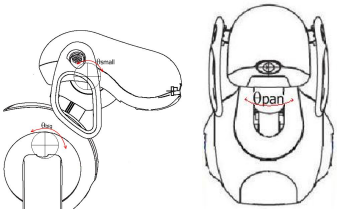


Fig. 4. Geometry of the head articulations

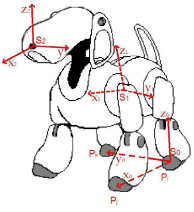


Fig. 5. Reference systems involved in the visual servoing

must be expressed in terms of the robot degrees of freedom. The Jacobian matrix gives the linear approximation to this functional dependence on each articulation positions changes. The coordinate reference systems of the robot are illustrated in figure 5. The Jacobian must be expressed in a single reference frame, therefore the need to compute the transformations between these reference systems. The transformation between the ground plane system S_0 and the body system S_1 is given by ${}_{S_1}I_{S_0} = T \cdot R$. Where T and R are a translation and rotation matrix, resp. The transformation between S_1 and S_2 is given by ${}_{S_2}I_{S_2} = T_2 \cdot R_2 \cdot R_1 \cdot T_1$. It is obtained composing the transformations that go from the body system S_1, to the head system S_2.

2.2 Observed Image Features

The camera reference system is fixed to the robot head. The observed image features, $c = (u, v)^T$, are determined by the ball position in the camera system, according to the projective transformation:

$$\begin{pmatrix} u \\ v \end{pmatrix} = f(b_2) = \frac{\lambda}{x_{b_2}} \begin{pmatrix} y_{b_2} \\ z_{b_2} \end{pmatrix}. \tag{1}$$

The observed image features are expressed in terms of b_2, the ball position in the system S_2, but we can compute them in terms of b_0, the ball position in S_0, and the transformation between S_0 and S_2.

$$\begin{pmatrix} u \\ v \end{pmatrix} = f({}_{S_2}I_{S_0}.b_0) \tag{2}$$

2.3 Feature Jacobian Matrix

To construct the Jacobian matrix that relates the variations of the diverse degrees of freedom of the robot with the variations in the image plane, we start with the following relation:

$$\begin{pmatrix} \delta u \\ \delta v \end{pmatrix} = \begin{pmatrix} \frac{-\lambda . y_p}{x_p^2} & \frac{\lambda}{x_p} & 0 & 0 \\ \frac{-\lambda . z_p}{x_p^2} & 0 & \frac{\lambda}{x_p} & 0 \end{pmatrix} \begin{pmatrix} \delta x_{b_2} \\ \delta y_{b_2} \\ \delta z_{b_2} \\ 0 \end{pmatrix} \tag{3}$$

We denote J_{cb} the Jacobian matrix of the equation (3). Then J_{cb} defines a linear transformation from variations of the positions of the ball in S_2 into variations of the image features.

$$\Delta c \simeq J_{cb} \cdot \Delta b \tag{4}$$

By deriving equation (2) we obtain the Jacobian matrix that relates the variations in the image ball position with the variations in the support points positions and in the head articulations

$$J_{br} = \frac{\delta({}_{S_2}I_{S_1} \circ_{S_1} I_{S_0})}{\delta r} b_0 \tag{5}$$

The dependence between the variations in the ball position and the variations in the head degree of freedoms and in the legs positions can be summarized by:

$$\Delta b \ \simeq \ J_{br} \cdot \Delta r \tag{6}$$

Dependence on the support points. First we observe that according to which part of the leg is in contact with the ground there are two possible Jacobian matrices, one for the foot tip (J_{p_i}) an another for the knee (J_{r_i}). We model the dependence of the foot and the knees coordinates on the articulation degrees of freedom as follows:

$$\Delta p_i \simeq Jp_i \cdot \Delta J_i \tag{7}$$

$$\Delta r_i \simeq Jr_i \cdot \Delta J_i \tag{8}$$

Let us denote ∂J_{3i}, ∂J_{1i} and ∂J_{2i} the instantaneous variations of the respective articulation in leg i. Composing with the support point Jacobians, we obtain the following jacobian matrix:

$$\begin{pmatrix} \partial P_1 \\ \partial P_2 \\ \partial P_3 \\ \partial P_4 \end{pmatrix} = \begin{pmatrix} M_1 & 0 & 0 & 0 \\ 0 & M_2 & 0 & 0 \\ 0 & 0 & M_3 & 0 \\ 0 & 0 & 0 & M_4 \end{pmatrix} \begin{pmatrix} \partial J_1 \\ \partial J_2 \\ \partial J_3 \\ \partial J_4 \end{pmatrix} \tag{9}$$

We denote this Jacobian matrix $J_{p\theta}$, where each M_i is: Jp_i, if the support point for the leg i is the foot. Jr_i, if the support point for the leg i is the knee. $Zero$ (the matrix with all the elements equal 0) if this leg has not a resting point on the ground plane. The dependence of the support ground points on the limb's degrees of freedom is summarized as follows:

$$\Delta P \ \simeq \ J_{p\theta} . \Delta J \tag{10}$$

Dependence on the robot articulations. The following matrix relates the variations in the generic support points, δr, with the variations in the true leg support points and the free leg support point, δp. We call δr_h and δp_h the head articulation variations in order to define r and p.

$$\begin{pmatrix} \delta p_h \\ \delta p_1 \\ \delta p_2 \\ \delta p_3 \\ \delta p_4 \end{pmatrix} = \begin{pmatrix} Id & 0 & 0 & 0 \\ 0 & M_{1i} & M_{1j} & M_{1k} \\ 0 & M_{2i} & M_{2j} & M_{2k} \\ 0 & M_{3i} & M_{3j} & M_{3k} \\ 0 & M_{4i} & M_{4j} & M_{4k} \end{pmatrix} \begin{pmatrix} \delta r_h \\ \delta r_i \\ \delta r_j \\ \delta r_k \end{pmatrix} \tag{11}$$

We call $J_{pr} = \begin{pmatrix} 0 & \alpha_i & \alpha_j & \alpha_k \end{pmatrix}$. This coefficient matrix defines the relation between the free leg motion and the motion of the remaining legs. The dependence between the variations in the legs articulations with the supporting points positions and the head articulations variations is summarized in the following equation

$$\Delta p \simeq J_{pr} \cdot \Delta r \tag{12}$$

Defining the Jacobian matrix $J_{r\theta}$, as the product $J_{pr}^{+} \circ J_{p\theta}$, we get the following dependence relation between the variations of support points positions and robot articulations,

$$\Delta r \simeq J_{r\theta} \cdot \Delta\theta \tag{13}$$

Full Jacobian matrix. Finally, to obtain the full Jacobian matrix $J_{c\theta}$ that models the dependence of the image features on the diverse degrees of freedom of the robot we must compose the transformations given in equations (4), (6), (12) and (10).

$$\Delta c = [(J_{cb} \circ J_{br}) \circ J_{r\theta}].\Delta\theta \tag{14}$$

3 Inverse Kinematics

To determine the desired velocity at each robot degree of freedom we should obtain the inverse of the $J_{c\theta}$ matrix in equation (14). However, this is not possible because the matrix is not invertible. The general solution is to use the seudoinverse of $J_{c\theta}^{+}$,

$$\dot{\theta} = J_{c\theta}^{+}\dot{c} + (I - J_{c\theta}^{+}J_{c\theta})n \tag{15}$$

In general, $(I - J_{xj}^{+}J_{xj})n \neq 0$, and all the vectors of the form $(I - J_{xj}^{+}J_{xj})n$ belong to the kernel of the transformation associated to $J_{c\theta}$. This solution minimizes the norm $\left\| \dot{x} - (J_{c\theta})\dot{\theta} \right\|$. As our objective is to center the ball in the image, we will not get into more details about the movements which minimize de error, so our solution will be

$$\dot{\theta} = J_{c\theta}^{+}\dot{c} \tag{16}$$

But this solution does not take into account the restriction of keeping the distances between support points constant. So, we need to determine how

these variations in the supporting points positions affect the distances between them.

$$l = \left(\|l_1\|, \|l_2\|, \|l_3\|\right)^t = \left(\|r_i - r_j\|, \|r_j - r_k\|, \|r_k - r_i\|\right)^t \tag{17}$$

Differencing l we get the Jacobian matrix J_{lr} that relates these changes.Finally this dependence is resumed in the following equation:

$$\Delta l \simeq J_{lr} \cdot \Delta r \tag{18}$$

Now we use the Jacobian matrix J_{rc} with J_{lr} and $J_{r\theta}$ to get the variations on the robot articulations that make the image features converge to the desired features, keeping constants the distances between the supporting points.

To mantain the distances constants, the vector Δr must belong to the kernel of the transformation associated to J_{lr}. The following equation ensures that Δr belongs to the kernel of J_{lr}.

$$\Delta r = [(I - J_{lr}^{+} J_{lr})\{(I - J_{lr}^{+} J_{lr}) J_{rc}^{+}\}^{+}]\Delta c \tag{19}$$

As our final objective is to get the articulations variations we add the seudoinverse of $J_{r\theta}$, also we add a velocity constant to control the advance velocity of the robot

$$\Delta\theta = J_{r\theta}^{+}(I - J_{lr}^{+} J_{lr})\{(I - J_{lr}^{+} J_{lr}) J_{rc}^{+}\}^{+}\{k_i \Delta c\}. \tag{20}$$

This equation allows us to determine the variations on the robot degrees of freedom to get the desired configuration of the image.

However, this equation is unrestricted and may drive the robot into unstable configurations, that is, to articulation configurations out of the region of stable poses in configuration space. Stable poses are characterized by the existence of a triplet of ground support points which fulfill the condition illustrated in figure 2. When this does not happen, or the projection point is too close to the triangle boundary, we restrict the visual servoing to the head degrees of freedom, using the transformation $S_0(I)S_2$ instead of $S_1(I)S_2$, to construct a reduced Jacobian M_h that relates the image features to the head degrees of freedom. Its seudoinverse gives the control for the head degrees of freedom. This approach has already been applied in [2,3]:

$$\left(d\theta\right) = M_h^{+}.dr \tag{21}$$

4 Conclusion

We have developed the visual servoing for the whole set of degrees of freedom of the Aibo 7 following a principled approach. From the geometrical description of the robot we have constructed the full Jacobian matrix that linearizes the functional dependence of the image plane viewed by the robot camera on the robot degrees of freedom. The seudoinverse of this Jacobian matrix provide the desired controls. The blind application of this control strategy may lead

the robot to unstable or unfeasible configurations for a standing pose. Therefore, we test the stability of the robot configuration. When it is compromised we restrict the visual servoing to the head. The implementation shows that the approach gives real time response when the seudoinverse is computed in the onboard processor of the robot.

References

1. Hutchinson, S., Hager, G.D., Corke, P.I.: A tutorial on visual servo control. IEEE Transactions on Robotics and Automation 12(5), 651–670
2. Quinlan, M., Murch, C., Moore, T., Middleton, R., Li, L., King, R., Chalup, S.: The, 2004 NUbots Team Report (2004), http://robots.newcastle.edu.au/publications/NUbotFinalReport2004.pdf
3. Röfer, Th., Burkhard, H.-D., Düffert, U., Hoffmann, J., Göhring, D., Jüngel, M., Lötzsch, M., Stryk, O.V., Brunn, R., Kallnik, M., Kunz, M., Petters, S., Risler, M., Stelzer, M., Dahm, I., Wachter, M., Engel, K., Osterhues, A., Schumann, C., Ziegler, J.: GermanTeam RoboCup 2003. Technical report (2003), http://www.robocup.de/germanteam/GT2003.pdf
4. http://www.tekkotsu.org
5. Sony Corporation: OPEN-R SDK Model Information for ERS-7 (2003)
6. Pomares, J., Torres, F.: Movement-flow-based visual servoing and force control fusion for Manipulation Tasks in unstructured environments. IEEE trans. Systems, Man and Cybernetics, Part C 35(1), 4–15
7. Garcia-Aracil, N., Malis, E., Aracil-Santonja, R., Perez-Vidal, C.: Continuous visual servoing despite the changes of visibility in image features. IEEE Trans. Robotics 21(6), 1214–1220

Incremental Evolution of Stigmergy-Based Multi Robot Controllers Through Utility Functions

P. Caamaño, J.A. Becerra, R.J. Duro, and F. Bellas

Integrated Group for Engineering Research, Universidade da Coruña, Spain
{pcsobrino, ronin, richard, fran}@udc.es

Abstract. This paper deals with the problem of jointly designing the behaviors of a group of robots so as to produce a particular desired collaborative behavior in stigmergy-based settings. In this line, the approach followed has its roots on traditional evolutionary behavior based robotics techniques, but instead of trying to evolve the whole controller for a particular complex behavior in one step, an incremental approach is used by combining the results from different evolutionary/learning processes in different settings for the construction of a complete controller architecture. The key aspects when trying to generalize the process of obtaining the controllers from single robots to multi robot systems are discussed in the framework of a set of simple collaborative tasks, focusing the discussion on the utility functions that guide the evolution.

Keywords: Multi Robot systems, Collaborative Behaviors, Evolutionary algorithms, Utility Functions, Neural Networks.

1 Introduction

One of the most relevant and complex topics when dealing with a group of robots is that of automatically obtaining controllers for all of them that, in a set of given circumstances, produce a collaborative behavior. These controllers may be handcrafted by the designer so as to obtain a particular cooperating behavior in a particular set of robots, but this approach is very brittle and limited. On the other hand, a very promising research line in the direction of obtaining robust cooperation schemes seeks the emergence of cooperation by reigning in the complexity of the interactions between simple controllers [1]. The basics for this process have been widely studied in the areas of Artificial Life [2] and swarm-based systems [3].

The main concept behind these approaches is similar to the one we are following here: cooperation among agents must emerge from their interaction by imposing conditions whereby cooperation provides a better way to fulfill the task. This may seem a trivial concept, but the high level of complexity of the search spaces in multi robot systems, due to the large number of possible interactions, imply that the emergence of stable collaboration is not trivial to obtain, especially when dealing with the process of designing emergent behaviors, as opposed to observing emergent behaviors. In the former, which is what we are interested in, the objective is a particular type of collective or cooperating behavior and, in most cases, there are lots

B. Apolloni et al. (Eds.): KES 2007/ WIRN 2007, Part II, LNAI 4693, pp. 1187–1195, 2007.
© Springer-Verlag Berlin Heidelberg 2007

of possible behaviors for the multi robot system that do not require collaboration and could result in stable states or solutions.

The approach presented here for the design of robot controllers that will display this kind of collaborative behavior, is based on an incremental and modular architecture developed within our group and widely tested in single robot tasks [4]. It is a bottom-up architecture based on the philosophy that a complex behavior can be obtained as the combination of simpler ones. In addition, it is very important to be able to reuse previous experience in the form of previously obtained behaviors that combined through selection and/or modulation may result in behaviors applicable to new circumstances. A similar procedure was presented by Stone and Veloso and applied to cooperative behaviors in [5]. They employ a *layered learning* approach to soccer playing robots. In this case, the authors used neural networks to initially learn low-level individual behaviors and then *decision trees* to learn the higher-level cooperative behaviors with successful results in a simulated environment.

The emergence of cooperation in our case results from the use of an evolutionary algorithm (EA) to obtain the modules which make up the controller. The use of evolution to obtain cooperative behaviors has been widely studied in the autonomous robotics field [6], providing very promising results but, in most cases, the processes display a strong influence of the designer in the controllers obtained due to the explicitness of the fitness function. Here we are more interested in establishing fitness values on what the task should be and not on how it should be carried out. Thus, the fitness function of the EA is a *utility function*, this is, a simple, usually energy based, quantitative measure is employed to guide evolution. Thus, the freedom of the system to provide an original solution is maximized, but at the cost of increasing the number of possible suboptimal behaviors.

A very relevant aspect that must be noted is that the kind of collaboration a multi robot system may show depends on the communication capabilities of the robots. Two main types of communications may be considered, as explained in [7]: explicit (a signal is intentionally shared between two or more robots) and implicit. This second type of communication is called *Stigmergy*, and arises in emergent systems in which the individual parts of the system end up communicating with one another through the use of features of their local environment they can modify. Some work on multi robot systems that use stigmergy with successful results in collaborative tasks has already been carried out [8][9][10]. This is the type of communication process we are addressing here in order to obtain collaboration.

2 Obtaining the Controllers

A simulation tool was developed to obtain the controllers. This tool is called *BDesigner* and permits the creation of the controllers through a graphical user interface. Details about the particular methodology to define the simulation environment and its elements can be found in [11]. What is important here is that the procedure for obtaining the controllers follows an *engineering approach* rather than a biological or observational approach where the researchers set some conditions and try to observe what comes out [12]. This is, we want to obtain systems that produce specific behaviors required for particular tasks or processes.

In this particular application, the robot controllers will consist of modules of Artificial Neural Networks (ANN). In general, these controllers have as inputs the sensorial information corresponding to each particular robot and must provide as outputs the actions that must be executed through its effectors. Obtaining complex behaviors using a single controller that manages all the sensorial inputs and provides outputs to all the effectors in a general real robot is, obviously, unapproachable. To solve, in part, this problem we have decided to define a complex behavior by dividing it into simpler behaviors and using an incremental architecture to efficiently combine them through the higher level modules modulating inputs and/or outputs of the lower level modules.

The approach followed to progressively construct complex behaviors and, what is more important, reuse behaviors that have been previously obtained, can be described as follows: first, simple low level behaviors, which take the control of effectors, are identified and obtained separately or reused from previously obtained libraries. The next step is to obtain the higher-level behavior controllers. These do not take the control of effectors but modify the way lower level behaviors act. Obviously, in many cases, the available low-level behaviors are not appropriate for the higher-level modules to be able to implement the desired global behaviors and thus it is necessary to simultaneously obtain another low-level behavior that complements those already present in the library and performs every task not covered by them. By iterating this process as required, a tree-like control architecture is produced. Higher-level modules modify the operation of lower level modules through modulation or selection by acting on their inputs or their outputs. A more formal description of the components of the architecture can be found in [4].

The parameters for the modules that make up the controllers of the robots in the form of ANNs are adjusted using an EA. This is, we must perform as many evolutionary processes as modules conform the controller, unless we resort to libraries of previously obtained behaviors, which is the most usual approach. In the present case, we have opted for a homogeneous controller structure for the set of robots that are to collaborate. All the robots have the same controller structure. Thus, each individual in the population of each evolutionary process represents a module used by all the robots in the simulation. As a general procedure, during the ”m” generations of evolution, each individual is executed in ”n” different initial situations (initial states for all the robots in the robot population) that last for ”p” time steps. This way, each individual is evaluated in n*m*p situations, providing a high level of robustness to the controllers.

3 Designing Cooperation

The system presented in the previous section has been successfully applied in settings involving a single real robot [4] and a group of boids [11], but in both cases with tasks that did not require collaboration. In the present work, our main objective has been to apply the system in order to study the key aspects of obtaining controllers for multi robot systems that display emergent collaboration.

To do this, we have designed an experiment consisting in a simulated environment that contains just two elements: a source of *food* (circle in the figures) and a *key* (square in the figures) that opens the source of food, allowing the robots to eat. These

two elements are placed in a square environment and are separated from one another (see Fig. 2 right for a representation of the environment). They are endowed with distance and angle sensors to the food and the key, and a sound sensor. In addition, they have one internal sensor corresponding to their individual hunger level (lack of energy). The effectors simply provide motion (linear and angular speed). This simulated world follows a set of simple rules:

- The food source is closed unless a robot is activating the key.
- When the key is activated, it beeps.
- A robot must have a minimum level of energy (maximum hunger level) to be able to activate the key.
- The robots spend energy as they move around (increasing their hunger level), but not if they stop.
- The robots spend energy when they are manipulating the key (increasing their hunger level).
- The robots obtain energy when they eat (decreasing their hunger level).

The objective of this example is to reach a stable state (based on robot behavior patterns) where the energy level of all the robots is maintained as high as possible (lowest possible hunger level). As commented in the introduction, the expected collaboration must be based on stigmergy because the robots cannot communicate directly. In addition, cooperation is the only way to stably maximize the energy.

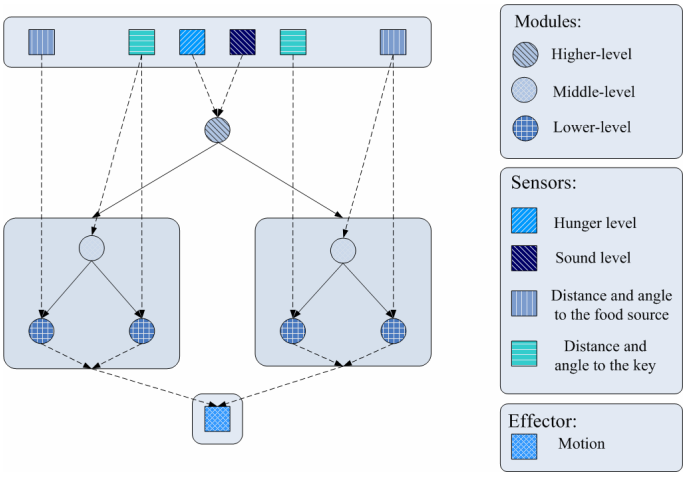


Fig. 1. Controller architecture for the collaborative example. Higher-level controllers modulate lower-level controllers.

The modules shown in Fig. 1 make up the controller of the collaborative behavior and have been obtained in 3 different evolutionary processes: one for the lower-level modules, one for the middle-level modules and one for the higher-level module. The evolution parameters as well as the structure of the ANNs are shown in Table 1. As we can see, all the processes use very similar parameters and, as commented before,

the ANNs that make up the controllers are simple. What is more interesting in the evolution is the utility function that will be discussed in depth in the next section.

Table 1. Parameters of the evolutionary processes needed to obtain the modules

Module	Lower-level	Middle-level	Higher-level
Population size	200	200	200
ANN size	2x4x2	2x4x1	2x4x1
Crossover %	5%	5%	5%
Mutation %	80%	80%	80%
Generations	1000	1000	100
Initial situations	10	10	10
Time steps in evaluation	200	200	200

The final controller, represented in Fig. 1, is made up of 4 lower-level modules, 2 middle-level modules and 1 higher-level module. This decomposition into modules has been achieved simply dividing the final behavior that each robot should show into simpler ones. This way, the robot must have two lower-level modules: one to *reach the nearest object* and another to *escape from the nearest object*. Both modules are represented by an ANN with 2 inputs (distance and angle to the nearest object) and 2 outputs (linear and angular speed). With these controllers for these simple behaviors, a middle-level module was created that, by modulating the outputs of the lower-level modules, provides a behavior that could be described as *reach the food source while escaping from the key*. This module is implemented through an ANN with 2 inputs (distance and angle to the food source) and 1 output (the modulation level over the outputs of the lower-level modules). It is important to note that this middle-level module controls the contribution of each lower-level module to the final linear and angular speed of the robot, but the modules are not modified at all. The other middle-level module shown in Fig. 1 provides a behavior that can be termed *reach the key while avoiding the food source* and is exactly the same ANN as in the other case, but exchanging the inputs corresponding to the distance and angle to the key. Finally, the higher-level module is in charge of deciding what global behavior the robot is going to perform in a given instant. It uses an ANN with 2 inputs (the sound in the environment and the hunger level) and 1 output (the modulation level over the two middle-level modules). Consequently, depending on the hunger level and on the state of the food source, the robot can decide between two behaviors: *reach the food source while escaping from the key* or *reach the key avoiding the food source.*

This example has allowed us to study the key aspects of the process for obtaining a collaborative behavior. Both the lower-level modules and the middle-level modules were obtained on single robot evolutionary processes as the behaviors of these modules were individual (reach or avoid a given object). The utility function in the case of the lower-level modules was simply the maximization (escape) or minimization (reach) of the distance to the target object at the end of the execution. In the case of the middle-level modules, the utility function was the maximization of the distance to the target (food or key) at the end of execution again, but taking into account that now collisions are penalized in the utility value. In both cases, these simple functions provided successful results in execution.

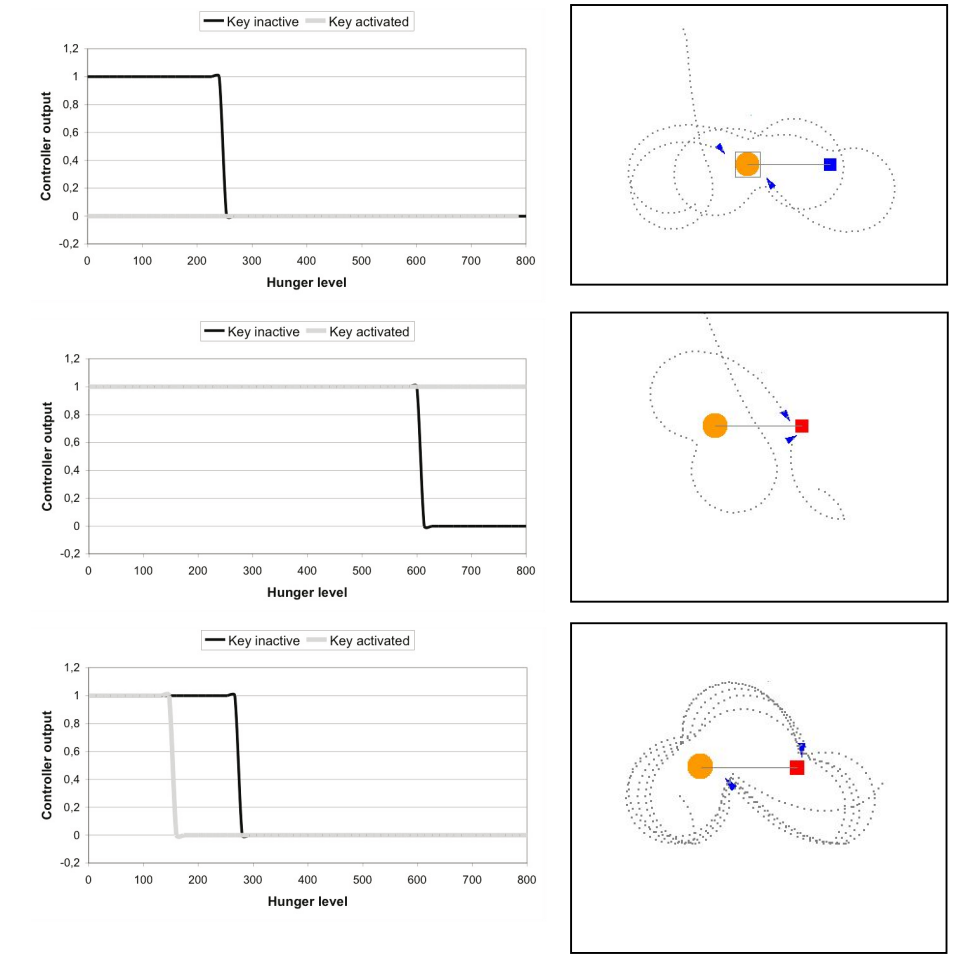


Fig. 2. Middle-level module selected by the higher-level module as a function of the hunger level when the key is activated and inactive for different energetic balances (left graphs). The right images show the final behavior obtained in execution (the robots are represented by triangles).

The utility function in the case of the higher-level module was much more complicated to establish. The first problem we met was that the utility should be local instead of global to provide more intermediate information to the final utility value, making the search process easier. But the main problem we encountered was how to establish the *energetic balance* required for obtaining the desired collaborative behavior. Fig. 2 shows the results obtained using the following energetic balances (from top to bottom in Fig. 2):

Instantaneous cost of moving	Instantaneous cost of activating key	Instantaneous food intake
1	10	20
1	1	20
1	4	20

In all cases shown in Fig. 2, the maximum hunger level is 800 units and that the threshold for activating the key is 600 units (that is, if it is hungrier than 600 units its energy is not enough to activate the key). The left graphs try to summarize the behavior of the robots depending on the hunger level and the state of the key. The lines show the response of the higher-level controller, this is, where the behavior of the robot changes from trying to *reach the key while avoiding the food source* (output to 1) to trying to *reach the food source while avoiding the key* (output to 0), in the cases when the key is inactive and activated. In the first case (Fig. 2 top), when the key is inactive (black line) the robots never try to reach the key (output to 0) and their behavior always tries to reach the food independently of the hunger level even though a robot with a low hunger level would be able activate the key. When the key is active and beeping (grey line), the behavior of the robots changes at a hunger level of 250 units from *reach the key while avoiding the food source* to *reach the food source while avoiding the key*. However, as initially the key is not active, both robots reach the food source (top right image of Fig. 2), after moving around the key without reaching it, and they just wait for someone to open it. Consequently, both robots "die" because none of them has enough energy to activate the key. The main reason behind this inappropriate solution is that the cost for activating the key is too high when compared to the cost of moving, so the robots prefer moving.

If we change the energetic balance to solve the previous problem as shown in the second case (Fig. 2 middle), we have the opposite behavior. Now the robots are mainly in the key and, at certain times, they move to the food source. This is not a bad behavior, the problem is that the hunger level used by the robots to leave the key is too close to the threshold of 600 units and often both robots "die" because they leave the key almost together and when they reach the food it is closed as shown in the right central image of Fig. 2. Finally, in Fig. 2 bottom we have represented the results for the third energetic balance. With these values, a successful collaborative behavior is obtained as expected and shown in the right bottom image. The robots alternatively go from the food source to the key. That is, one opens the key while the other eats and after a while they exchange places.

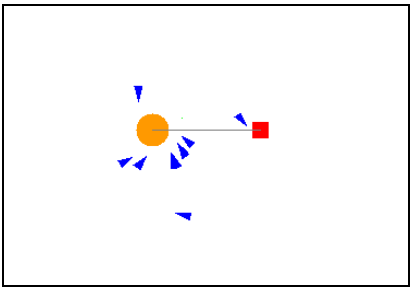


Fig. 3. Execution of the example using 10 robots (third energetic balance)

The main objective of this discussion is to highlight that the key aspect in obtaining a collaborative behavior through energy based utility functions and stigmergy is the energetic balance established in the environment. This problem did not arise with the lower and middle-level controllers because it is a feature of the collaborative tasks, and does not depend on the complexity of the individual task [4] or on the number of robots [11]. The key aspect is that, in a collaborative task, all the robots contribute to the global utility but some of them adding utility and others subtracting. As a consequence, there are a large number of possible stable states in the multi robot system corresponding to local optima for the number of iterations considered in the utility function, but not to a stable collaborative behavior.

Fig. 3 represents the execution of this example using 10 robots with the final energetic balance of the study. The obtained result is successful again and the robots cooperate in the sense that one is activating the key while the others are eating. When its hunger level increases, the one in the key is replaced by a more energetic one.

4 Conclusions

We have presented a procedure to automatically obtain the behavior controllers for collaborating robots in a multi robot system. It is based on an incremental and modular architecture that uses evolutionary processes based on utility functions to obtain the modules. The system has been tested in a stigmergy-based collaborative task where cooperation is required to achieve an energetically efficient state. From this example we have extracted the conclusion that the main challenge in this type of processes is establishing the energetic balance that makes up the utility. We are now studying these types of problems to establish a methodology to achieve this objective.

Acknowledgements. This work was partially funded by the MEC of Spain through projects DEP2006-56158-C03-02 and DPI2006-15346-C03-01.

References

1. Nitschke, G.: Emergence of Cooperation: State of the Art. Artificial Life 11(3), 367–396 (2005)
2. Langton, C.: Artificial Life: An Overview. MIT Press, Cambridge, MA (1995)
3. Sahin, E., Spears, W.: Swarm Robotics - SAB 2004 International Workshop Revised Selected Papers. LNCS, vol. 3342. Springer, Heidelberg (2005)
4. Becerra, J.A., Santos, J., Duro, R.J.: Multimodule Artificial Neural Network Architectures for Autonomous Robot Control Through Behavior Modulation. In: Mira, J.M., Álvarez, J.R. (eds.) IWANN 2003. LNCS, vol. 2687, pp. 169–176. Springer, Heidelberg (2003)
5. Stone, P., Veloso, M.: Towards Collaborative and Adversarial Learning: A Case Study in Robotic Soccer. Evolution and learning in Multiagent Systems 48(1), 83–104 (1998)
6. Quinn, M.: A Comparison of Approaches to the Evolution of Homogeneous Multi-Robot Teams. In: Proceedings of the 2001 Congress on Evolutionary Computation, pp. 128–135 (2001)
7. Arkin, R.C.: Behavior-based Robotics. MIT Press, Cambridge (1998)

8. Arkin, R.C.: Cooperation without communication: Multiagent schema-based robot navigation. Journal of Robotic Systems 9, 351–364 (1992)
9. Steele, F., Thomas, G.: Directed stigmergy-based control for multi-robot systems. In: Proceedings of the ACM/IEEE int. conf. on Human-robot interaction, pp. 223–230 (2007)
10. D'Angelo, A., Pagello, E.: Making Collective Behaviours to work through Implicit Communication. In: Proceedings of ICRA 2005, pp. 81–86 (2005)
11. Caamaño, P., Prieto, A., Becerra, J.A., Duro, R., Bellas, F.: Evolutionary Tool for the Incremental Design of Controllers for Collective Behaviors. In: Proc. IWINAC 2007. LNCS, vol. v, pp. 4527–4528. Springer, Heidelberg (2007)
12. Terzopoulos, D., Tu, X., Grzeszczuk, R.: Artificial fishes with autonomous locomotion. In: perception, behavior, and learning in a simulated physical world. Proc. Art. Life IV, pp. 17–27. MIT Press, Cambridge (1994)

Lattice Independence and Vision Based Mobile Robot Navigation

I. Villaverde, M. Graña, and J.L. Jimenez*

Dept. CCIA, UPV/EHU, Apdo. 649, 20080 San Sebastian, Spain
ccpgrrom@si.ehu.es

Abstract. Strong Lattice Independence implies Affine Independence. Affine Independent sets of vectors define a convex polytope and if this polytope is a good approximation to the convex hull of a set data points, we can use them to represent the data points through their convex coordinates. This representation can be used as a feature extraction or dimensionality reduction method. Morphological Associative Memories (MAM) have been proposed for image denoising and pattern recognition. Recent works show that, by construction, Autoassociative Morphological Memories (AMM) are composed of lattice independent vectors. After a transformation these vectors can be shown to be a good approximation to the data convex hull, and therefore as a candidate set of points for convex coordinate representation of the data. In this paper we present some results on the task of visual landmark recognition for a mobile robot self-localization task improving previous results using AMM.

Keywords: Lattice, Morphological Neural Networks, Robot Navigation, Machine Vision.

1 Introduction

Navigation is the ability of an agent to move around its environment with a specific purpose [2]. It implies some knowledge of the environment, be it topological or not. A needed ability for navigation is self-localization: the capacity of the robot to ascertain, more or less accurately, where it is from the information provided by its sensors. This knowledge makes possible other navigation related tasks like planning. The basic self-localization procedure is odometry: self-sensing and keeping track of motion commands. However the uncertainties related to the environment and the robot internal status call for sensor based external confirmation of the position internal estimation. Self-localization based on low range external sensors has been formulated in a probabilistic framework [4]. Vision based [3] and mixed [18,28] systems are proposed to increase the sensing range and self-localization robustness. Visual self-localization methods usually are based on landmark recognition [19,1,27,14,15,17,11,29,21]. For instance, the system described in [1] computes for each stored view a graph model representation of salient points in the image as measured by the information content of a neighborhood of the point in a gradient image. To recognize the view, the salient points

* The Spanish Ministerio de Educacion y Ciencia supports this work through grant DPI2006-15346-C03-03.

B. Apolloni et al. (Eds.): KES 2007/ WIRN 2007, Part II, LNAI 4693, pp. 1196–1203, 2007.
© Springer-Verlag Berlin Heidelberg 2007

in the current image are compared to the stored models. Model matching as performed in [1] is not invariant or robust against translations and rotations. Therefore each view is only recognized when the robot is within a small neighborhood of the physical position and orientation where the landmark was detected originally. Recognition in this case is not continuous, unless the stored views built up a dense map. Our approach based on Morphological Autoassociative Memories (MAAM) [24,22], falls within the class of holistic approaches. The stated goal is to recognize, with some degree of robustness, several predetermined robot placements and orientations based on the recognition of the visual information captured by the robot. We can associate to each position an area of the physical environment where the robot recognizes this position, like in [1],[15]. The robot is supposed to wander looking forward, taking images at a steady rate. Each image corresponds to a view of the world, characterized by a physical position and orientation. Images are analyzed continuously and when a scene is recognized a certain spatial position is assumed for the robot. Robustness implies that the recognition must cope with some variations in lighting and small rotations and translations of the camera due to the uncertainty of the robot position, which, in its turn, is due to the uncertainties in the motion of the robot. This paper follows the work done in [34]. The approach we follow is to characterize the data by a convex region that encloses them or most of them. The features extracted are the relative coordinates of the data points in this region, the convex coordinates. These convex coordinates are the result of the linear unmixing relative to the vertices of this convex region. Therefore the dimensionality reduction depends on the degree of detail of the definition of this convex region: the number of vertices that describe it.

The Morphological Associative Memories (MAM) [24], [22], [25] were proposed as the morphological counterpart of the well known Hopfield Associative Memories [12]. AMM's are constructed as correlation matrices computed by either Min or Max matrix product. Dual constructions can be made using the dual Min and Max operators. The AMM selective sensitivity to specific types of noise (erosive and dilative noise) is of special interest to us. It was established that AMM are able to store and recall morphologically strongly independent sets of patterns. The notion of morphological independence and morphological strong independence was introduced in [25] to study the construction of AMM robust to general noise. Following these works, studies on the fixed points of AMM had lead to the more precise definition of the notion of lattice independence[23] and strong lattice independence, including the relation between lattice independence and affine independence, which is the kind of independence needed for convex coordinate definition. A key result are the conditions for a subset of column vectors of an AMM matrix to be strongly lattice independent and, therefore, defining a convex region. A simple transformation makes this convex region an approximation to the data convex hull.

The early application focus of our approach was the unsupervised analysis of hyperspectral images [6,7] to obtain salient image regions that may deserve further analysis and search for labeled data. The approach we favored is that of linear filtering for target detection [16] and the "spectral unmixing" [13] model. In [8,9] we have done the next step assuming that the convex coordinates of the pixel spectra can be used as feature vectors for classification, with surprising good results. Here we try to extend this

approach to the visual recognition of landmark views for mobile robot self-localization. In the following sections we will review the basic results relative to autoassociative morphological memories and lattice independence, some experimental results and, finally, some conclusions and directions for further work.

2 Lattice Independence and AMM

The work on Morphological Associative Memories stems from the consideration of an algebraic lattice structure $(\mathbb{R}, \vee, \wedge, +)$ as the alternative to the algebraic framework $(\mathbb{R}, +, \cdot)$ for the definition of Neural Networks computation [24] [22]. The operators $\vee$ and $\wedge$ denote, respectively, the discrete $\max$ and $\min$ operators (resp. $\sup$ and $\inf$ in a continuous setting), which correspond to the morphological dilation and erosion operators, respectively. Given a set of input/output pairs of pattern $(X, Y) = \left\{\left(\mathbf{x}^\xi, \mathbf{y}^\xi\right); \xi = 1, .., k\right\}$, an heteroassociative neural network based on the pattern's cross correlation [12] is built up as $W = \sum_\xi \mathbf{y}^\xi \cdot \left(\mathbf{x}^\xi\right)'$. Mimicking this construction procedure [24], [22] propose the following constructions of Heteroassociative Morphological Memories (HMM's):

$$W_{XY} = \bigwedge_{\xi=1}^{k} \left[\mathbf{y}^\xi \times \left(-\mathbf{x}^\xi\right)'\right] \text{ and } M_{XY} = \bigvee_{\xi=1}^{k} \left[\mathbf{y}^\xi \times \left(-\mathbf{x}^\xi\right)'\right], \tag{1}$$

where $\times$ is any of the $\boxtimes$ or $\boxtimes$ operators. Here $\boxtimes$ and $\boxtimes$ denote the $\max$ and $\min$ matrix product, respectively defined as follows:

$$C = A \boxtimes B = [c_{ij}] \Leftrightarrow c_{ij} = \bigvee_{k=1..n} \{a_{ik} + b_{kj}\}, \tag{2}$$

$$C = A \boxtimes B = [c_{ij}] \Leftrightarrow c_{ij} = \bigwedge_{k=1..n} \{a_{ik} + b_{kj}\}. \tag{3}$$

If $X = Y$ then the HMM memories are Autoassociative Morphological Memories (AMM). Conditions of perfect recall by the HMM's and AMM's of the stored patterns proved in [24,22] encouraged the research on them, because in the continuous case, the AMM's are able to store and recall any set of patterns: $W_{XX} \boxtimes X = X = M_{XX} \boxtimes X$, for any X. This result holds when we deal with noise-free patterns. Research on robust recall [20,22,25,30] based on the so-called kernel patterns lead to the notion of morphological independence, in the erosive and dilative sense, and finally to the definition of lattice independence and strong lattice independence. We gather some results from [23] that set the theoretical background for the approach to endmember induction applied.

Definition 1. *Given a set of vectors* $\left\{\mathbf{x}^1, ..., \mathbf{x}^k\right\} \subset \mathbb{R}^n$ *a* linear minimax combination *of vectors from this set is any vector* $\mathbf{x} \in \mathbb{R}^n_{\pm\infty}$ *which is a* linear minimax sum *of these vectors:*

$$x = \mathcal{L}\left(\mathbf{x}^1, ..., \mathbf{x}^k\right) = \bigvee_{j \in J} \bigwedge_{\xi=1}^{k} \left(a_{\xi j} + \mathbf{x}^\xi\right),$$

where J *is a finite set of indices and* $a_{\xi j} \in \mathbb{R}_{\pm\infty}$ $\forall j \in J$ *and* $\forall \xi = 1, ..., k$.

Definition 2. *The* linear minimax span *of vectors* $\{\mathbf{x}^1, ..., \mathbf{x}^k\} = X \subset \mathbb{R}^n$ *is the set of all linear minimax sums of subsets of* X*, denoted* $LMS\left(\mathbf{x}^1, ..., \mathbf{x}^k\right)$.

Definition 3. *Given a set of vectors* $X = \{\mathbf{x}^1, ..., \mathbf{x}^k\} \subset \mathbb{R}^n$*, a vector* $\mathbf{x} \in \mathbb{R}^n_{\pm\infty}$ *is* lattice dependent *if and only if* $x = \mathcal{L}\left(\mathbf{x}^1, ..., \mathbf{x}^k\right)$ *for some linear minimax sum of vectors from* X*. The vector* $\mathbf{x}$ *is* lattice independent *if and only if it is not lattice dependent on* X*. The set* X *is said to be* lattice independent *if and only if* $\forall \lambda \in \{1, ..., k\}$, $\mathbf{x}^\lambda$ *is lattice independent of* $X \setminus \{\mathbf{x}^\lambda\} = \{\mathbf{x}^\xi \in X : \xi \neq \lambda\}$.

Definition 4. *A set of vectors* $X = \{\mathbf{x}^1, ..., \mathbf{x}^k\} \subset \mathbb{R}^n$ *is said to be* max dominant *if and only if for every* $\lambda \in \{1, ..., k\}$ *there exists and index* $j_\lambda \in \{1, ..., n\}$ *such that*

$$x^\lambda_{j_\lambda} - x^\lambda_i = \bigvee_{\xi=1}^{k} \left(x^\xi_{j_\lambda} - x^\xi_i\right) \forall i \in \{1, ..., n\}.$$

Similarly, X *is said to be* min dominant *if and only if for every* $\lambda \in \{1, ..., k\}$ *there exists and index* $j_\lambda \in \{1, ..., n\}$ *such that*

$$x^\lambda_{j_\lambda} - x^\lambda_i = \bigwedge_{\xi=1}^{k} \left(x^\xi_{j_\lambda} - x^\xi_i\right) \forall i \in \{1, ..., n\}.$$

Definition 5. *A set of lattice independent vectors* $\{\mathbf{x}^1, ..., \mathbf{x}^k\} \subset \mathbb{R}^n$ *is said to be* strongly lattice independent *if and only if* X *is max dominant or min dominant or both.*

Definition 6. *The fixed point set of* W_{XX} *and* M_{XX} *is the set of all fixed points of both transformations* $F(X) = \{\mathbf{x} \in \mathbb{R}^n : W_{XX} \boxtimes \mathbf{x} = \mathbf{x}\}$.

Theorem 1. *For each* $k \in \{1, ..., k\}$ *there exists a set of vectors* $X \subset \mathbb{R}^n$ *such that* $\dim F(X) = k$ *and* X *is strongly lattice independent.*

Corollary 1. *If* $X = \{\mathbf{x}^1, ..., \mathbf{x}^k\} \subset \mathbb{R}^n$ *is strongly lattice independent then* $k \leq n$.

Theorem 2. *If* $X = \{\mathbf{x}^1, ..., \mathbf{x}^k\} \subset \mathbb{R}^n$ *is strongly lattice independent then* X *is affinely independent.*

Theorem 3. *Let* $X = \{\mathbf{x}^1, ..., \mathbf{x}^k\} \subset \mathbb{R}^n$ *and let* W *be the set of vectors consisting of the columns of the matrix* W_{XX}*. If* $V \subset W$ *denotes the smallest set of vectors with the property that* $W_{VV} = W_{XX}$*, then the set* V *is strongly lattice independent.*

The key theorem 2 provides the relationship between the lattice independence and the affine independence that characterizes the sets of vertices of a convex region, the one we need to compute the convex coordinate features of the data. The theorem 3 provides the key for the induction of the endmembers from the data set. A similar results seems to hold for the rows of W. However, there are several unsolved problems:

– The number of strongly lattice (and affinely) independent vectors is quite high, close to the space dimensionality. One of the goals of the approach we follow is to obtain some dimensionality reduction. We have devised below an approach to select a smaller set of endmembers, but we do not guarantee the optimality of this selection.

- We assume that a subset of a set strongly lattice independent is also strongly lattice independent, butwe have not proved it yet.
- The convex polytope defined by set V of strongly lattice independent vectors is not related to the data points. We need it to be covering the data point set as much as possible, approaching the data convex hull, in order to have meaningful convex coordinates.

The steps we follow in our algorithm are:

1. Compute the min auto-associative memory W_{XX} from the data
2. Compute the conjugate of each element in the matrix, obtaining the conjugate matrix $-W_{XX}$.
3. We consider that the rows of the matrix $-W_{XX}$ as the strong lattice independent set. We add the minimum of each dimension in the data set to the corresponding dimensions of the strong lattice independent vectors.
4. Select the vectors with minimum mutal information to obtain a set with smaller cardinality than the data dimensionality.

3 Experimental Results

The experimental setup is as follows. First a path was defined from our laboratory to the stairs hall on 3d floor of our building. The mobile robot platform was guided manually seven times following this path. In each of those trips, the odometry was recorded and the images taken from the camera, at an average of 10 frames per second, were also recorded.

In the computational experiments that follow, the first trip was used to train the system parameters, and the six other trips were used as test sequences, simulating a real trip. The task to perform is to recognize a given set of map positions from their landmark views. The positions were selected on the floor plane, selecting places of practical relevancy, like doors to other laboratories, and the corresponding landmark views were selected from the first trip image sequence based on odometry readings.

Classes are composed for each of the selected landmark position, assigning the images in the sequences to the closest map position according to its odometry reading. Therefore the task becomes the classification of the images into one of the map classes, where each class is composed of images taken in robot path positions before and after the map position for which it is the closest map position. The clasification was done using k-NN. Each of the images of the training trip was assigned to a class, using them as a cluster of images representing each selected position. For the test trips, each image was classified on those classes using 3-NN.

From the image sequence of the first trip a PCA transformation consisting of 230 eigenvectors was computed. All the following computations were done on the PCA coefficients of the images. Table 1 contains the landmark recognition for the consecutive walks over the path, varying the number of PCA selected. The degradation of the results as the time between the training walk and the test walk increase. Table 2 shows the results of the convex coordinate approach with varying number of endmembers selected applying the algorithm described at the end of the previous section. The best results are

obtained when selecting 10 and 15 extrema, which give results comparable to the PCA. Contrary to the results in [34] the algorithm tested in this paper does not have any random component, so there is no need to repeat the computation. Also the number of extrema is a parameter of the algorithm.

Table 1. Landmark recognition success rate based on the PCA representation of the navigation images for several sets of eigenvectors selected, using 3-nn

Set	Train	Pass 1	Pass 2	Pass 3	Pass 4	Pass 5	Av.
PCA 197	0,95	0,84	0,76	0,65	0,79	0,77	0,82
PCA 150	0,95	0,84	0,76	0,65	0,78	0,76	0,81
PCA 100	0,95	0,85	0,76	0,65	0,79	0,76	0,82
PCA 50	0,95	0,85	0,76	0,65	0,78	0,78	0,81
PCA 30	0,96	0,87	0,77	0,64	0,78	0,78	0,82
PCA 10	0,96	0,86	0,78	0,66	0,76	0,73	0,81
Av.	0,96	0,85	0,76	0,65	0,78	0,76	0,82

Table 2. Landmark recognition success rate based on the convex coordinates representation of the navigation images for several numbers of endmembers extracted and using 3-nn

	#extrema	Train	Pass 1	Pass 2	Pass 3	Pass 4	Pass 5	Av.
	5	0,96	0,79	0,74	0,64	0,71	0,61	0,80
	10	0,96	0,80	0,76	0,61	0,80	0,72	0,81
	15	0,96	0,80	0,74	0,66	0,79	0,69	0,82
	20	0,96	0,80	0,76	0,65	0,81	0,67	0,79
	25	0,96	0,78	0,72	0,62	0,74	0,68	0,78
	30	0,96	0,81	0,73	0,60	0,75	0,69	0,77
Av.		0,96	0,79	0,74	0,64	0,78	0,70	

Although these results are much less spectacular than the ones reported in [8,9], they confirm the trend that points to the usefulness of convex coordinates as a feature selection algorithm. We are trying to recompute the results using the original images as inputs to the extrema selection algorithm, in the hope of improving the results of our algorithm relative to the PCA. When applied to the PCA coefficient vectors, our algorithm may be greatly restricted by the data transformation already performed.

4 Conclusions and Further Work

We claim that we can use the convex coordinates of the data points based on the vertices of a convex region covering the data as features for pattern classification. The convex region vertices are selected using the rows of the W AMM of the data. Mobile robot self-localization is stated as a classification of images taken from the camera. We use the PCA and the convex coordinates based on the MAM detected endmembers as the features for classification. We found that our approach improves the PCA approach in some runs, while on average performs similar to most PCA eigenvector selections tried.

References

1. Balkenius, C., Kopp, L.: Robust Self-Localization Using Elastic Templates. In: Lindberg, T. (ed.) Proceedings of Swedish Symposium on Image Analysis (1997)
2. Chatila, R.: Deliberation and Reactivity in Autonomous Mobile Robots. Robotics and Autonomous Systems 16, 197–211 (1995)
3. DeSouza, G.N., Kak, A.C.: Vision for Mobile Robot Navigation: A Survey. IEEE Transactions on Pattern Analysis and Machine Intelligence 24(2), 237–267 (2002)
4. Fox, D.: Markov Localization: A Probabilistic Framework for Mobile Robot Localization and Navigdation, Ph. D. Thesis, University of Bonn, Germany (December 1998)
5. Fukunaga, K.: Introduction to statistical pattern recognition. Academic Press, Boston, MA (1990)
6. Graña, M., Gallego, J.: Associative Mophological Memories for endmember induction. In: Proc. IGARSS'2003, Tolouse, France (2003)
7. Graña, M., Sussner, P., Ritter, G.: Associative Morphological Memories for Endmember Determination in Spectral Unmixing. In: Proc. FUZZ-IEEE'03 (2003)
8. Graña, M., d'Anjou, A.: Feature Extraction by Linear Spectral Unmixing. In: Negoita, M.G., Howlett, R.J., Jain, L.C. (eds.) KES 2004. LNCS (LNAI), vol. 3213, pp. 692–697. Springer, Heidelberg (2004)
9. Graña, M., d'Anjou, A., Albizuri, X.: Morphological memories for feature extraction in hyperspectral images. In: Verleysen, M. (ed.) ESANN 2005, pp. 497–502. dFacto press (2005)
10. Graña, M., Villaverde, I., Moreno, R., FX, A.: Convex Coordinates From Lattice Independent Sets for Visual Pattern Recognition. In: VG, K., GX, R. (eds.) Computational Intelligence Based on Lattice Theory, Springer, Heidelberg (2007) (in press)
11. Gross, H.M., Koening, A., Boehme, H.J., Schroeter, Ch.: Vision-based Monte Carlo Self-localization for a Mobile Service Robot Acting as Shopping Assistant in a Home Store. In: Proceedings of the IEEE Intl. Conference on Intelligent Robots and Systems, IEEE Computer Society Press, Los Alamitos (2002)
12. Hopfield, J.J.: Neural networks and physical systems with emergent collective computational abilities. Proc. Nat. Acad. Sciences 79, 2554–2558 (1982)
13. Keshava, N., Mustard, J.F.: Spectral unimixing. IEEE Signal Proc. Mag. 19(1), 44–57 (2002)
14. Livatino, S., Madsen, C.: Optimization of Robot Self-Localization Accuracy by Automatic Visual-Landmark Selection. In: SCIA. Proceedings of 11th Scabdinavian Conference on Image Analysis, pp. 501–506 (1999)
15. Livatino, S., Madsen, C.: Autonomous Robot Navigation with Automatic Learning of Visual Landmarks. In: SIRS99. International Symposium of Intelligent Robotic Systems (1999)
16. Manolakis, D., Shaw, G.: Detection algorithms for hyperspectral imaging applications. IEEE Signal Proc. Mag. 19(1), 29–43 (2002)
17. Marando, F., Piaggio, M., Scalzo, A.: Real Time Self Localization Using a Single Frontal Camera. In: SIRS01. International Symposium of Intelligent Robotic Systems (2001)
18. Ohya, A., Kosaka, A., Kak, A.C.: Vision-Based Navigation by a Mobile Robot with Obstacle Avoidance Using Single-Camera Vision and Ultrasonic Sensing. IEEE Journal of Robotics and Automation 14(6), 969–978 (1998)
19. Olson, C.F.: Landmark Selection for Terrain Matching. In: Proceedings ICRA2000 (2000)
20. Raducanu, B., Graña, M., Albizuri, X.: Morphological scale spaces and associative morphological memories: results on robustness and practical applications. J. Math. Imaging and Vision 19(2), 113–122 (2003)
21. Reuter, J.: Mobile Robot Self-Localization Using PDAB. In: ICRA. Proceedings of International Conference on Robotics and Automation (2000)

22. Ritter, G.X., Diaz-de-Leon, J.L., Sussner, P.: Morphological bidirectional associative memories. Neural Networks 12, 851–867 (1999)
23. Ritter, G.X., Gader, P.: Fixed points of lattice transforms and lattice associative memories. In: Hawkes, P. (ed.) Advances in Imaging and Electron Physics, vol. 144, pp. 165–242. Elsevier, San Diego, CA (2006)
24. Ritter, G.X., Sussner, P., Diaz-de-Leon, J.L.: Morphological associative memories. IEEE Trans. on Neural Networks 9(2), 281–292 (1998)
25. Ritter, G.X., Urcid, G., Iancu, L.: Reconstruction of patterns from moisy inputs using morphological associative memories. J. Math. Imaging and Vision 19(2), 95–112 (2003)
26. Ritter, G.X., Urcid, G.: Lattice algebra approach to single-neuron computation. IEEE Trans Neural Networks 14(2), 282–295 (2003)
27. Rizzi, A., Duina, D., Inelli, S., Cassinis, R.: Unsupervised Matching of Visual Landmarks for Robotic Homing using Fourier-Mellin Transform. Robotics and Autonomous Systems 40, 131–138 (2002)
28. Saffiotti, A., Wesley, L.P.: Perception-Based Self-Localization Using Fuzzy Location. In: Dorst, L., Voorbraak, F., van Lambalgen, M. (eds.) RUR 1995. LNCS, vol. 1093, pp. 368–385. Springer, Heidelberg (1996)
29. Sekimori, D., Usui, T., Masutani, Y., Miyazaki, F.: High-Speed Obstacle Avoidance and Self-Localization for Mobile Robots Based on Omni-Directional Imaging of Floor Region. IPSJ Transactions on Computer Vision and Image Media 42(SIG13-012) (2002)
30. Sussner, P.: Observations on Morphological Associative Memories and the Kernel Method. In: Proc. IJCNN'2001, Washington DC (July 2001)
31. Sussner, P.: Generalizing operations of binary autoassociative morphological memories using fuzzy set theory. J. Math. Imaging and Vision 19(2), 81–94 (2003)
32. Villaverde, I., Ibañez, S., Albizuri, F.X., Graña, M.: Morphological neural networks for real-time vision based self-localization. In: Abrham, A., Dote, Y., Furuhashi, T., Köpen, M., Ohuchi, A., Ohsawa, Y. (eds.) Soft Computing as transdisciplinary Science and Techonology, Proc. WSTST'05, Advances in Soft Computing, pp. 70–79. Springer, Heidelberg (2005)
33. Villaverde, I., Graña, M., D'Anjou, A.: Morphological Neural Networks for Localization and Mapping. In: CIMSA'06. Proceedings of the IEEE International Conference on Computational Intelligence for Measurement Systems and Applications, La Coruña (Spain), pp. 12–14. IEEE Computer Society Press, Los Alamitos (2006)
34. Villaverde, I., Graña, M, D'Anjou, A: Morphological Neural Networks and Vision Based Mobile Robot Navigation. In: Kollias, S., Stafylopatis, A., Duch, W., Oja, E. (eds.) ICANN 2006. LNCS, vol. 4131, pp. 878–887. Springer, Heidelberg (2006)

ROBMAT: Teleoperation of a Modular Robot for Collaborative Manipulation

Juan A. Escalera, Manuel Ferre, Rafael Aracil, and José Baca

Dpto. Automática, Ing. Electrónica, e Inf. Industrial,
Universidad Politecnica de Madrid
C/. José Gutierrez Abascal,2, Madrid - Spain,
{jescalera, mferre, aracil, jbaca}@etsii.upm.es
http://www.disam.upm.es

Abstract. The present paper describes the architecture of a modular robot called RobMAT. A module description which includes both hardware and software features are shown. The module has 3 DoF which gives great functionality to the module. The connection between modules forms molecules. The simplest molecule is called base with 5 DoF. Channels of communication between modules (CAN bus) and molecules (Bluetooth) are also described. A method in representing molecule configurations based on Graph Theory is explained. This method allows developing algorithm for dynamically changing-robots such as modular robots. It also explains how the molecule is allowed to act as a whole synchronization algorithm. Finally a collaborative manipulation task, which is teleoperated, is shown. Elements involve in teleoperation are also described.

Keywords: Modular, robot, self-configurable, teleoperation, human-interface.

1 Introduction

Research on modularity and re-configurability in robotics is currently on the rise. These features add three important qualities in the field of robotics: fault tolerance, flexibility and low cost. Modular robots can be thought as sets of simple elements that can be joined together to form a complex structure. The performance of the whole is greater than the sum of each module?s performance. A primary objective in the design of a reconfigurable modular robot is to allow the robot to assume any geometric shape. The robot as explained has been considered to build systems that are adaptable to a number of tasks and unknown environments. Such interest is focused on modular robots because they allow making a wide variety of specialized robots from a set of standard components. This aim is different from other types of shape-changing robots that are capable of forming limited number of shapes. Developing functional reconfigurable robots poses a significant challenge. Hardware must be designed and built in a manner that the robot will be capable of reconfiguration, perform autonomous operation and handle algorithms that can confer the robot the ability to change shape.

B. Apolloni et al. (Eds.): KES 2007/ WIRN 2007, Part II, LNAI 4693, pp. 1204–1213, 2007.
© Springer-Verlag Berlin Heidelberg 2007

Where would reconfigurable modular robots be used? This kind of robot is most useful in an unknown complex environment. For example, a building damaged by an earthquake contains a variety of obstructions and may not be suitable for any particular standard robot. A reconfigurable robot, with the ability of locomotion over a variety of terrains, through gaps and over obstacles, can perform well in such situation.

The adaptability of these modular robots means that they can travel through narrow passageways (in a snake-like configuration), be reshaped into a legged robot to walk over rubble and even climb stairs or on desks. Another application is space/planetary exploration, where unpredictable terrains on a planet must be explored by a robot before humans are sent [1].

The classification of modular robots is widely accepted in lattice and chain types. Lattice-based robots are characterized by a set of discrete locations that can be reached by module movements to dock with another module.([5,6,7]. This in contrast to chain-based robots which can reach all positions into the corresponding joint range. [8,1,2].

There is another classification depending on the amount of module classes. If there is just one type of module robot, it is termed homogeneous or a unit modular robot. Whereas, if there are different module classes, the robot will then be heterogeneous. Up to now, most are developed as modular unit, because of its ease for modelling. The main advantage is that their control is easy to implement. On the other hand, heterogeneous systems have different types of modules that are generally specialized in performing certain tasks more accurately, such as manipulation or movement. Such characteristics enhance robot performance but control becomes more complex.

This paper has been organized as follows. Section 2 describes RobMAT control architecture, which explains module and communication channels. It is also described molecule as union of modules. Section 3 is focused on the representation of modular robots. Molecules that are made up of identical 3 DoF modules and are analyzed by Graph Theory. Main experimental results are described in section 4. The RobMAT system has been used for the execution of tasks outside the laboratory. Advanced teleoperation tools have also been used for controlling the modular robot remotely. They are based on stereoscopic vision human interfaces that both allow a remote control of the modular robot for unstructured environments. These experiments highlight the main features of a multi-robot system based on modular robots that change their configuration according to the characteristics of each task step. Finally, main conclusions from this work about modular robots are summarized in the last section.

2 Architecture of the RobMAT System

RobMAT architecture is based on simple components, **modules**, which joins each other to form **molecules**. Being all of molecules the **colony**. Module is the simplest part that is capable of moving and communicating. A molecule is an autonomous robot which is made up of modules. Therefore molecules can have

different configurations. Each molecule has a master module and the rest of the modules have a slave role. Master and slave roles will be described in the next sections. Finally, the colony is remotely commanded by a human operator via an advance teleoperation interface. Figure 1 shows the main components of the RobMAT architecture.

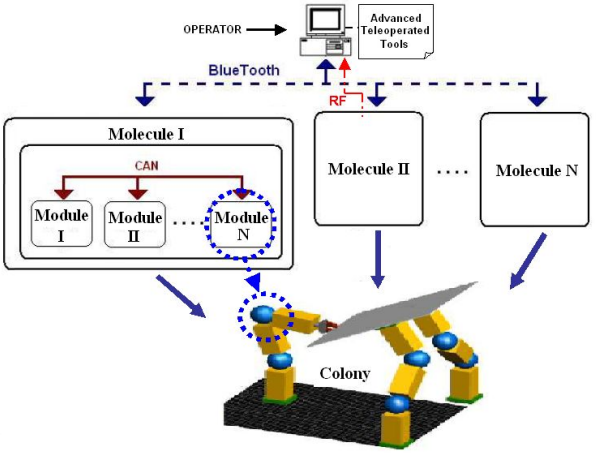


Fig. 1. Main components of RobMAT architecture

2.1 Description of the RobMAT Module

The module designed for the RobMAT robot attempts to reach a balance between complexity of design and degree of functionality. The goal is to obtain very functional modules, in spite of the fact that such objective requires more complex designs. However, the module has always been considered to be very simple compared with the degree of complexity that the molecule as a whole may have. Taking such into account, a module is created so as to be able to form very functional molecules comprised of only a few modules.

The module has an actuated central part that provides 3 d.o.f. The axis of each d.o.f. intersect in one point and thus the atom has an actuated spherical joint, figure 2a. two prisms Next to the joint are that hold the electronics control and the batteries that would feed the module. The sensors, essential for the module to know about its internal state and to interact with its surroundings, are distributed all over the module depending on their function and the best location.

The electronics control are found within each module and have been designed specifically for this application. The control hardware is made up of two boards joined by a 100-ways SMD connector. One of the boards is the data processing and storage unit, while the other is in charge of driving the actuators and interface functions. The kernel of the processing board is a DSP by Texas Instruments, model TMS320F2812. The main features of this board are the following;

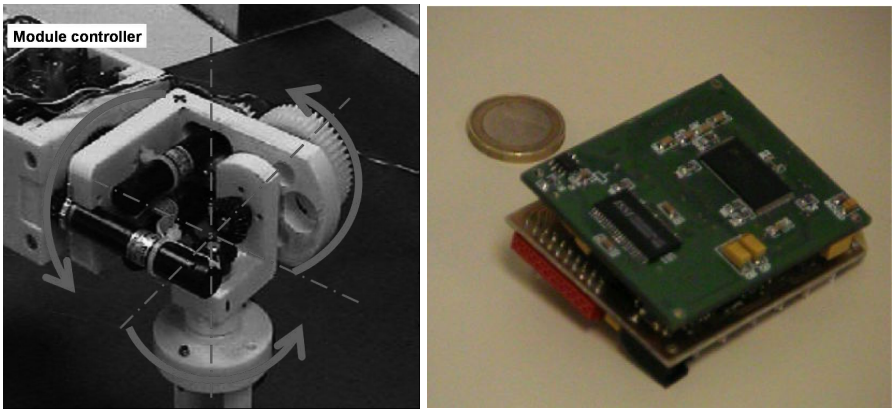


Fig. 2. Module Design. Actuated axes and electronic boards

150 MIPS, 56 General Purpose IO, 16 ADC, 16 PWM, CAP/QEP, 2 SCI, CAN, SPI, McBSP, 4 16bits Timers, 1 Watchdog, Flash memory 768KB, RAM 48KB.

The other board performs tasks related to the interface of the module actuators, sensors and communication channels. Its main features are the following; 3 H-bridge to drive the motors, current measure for each motor, Bluetooth communication device, external interface, voltage converters. Figure 2b shows the two boards, which have a size of 50x65mm.

As mentioned above, the boards developed include various communication channels. Communication has always played an important role in the hardware of modular robots, since their communication capacity is a determining factor in the performance of tasks. The robot really uses only two communication channels. On the one hand, a communication network is implemented on a CAN bus. Communication between atoms belonging to the same molecule takes place through this channel, which transmits the messages of synchronization and cooperation, as well as the diffusion of commands received through the Bluetooth link. The speed of transmission of this bus is 1Mbit/s. The other communication channel used is the Bluetooth link. This allows for the exchange of messages between molecules, or between a computer and a molecule. The speed of transmission is lower than that of the CAN bus, approximately 9600 bit/s. For this reason the messages transmitted are of high level and are less frequent.

2.2 Base Molecule

A molecule is the structure formed by various modules joined together. The molecule with the fewest atoms and which has sufficient autonomy [4] is called the base molecule. Most of the demands made from an autonomous robot are met in each module, but the capacity for proper displacement requires at least two atoms. The base molecule needs to be fixed on the floor to keep balance. For this reason, a small tool is added to the base molecule in order to overcome

this problem. Namely an electromagnet on either end side allows the molecule to fix on metallic surfaces.

Modules that compound a base molecule are rigidly joined, therefore base molecule is the minimum element to form more complex molecules. Each molecule has a connector which allows docking between them. There are two different connectors on each molecule. One of them is made up of an electromagnet and four rods that give stiffness to union. Two of them also transmit the signals of CAN bus. The other one is simpler. It is just a small sheet metal where electromagnet is fixed. There are two contacts on module surface to transmit CAN bus signal to another molecule connector rods.

All modules are identical. However, only one activates its Bluetooth device in order to receive and send messages from/to the molecule. This master module is designated from a PC where the user is. The master module mainly dispatches servo motors references to every module in the molecule.

3 Molecules Performance

Self-reconfigurable modular robots have the ability to change their configuration whenever they need. Such property gives great flexibility to these modular systems. However, variable configuration involves new parameters in control algorithms. In some cases, new algorithms will even be added to cope with their configuration. Therefore, performance of these systems is far from optimum. Further research has to be done in order to develop algorithms that are suitable to these kinds of changing-robots.

3.1 Modelling Reconfigurable Modular Robots: Graph Theory

All mechanism is composed of joints and links which are connected to each other. According to Graph theory [13], a mechanism is represented by a graph with vertexes and edges. Vertexes represent links and edges represent joints [9]. Each module of a modular robot can be considered as mechanism. Therefore, kinematical chains in modular robots can be shown as described above. These chains will be open or closed, or a combination of both. Namely, modular robots can contain truss-like substructures, open chains, and closed-loop mechanisms. Figure 3 shows a robot configuration formed by 4 modules and its corresponding graph representation.

A graph can be more complicated with directed edges or different kinds of vertexes and edges. However, graph information can always be arranged in matrices. There are two meaningful matrices: adjacency (A) and incidence (IM). For kinematical modelling, the incidence matrix is quite useful. This is a NxM matrix where N and M are the number of vertexes and edges respectively. Entry a_{ij} is equal to 1 when $vertex_i$ and $edge_j$ are incident[1]. If this is not the case, the value is zero.

[1] If vertex v is an endpoint of edge e then, e is incident on v.

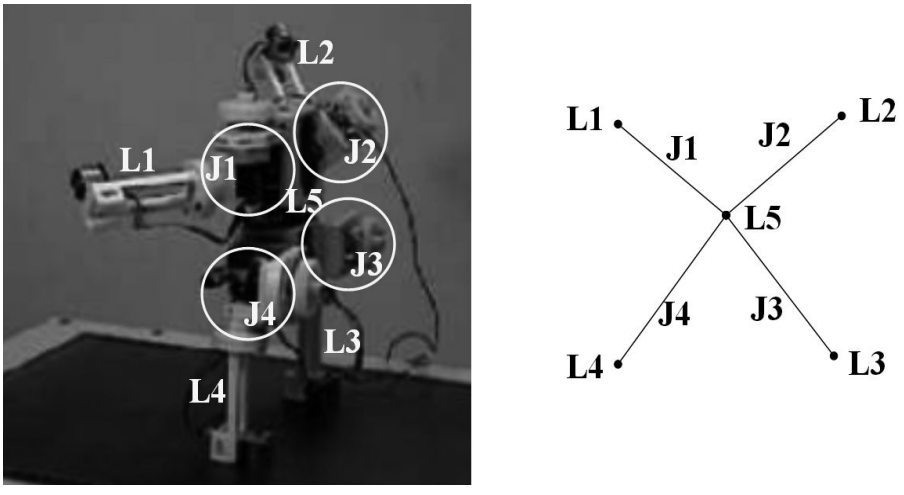


Fig. 3. Robot composed of 4 modules (left) and a graph which shows its corresponding configuration (right)

Currently the base molecule is the smallest element that can be docked to other elements during the execution of tasks. Moreover, all possible molecules that can be formed by RobMAT modules are combinations of base molecules. Nevertheless, from a modelling point of view, modules are major units. Each module is represented as a vertex on the graph and the connections between modules as edges. Since each module may have several ports (connection places) where other modules can dock with, more information has to be added on the graph. Namely, each module port is numbered following the same pattern in every module. Then port number is shown at each edge end side in the graph. Gathering this information on a graph, any RobMAT molecule will be represented by one type of vertex (module), one type of edge (port) with a pair of number at each end side. The matrix that is yielded by the graph can be used for getting both the kinematical [10] and dynamical model among other.

As for the applications described above, these matrices are used so as to know the connection between modules. Consequently, the master module could send reference command to proper module. More advanced uses of these matrices will be presented in future works.

3.2 Molecules Synchronization Algorithm

A molecule has robot behaviour thanks to the module synchronization. This global behaviour is required to move the molecule as a whole. This is the key point in order to implement a collaborative behavior for multi-robots. For example, it allows starting and finishing all joint movements in a module simultaneously.

One of the problems in distributed computational systems is the lack of a global clock [11]. This is a handicap to carry out concurrent actions in a

coordinated way. From the computational point of view modular robots are a set of processing units joined by a communication bus. We propose a discrete time closed-loop method which keeps all system clocks in the same phase and period using a single short message in every cycle. The period of that cycle can be much longer (seconds) than a control cycle period (milliseconds, typically the period of timer interruption.).

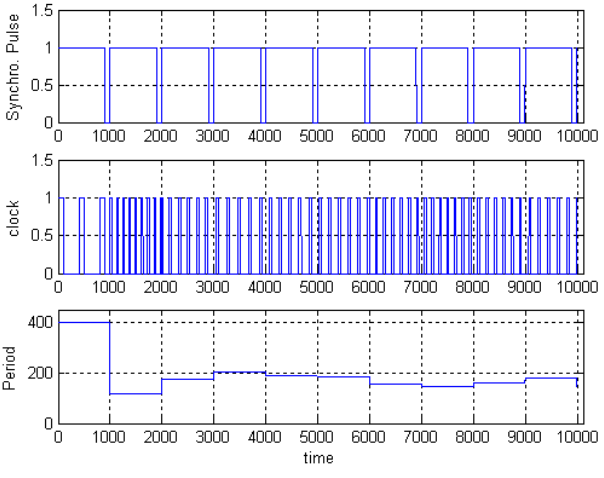


Fig. 4. Signals at the process of synchronization

The method needs a periodical signal which acts as a trigger for the closed loop. This signal is generated by the master module of molecule every N control cycles. Namely the signal is a high priority short CAN message. It is supposed, with neglected error, that all modules receive the message at the same time. Each time that message is received, a local timer (ticks counter) is reseted and its previous value is used to correct the local timer period. Figure 4.

The result is N slots of both same length and phase in every module. This allows molecule to get tightly coordinated movements. Also, once all module clocks are adjusted a master fault can be detected analyzing the synchronization message.

4 Teleoperated Task with the Modular Robot RobMAT

The most basic tasks are manipulation and displacement of a base molecule. They represent the simplest functionalities of the RobMAT system. Manipulation is carried out by fixing a robot end and controlling the joint movements. According to features of the task, it could be needed to use an additional tool. The base molecule has five DoF. Therefore, it can execute most of the manipulation operations. In case of complex tasks two base molecules could be linked to increase the molecule DoFs, consequently molecule dexterity is also increased.

Base molecule displacement is done by executing a set of simple steps in order to achieve the goal position.

4.1 Elements Involves in RobMAT Teleoperation

There is a PC from which teleoperation commands are sent. A 3DoF Joystick is plugged to the PC and a user drives it. The joystick produces the three joint references of the module. Changing between modules is carried out by joystick buttons. The references are sent from PC to master module via Bluetooth.

One molecule carries two wireless mini-cameras so that remote environment can be seen by the user. The cameras are arranged in such a way that a 3D image is generated. The user, assisted with shutter glasses, has a 3D perception of the remote scene. Beside joystick and 3D image, video image blending system [12] is used. This gives extra information to the scene in order to make teleoperation easier. In figure 5a is shown the user during a teleoperation task.

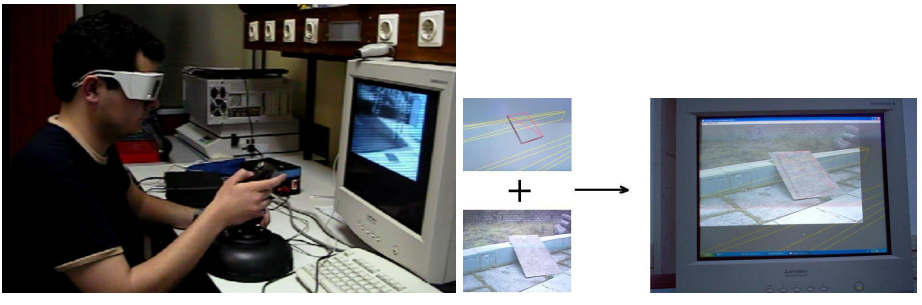


Fig. 5. a) Operator during a teleoperated task using a stereoscopic display. b) Video blending: a simulation of robot environment is merged with real video.

4.2 Execution of Tasks in Unstructured Environments

A feasible complex task is described. Initially three base molecules are linked as a six-legged configuration. A movement gait is implemented so that the molecule can move forward. Once the goal is reached, base molecule that carries mini-cameras undocks from the other two. It is then teleoperated to a new position where a remote environment 3D image is obtained. Then the four-legged molecule is led to its workspace. Once there, the two base molecules undock and each molecule begins a manipulation task on a metallic floor.

Commands of undocking , appointing master modules, molecules motion are sent from the PC. Figure 6 shows a sequence of this collaborative task.

The video image blending system allows the operator to calibrate remote environments. That is, the operator superimposes a video image from the remote working environment onto a simulation of this environment. When objects of both images match, a rough estimation of the spatial location of objects is



Fig. 6. Cooperative work carried out in an unstructured environment

obtained within the remote environment. This is a powerful tool for working with unstructured environments. Figure 5b shows how video image blending works.

Others molecules are in charge of executing manipulation task. A molecule gets a part and puts it closer to other molecules. This is an example of a collaborative manipulation that requires distributing and synchronizing the task execution commands between two molecules. Synchronization among molecules is carried out using wireless network.

4.3 Conclusions

This paper has described the architecture of a self-configurable modular robot. The module has three DoF which forms an actuated spherical joint. This arrangement gives great functionality to the module. However, module functionality is still simpler than what it could be when joining modules to each other. These unions or molecules can carry out a number of tasks. For example, the simplest molecule or base molecule is made up of two modules able to manipulate and move around. Moreover, custom electronic boards have been developed to control each module. A powerful kernel device gives both enough computational capacity and signal interface. Two communication channels allow modules and molecules to collaborate. Graph theory is a suitable tool for modelling such dynamically changing robots. Graph information and a synchronization algorithm, which achieves a global clock for all modules, permit the molecule to behave as a whole. Finally, teleoperation elements are described. 3D visual feedback of remote environment helps in teleoperation. Validating all RobMAT elements is described as a complex task with manipulation that requires reconfiguration.

References

1. Yim, M., Duff, D., Roufas, K.D.: Walk on the Wild Side. IEEE Robotics & Automation Magazine 9, 49–53 (2002)
2. Kurokawa, H., Kamimura, A., Yoshida, E., Tomita, K., Murata, S., Kokaji, S.: Self-Reconfigurable Modular Robot (M-TRAN) and its Motion Design. In: Proceedings of 7th International Conference on Control, Automation, Robotics and Vision (2002)
3. Shen, W.-M., Krivokon, M., Chiu, H., Everist, J., Rubenstein, M., Venkatesh, J.: Multimode Locomotion for Reconfigurable Robots 20, 165–177 (2006)
4. Escalera, J.A., Saltaren, R., Aracil, R., Ferre, M., Garca, C.: Base Molecule Design and Simulation of a Modular Robot RobMAT. In: Proceeding of 16th IFAC World Congress, Prague, Czech Republic (2005)
5. Kotay, K., Rus, D., Vona, M., McGray, G.: The Self-reconfiguring Robotic Molecule: Design and Control Algorithms. In: Proceedings of IEEE International Conference on Robotics and Automation, IEEE Computer Society Press, Los Alamitos (1998)
6. Lipson, H., White, P., Zykov, V., Bongard, J.: 3D Stochastic Reconfiguration of Modular Robots. In: Presentation at the Workshop on Self-reconfigurable Robotics at the Robotics Science and System (2005)
7. Jorgensen, M.W., Ostergaard, E.H., Lund, H.H.: Modular ATRON: Modules for a Self-reconfigurable robot. In: Proceedings of IEEE/RJS International Conference on Robots and Systems (2004)
8. Castano, A., Behar, A., Will, P.M.: The Conro Modules for Reconfigurable Robots. IEEE/ASME Transactions on Mechatronics 7, 403–409 (2002)
9. Freudenstein, F., Dobrjanskyj, L.: On a Theory for the Type Synthesis of Mechanisms. In: Proceedings of the 11th International Congress of Applied Mechanics, Munich, Germany, pp. 420–428 (1964)
10. Chen, I-M., Yang, G., Kang, I.-G.: Numerical Inverse Kinematics for Modular Reconfigurable Robots. Journal of Robotic Systems 14, 213–225 (1999)
11. Chow, A., Johnson, T.: Distributed Operating Systems and Algorithm Analysis, Ed. Addison-Wesley, Reading (1997)
12. Ferre, M., Aracil, R., Navas, M.: Stereoscopic Video Images for Telerobotic Applications. Journal of Robotic Systems 22, 131–146 (2005)
13. Gross, J.L., Yellen, J.: Graph Theory and Its applications. Chapman & Hall/CRC (2006)

Data Mining of Time-Series Medical Data by Formal Concept Analysis

Kenji Sato[1], Yoshiaki Okubo[2], Makoto Haraguchi[2], and Susumu Kunifuji[1]

[1] School of Knowledge Science Japan Advanced Institute of Science and Technology
1-1 Asahidai, Nomi-City, Ishikawa 923-1292 Japan
{k-satou,kuni}@jaist.ac.jp
[2] Division of Computer Science
Graduate School of Information Science and Technology
Hokkaido University
N-14 W-9, Sapporo 060-0814, Japan
{yoshiaki,mh}@ist.hokudai.ac.jp

Abstract. This paper presents a method for clustering time-series medical data based on Formal Concept Analysis. We made a prototype system, and verified it by applying it to several medical cases. Our clusters can be explicitly provided with more convincing meanings, with the help of FCA. We plan to continue, and to verify the practicality of this method by applying it to hundreds of medical cases.

Keywords: Clustering, Medical case search system, Formal Concept Analysis, Time series Data.

1 Introduction

Case search systems has been created for referring to past examples of the same or a similar set of symptoms when patients are diagnosed. However, the prior systems for researching past cases have only targeted text information of on the clinical records. Medical cases consist of both text information and time series data related to test results. We propose using Formal Concept Analysis (FCA)[1] as a method finding similarities in this time series data for classifying of paying attention to the similarity of the time series data for classifying medical cases. Recently, the research for supporting medical diagnosis has been advanced by data mining of the time series data of medical test by Active Mining [3]m for example. However, treatment must often be given before data about the characteristics a particular disease is available. Therfore, We are not specifically concerned with a particular disease, and we propose FCA as the way of uniformly data mining the time series data of medical tests.

2 Preliminaries[2]

FCA is a theory of data analysis that identifies conceptual structures among objects (individuals).

B. Apolloni et al. (Eds.): KES 2007/ WIRN 2007, Part II, LNAI 4693, pp. 1214–1221, 2007.
© Springer-Verlag Berlin Heidelberg 2007

Let $\mathcal{O}$ be a set of objects (individuals) and $\mathcal{F}$ a set of features (attributes). A pair of $\mathcal{O}$ and $\mathcal{F}(< \mathcal{O}, \mathcal{F} >)$ is called a formal context. Under the context $< \mathcal{O}, \mathcal{F} >$, we can show each object in $\mathcal{O}$ as the set of all features in $\mathcal{F}$ that the object has.

Given a context $< \mathcal{O}, \mathcal{F} >$, for a set of objects $O \subseteq \mathcal{O}$ and a set of features $F \subseteq \mathcal{F}$, two functions $\varphi : 2^{\mathcal{O}} \rightarrow 2^{\mathcal{F}}$,$\psi : 2^{\mathcal{F}} \rightarrow 2^{\mathcal{O}}$ are defined respectively as follows.

$\varphi(O) = \{f \in \mathcal{F} | \forall o \in O, f \in o\} = \cap_{o \in O} o$
and
$\psi(F) = \{o \in O | F \subseteq o\}$.

φ computes the feature set shared by every object in O. On the other hand, ψ returns the set of objects with F.

Where $O \subseteq \mathcal{O}, F \subseteq \mathcal{F}$,$\varphi(O) = F$ and $\psi(F) = O$, a formal concept (FC) under the context is defined as a pair of the object set and the feature set,(O, F). O and F are called the extent and intent of the concept, respectively. Obviously, $\psi\ (\varphi(O)) = O$ and $\varphi\ (\psi(F)) = F$ from the definition. In other words, a formal concept is defined as a pair of closed sets under the mappings.

Let $\mathcal{S}$ be a set of combinations of daily data. A patient p can be represented as a set of combinations of daily data in $\mathcal{S}$ appearing in the patient's test results, that is, $p \subseteq \mathcal{S}$. For a set of patients $\mathcal{P}$; therefore, we can consider a formal concept (P, S) under the context $< \mathcal{P}, \mathcal{S} >$. The extent P is regarded as a cluster of patients and the intent is the combinations of daily data shared by the patients in $\mathcal{P}$.

Let $G = (V, E)$ be an undirected graph where V is a set of vertices and $E(\subseteq V \times V)$ a set of edges. For a set of vertices $V' \subseteq V$, the graph G' defined as $G' = (V', E \cap (V' \times V'))$ is called a subgraph of G (induced by V'). If a subgraph $G' = (V', E')$ of G is complete, then G' is called a clique of G, and the clique is often denoted simply by V'.

3 Finding Time Series Clusters Based on Formal Concept Analysis

3.1 Time Series Clusters as Formal Concepts

3.1.1 Characteristics of Patient Test Data

Tests are assigned a positive or negative numerical value and, the signs vary. Forty-one kinds of test data in date order are given for each patient. The number and period of dates are different for each patient.(Table.1)

Table 1. Data of each patient

	$test_1$	$test_2$	...	$test_m$
$date_1$	v_{11}	v_{12}	...	v_{1m}
$date_2$	v_{21}	v_{22}	...	v_{2m}
...	...	...	...	...
$date_n$	v_{n1}	v_{n2}	...	v_{nm}

A gap caused by losts values exists in the time series of the test data. The length of the gap can vary also too. A gap period can be coped with by a complementary period.

On the other hand, the test data included a large number of disconnection values. The following are examples of disconnection values that were detected.

1. The cause of the disconnection value is a test error.
2. The cause of the disconnection value is very unusual biologically, but it cannot be concluded to be a measurement mistake.
3. The cause of the disconnection value is not unusual; it cannot be concluded to be a measurement mistake, and it is seen as biological.

In each of these cases, the method classifies the disconnection value, and the copes with it accoudingly.

3.1.2 Classification of Daily Data

The medical test data is classified into seven stages from A to G each corresponding one of the characters.(Table.2) Numerical value data are sorted in a maximum of seven steps based on a normality value range.

Table 2. A part of the classified standard table of the medical test values

A	B	C	D	E	F	G
3-	2-	-	0	+	2+	3+
—	–	–	0	+	++	+++
↓↓↓	↓↓	↓	±	↑	↑↑	↑↑↑
			-	+		

By a sign's like this; for each patient's inspection data, for example, is shown as sign line

```
DDDD__EEEFFE_DDCCDED (We decided that '_' would indicate a lost value.).
```

3.1.3 Combination of Daily Data - Single Pair Triple ... etc

Our purpose is seeing the relation of the change that occurs in a series medical tests over time in addition to the change in each test. We set a precise date for the system to respond and show the features of a disease. (For example, it is understood that test2 is often a low day value while test1 is a high price gradually.)

The following is the example when the days are set to 1.

1. single feature: First, to show a patient's situation, we consider the classification sign (A through-G) of a daily test.

$test_i$=v_{ij}, (j=1...m)

(example) GOT=E (The value of the inspection GOT is show to rank E in one day.)

2. pair feature: Two arbitrary tests are assumed and the one that the value on the choice same day was arranged.

$test_k_test_l = v_{kjj}v_{lj}, (j = 1...m)$

(example) $GOT_GPT = EF$ (It is shown that the value of the inspection GOT was shown to rank E and the inspection GPT was rank F in one day.)

3. 3-combination

Three arbitrary inspections are assumed and the one that the value on the choice same day was arranged is assumed to be a value.

$test_k_test_l_test_m = v_{kjj}v_{lj}v_{mj}$, (j=1...m)

(example) $GOT_GPT_LDH = EFD$

4. n-combination

Combinations of higher-order are considered further as necessary.

3.2 Formal Concept-Based Cluster

Let S be a set of combinations of daily data and P a set of patients, where each patient is represented as a set of combinations of daily data in S appearing in time-series data of the patient.

Regarding $\mathcal{C} =< \mathcal{P}, \mathcal{S} >$ as a context in FCA , a formal concept (P, S) under $\mathcal{C}$ can be considered to be a cluster of patients. We can explain why patients in P are grouped together. Each patient in P shares the set of combinations of daily data S and any other patient never contains S. That is, P can form a meaningful grouping (cluster) of patients. Thus, by restricting our clusters to being formal concepts under $\mathcal{C}$, we can consider the meanings based on their intents. We call these clusters Formal Concept-Based clusters (FC-clusters in short).

The meaningfulness of FC-cluster is influenced by both of its intent and extent. For example, a cluster with smaller intent might be unconvincing because the evidence seems to be weak for the grouping , though its extent tends to be larger. Oppositely, although a cluster with larger intent might have more convincing evidence, its extent tends to be smaller.

3.3 Problem Definition

Our problem is defined as follows:

[Given]
S: a set of combinations of daily data
P: a set of patients each of which is represented as a subset of S
w_p: an evaluation function for sets of patients
w_s: an evaluation function for sets of combinations of daily data
δ: a threshold for the minimum evaluation value for intent

[Find]

the set of formal concepts (S, P) such that the evaluation value of $P, w_p(P)$, is in the top N among δ-valid formal concepts under the context $< S, P >$.

4 Finding Clusters by Top-N Formal Concept Search

4.1 Graph Representation of Time-Series Data Set

Top-N δ-valid FC-clusters are obtained by discovering certain cliques in a weighted undirected graph.

Given a context $C =< S, P >$, an evaluation function for sets of combinations of daily data w_s and a threshold δ for the quality of intents,

First, we make a weighted undirected graph $G = (P, E)$, where E is defined as

$E = \{(p_i, p_j) | p_i, p_j \in P(i \neq j) \wedge w_s\ (p_i \cap p_j) \geq \delta\ \}$.

If a pair of patients shares a set of combinations of daily data in which the evaluation value is greater than or equal to δ , they are connected by an edge.

From the definition of the graph G, any extent of δ-valid FC is a clique in G.

From the graph G , we try to extract δ-valid FCs with extents that have Top-N evaluation values.

4.2 Branch-and-Bound Depth-First Search

We find Top-N formal concepts by adopting branch-and-bound depth-first search strategy. [4]

We maintain a list of formal concepts that stores the Top-N FCs among those already found during the search.

The list holds tentative Top-N clusters.

For each clique (a set of patients) Q , $\varphi(Q)$is computed and then evaluated.

If $w_s\ (\varphi(Q)) \geq \delta$, the set of combinations of daily data $\varphi(Q)$can become the intent of a δ-valid FC with its extent $\psi(\varphi(Q))$.

Then, the tentative Top-N list is adequately updated for the obtained $FC(\psi(\varphi(Q)), \varphi(Q))$.

The procedure is iterated in depth-first manner until no Q remains to be examined.

Starting with the initial Q of the empty set, Q is expanded step by step.

5 Experimental Results

In this section, we present our experimental results. We processed the time series data of 41 kinds of medical tests in about 1200 medical cases to observe the characteristics of FC-clusters.

5.1 Preprocessing

We removed a particular value and factorized a consecutive value. Moreover, we made amends for a missing value by linear interpolation. Next, we made the

time series into a sign row by classifying the inspection value into seven stages or less (A- through G). In addition, we defined intent(feature) by continually combining the medical test values.

5.2 The Trial of FCA and the Analysis of That Result

It had the possibility that a diagnosis was being classified delicately with a factor except for the result of the inspection. So, we tried FCA to a part of the data, and analyzed relations between extent and intent.

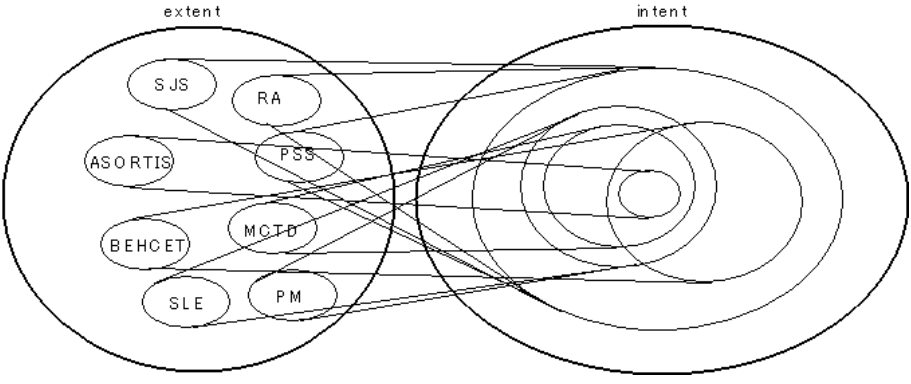


Fig. 1. Mapping from extent to intent

Let $\overline{D}$ be a set of intents mapped from the extent which has diagnosis D. We observed that $\overline{SJS} = \overline{RA} = \overline{PSS}$, $\overline{SLE} = \overline{PM}$, $\overline{BEHCET} = \overline{DM} = \overline{MRA}$, $\overline{MCTD} = \overline{APS}$, $\overline{AORTITIS} = \overline{PN} = \overline{PNc} = \overline{ITP} = \overline{Sarcoidosis}$ and we could get Table.3 by analyzing the intersection of sets of intents of FCs.

Table 3. The Intersection of Intents

	number of features	SJS,RA, PSS	SLE, PM	BEHCET,DM, MRA	MCTD, APS	AORTITIS,PN,PNc, ITP,Sarcoidosis
SJS,RA,PSS	168	-	99	111	66	33
SLE,PM	99	-	-	42	66	33
BEHCET,DM,MRA	111	-	-	-	33	33
MCTD,APS	66	-	-	-	-	33
AORTITIS,PN,PNc, ITP,Sarcoidosis	33	-	-	-	-	-

For example, $|\overline{SJS}| = 168$ and $|\overline{MCTD}| = 66$ and $|\overline{SJS} \bigcap \overline{MCTD}| = 66$ then $\overline{SJS} \supset \overline{MCTD}$.

5.3 Grouping of the Diagnosis

From Table.3, we classified the diagnosis into five groups(Table.4).

Table 4. The five groups of diagnosis

Group	dianosis
Group A	SJS, RA, PSS
Group B	SLE, PM
Group C	BEHCET, DM, MRA
Group D	MCTD, APS
Group E	AORTITIS, PN, PNc, ITP, Sarcoidosis

5.4 FCA

We tried to extract the Top-10 δ-valid FC-clusters from the graph. Some of the obtained clusters are shown in Table.5.

For comparison, we extracted several clusters based on patient diseases.

As one example, we obtained a cluster1 consisting of the patients with the diagnosis Group A, Group A and AIHA, Group E and AGA, Group A, Group A, Group B, Group A, Group A and Group A, Group B, Group B, Group A and paniculitis.

From Table.3, intent of Group A contains intent of Group B and intent of Group E contains intent of Group B. We collated the obtained result with the standard diagnosis. As a result, we found that 80% or more fulfilled the diagnostic criteria.

Table 5. Part of the FCA results

cluster	diagnosis
cluster1	Group A, Group A and AIHA, Group E and AGA, Group A, Group A, Group B, Group A, Group A and Group A, Group B, Group B, Group A and paniculitis
cluster2	Group A, Group A and AIHA, Group E and AGA, Group A, Group A, Group B, Group A and Group A, Group B, Group B, Group A and paniculitis
cluster3	Group A, Group A and AIHA, Group E and AGA, Group B, Group A, Group A, Group B, Group A and Group A, Group B, Group B, Group A and paniculitis

6 Concluding Remarks

In this paper, we discussed a method of mining time-series medical data, the method is based on Formal Concept Analysis. We made a prototype system, and verified about several cases.

Our clusters can be explicitly provided with more convincing meanings, with the help of FCA. By restricting our clusters (cliques)to formal concepts, we can consider their clear conceptual meanings as being their intents (the set of shared combinations of daily data). We designed an algorithm for finding Top-N FC-clusters.

This algorithm can be viewed as an extension of our previous one. From these observations, we expect that our method based on Formal Concept Analysis is a promising approach to finding meaningful patient clusters.

We plan to continue testing our method, and to verify the practicality of this method by applying it to the other syndrome.

References

1. Ganter, B., Wille, R.: Formal Concept Analysis: Mathematical Foundations. Springer, Heidelberg (1999)
2. Haraguchi, M., Okubo, Y.: An extended branchand-bound search algorithm for finding top-n formal concepts of documents. In: Proc. of the 4th Workshop on Learning with Logics and Logics for Learning, pp. 41–47 (2006)
3. Motoda, H. (ed.): Active Mining. IOS Press, Amsterdam, Holland (2002)
4. Tomita, E., Seki, T.: An efficient branch and bound algorithm for finding a maximum clique. In: Calude, C.S., Dinneen, M.J., Vajnovszki, V. (eds.) DMTCS 2003. LNCS, vol. 2731, pp. 278–289. Springer, Heidelberg (2003)

A Case Study on Project-Management Training-Support Tools for Japanese/Chinese/Indian Offshore Development Engineers

Sadamitsu Toyoda[1,2], Motoki Miura[1], and Susumu Kunifuji[1]

[1] School of Knowledge Science, Japan Advanced Institute of Science and Technology
1-1 Asahidai, Nomi, Ishikawa, 923-1292, Japan
[2] The SANNO Institute of Management, 6-39-15 Todoroki, Setagaya,
Tokyo,158-8630, Japan
TOYODA_sadamitu@hj.sanno.ac.jp, {miuramo,kuni}@jaist.ac.jp

Abstract. We developed project-management training-support tools. The tools consist of a Web-based training (WBT) (e-learning) system and software designed to teach bridge software engineers (SEs) methods for solving problems. We asked a group of bridge SEs to participate in an experiment to determine whether training with our tools would improve their project-management and/or problem-solving skills. The results suggest that WBT (e-learning) is an effective method for acquiring explicit knowledge. However, the results also suggest that the software designed to teach the SEs methods for solving problems is ineffective in improving their problem-solving skills.

Keywords: Bridge Software Engineers, e-learning, Project Management, Problem-Solving Support Tool.

1 Introduction

Software development using foreign resources is called "offshore development," and software engineers who engage in this software development as go-betweens are called "bridge software-engineers" (bridge SEs). Offshore development originated in the United States because that is where the outsourcing trend started. India, where English is used as a common language, is a major force of development activity and is growing [1]. Japan, with the IT recession of the late 1990's, started taking on this practice, and China and India, whose strengths lie in the programming phase due to low cost and a sufficient supply of SEs, are emerging as powerhouses of development activity.

However, there are numerous problems between Japanese companies, which place orders, and Chinese/Indian companies, which accept them. From the S-open Offshore Study [2], the major problems identified on both sides are 1) communication, 2) management of specifications and modifications, and 3) project management (Table 1). Deeply seated hurdles such as differences in language, culture, business practices, and employment issues may be difficult to overcome, but for the strong

B. Apolloni et al. (Eds.): KES 2007/ WIRN 2007, Part II, LNAI 4693, pp. 1222–1229, 2007.
© Springer-Verlag Berlin Heidelberg 2007

growth of offshore development among Japan, China, and India, these problems need to be solved.

Table 1. Differences in software development among Japan, China, and India

Difference	Japanese Domestic Development	Offshore Development in China	Offshore Development in India
Language/ Communication/ Culture	Japanese/ Must-Obey-the-Client business practice/ team-oriented	Chinese/ individual-skill- (rather than team-) oriented	English/ classified society/ debate culture
Specification and modification management	It is customary to proceed to the following phase even if specifications are unclear. Modification is an ordinary practice.	Do not proceed unless requests or specifications are defined. Specifications are not easily modified.	Do not proceed unless requests or specifications are defined. Specifications are difficult to modify.
Project management (PM) (quality/cost/due date/employment turnover ratio)	Lack of PM is a concern. Quality and due date are stable. Turnover ratio is low.	Management level varies due to lack of experience. Due date is emphasized rather than cost. Turnover ratio is high.	Many are classified as Quality CMM level 5. Have a strong sense of American-style management. Employment stability is low.

As shown in Table 1, some of the problems that make offshore development difficult are differences in culture, in project management, and in how to provide solutions. To solve these problems, bridge SEs (who play the role of go-betweens) and their skills are essential for successful international distributed software development [3].

In addition, bridge SEs requires the following specialized skills to gain the confidence and trust of the people on the order side.

1. They must be able to communicate effectively with the other international development team members using correct project-management terminology in a common language.
2. They must be able to think about problems from the perspective of the order side and use the appropriate skills to solve them.

We think that the required project-management training systems are based on differences in culture, so we developed training support tools as a means of improving the level of bridge SEs' project-management and problem-solving skills. The tools are based on the concept of sharing project-management knowledge and problem-solving methods [4]. We asked Japanese and Chinese SEs to use these tools, after which we evaluated whether the level of their project-management skills had improved.

2 Project-Management Training Support Systems

Our project-management training support systems were designed to improve the level of bridge SEs' project-management skills. Bridge SEs act as go-betweens in international distributed software development (Figure 1).

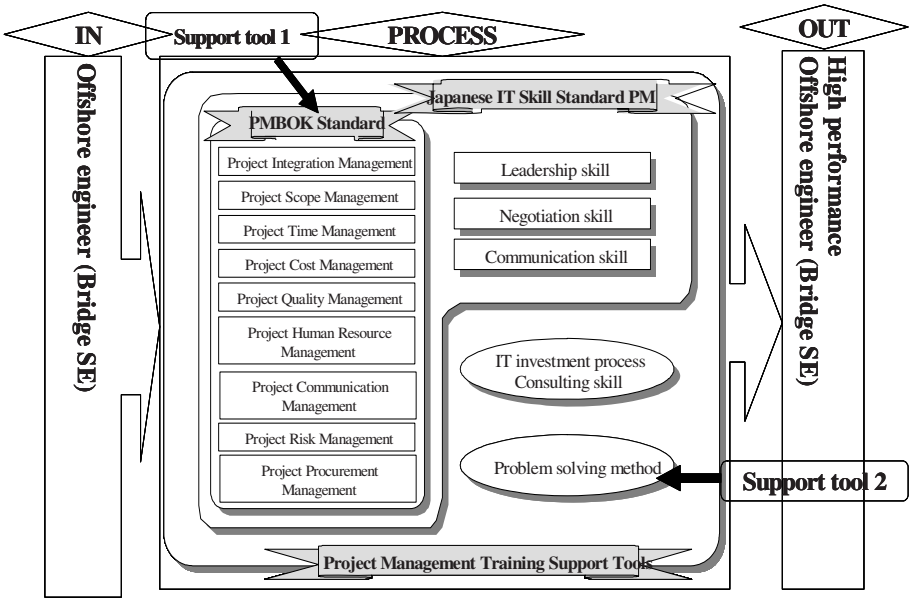


Fig. 1. Project-management training support systems

2.1 Support Tool 1: "Project-Management Basic Course" (e-Learning)

We prepared a Web-based training (WBT) system[1] whose major features include nine areas of common knowledge conforming to an international project management standard, PMBOK® [5], and a group of five integrated project-management processes. At any time, bridge SEs can access the WBT on a PC connected with those of the other international development team members over the Internet to move work.

The e-learning component is comprised of the following nine areas of common knowledge. Project integration management consists of project chartering, monitoring, and controlling; project time management consists of defining and scheduling project-related tasks and activities; project cost management consists of methods for controlling project cost; and project quality management consists of quality planning, assurance, and control. These tools comprise the visual content for the beginner's class.

[1] The author is a co-developer of the course at The SANNO Institute of Management. He is mainly responsible for the primary design, monitor/control, and ending phase of the course.

2.2 Support Tool 2: "Software Designed to Teach Bridge SEs Methods for Solving Problems" (Figure 2)

We developed a system[2] that helps bridge SEs improve their problem-solving skills. People often describe and analyze the main points of problems. Our system provides functions to support problem analysis. The user can add labels to the problem node of the system. Labeling helps the user learns to identify and solve problems.

The software features six nodes of problem solving (fact, definition, structure, solution method, courses of action, and expected achievement) and links (the category of events that take place within each node). These nodes and links support the acquisition of knowledge and skill on the basis of the solution process. The acquisition of project-management knowledge is supported within each node.

Steps[3] in the learning process of users

1. Users read a prescribed case study about a project-related problem.
2. The users input selected facts with the toggle under the "Node 1" tab of the software.
3. Users input the meanings of the facts under consideration with the toggle under the "Node 2" tab of the software.
4. Users input the structure of the problem with the toggle under the "Node 3" tab of the software.
5. Users input the methods required for solving the problem with the toggle under the "Node 4" tab of the software.
6. Users input the risk assessment with the toggle under the "Node 5" tab of the software.
7. Users input the result of the assumption with the toggle under the "Node 6" tab of the software.

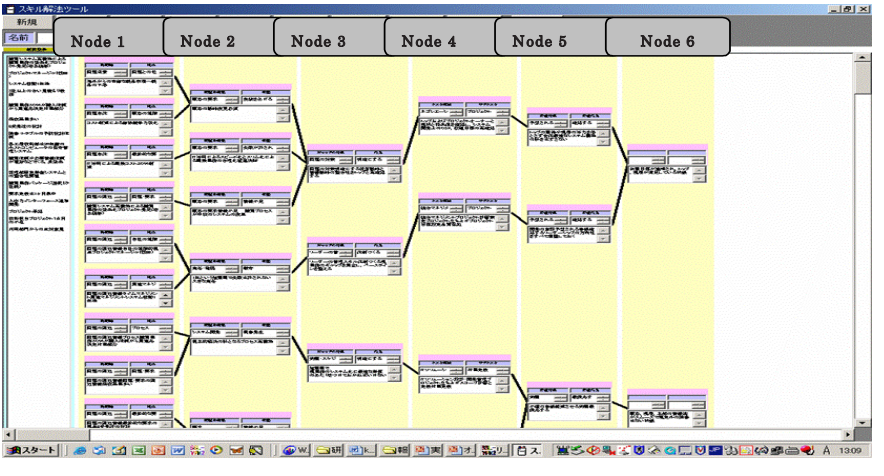


Fig. 2. Support tool for teaching SEs methods for solving problems

[2] The author is a course co-developer at Gakken and The SANNO Institute of Management. He is mainly responsible for designing user manuals as well as developing text for them.

[3] Each step in the process is related to a corresponding toggle and link.

3 Experiment

To determine whether these tools improved the level of the bridge SEs' project-management and problem-solving skills, we performed an experiment at a company with bridge SE employees in Tokyo.

3.1 Items Evaluated

We used the e-learning support tool (WBT) to determine whether the SEs acquired knowledge of common terminology, and if so, to evaluate the level of knowledge they acquired.

Ideally, the SEs should learn from the case studies how to solve problems that occur during a typical software development project. After the SEs used the software, we determined whether the level of SEs' problem-solving skills had improved, and if so, we evaluated the level of improvement.

3.2 Experimental Process

We asked six Chinese and three Japanese bridge-SEs, who wanted to learn project-management skills, to participate in our experiment. We asked them to answer a questionnaire prior to the experiment. We then asked them to study using the e-learning (WBT) and problem-solving-based learning tools, and we tested them after the experiment to re-evaluate their skill level. Time Period: February 1–August 31, 2006. Location: Tokyo. Method of evaluating the performance of support tools.

Support tool 1: Evaluating the level of project-management knowledge the SEs acquired. The experiment started February 1, 2006 and ended May 23, 2006. We analyzed the change in the level of knowledge the SEs acquired as follows. We pretested the SEs on February 23, 2006. They were then asked to train in e-learning with the WBT system between February 1 and April 23, 2006. We posttested them on May 23, 2006 to determine whether they had increased their level of knowledge, and if so, we evaluated the level of increase.

Support tool 2: Evaluating the level of problem-solving skills the SEs acquired. The experiment started April 20, 2006 and ended Jun 15, 2006. We analyzed the change in the level of problem-solving skills the SEs acquired as follows. We pretested the SEs on February 23, 2006. They were then asked to participate in group-training consisting of lectures on our software-based method for learning problem-solving skills and two case studies. This phase lasted for 16 hours. They were then asked to train individually using the software on a PC once a month for a total of 16 hours. We posttested them on August 21, 2006 to determine whether they had increased their level of problem-solving skills, and if so, we evaluated the level of increase.

4 Results and Considerations

4.1 Web-Based Training (e-Learning)

Table 2 shows experimental results of change in SEs level of knowledge. We pretested seven SEs. They were then asked to participate in an e-learning training period lasting two months. One month after completing the training period, six of the SEs were asked to take and unannounced test.

Table 2. Experimental Results

Knowledge/ Process	Before training (01, Feb. 2006)	After training (23, May 2006)	Knowledge acquired
9 areas of common knowledge	7%	83%	74%
5 processes	20%	93%	73%

The results of the posttest showed that the SEs increased their level of project management-knowledge. Before the training period, some of the SEs did not know the standard project-management terminology of the distributed software development industry. However, they acquired that knowledge in the nine areas of project-management knowledge: integration, scope, time, cost, quality, organization, communication, risk, and procurement. They also acquired that knowledge in the five project-management processes: start-up, planning, execution, monitor and control, and closing.

After the training, the SEs were able to organize their personal knowledge, skills, and project-management methods that they learned and were able to use them in daily work [6].

4.2 Software Designed to Teach Bridge SEs Methods for Solving Problems

Figure 4 shows difference between the average 7-point scores for all the SEs pre- and post-questionnaire. We designed questionnaire that consists 60 questions in the sequence of 1-60. The main objective of the questionnaire is to ask SE's about the different aspects of the problems they face and their problem solving behavior. They were asking to input 7 if the answer applies very well and 1 if the answer doesn't apply at all. The numbers nearing 7 shows that the ability to solve problems are improving.

We asked the SEs to answer questionnaires about their problem-solving skills before and after the training, and we determined the change in the level of their problem-solving skills by comparing their before- and after-training responses.

We found that the lecture on our method for solving problems, the case studies, and the software were ineffective. The SEs improved their overall understanding of solving problems through the lectures and case studies; however, their overall level of problem-solving skills did not improve significantly. This is partly due to the software's maneuverability and the way Japanese people understand and interpret definitions. Some possible reasons for this are as follows.

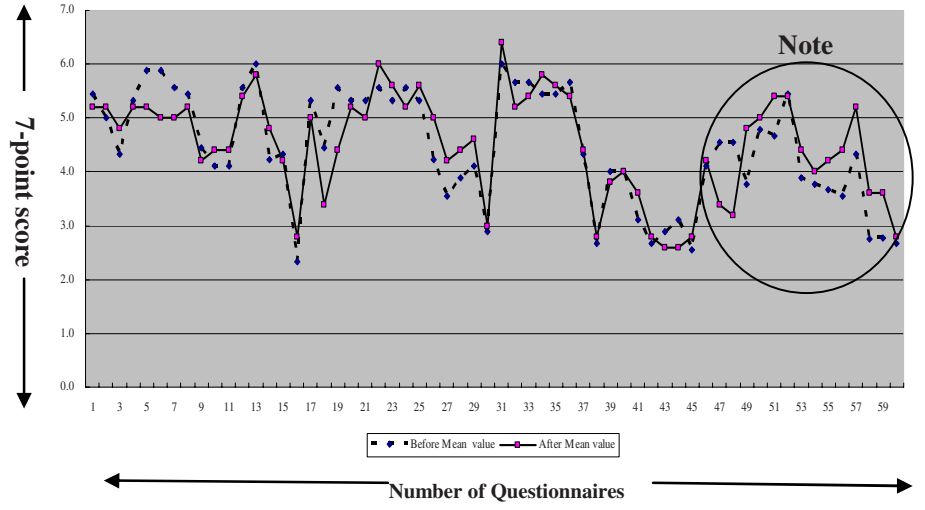


Fig. 3. Difference of Problem Solving Style

A. The SEs already had enough skills a sufficient level of problem-solving skills.

B. The SEs could not motivate themselves to learn a new method for solving problems because they were reluctant to substitute our method for the ones they had always used.

C. The case studies were too difficult for the SEs to follow, and the software was not used effectively, so the SEs could not master our problem-solving method.

D. Because all the SEs scored high on the pretest, the difference in level of problem-solving skills acquired among the Chinese and the Japanese SEs who participated in our study was statistically insignificant.

We found that the software maneuverability and the material content of the case studies need to be improved.

However, when we consider the significant differences between the pretest and posttest results for some of the SEs, as shown in Figure 4 in the encircled area denoted by "Note 1," we can see a significant improvement in the SEs' level of attention paid to detail, in their ability to analyze and solve problems, and in compiling important facts. This may mean that Japanese SEs have changed how they develop software in that they have become more detail-oriented.

5 Conclusion

We performed a case study on project-management training-support tools for Japanese/Chinese offshore development engineers.

We demonstrated that these tools are effective for increasing the level of project management knowledge of bridge SEs, who engages in offshore software development. For bridge SEs, gaining systemized knowledge of the de facto standard

of international project management on top of their current knowledge of project management for offshore software development is useful for managing the progress of projects and for building up organizational assets [7].

By attending the lectures, the bridge SEs' improved their understanding of the methods of problem solving and of the concepts of the analytical and solution-action approaches, but they did not improve their level of problem-solving skills by using the software.

Bridge SEs who engages in offshore development communicates and work with those who speak different languages and have different cultural and business practices. These SEs use the knowledge and skills they acquire in their daily work. However, such experiential knowledge is tacitly held by individuals and is not shared throughout their organization. We demonstrated that the tools we developed are somewhat effective in solving this problem, but they need to be improved. The following three issues are especially important in that regard [8].

1. We must improve the credibility of the experiment by testing SEs from a variety of other countries with more international project management experience.
2. We also need to improve the tools and the case study texts that were ineffective in order to improve the level of the SEs' international project management skills.

We want to develop an effective support tool that can be used as a developer-client interface in a project-management environment to improve SEs' communication and negotiation skills.

References

1. Anandasivam, G., Traids, M., Mayuram, S.K.: The Role of Software Processes and Communication in Offshore Development. COMMUNICATION OF THE ACM2002 4ve, 193–199 (2002)
2. S-open Off-shore Study: Offshore Software Development, Nikkei Business Publications Inc., 94–103 (2004)
3. Japan Information Processing Development Center: White Paper 2006, Edition, The BCN, pp. 113–123 (2006)
4. Harold, K.: Strategic planning for management using a project management maturity model, pp. 59–72. John Wiley & Sons, England (2001)
5. Project Management Institute: PMBOK® Guide - Third Edition, Project Management Institute, pp. 40–44 (2004)
6. Harold, K.: Advanced Project Management, pp. 471–480. John Wiley & Sons, England (2004)
7. Ikujiro, N., Hirotaka, T.: The Knowledge-Creating Company - How Japanese Companies Create the Dynamics of Innovation, pp. 38–45. Oxford University Press Inc., Oxford (1995)
8. Roger, C.S.: Dynamic Memory, pp. 20–40. Cambridge University Press, Cambridge (1982)

A Modelling Framework for Sharing Knowledge

Reyes Grangel, Ricardo Chalmeta, and Cristina Campos

Grupo de Investigación en Integración y Re-Ingeniería de Sistemas (IRIS), Dept. de Llenguatges i Sistemes Informàtics, Universitat Jaume I, 12071 Castelló, Spain
{grangel, rchalmet, camposc}@uji.es

Abstract. Enterprise Modelling can be used successfully for different purposes, which includes capturing enterprise knowledge. However, one of the weaknesses of Enterprise Modelling is the lack of strong links with software generation. Model Driven Engineering attempts to solve this situation by promoting the use of models and their transformations in the software development process. In this context, the use of enterprise models that are able to capture knowledge and help to implement Knowledge Management Systems would be an important step forward.

In this paper, we present a proposal for Enterprise Modelling focused on enterprise knowledge. It starts from the CIM level and follows a model-driven approach. The modelling proposal provides a conceptual framework that allows enterprises to share knowledge by using a defined UML2 Profile for Modelling Enterprise Knowledge.

Keywords: Enterprise Modelling, Knowledge Representation, Knowledge Modelling, Model Driven Architecture, UML.

1 Introduction

Enterprise Modelling can be used successfully for different purposes, such as capturing enterprise knowledge [1,2,3]. However, enterprise models are not normally used in these processes, due to the fact that one of the weaknesses of Enterprise Modelling is the lack of strong links to software generation [3,4,5].

Model Driven Engineering (MDE) approaches are a new paradigm in the context of Software Engineering. Such perspective attempt to improve the software development process by focusing on models as the primary artefacts and transformations as the primary operation carried out on models (which are used to map information from one model to another). As an example, Model Driven Architecture (MDA) defined by the OMG [6]. The main purpose of this approach is to separate the functional specification of a system from the details of its implementation on a specific platform. This architecture therefore defines a hierarchy of models from three points of view [3,6,7], namely: **Computation Independent Model (CIM)**, **Platform Independent Model (PIM)**, and **Platform Specific Model (PSM)**.

Enterprise models can be considered to be CIM models. Thus, new proposals for Enterprise Modelling that are focused on following a model-driven approach

B. Apolloni et al. (Eds.): KES 2007/ WIRN 2007, Part II, LNAI 4693, pp. 1230–1237, 2007.
© Springer-Verlag Berlin Heidelberg 2007

are needed in order to improve the connection between the CIM level and software generation. Based on this initial analysis, the problem dealt with in this paper is that of how to improve existing Enterprise Modelling proposals for capturing knowledge by following a model-driven approach, such as MDA, in order to implement a Knowledge Management System. The objective of this paper is, therefore, to present a proposal for Enterprise Modelling that is focused on knowledge, and to show how it is possible to capture this knowledge in models at the CIM level so that it can then be transformed into the PIM level.

In this section, the problem dealt with by the research presented in this paper is described. Section 2 outlines the main problems in the context of Enterprise Modelling, which concern the complexity of knowledge modelling and finally existing UML approaches to model enterprises and knowledge are examined. In section 3, the framework proposed for knowledge modelling is presented. This section also includes an explanation of the principles of the proposal and the mechanism used for modelling from technological point of view, that is to say, the defined UML2 Profile. Finally, section 4 outlines the main conclusions.

2 Modelling at the CIM Level

Modelling a system at the CIM level involves developing a model of it at a conceptual level by using the abstraction mechanisms of a specific language or formalism. This provides a defined set of constructs and rules in order to suppress certain details so that a simplified model can be established independently from the computation viewpoint.

2.1 Enterprise Modelling Perspective

Enterprise Modelling refers to the externalisation and expression of enterprise knowledge [1], which provides a holistic view of an enterprise and considers all its dimensions, i.e. process, decision, information, and so forth [8]. Nowadays, there are a great number of languages, standards, methodologies and their corresponding tools, which are classified as traditional Enterprise Modelling Languages (EMLs) in [3], and they cover different dimensions of the enterprise defined in GERAM [9], even overlapped. Moreover, other EMLs exist that have been created in order to make different kinds of exchanges easier, since interoperability problems are increasing among systems that use different EMLs [8]. This last category, among them UEML [4,10,11] and POP* [12,13], provide common exchange formats to smooth the exchange of enterprise models at horizontal level.

However, one of the main weaknesses of Enterprise Modelling is the lack of strong links between enterprise models and software generation. To solve this gap, one solution, as pointed out in [14], is that the role of enterprise models should be that of facilitating the design, analysis and operation of the enterprise according to models, i.e. it should be driven by models (model-driven). Nowadays, the model-driven approach is followed by numerous projects such as MODELWARE [15], ATHENA [12], and INTEROP [10] in the European Union, and Model Driven Architecture (MDA) [6], which is carried out by the OMG.

MDA is an emerging paradigm. A lot of work is being carried out within the OMG framework in relation to PIMs, PSMs, QVT, and so forth, but the characterisation of CIMs and the features that an enterprise model must satisfy in order to be considered a CIM and generate appropriate software are still in progress [3]. This gap is specially remarkable when the purpose of modelling is to capture and to make enterprise knowledge explicit.

2.2 Knowledge Complexity and Representation

But, what do we understand by 'enterprise knowledge'? First of all, there is no universally accepted definition of exactly what knowledge is. Some authors define it, for example, as the information individuals possess in their minds [16]. This definition is argued by saying that data (raw numbers and facts) exist within an organisation. After processing these data they are converted into information and, once it is actively possessed by an individual, this information in turn becomes knowledge. There are also other approaches to defining knowledge that are less dependent on the information technologies. One of the most cited is the approach proposed by Nonaka [17], who defines knowledge as the justified belief that increases the capacity of an entity to take effective action. Following this line of reasoning, knowledge can be seen from five different perspectives [18]: (1) as a state of mind, (2) as an object, (3) as a process, (4) as a condition for access to information, or (5) as a capability. In this context and based on our own empirical observations, we define knowledge as the awareness that enables us to possess the skill or the capacity required in a particular situation (1) to deal with and resolve complex issues in an efficient and creative manner, and (2) to take advantage of opportunities by making the most appropriate decisions; and, enterprise knowledge as the network of connections among data and information that gives the people involved in an enterprise and insight into its workings and enables them to act and to make decisions that add value to the enterprise [19].

A key factor for achieving correct Knowledge Management in an enterprise is the development and implementation of a special kind of Information System, called a Knowledge Management System (KMS). That is to say, a technological system that allows enterprise knowledge to be created, codified, stored and distributed within the organisation [20]. One of the weak points of these kinds of systems is the need to link the conceptual framework with the technological level, especially for knowledge representation [19].

2.3 UML for Knowledge Modelling

The Unified Modeling Language (UML) has become a standard visual language for object-oriented modelling that has been used successfully for modelling information systems in very different domains [21]. However, UML is a general-purpose modelling language that can also be useful for modelling other types of systems such as, for example, an enterprise [22,23]. Other works, such as [24], point out the possibility of using UML as a language for Enterprise Modelling. However, the benefits of model-driven approaches and the new specification of

UML2 provided by the OMG suggest the need to provide more practical examples for Enterprise Modelling with UML based on these recent works [25], and especially for Enterprise Knowledge Modelling. Furthermore, despite the fact that the weakness of the stereotype mechanism is pointed out in [26], the new specification of UML2 [21] provides profiles that are more complete than version 1.5 [27]. It will therefore be possible to customise UML in a better way [28].

The objective of the research presented in this paper was to consider the possibility of using UML as a knowledge representation language on the basis of two positive factors: first, that it is a visual language which has become a standard object-oriented language and thus there are a lot of tools available on the market; and, second, that it is commonly used by engineers in enterprises for software development. To make this possible, the capacity of UML2 to extend the language to a specific domain was used. A UML2 Profile for Modelling Enterprise Knowledge was then defined in an attempt to achieve a common understanding within the context of Enterprise Modelling. This profile takes into account enterprise dimensions and previous works leading to initiatives such as UEML and POP*.

3 Proposal for Enterprise Knowledge Modelling

3.1 Principles of the Proposal

The objective of the proposal presented in this paper is to represent Enterprise Knowledge at the CIM level and, thus, the result of using it in enterprises is a graphical model of the enterprise knowledge map that allows enterprises to share knowledge. In general terms, this proposal is based on Model Driven Engineering (MDE), which promotes the design and development of computer systems using different types of models built at different levels of abstraction. More particularly, it is also supported by the MDA defined by the OMG, which is an instantiation of MDE. According to this approach, the process of developing a computer system is based on the separation of the functional characteristics of the system from the details of its specification on a specific platform. The proposal follows this premise as a fundamental concept together with the following principles:

- **Model-driven approach.** The separation of the functional specification of a system from the specific kind of technology that will be used to develop it, improve interoperability, portability, maintenance and usability of the resulting system, in this case, a KMS.
- **Proposal focused on Enterprise Modelling.** Traditionally, Enterprise Modelling has been able to make enterprise knowledge explicit by modelling enterprise processes, products, organisation, etc. If correctly managed these models help enterprises in knowledge management, since graphical models provide a better understanding of enterprise functions and allow decisions to be made in a more efficient way. In this proposal, the traditional enterprise dimensions were adapted so as to take enterprise knowledge into account as a dimension in itself, which can be represented at the CIM level and then transformed at different levels of abstraction.

– **User-oriented modelling framework.** This proposal is carried out at the CIM level, as a starting point from which a KMS based on models has to be built. Bearing in mind that one of the prerequisites of Enterprise Modelling is to establish the objective and the scope of modelling, this proposal attempts to provide the simplicity that is needed at this level. In fact, the enterprise should be represented so that it can be understood by users not specialised in modelling and, at the same time, with a sufficient level of detail to allow the definition of requirements for the future computer system. Thus, the proposal presents a set of models oriented towards improving user understanding, but they also set down the bases for future transformation of the models developed at the CIM level towards lower levels.

3.2 Modelling Framework of the Proposal

In order to achieve the objectives detailed in the previous section, the proposal for Enterprise Knowledge Modelling makes use of the following components:

– A metamodel of Enterprise Knowledge at the CIM level, as well as diverse metamodels for representing the other enterprise dimensions such as process, product and so on.
– A UML2 Profile for Enterprise Knowledge Modelling.
– A guide that can help enterprises to use this profile with the objective of obtaining their knowledge map.

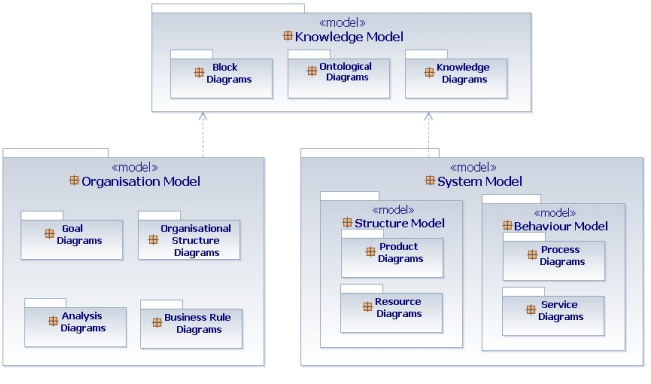


Fig. 1. Framework for Modeling Enterprise Knowledge at the CIM level

This general structure defines a framework for modelling enterprise knowledge on the CIM level at two levels of abstraction, which are required due to the great complexity of this level (see Fig. 1):

1. **Knowledge Model.** This corresponds to the top level of the model at the CIM level; the enterprise is represented from a holistic point of view, thus providing a general vision of the enterprise and its business that will later

be detailed in a local way in successive lower levels. Since this proposal is focused on representing enterprise knowledge, at this level the following elements are modelled: the conceptual blocks of knowledge defined by the enterprise, together with target knowledge defined for each block, the ontological categories that make it possible to connect to target knowledge, and the variables and knowledge sources needed for the extraction procedure.

2. **Business Model.** Here, the vision of the business is represented according to three types of models, and the main objective of these models is to represent a company from an organisational, structure and behavioural point of view, respectively. The **Organisational Model** is used to model objectives, organisational structures and business rules of an enterprise; the **Structure Model** represents enterprise products and resources; and, finally, the **Behaviour Model** shows how enterprise activities are carried out and information flows among them, that is, representing processes and services.

From a technological point of view (see Table 1), this proposal was implemented using the capacity of UML2 to extend a metamodel, that is to say, using a UML2 Profile. The UML2 Profile was defined for Enterprise Knowledge Modelling at the CIM level, following an MDA approach. This Profile is developed from the principles and the conceptual framework defined above. The profile provides the constructs needed to perform the models proposed earlier and it was implemented using IBM Rational Software Modeler. Finally, this modelling proposal can be applied in an enterprise following the KM-IRIS Methodology for the Implementation of KMS described in [29].

Table 1. Framework for Modeling Enterprise Knowledge from technological viewpoint

Abstraction Level	Metamodel	UML Profile	Model	Diagram
CIM-Knowledge	Knowledge	UML Profile for KM	Knowledge	Blocks Ontological Knowledge
CIM-Business	Organisation	UML Profile for GM UML Profile for OSM UML Profile for AM UML Profile for BRM	Organisation	Goals Organisational Structure Analysis Business Rules
	Structure	UML Profile for SM	Structure	Product Resource
	Behaviour	UML Profile for BM	Behaviour	Process Service

4 Conclusion

This research work intends to offer a systematic view of what Enterprise Knowledge is and how it can be modelled from a Model-Driven Engineering approach

in order to provide a conceptual framework of sharing knowledge. The idea is not to define yet another Enterprise Modelling Language, but to adapt and to extend an existing standard modelling language like UML, while also taking into account the work carried out in the context of Enterprise Modelling, such as UEML and POP*.

The model-driven approach followed by the proposal is a promising beginning, but to take advantage of it, further research is needed. This modelling framework was applied in a real Case Study on a audit company to test its feasibility, and a first set of improvements was obtained. However, further research will be needed in order to improve the proposal through the feedback from applications in other domains, and providing a method of transforming CIM models of Enterprise Knowledge proposed at the CIM level into PIM models.

Acknowledgments. This work was funded by CICYT DPI2003-02515 and DPI2006-14708, and the EC within the 6th FP, INTEROP NoE (IST-2003-508011). The authors are indebted to TG2 [10].

References

1. Vernadat, F.B.: Enterprise Modeling and Integration: Principles and Applications. Chapman and Hall, Sydney, Australia (1996)
2. Vernadat, F.B.: Enterprise modeling and integration (EMI): Current status and research perspectives. Annual Reviews in Control 26(1), 15–25 (2002)
3. Grangel, R., Chalmeta, R., Campos, C., Coltell, O.: Enterprise Modelling, an overview focused on software generation. In: Panetto, H. (ed.) Interoperability of ESA Workshops of the INTEROP-ESA International Conference EI2N, WSI, ISIDI and IEHENA 2005, Hermes Science Publishing (2005)
4. UEML: Deliverable 1.1. Unified Enterprise Modelling Language Project (IST-2001-34229) (2004), `http://www.ueml.org`
5. ATHENA: Deliverable DA1.3.1. Report on Methodology description and guidelines definition (2005), `http://www.athena-ip.org`
6. OMG: MDA Guide Version 1.0.1. Object Management Group. Document number: omg/2003-06-01 edn. (2003)
7. Berrisford, G.: Why IT veterans are sceptical about MDA. In: 2nd European Workshop on MDA with an emphasis on Methodologies and Transformations, Kent, pp. 125–135. Computing Laboratory, University of Kent (2004)
8. Chen, D., Doumeingts, G.: European initiatives to develop interoperability of enterprise applications–basic concepts, framework and roadmap. Annual Reviews in Control 27(2), 153–162 (2003)
9. IFIP-IFAC: Generalised enterprise reference architecture and methodology (GERAM). Technical Report Version 1.6.3 (1999), `http://www.cit.gu.edu.au/~bernus/taskforce/geram/versions`
10. INTEROP: Interoperability Research for Networked Enterprises Applications and Software NoE (IST-2003-508011) (2007), `http://www.interop-noe.org`
11. Opdahl, A., Henderson-Sellers, B.: Template-Based Definition of Information Systems and Enterprise Modelling Constructs. In: Ontologies and Business System Analysis, ch. 6, Idea Group Publishing (2005)

12. ATHENA: Advanced Technologies for interoperability of Heterogeneous Enterprise Networks and their Applications IP (IST-2001- 507849) (2007), http://www.athena-ip.org
13. Grangel, R., Chalmeta, R., Schuster, S., Peña, I.: Exchange of Business Process Models using the POP* Meta-model. In: Bussler, C., Haller, A. (eds.) BPM 2005. LNCS, vol. 3812, Springer, Heidelberg (2006)
14. Fox, M.S., Gruninger, M.: Enterprise Modelling. AI Magazine 19, 109–121 (1998)
15. MODELWARE: Modeling solution for software systems Project (IST-2004-511731) (2007), http://www.modelware-ist.org/
16. Dretske, F.I.: Knowledge and the Flow of Information. MIT Press, Cambridge, MA (1981)
17. Nonaka, I., Takeuchi, H.: The Knowledge Creating Company: How Japanese Companies Create the Dynamics of Innovation. Oxford University Press, Nueva York (1995)
18. Alavi, M., Leidner, D.: Knowledge Management and Knowledge Management Systems: Conceptual Foundations and Research issues. Management Information Systems Quarterly 25(1), 107–136 (2001)
19. Grangel, R., Chalmeta, R., Campos, C.: Defining of Target Knowledge in Virtual Enterprise. In: Abramowicz, W., Mayr, W. (eds.) Business Information Systems - 9th Int. Conf. on BIS 2006, pp. 256–266. Gesellschaft für Informatik (GI) (2006)
20. Day, R.: Totality and representation: A history of knowledge management through European documentation, critical modernity, and post-Fordism. Journal of the American Society for Information Science and Technology 52(9), 725–735 (2001)
21. OMG: UML 2.1.1 Superstructure Specification. Object Management Group. version 2.1.1 formal/07-02-05 edn. (2007)
22. Eriksson, H., Penker, M.: Business Modeling with UML: Business Patterns at Work. J. Wiley, New York (2000)
23. Marshall, C.: Enterprise Modeling with UML. Designing Successful Software Through Business Analysis. Addison-Wesley, Reading (2000)
24. Panetto, H.: UML Semantics Representation of Enterprise Modelling Constructs. In: ICEIMT, pp. 381–387 (2002)
25. Grangel, R., Bourey, J-P., Chalmeta, R., Bigand, M.: UML for Enterprise Modelling: basis for a Model-Driven Approach. In: Doumeingts, G., Müller, J., Morel, G., Vallespir, B. (eds.) Enterprise Interoperability. New Challenges and Approaches (I-ESA'06), pp. 91–102. Springer, Heidelberg (2007)
26. Berio, G., Petit, M.: Enterprise Modelling and the UML (sometimes) a conflict without a case. In: Proceedings of 10th ISPE International Conference on Concurrent Engineering: Research and applications, pp. 26–30 (2003)
27. OMG: OMG Unified Modeling Language Specification, version 1.5. Object Management Group. formal/03-03-01 edn. (2003)
28. Fuentes, L., Vallecillo, A., Troya, J.: Using UML Profiles for Documenting Web-Based Application Frameworks. Annals of Software Engineering 13, 249–264 (2002)
29. Chalmeta, R., Grangel, R.: Methodology for the Implementation of KMS. evaluation: Journal of the American Society for Information Science and Technology (2007)

An Assistant Interface for Finding Query-Related Proper Nouns

Tomoya Iwakura, Kanji Uchino, and Seishi Okamoto

Fujitsu Laboratories Ltd.
1-1, Kamikodanaka 4-chome, Nakahara-ku, Kawasaki 211-8588, Japan
{iwakura.tomoya,kanji,seishi}@jp.fujitsu.com

Abstract. Searching information on the Web is now one of the important acts in dairy work because we can obtain several types of information from the Web. For example, we use web search engines to obtain information related to proper nouns, such as famous persons, shops, restaurants, sightseeing places and so on. However, we often obtain irrelevant information with what we want to obtain because most web search engines just return web pages including a posted query. This paper describes a web search engine interface for users finding proper nouns and their relevant information. Our interface selects proper nouns related to a query from a search result with Natural Language Processing technologies. Our interface provides users the following information in addition to original search result: 1) Lists of proper nouns related to a posted query. 2) Links to each proper noun to results includi ng the proper noun. 3) Links for searching new information by using proper nouns as a new query. The experimental results with four subjects have shown that our interface contributes to find target proper nouns.

Keywords: Search Engine Interface, Named Entity Extraction.

1 Introduction

Finding information on the Web is now one of the important roles for business, research fields, private activities, and so on. To find information on the Web, we use web search engines. A result of a survey about frequencies of using web search engines showed that 95.4% people uses search engines more than one time in each day, and 44% people uses search engines more than four times [1].

One of the purposes using web search engines is for finding proper nouns and their relevant information. For example, if we go a trip to *Kamakura*, where one of the famous sightseeing spots in Japan, we would post a query like "sight spot and *Kamakura*" for finding good spots to be visited around the Kamakura. However, we often obtain irrelevant information with what we want to obtain because there are several web pages satisfying a query.

In this paper, we describe a search engine interface for users finding proper nouns and their relevant information. To find proper nouns related to a posted query, our interface uses Natural Language Processing (NLP) technologies for selecting them from search results. Furthermore, our interface provides lists of proper nouns, links to results including proper nouns, and links for finding new information with proper nouns as new query to users of our interface.

B. Apolloni et al. (Eds.): KES 2007/ WIRN 2007, Part II, LNAI 4693, pp. 1238–1245, 2007.
© Springer-Verlag Berlin Heidelberg 2007

This paper is organized as follows. Next section describes an overview of our interface. Section 3 describes NLP technologies used in our interface, and Section 4 describes our method for finding proper nouns related to each search query. Section 5 reports experimental results. Section 6 shows related works, and we conclude in Section 7.

2 Search Engine Interface

We describe our search engine interface as follows. Let q be a search query and $R(q)_i$ (for $1 \leq i$) be i-th web page in a search result obtained with q. We denote a set of m web pages obtained by posting q to a web search engine as $\mathbf{R}(q) = \{R(q)_1, ...R(q)_m\}$. We assume that $R(q)_i$ consists of the title, the link to the web page, and the summary of $R(q)_i$ which describe the part of the content. We denote the summary of $R(q)_i$ as $S(q)_i$.

In addition to $\mathbf{R}(q)$, our interface provides users information related to proper nouns. Our interface extract proper nouns related to a query, which we call relevant proper nouns, from $\mathbf{S}(q) = \{S(q)_1, ..., S(q)_m\}$ with NLP technologies. Our interface selects relevant proper nouns as follows.

Step 1: Receive a query q and the class c of proper nouns to be selected from a user
Step 2: Search web pages by positing q to a web search engine and obtain $\mathbf{R}(q)$[1]
Step 3: Extract proper nouns $\mathbf{pn} = \{pn_1, ..., pn_n\}$ with their classes from $\mathbf{S}(q)$: We present a NLP technology for extracting proper nouns in Sec. 3.
Step 4: Select proper nouns satisfying one of the conditions and having class c as relevant proper nouns for q: The relevant proper nouns denoted as $\mathbf{rpn} = \{rpn_1, ..., rpn_l\}$ and rpn_k (for $1 \leq k \leq l$) denotes the k-th relevant proper noun to q. We present a selecting method for relevant proper nouns in Sec. 4.
Step 5: Create information by adding them into original search result $\mathbf{R}(q)$

Fig. 1 shows snapshots of our search engine interface. Our interface adds the following information at the Step. 5.

- Lists of candidate proper nouns related to a posted query: Our interface provides users relevant proper nouns $rpn_k \in \mathbf{rpn}$, for ($1 \leq j \leq l$), in ascending order. Let rpn be a relevant proper noun in $\mathbf{rpn}$ and $S_(q)$ be a summary, then we denote $rpn \subseteq S_(q)$ if $S_(q)$ contains rpn. We present $rpn_k \in \mathbf{rpn}$ with summaries satisfying $rpn_k \subseteq S_(q)_i$, for $1 \leq i \leq m$. Underlined strings next to "Candidate 1" and "Candidate 2" in top window of Fig. 1 are person names judged as relevant proper nouns (i.e., rpn_1 and rpn_2). The strings between "Candidate1" and "Candidate 2" are summaries satisfying $rpn_1 \subseteq S_(q)_i$ for $1 \leq i \leq m$.
- Lists of all the proper nouns extracted from summaries: Our interface lists all the proper nouns in $\mathbf{pn}$ by their class. Rightmost boxes in top window of Fig. 1 is a list of proper nouns that class is PERSON (i.e., $\{pn | pn \in \mathbf{pn} \wedge class(pn) = c\}$, where $class(pn)$ is the class of a proper noun pn).

[1] We use Nifty@Search for searching web pages. We experimentally use top 40 results listed in the search results.

– Links for moving search results: Our interface creates hyper links for moving results including a proper noun. For example, if a proper none pn is extracted from $S(q)_i$ that is the summary of i-th result in $\mathbf{R}(q)$, a link to move to the place presenting $R(q)_i$ is created. The link represented i is put to the next of pn. User can move i-th result $R(q)_i$ by clicking i that is the next of the proper noun. Underlined strings on the windows in top window of Fig. 1 are proper nouns and the next underlined integers are hyperlinks. The bottom window in Fig. 1 is a snapshot after clicking the link 2 next to "Ronaldo" in the top window.
– Links for next search: Our interface assigns hyper links to extracted proper nouns for finding new pages by using proper nouns as a new query. For example, if we click "Ronaldo" in the top window in Fig 1, our interface posts "Ronaldo" to a web search engine and return the search results with those information added by our

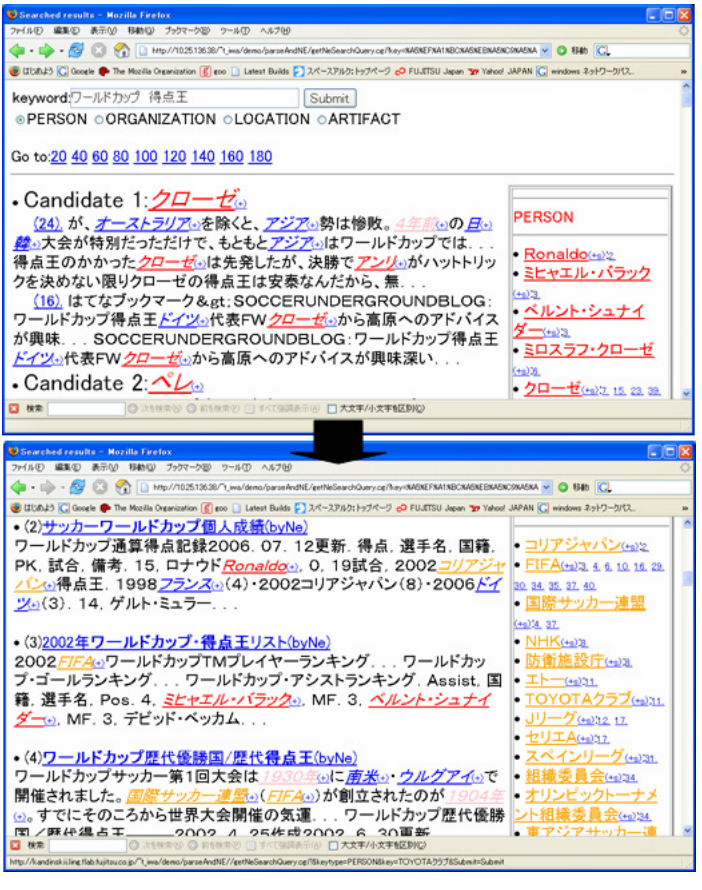


Fig. 1. Snapshots of our search engine interface: The search results were obtained with the query of "World Cup & goal-scoring leader". The right-hand side in the top window presents lists of PERSON related to the query. The bottom window is a snapshot after clicking a link to move a result.

Table 1. NE classes defined by IREX and the examples

NE Class	ARTIFACT	LOCATION	PERSON	ORGANIZATION
Example	Nobel Prize in Chemistry	Japan	the Ministry of Foreign Affairs	Kouichi Tanaka

interface. If we click the "(+)" next to "Ronaldo", our interface posts "Ronaldo" with "World Cup & goal-scoring leader", which is previous search query, to a web search engine.

3 Natural Language Processing Technologies

Our interface consists of a Named Entity (NE) extractor and a Japanese dependency parser for finding relevant proper nouns.

3.1 Named Entity Extractor

One of the problems for finding proper nouns in text is that we have to correctly discriminate the meaning of words. For example, most Japanese location names are used as Japanese family names.

In order to correctly find proper nouns with their classes from text, we use a Named Entity (NE) extractor extracting proper nouns and classify the word chunks into their classes, such as persons, locations, organizations. Table 1 lists NE classes and their examples used in our interface. In this paper, we extract proper nouns and their class with an NE extractor. We briefly present our NE extraction method as follows.

To extract NEs, we use a method that classifies words into NE chunk labels by using context information because NEs consist of one word or more than one word. We used Start/End (SE) representation that uses the five tags, S, B, I, E and O, for representing word chunks becoming NEs [2]. S is used for each word becoming a chunk by itself. B is used for each word becoming the start of a chunk consisting of more than one word. E is used for each word becoming the end of a chunk consisting of more than one word. I is used for each word becoming the inside of a chunk consisting of more than two words. O is used for each word existing the outside of any chunk.

We used the five tags with NE class labels as NE chunk labels for representing NEs in text. For example, the person names and the outside of person names in

... Mr. Michael W. White said to Mr. Brow ...

are represented by using NE chunk labels as follows.

... Mr./O Michael/B-PER W./I-PER White/E-PER said/O to/O Mr./O Brown/S-PER ...

B-PER, I-PER and E-PER indicates the beginning, inside, and end of a person name, respectively. S-PER indicates a person name consisting of only one word. In this paper, we defined 17 labels $(4 * 4 + 1)$ for extracting 4 types of NEs.

When extracting NEs from a sentence s consists of N words $\{w_1, ...w_N\}$, an NE extractor classify each word in s into an NE label with following features.

- Words and Part of speech (POS) tags: When classifying i-th word in s, we use the POS tag p_i of w_i, and the surrounding words (w_{i-2}, w_{i-1},w_{i+1} w_{i+2}) and their POS tags (p_{i-2}, p_{i-1},p_{i+1} p_{i+2}). In addition to them, we use character types of $\{w_{i-2}, ..., w_{i+2}\}$, such as 'Hiragana', 'Katakana', 'Chinese letter', 'capitalized alphabet', 'digit', 'sign' and these combination. We used a morphological analyzer ChaSen for word segmentation and assigning POS tags into each word. [2]
- Dictionaries: We used word information consisting of candidate NE labels of words and candidate NE labels of co-occurring words. We collected those information from news articles and web pages by using several NE extractors [3].

We created the NE extractor from a training corpus for NE extraction with a boosting algorithm [4]. Please refer to the paper [4] for the detailed explanation of boosting algorithms[3].

3.2 Japanese Dependency Analyzer

We use a Japanese dependency analyzer for finding proper nouns related to each search query. Japanese dependency parsers are used to find a relation between two bunsetsu. Each bunsetsu represents Japanese base phrase information. Fig 2 shows an example of a result parsed by a Japanese NE extractor and a Japanese dependency parser. We have developed a Japanese dependency parser based on the algorithm proposed by [5]. Please refer to the paper [5] for a detailed explanation. We use the boosting algorithm [4] for creating the Japanese dependency parser.

4 Finding Proper Nouns Related to Queries

This section presents how our search engine interface selects relevant proper nouns, that are proper nouns related to query. A query may consist of terms more than one. In our interface, we find NEs related to a term in a query as relevant proper nouns. We select relevant proper nouns with the following conditions.

(a) Finding sub-trees consisting of two bunsetsu, the mother bunsetsu has two daughter bunsetsu. The one bunsetsu includes NE, the other includes a term in a query.
(b) Finding sub-trees consisting of two bunsetsu, the mother bunsetsu includes NE, and the daughter bunsetsu includes a term in a query.
(c) Finding sub-trees consisting of two bunsetsu, the mother bunsetsu includes a term in a query, and the daughter bunsetsu includes NE.
(d) Finding bunsetsu including NE and a term in a query

Fig. 3 shows images of these conditions. By using dependency information between two bunsetsu, we can find NEs related to a query if NEs appear far from a term in the query. For example, when we find an NE related to " (birth day)" from the sentence

[2] http://chasen.naist.jp/hiki/ChaSen/

[3] Although the boosting algorithm is designed for multi labeling, we deal with extraction of several NE types by One-vs-Rest method that prepare N classifier for a N class problem. The reason is that we can train classifiers for each classifier independently in One-vs Rest method.

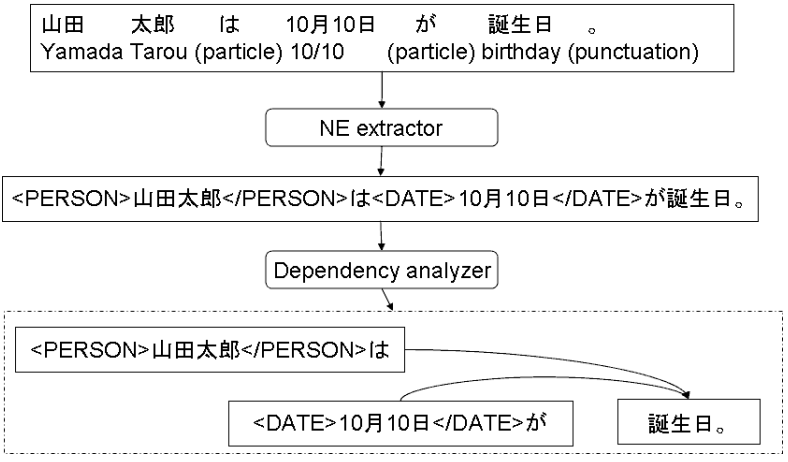


Fig. 2. An example for parsed results. The sentence, "1010", means "Yamada Taro's birthday is October 10th." Strings bracketed by ⟨⟩ indicate NE, and strings in ⟨⟩ indicate class of NE. Each box under in the dashed box indicates bunsetsu. Each line in boxes of bunsetsu indicates a dependency relation.

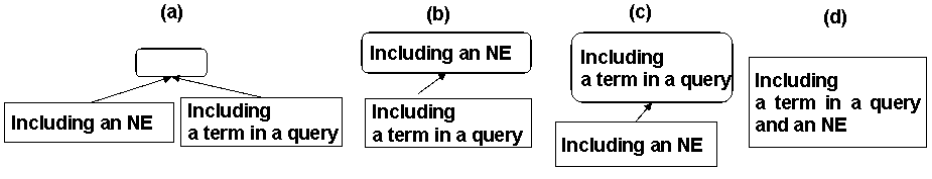


Fig. 3. Conditions for selecting NEs related to a query. Ellipses and boxes indicates mother bunsetsu and daughter bunsetsu.

described in Fig. 2, " (Yamada Tarou)" and "*1010* (10/10)" are extracted by condition (b) as NEs related to " (birthday)".

We select relevant proper nouns as follows. Let pn be a proper noun and $|pn|$ be the number of summaries including pn while satisfying one of these conditions (a) to (d). Our interface uses $|pn|$ as the criterion for selecting proper nouns related to a posted query. If $|pn_j|$ for $pn_j \in \mathbf{pn}$ is k-th largest, then we select pn_j as k-th relevant proper noun rpn_k. For example, if a set of proper nouns $\{pn_1, pn_2, pn_3\}$ are given with $|pn_1|$=1,$|pn_2|$=2, and $|pn_3|$=0, then the relevant proper nouns selected from them are $rpn1 = pn_2$ and $rpn2 = pn_1$.

5 Experiments

We evaluated effectiveness of our search engine interface by comparing the times for solving the following questions.

Table 2. Experimental results: Times for finding the answer for each question. Time limit for finding the answer for each question is 10 minutes. Av. time indicates average times for solving the four questions. "-" indicates that a subject knew the answer in advance.

Interface	Subject\Question	Q1	Q2	Q3	Q4	Av. time
Our interface	A	0:45	1:12	-	2:38	1:32
	B	4:42	2:31	6:06	1:30	3:45
Nifty@Search	C	10:00	0:39	0:30	8:13	4:50
	D	3:48	6:30	9:41	10:00	7:30

- Q1: Which food company provides a telephone service for folk stories?
- Q2: What is the address of the office of Japanese Prime Minister?
- Q3: There is a story that central character is a bear. What is the name of the bear? The name of the bear is the same as a station name in British.
- Q4: Who composed a music in baroque period that Japanese title is "Tameiki (Les Soupirs)"?

We evaluated our interface with four subjects. Two subjects used our interface and the others used Nifty@Search for solving these questions. We set the time limit for solving each query to 10 minutes. The times for processing search results by NLP technologies are included in the evaluation. Before starting experiments, all the subjects solve a review exercise.

Table 2 lists times for solving each question by each subject. These results showed that two subjects, A and B, using our search engine interface could answer all the questions. Subject C failed to answer Q1, and Subject D failed to answer Q4. All subjects posted similar queries for solving Q1 and Q4. However, Web pages including those answers for Q1 and Q4 were ranked around 20 to 30 in original results returned by the web search engine. By contrast, our interface listed answer proper nouns as related to a query. As a result, subjects using our interface could find answers from proper nouns listed as related to the posted query.

Subject C showed the best performance on Q2. Q2 is the easiest question because all the subjects answered correctly. A reason why subject C showed better performance seems to be response times of our interface. Subject A, B and C used almost the same queries for solving Q2. However, our interface requires about 5 to 20 seconds for returning results to users compared to only about 3 second for Nifty@Search. Thus, subjects using our interface have to wait for obtaining search results. Some reasons requiring much time are processing times required for NE extraction, dependency analysis, and parsing for html files, and so on. These showed that our interface would have disadvantage for finding information that are ranked in higher by search engines.

6 Related Work

Several methods are researched for finding information what users want to know from the web. The first is Question Answering (QA) technologies [6]. QA systems assume that each query is a sentence or sentences written in natural language, and find answers

related to the query by the following steps. The first is identification of types of query sentences, such as for finding person names, organizations, and so on. The second is finding answer corresponding to types of queries.

The another method is discrimination of multi-sense information that users want to know. Pedersen and Kulkarni proposed a method for discriminating person names in Web contexts [7]. They used unsupervised clustering techniques to discriminate among entities named in Web pages.

Meta search engines are also interfaces for searching information on the Web [8]. Meta search engines post a query to multiple search engines and gather the results returned by all the search engines. Next, they merge all the results and present the merged results users as search results.

Our interface has following characteristics. The first is that our approach focuses on adding information to existing search engines. Our interface finds NEs related to a query from search results obtained with web search engines. Thus, users can use our interface like using existing search engines. Furthermore, users can find a specific class of NEs by checking a radio button corresponding to the NE class. The other is that our interface adds hyperlinks for finding further information and moving search results appearing NEs in addition to NEs related to a query.

7 Conclusion

We have presented a web search engine interface for finding proper nouns from the Web. The experimental results solving four questions with four subjects have shown that our interface contributes to find objective proper nouns. Future work should evaluate our search engine interface with more subjects and questions.

References

1. Impress R&D: Internet Hakusyo (in Japanese) (2004)
2. Uchimoto, K., Ma, Q., Murata, M., Ozaku, H., Utiyama, M., Isahara, H.: Named entity extraction based on a maximum entropy model and transformati on rules. In: Proc. of the ACL 2000, pp. 326–335 (2000)
3. Iwakura, T., Okamoto, S.: Improving named entity extraction accuracy using unlabeled data and sev eral extractors. In: Gelbukh, A. (ed.) CICLing 2007. LNCS, vol. 4394, Springer, Heidelberg (2007) (to appear)
4. Schapire, R.E., Singer, Y.: Boostexter: A boosting-based system for text categorization. Machine Learning 39(2/3), 135–168 (2000)
5. Sassano, M.: Linear-time dependency analysis for japanese. In: Proc. of Coling 2004, COLING, pp. 8–14 (2004)
6. Voorhees, E.M, D.H.: Overview of the trec 2005 question answering track. In: The Fourteenth Text REtrieval Conference (2005)
7. Pedersen, T., Kulkarni, A.: Unsupervised discrimination of person names in web contexts. In: Gelbukh, A. (ed.) CICLing 2007. LNCS, vol. 4394, pp. 299–310. Springer, Heidelberg (2007)
8. Selberg, E., Etzioni, O.: Multi-service search and comparison using the MetaCrawler. In: Proc. of the 4th International World-Wide Web Conference (1995)

Community Support System for Extensive Readers: Case Study in a Domain 'Knowledge Science'

Kohei Ishikawa[1,2], Kanayo Ogura[1], and Kozo Sugiyama[1]

[1] Japan Advanced Institute of Science and Technology
[2] ABeam Consulting
koheiis1@goo.jp, {k-ogura, sugi}@jaist.ac.jp

Abstract. Reading a large number and many kinds of books, or "extensive reading", is desirable for people who want to learn an interdisciplinary field. However, it is difficult for learners to continue the extensive reading individually for a long term due to the publication of too many books: 17,000/ year in Japan (see Fig.1). If a community of such learners united by shared interests can be formed, it is expected that each learner can be affected mutually in reviewing a lot of materials and finding books suitable for one's interests. The purpose of this research is to propose a method regarding how to form and activate extensive readers' community and to develop a proto-type community site supporting a community for extensive reading by using information sharing techniques.

Keywords: Extensive reading, Community support, Book map.

1 Introduction

As an interdisciplinary domain, we adopt "knowledge science" that is a newly developed domain aiming at the extensive integration of existing disciplines such as management, informatics, humanities, etc. from a viewpoint of 'knowledge creation'. A physical book titled "Knowledge Science [1]" has been published and its



Fig. 1. Difficulty in continuing extensive reading individually without navigation or guiding

B. Apolloni et al. (Eds.): KES 2007/ WIRN 2007, Part II, LNAI 4693, pp. 1246–1252, 2007.
© Springer-Verlag Berlin Heidelberg 2007



Fig. 2. Physical version and Web-version of 'Knowledge Science'

Web-version [2] developed with HTML has also been opened to the public (see Fig.2). Both of them have hypertext structures.

2 Research Purposes and Proto-type System

There are three purposes in supporting a community of extensive readers:

- Learners can know interests, ways of reading, understanding, and comprehension, and interpretation of other learners.
- Learners can find books that stimulate their favorites.
- Learners can continue extensive reading and increase the amount of reading.



Fig. 3. Facilities of our proto-type system

To attain these purposes, we have implemented a community site with the following facilities (see Figs.3-7) based on extensive surveys [3,4,5].

- Book Link: a facility for reviewing registered books and adding comments about them.
- Weblog: a facility for activating communications among learners.
- Bulletin Board: a facility for the user to discuss about specified topics.
- Book Map: a facility for the user to make a map of relationships among books that expresses his/her conceptual understandings.



Fig. 4. Snapshot of the screen for Book Link



Fig. 5. Snapshot of the screen for Weblog



Fig. 6. Snapshot of the screen for Bulletin Board

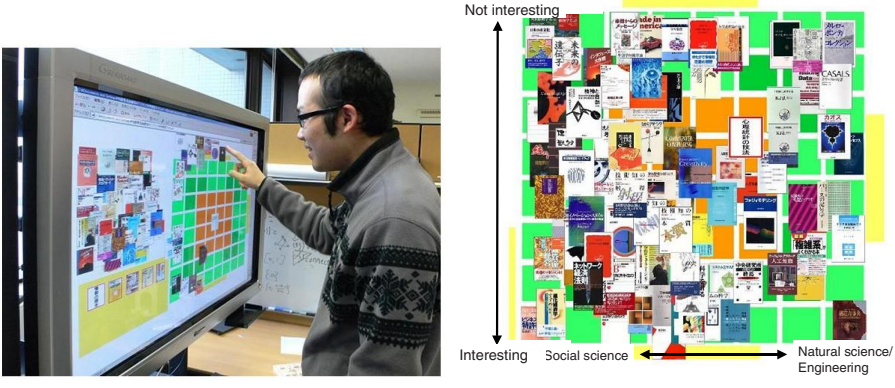


Fig. 7. The user and a screen for Book Map

3 Evaluation Experiments

Evaluation experiments of this system are carried out. Community Site supports communication among users and Book Map supports concept arrangements. In order to evaluate the effectiveness of Community Site and Book Map separately, we distinguish and compare two types of system utilization: using only Community Site and using both Community Site and Book Map.

Subjects for the experiments include 10 graduate students in the School of Knowledge Science, Japan Advanced Institute of Science and Technology, who are

divided into two groups A and B, each of them consisting of 5 members. Members of Group A (two extensive readers and three non-extensive readers) use only Community Site for a month, whereas members of Group B (three extensive readers and two non-extensive readers) use only Community Site for the first 15 days and both Community Site and Book Map for the last 15 days. Each student is requested to register at least 5 books, and to use the site at least for 5 minutes a day. The maximum numbers for registrations, login time and login length are not limited: users can utilize the site freely.

We made questionnaires (5 level rating from 1.0 (worst) to 5.0 (best)) about the following questions with members of both groups in order to compare the facilities for A and B:

1. Can users come to understand own concept and knowledge systematically?
2. Can learners have interests in book information from other users?

The result is shown in Table 1 where the evaluation average by group B is better than that by group A in both questions. This means that the facility of Book Map is effective.

Table 1. Comparison of answers for two questions for A and B groups

Question	Group A (average)	Group B (average)
Can users come to understand own concept and knowledge systematically?	3.2	4.6
Can learners have interests in book information from other users?	3.4	4.4

Then we made another questionnaires for group B and the analysis of video records when users of group B were being arranged book maps to identify effective items of book mapping.

1. Through seeing books maps others formed, can users systematically learn the book information that another user is reading?
2. Through forming book maps, does it encourage understating own concept and knowledge systematically? (Question to all members of group B)
3. Through forming book maps, does it encourage understating own concept and knowledge systematically? (Question to only extensive reading users of group B).

Table 2 shows the evaluation averages for three questions. We can see that average values for every item are high, especially for extensive readers. This means that the facility for Book Map is remarkably effective for supporting the community.

Table 2. Effectiveness of book mapping

Question	Average
Through seeing book maps that others formed, can users systematically learn the book information that another user is reading?	4.4
Through forming book maps by oneself, does it encourage understating own concept and knowledge systematically? (Question to all members of group B)	4.4
Through forming book maps by one self, does it encourage understating own concept and knowledge systematically? (Question to only extensive reading users of group B)	5.0

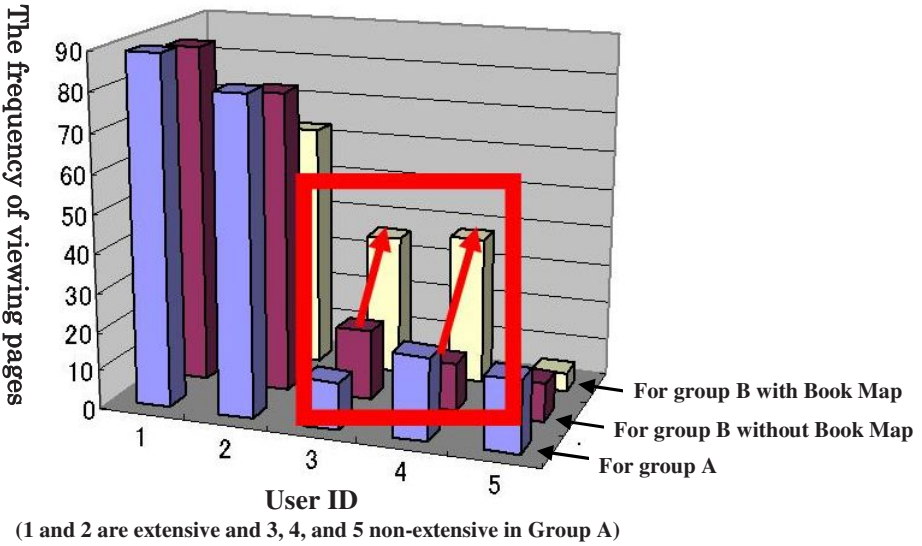


Fig. 8. Comparison of frequency for page-viewing between A and B, and extensive users and non-extensive users

The Fig 8 shows the frequencies (average viewing times per week) of viewing pages in the community site by users for group A and group B. For group B, frequencies before using the facility of Book Map and after using it are distinguished. Users 1 and 2 are extensive readers, and 3, 4, and 5 are non-extensive readers for group A, while users 1, 2, and 3 extensive, and 4 and 5 non-extensive for group B. From Fig 8, we can see the followings:

Extensive readers of both groups A and B have high frequencies of viewing pages while non-extensive readers of both A and B have low frequencies in general.

It is very interesting that users of group B whose frequencies before using Book Map are low, become frequently-viewing users after using Book Map (see a part enclosed by a red-line square in the figure).

4 Conclusions

As the results, we have the following conclusions:

It is attained that the community support system can support for extensive readers to use it continuously. (According to the result of the questionnaire, the system can activate communications and encourage the user to continue using the system. In addition, it is suggested that Book Map function raises interests about other users' concerns in the community.)

2. Through the use of our system, the difficulty for extensive readers to understand comprehension and interpretation of books by other users can be reduced. (According to the results of the questionnaire, the ability to see another user's book evaluation helps them to find new reading materials.)

References

Sugiyama, K., Shimojima, A., Nagata, A.: Knowledge Science. Kinokuniya-shuppan(2002) (in Japanese)

Knowledge Science Web Version: http://www.kousakusha.com/ ks/index.html

Stahl, G.: Group Cognition: Computer Support for Building Collaborative Knowledge (Acting With Technology). MIT Press, Cambridge (2006)

Barab, S.A., Kling, R., Gray, J.H.: Designing Virtual Communities in the Service of Learning (Learning in Doing). Cambridge University Press, Cambridge (2004)

Sawyer, R.K.: The Cambridge Handbook of the Learning Sciences. Cambridge University Press, Cambridge (2006)

Design and Evaluation a Knowledge Management System by Using Mathematical Model of Knowledge Transfer

Kouji Aoyama, Takanori Ugai, and Jun Arima

Fujitsu Laboratories Ltd.
4-1-1 Kamiodanaka Nakaharaku Kawasakishi Kanagawa 211-8588 Japan
{aoyama.kouji, ugai, arima.jun}@jp.fujitsu.com

Abstract. Based on our experiences, we have proposed a mathematical model for knowledge transfer in order to make knowledge management mechanism or system take root in the organization and to obtain guidelines to make it work. We are developing a know-how sharing system designed based on the insight obtained from the proposed model. We derived and applied the two ideas as design guidelines based on the analysis using the proposed model: one is mutual reviewing to increase the sense of participation, and another is establishment of the criteria to evaluate the background information about the knowledge to be shared. In this paper, we explain the proposed mathematical model and the system design based on the model. And we describe the evaluation on the prototype system. It shows that the mathematical model could derive guidelines to make the KM system work well.

Keywords: knowledge management, knowledge transfer, knowledge sharing.

1 Introduction

In recent years Knowledge Management (KM) is getting greater and greater attention. It has become a common practice to assess corporate knowledge as a part of enterprise value. Utilization of knowledge has become important in beating the competitors. This is why KM systems have been introduced in many enterprises. However, we can find only a few successful cases. We have also experiences of failures. For example, a system could not accumulate enough volume of knowledge because of low sense of participation of users. And another system could not utilize accumulated knowledge because the users didn't understand the knowledge well enough to use it. On the other hand there is no generally effective formal KM analysis technique, which can be used 1) to judge whether the KM mechanism and system will work well before introducing them into the organization; 2) to compare different KM mechanisms; and 3) to analyze why and how it is not functioning well if an introduced KM mechanism is not working well. As a result, we have repeated similar failures.

Jun Ma, et al, [1] studied influences of organizational structure on knowledge transfer in organization and showed that knowledge transfer happens more in the

B. Apolloni et al. (Eds.): KES 2007/ WIRN 2007, Part II, LNAI 4693, pp. 1253–1260, 2007.
© Springer-Verlag Berlin Heidelberg 2007

organization with a flat structure than in the organization with a hierarchical structure. However, their work only treats the organizational structure so it is not enough to use to analyze a KM mechanism in an organization and to compare KM mechanisms. C. Loebecke, et al. [2] established a model to show how knowledge transfer takes place in co-opetition relationship based on game theory and made some suggestions about its management. But the model explains a specific situation where the relationship is co-opetition, and the model doesn't cover the many practical enterprise situations.

In order to solve the above problems we have proposed a mathematical model for knowledge transfer [3]. The model is based on our experiences applying KM system to enterprise environment, and one of the purposes is to obtain guidelines to make a KM mechanism or system take root in the organization and to make it work well. The model shows a knowledge transfer takes place depending on the balance between profit and cost on the provider and the recipient. Furthermore, we are developing a know-how sharing system designed based on the insight obtained from the proposed model. This system is intended to promote the circulation of knowledge among system engineers (SE). We derived the following two ideas as design guidelines based on the analysis with the proposed model and applied them to the system:

1) Mutual reviewing to increase the sense of participation;
2) Establishment of the criteria to evaluate the background information about the knowledge to be shared.

In this paper, the proposed mathematical model for knowledge transfer is explained in chapter 2 and the conditions where the KM mechanism works successfully are shown based on the model. In chapter 3, we explain the system design based on the mathematical model and explain about the evaluation on the prototype system. It will show that the mathematical model could derive guidelines to make the KM system function successfully.

2 Mathematical Model of Knowledge Transfer

There are various methods for knowledge transfer. For instance, the methods can include written or oral reports or those employing visuals, site tours and visits, personnel rotation, education, training and standardization, etc [4]. Whatever the method used, in considering knowledge transfer, we can divide the parties involved into two sides: the side which sends the knowledge and the one which receives the knowledge. In this paper, we define knowledge transfer by separating the parties into the one which provides the knowledge (Sender) and the one which receives the provided knowledge (Receiver).

The knowledge transfer is defined as a relationship between the sender and the receiver and we introduce the profit, the cost, and the barrier as factors that influence the transfer. Specifically, they are defined as follows (Table1).

Table 1. Three factors concerning knowledge transfer

	Profit	Cost	Barrier
Sender	Appreciation, reward, sense of contribution, accomplishment, satisfaction	Document creation time, explanation time, support time	Competitive relationship, level of trust, sense of uneasiness, success or failure experience, exhibitionism, sense of fellowship
Receiver	Improvement of work efficiency, improvement of success probability, satisfaction	Time spent in participating the meeting, conversion time	Competitive relationship, level of trust, sense of uneasiness, success or failure experience, exhibitionism, sense of fellowship

Profit: Direct benefit obtained by providing and receiving knowledge. The sender's profit (Ps) includes incentives, the improvement of the sender's own skills, etc. The receiver's (Pr) profit includes improvement of work efficiency and improvement in success probability as a result of receiving the knowledge, etc.

Cost: Cost required for providing and receiving knowledge. The sender's cost (Cs) includes the time required for creating documents, communicating verbally, the work to provide information beneficial for the receiver, etc. The receiver's cost (Cr) includes the work to interpret or convert the received knowledge in order to use it, etc.

Contextual difference (d) is considered as a factor that greatly influences the cost. According to Ishizuka [5], there are many indications that whether the transferred knowledge can fully demonstrate its power at its destination is dependent on the context of the individuals or organizations by which the knowledge was generated. Context means the backgrounds of the knowledge, such as tacit knowledge, situation, environment, logical thread or organizational culture, etc. Therefore the cost can be regarded as an increasing function of the contextual difference between the sender and the receiver. This means the greater the contextual difference is, the more work it takes, for example, to explain in order to close the gap, hence the greater cost. For instance, if you compare the case where one tries to convey the knowledge about a sales activities in one industry sector to a member belonging to the same team with the case where one tries to convey it to a team engaged in sales activities to a different customer in a different industry sector, the latter clearly requires more explanation for knowledge transfer compared with the former.

On the other hand, when the contextual difference between a sender and a receiver is large, the profit obtained from the knowledge tends to fall in general. This is because as the contextual difference becomes larger, it becomes less likely for a situation conducive to the utilization of the knowledge to be realized. In view of this,

in this paper, we consider the profit of the receiver (Pr) as the attenuation function of the contextual difference (d) between the sender and the receiver.

Based on this model which consists of the profit and the cost, if the profit is larger than the cost, knowledge transfer will actively take place while there is likely to be little knowledge transfer if the profit is smaller than the cost. However, generally, the condition alone does not seem sufficient to promote knowledge transfer. For example, simply giving incentives to increase profit does not drive people to actively provide knowledge to others. It is suspected that a factor other than the profit and the cost is involved. In this paper, we assume the factor as the barrier and define it as follows:

Barrier: something that influences knowledge transfer in the relationship with others or in the environment. It can give either positive or negative influence on the motivation to provide/receive knowledge. The barriers on the sender side (Bs) include factors that affect the motivation to provide knowledge, such as the level of trust in the knowledge receiver, (competitive/cooperative) human relationship with the receiver, etc. The barriers on the receiver side (Br) include factors that affect the motivation to acquire knowledge, such as the level of trust in or preconceived notion about the knowledge provider as a result of past success/failure experiences, the availability of the search on the system, the trouble it takes to listen to the knowledge provider, etc.

McClelland[6] classified human motivation into four types: achievement, power, affiliation and avoidance motives. They are respectively included in sense of achievement (Profit), exhibitionism (Barrier), sense of fellowship (Barrier), and failure experience (Barrier), etc.

Using the three factors including seven parameters that are Ps, Cs, Bs, Pr, Cr, Br, and d, we have modeled some aspects of knowledge transfer as follows:

- Case where knowledge is offered: $Ps - Cs(d) > Bs$
- Case where knowledge is received: $Pr(d) - Cr(d) > Br$

The left-hand side is called the sender's profit/loss and the receiver's profit/loss respectively. The above expressions mean, if more profit than barrier remains after the cost is subtracted from the profit, knowledge transfer will occur. We gave a formal definition of the plus (+)/minus (-) operators and the comparison operators (<,>) in Ugai[7]. But we treat the operators in the informal intuitional manner in this paper.

3 Design and Evaluation of the KM System Based on the Knowledge Transfer Model

3.1 Outline and Assumption for the Know-How Sharing System

We are developing a know-how sharing system and the way to operate it, with the aim being to accumulate system development tips and ideas invented by SE in order to allow a SE to utilize other SE's knowledge in their own work. This system handles knowledge about project management and communication. Some of the knowledge required for system development is difficult to be broadly transferred, because the

contextual difference is large depending on the packages or in the industry sectors of the customer's business. However, we believe knowledge about project management and communication can be utilized widely and it is possible to achieve major effects by sharing such knowledge across the organization. Furthermore, even if it is possible to get general knowledge of such kind through off-the-shelf books, we believe small tips found inside the organization will prove particularly effective.

3.2 Derivation of Design Guidelines Based on the Mathematical Model

We consider design by applying the mathematical model to the know-how sharing system. The sender of the knowledge is a SE and the receiver of the knowledge is another SE. And, knowledge transfer is done mutually and a sender of one piece of knowledge can be a receiver of another piece of knowledge. The contextual difference (d) is large, because SE's work is different depending on industry sectors of the customers' business.

The profit for the sender (Ps) is the sense of contribution resulting from the provision of the knowledge. The cost of the sender (Cs) is the time to convey the knowledge to the recipient who has a different context, while the barrier of the sender (Bs) is a sense of uneasiness that the knowledge can end up just being thrown to the recipient. On the other hand, the profit of the receiver (Pr) is the improved work efficiency thanks to the acquired knowledge. The cost of the receiver (Cr) is the time to interpret the knowledge generated in a different context, while the barrier of the receiver (Br) is the sense of uneasiness based on the past experiences.

Based on the above, we can conclude that the knowledge sharing system has the following issues:

- Because there is a large contextual difference (d) between SE, the cost to convey the context is heavy both on the sender and the receiver sides(Cs and Cr are large).
- The barrier of the sender (Bs) and reciever (Br) are large because of the past experiences.

Because of these reasons, for the knowledge sender, it is difficult for the remainder after the cost is subtracted from the profit to be greater than the barrier (Ps-Cs<Bs) and it can be concluded that there is not likely be a lot of knowledge provision. And for the knowledge reciever, it is also difficut (Pr-Cs<Bs). So, the knowledge transfer will not occur actively.

In order to solve the issue above, we considered the following two approaches as the guidelines to lower the cost and the barrier while increasing the profit for both the sender and the reciever:

- Mutual reviewing: Mutual review method where a SE reviews the knowledge provided by other SE while contributing SE's own knowledge (tips/ideas).
- Establishment of the criteria to evaluate the context: We assume the usage of the system is that SE read accumulated knowledge, and SE discuss based on it, and improve their work or add new tips. In order to do that it is important for the description of the knowledge to be understandable just by reading. Therefore, we decided to focus on making the knowledge description graphic ("vivid") and not to require the described knowledge to be useful in practice.

3.3 Re-analysis with the Use of the Mathematical Model for Design Guidelines

We re-analyze the profit, cost and barrier of both the sender and receiver of knowledge in the case where the design guidelines derived from the analysis based on the mathematical model. Since the reviewer comes between the sender and the receiver of the knowledge, it can be interpreted that knowledge transfer takes place in two stages and aforementioned three factors in that case become as shown in Table 2.

Table 2. Three factors after the design guidelines are introduced

	Profit	Cost	Barrier
SE (Sender)	early feedback, learning more knowledge from discussion, sense of contribution	Making a description in vivid	Sense of uneasiness about whether the knowledge will prove useful
reviewers	Sense of contribution	Time for reviewing	Give-and-take relationship
system	-	-	-
SE (Receiver)	Efficiency improvement	Interpreting time for knowledge	Sense of uneasiness about whether the knowledge will prove useful

Profit: Mutual reviewing adds profit for sender. So, the sender's profit (Ps) includes early feedback about the knowledge through the review and learning more knowledge through discussion with the reviewer. Furthermore, the sender can get specific evaluation based on the context evaluation criteria involved in the reviewing process. On the other hands, the reviewers' profit is the sense of contribution obtained by deepening the knowledge provided by the sender through the review and the dialogue. And the receiver's profit (Pr) is greater efficiency achieved by acquiring the knowledge.

Cost: The sender's cost (Cs) is making a description in vivid. And the reviewers' cost includes the reviewing time spent interpreting the provided knowledge and pointing out things that are difficult for the receivers to understand. And the receiver's cost (Cr) is time of interpreting the knowledge. However, the reviewer lowered the cost by the evaluating the context, so the evaluation on the context reduces the receiver's cost (Cr).

Barrier: The sender's barrier (Bs) is sense of uneasiness about whether the sender's knowledge will probe useful. Regarding the barrier for the reviewer, the system of mutual review is playing a role to lower the barrier. A reviewer can also become a sender. That means there is a case that the reviewer requests the sender to review. In that sense, the sender and reviewer are in a give-and-take relationship and that serves

to motivate the reviewer to do reviewing. And the receiver's barrier (Br) is also sense of uneasiness about whether the sender's knowledge will probe useful. Regarding the sense of uneasiness on both the sender and the receiver, the mutual reviewing gives some authority to the knowledge, serving to lower the barriers on both sides. So the mutual reviewing reduces the barrier of sender (Bs) and receiver (Br).

As above, the analysis based on the mathematical model shows that the mutual reviewing mechanism plays a role to increase the profit of sender (Ps), and lower the barrier of sender (Bs), hence it is inferred that knowledge provision will be promoted (Ps-Cs>Bs). Furthermore, regarding the knowledge reception, since the reviewing process serves to lower the cost (Cr) and the barrier (Br) of the receiver, it can be reasoned that it will promote knowledge transfer (Pr-Cr>Br). Because of the above, it can be envisioned that the know-how sharing system will work successfully by introducing the two design guidelines, which are mutual reviewing and the establishment of context evaluation criteria.

3.4 Evaluation on the Prototype System

We evaluated the guidelines derived from the mathematical model by using as the design guideline for a prototype system of know-how sharing system. We made an evaluation experiment involving about 20 engineers for 3 month up to September 2006. The prototype system has the functions that making entries tips and ideas about system integration, selecting reviewers, making entries reviewing comment, and integrating by using Puki-Wiki.

As a result, 80 cases of tips and ideas were accumulated. The provision of knowledge was done on the trial basis without any incentives except for a request for entries only once a week and done as an activity separate from the participants' main business. Still, within 24 hours of registration of each knowledge, approximately 80% of them received feedback of 3 or more comments as a result of reviewing. Furthermore, 11 people done knowledge provision in 20 people, and 17 people reviewed. So, we consider the guidelines worked successfully.

However, in this trial, we only evaluated whether SE provided their knowledge under the experimental environment. This is because there were no opportunities to utilize the knowledge with the trial period being 3 months and the trial small in scale.

As described above, we performed analysis, using the mathematical model for knowledge transfer, on the know-how sharing system design, studied guidelines to make the KM system work successfully and conducted evaluation of the prototype. As a result, it was learned that the design guidelines derived from the mathematical model were functioning successfully for knowledge .provision. This has allowed us to demonstrate the mathematical model could derive guidelines to make the KM system function successfully.

4 Summary

In this paper, we described the outline of the mathematical knowledge transfer model we have proposed, explained the design and prototyping of the KM system derived from analysis based on the model, and conducted evaluation on the prototype system.

We have successfully demonstrated that it is possible to derive the design guidelines using the model and that the guidelines are valid.

In our model, knowledge transfer is defined using three factors (profit, cost and barrier). Knowledge transfer actively takes place when the remainder after subtracting the cost from the profit is greater than the barrier. Furthermore, the contextual difference between the provider and the recipient of the knowledge affects the cost and the profit on the recipient side and serves as a parameter of the knowledge transfer.

Based on the analysis with the model, we derived two operation guidelines: 1) mutual reviewing to increase the sense of participation; 2) establishment of the criteria to evaluate the background information about the knowledge to be shared. Then we evaluated a prototype system which is know-how sharing system among SE and found out that the knowledge provision was successfully performed. As a result, it is shown that the mathematical model could derive guidelines to make the KM system function successfully.

However, the trial period was relatively short and the number of involved people was small. So we could not see opportunities to utilize the knowledge in the trial. In future, we will continue to run trials based on the prototype system, continuously observe how to use the transferred knowledge. And we are also planning to evaluate the KM system involving a greater number of people. Through the evaluation of the system we will verify the design based on the mathematical model.

References

1. Ma, J., Xi, Y., Chen, Y.: Simulate the Effect of Organization Structure on Knowledge Transfer. In: ISCIT2005 (2005)
2. Loebeck, C., Van Fenema, P.C., Powell, P.: Co-Opetition and Knowledge Transfer. DBAIS 30(2) (1999)
3. Aoyama, K., Ugai, T., Arima, J.: Mathematical model of knowledge transfer. In: National Conference of JASMIN 2006, Autumn, pp. 54–54. JSMIN (2006)
4. Garvin, D.A.: Building a Learning Organization, Diamond Harvard Business Review (March 2003)
5. Ishizuka, H.: Solutions for Knowledge Transfer Obstacles, Faculty of information and communication, Bunkyou University. Research information 33 (July 2005)
6. McClelland, D.C: Human Motivation. Cambridge University Press, Cambridge (1988)
7. Ugai, T., Aoyama, K., Arima, J.: A Mathematical Model of Knowledge Transfer and Case Studies. In: Seventh International Conference on Management (July 2007) (to be appeared)

Development of a Mimamori-Care System for Persons with Dementia Based on the Real World-Oriented Approach

Kenichi Nakagawa[1], Taro Sugihara[1], Hitoshi Koshiba[1], Ryozo Takatsuka[1], Naotaka Kato[2], and Susumu Kunifuji[1]

[1] Japan Advanced Institute of Science and Technology, Ishikawa, Japan
{nakagawa, sugihara, hkoshiba, ryozo-t, kuni}@jaist.ac.jp
[2] Industrial Research Institute of Ishikawa, Ishikawa, Japan
kato@irii.go.jp

Abstract. We have been developing a mimamori-care system to monitor persons with dementia (PWD) in group homes based on the real world-oriented approach. This approach is devised by the expansion of a creativity-supported framework between researchers and caregivers. A new finding was acquired from a joint collaboration of the opinion of caregivers and the insight of researchers. Initially, several technical problems were solved in the virtual group home laboratory at Japan Advanced Institute of Science and Technology. Subsequently, we introduced the system into a real group home, and at present, we continue to improve the system by analyzing the daily events at the site. Some useful effects of this system in dementia care have already been revealed; hence, caregivers can intervene in the problems faced by PWDs on time, and they can support PWD's independence without engaging in excessive care.

Keywords: groupware design, ubiquitous, dementia care, support system.

1 Introduction

Presently, Japan is moving towards a super-aged society. There are approximately two million PWD. Group homes are expected to provide facilities for PWD, and the number of these facilities has increased twenty-fold between 2000 and 2006. Despite the urgent need for care support systems and eligible care workers in group homes, neither qualified persons nor any type of system has caught up enough with the rapid social change to support these facilities.

We endeavor to develop a mimamori-care system to monitor the behaviors of the residents of group homes and to protect them from accidents. "Mimomori" is a Japanese word meaning watching someone or monitoring something. "Mimomori-care" implies not only watching PWD but also supporting their independence to ensure adequate dementia care. A watch system for the elderly had been developed in the past but was not well accepted by PWD and caregivers due to privacy issues. Hence, it is necessary to solve these privacy issues and win their trust in order to induce them to use digital equipment. Further, it is also important to clarify the essence of dementia

B. Apolloni et al. (Eds.): KES 2007/ WIRN 2007, Part II, LNAI 4693, pp. 1261–1268, 2007.
© Springer-Verlag Berlin Heidelberg 2007

care. Hospitality and humanity constitute two pillars of caregiving. Therefore, the mimamori-care system should include these two components in the care of PWD as well as to assist caregivers. In order to solve the traditional problems and to ensure reliable and warm hospitality, we began working on a new concept; this system adopts the real world-oriented approach and is based on ubiquitous technology. The advancement of ubiquitous technologies such as IC tags has narrowed the gap between the real and virtual worlds, and different applications with IC tags are developed in various fields. We look forward to the development of this technology and anticipate that it will play a vital role in improving the interface. In the past, many technologies were developed considering only those people who can operate personal computers (PC). Caregivers are after all not expert PC users; therefore, it is imperative that they find it easy to handle operational systems. A few universities of welfare course implemented PC in their lectures. However, general education regarding dementia care assumes medical, psychological and philosophical approach. Since caregivers are taught that they should not rely on machines and that communication with patients is important, such ideas are deeply rooted in their beliefs.

Many researches and developments related to ubiquitous technology have been conducted. However, according to the valley of death [1] and the Darwinian Sea [2] metaphors, laboratory results cannot be implemented directly in the real world. There are methodologies such as human-centered design [3] and user-centered design [4] that reflect users' needs, improve usability, and facilitate modification in terms of custom-made design. However, dementia care is a difficult field. It depends heavily on contexts such as the person involved, the type of care work, the equipment used, and the environment. The methodology used in the past is insufficient. Moreover, it is critical to understand the essence of dementia care with respect to factors such as the independence of PWD and their quality of life. We define this process as "the real world-oriented approach" and develop its design, development, introduction, and operation.

In this research, we implement our system in a real group home facility. This paper describes the methods used to coordinate caregivers and engineers in developing a mimamori-care system. Subsequently, the effective outcomes of this research are revealed. This system mainly supports the following three points. (1) The quality of life (QOL) of PWD: Each PWD has his/her own personhood; this system is useful for person-centered care [5]. (2) Quality improvement of mimamori-care: Excessive care undermines the independence of PWD. The system maintains the balance between independence and care. (3) Efficient improvement of care work: Caregivers are engaged in both nighttime and daytime caregiving. This system is not intended as a substitute for caregivers. However, it contributes the ability to make better judgments and improvements in care.

2 Prototype System

This chapter describes the outline of the monitoring system. Usually, in an outline, newly used technologies are often described in a technical viewpoint. However, in designing the present system, we emphasized and implemented the viewpoints that the caregivers presented us with. Unless the technology and its users are mutually coordinated, the technology does not become an advantage for the users. Hence, there

should be a correspondence between the problem faced by the dementia caregivers and the functions implemented by the system.

2.1 Construction of the Virtual Group Home [6]

Both caregivers and PWD live together in group homes. Caregivers protect the lives of PWD by caregiving for them and helping them around the clock. In the experiment as well, the caregivers requested the safety and durability of the system. Therefore, the virtual group home was first constructed in a graduate school where its safety and efficiency was verified. The experimental environment in the graduate school allowed the embedding of many sensors in the surroundings. Moreover, the partition and the interior of the home could be remodeled freely. In accordance with the systematic approach, we verified the effectiveness of the virtual group home by constructing a lively environment at the real group home.

2.2 Mimamori-Care System

Usually, PWD wander around group homes unsteadily and are a source of concern and anxiety for caregivers; the events may sometimes lead to a matter of life and death for PWD. Therefore, it is necessary to know the action and position of PWD at all times. This system combines the functions of position detection and image capture by means of IC tags and cameras. Thus, caregivers can easily confirm the position and behavior of all the PWD through several display screens. Figure 1 shows the schematic of the "mimamori-care" system. All the equipments used are commercially available. We developed software that allows the caregivers to communicate with each other. Originally, the intended functionality was to realize a cooperative screen sharing system by which two or more caregivers could contact each other through a screen, particularly during the daytime. The same screen is not displayed. Display information is edited by the profile and context using IC tags and sensors; therefore, only the necessary information is displayed. The Web context awareness technology [7] we had developed earlier was used for this processing. The server was developed in Visual Basic 6 using the API provided by the IC tag vendor. A Web browser was used to display the information to the client.

It is important that all the caregivers are notified of the collected information and that the priority of the work is stated. Before introducing the system, our observation

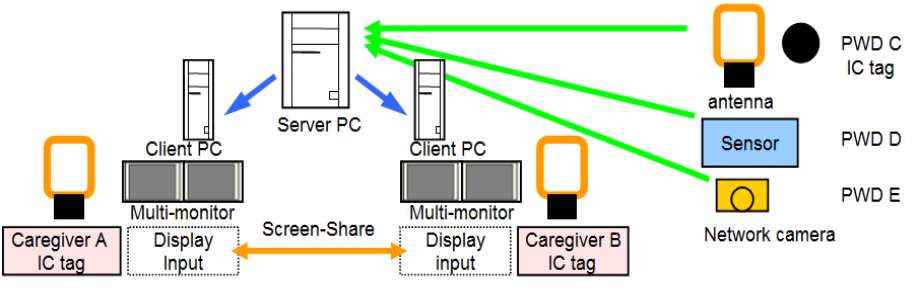


Fig. 1. Structure of the mimamori-care system

of the caregivers' activities revealed that their work involves a considerable amount of walking. To support their work, we embedded cameras in common spaces except bathroom and restroom and monitors in the walls, in furniture, as well as in other different objects and places. We thereby created an ambient environment. We considered environments in which water is used, such as bathrooms, kitchens, and restrooms, and adopted waterproof monitors in such environments. We coordinated all these monitors with computers. Since the caregivers were inexperienced with computers, we designed touch-screen type monitors with a few buttons, as illustrated in Figure 2, to enable even novice users to manipulate them appropriately.

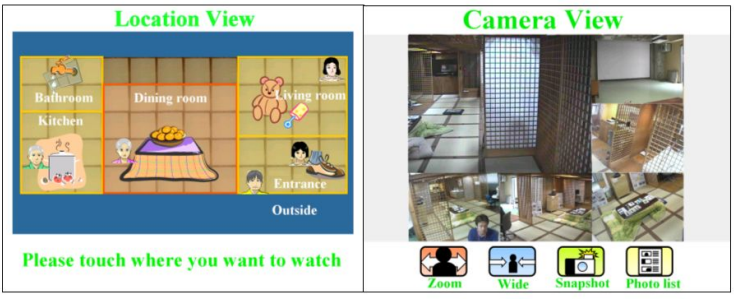


Fig. 2. Image of the prototype system inside the virtual group home

2.3 Related Works

In Japan, one of the most well-known researches using radio frequency identification (RFID) is a tracking system for elementary school pupils [8]. These systems typically use active RFID by which pupils can be automatically identified at the school gate. On the other hand, researchers also used Global Positioning System (GPS) to verify the location of wandering the elderly [9]. The problem, however, occurred when a resident removed the device due to its stumbling block. In the mimamori-care system, we used a 13.56 MHz passive RFID tag, which can be used permanently and does not require batteries. The RFID we used effectively communicates within the range of 70 cm. As our target is a typical house, this range does not pose a problem. By installing antennas in passages and on doors, we can effectively use this RFID. Using the above-mentioned RFID, we experimented whether movements could be detected in each room.

For the nursing action in hospitals, the e-nightingale project, which focused on recording with wearable sensors, was proposed [10]. Our target is not to record the nursing action in hospitals but to support PWDs' self-reliant life in group home.

3 Experiment

Part of the system was already in operation at a caregiving facility from November 2006. Before implementing it in the real group home, we conducted a six-month trial in a virtual group home. We attached importance to the following points.

3.1 Privacy and Ethics

Some critics insist that RFID and cameras are invasive and are privacy incursions. However, there is an obvious difference in "watching" or "monitoring" intended by suspicious persons and mimamori-care intended for PWD. With regard to watching by suspicious persons, the person being watched is usually a stranger or an unauthorized individual. For crime prevention, "watch systems" that transmit photographic evidence to remote security companies are already in operation. On the other hand, the mimamori-care system monitors a person at the same site to avoid accidents. At group homes, the caregivers and PWD are always together. Therefore, mutual trust is necessary. Moreover, the surveillance camera records the images as evidence. Further, some monitoring systems record and store videos of the PWD as well as the caregivers [3]. However, the system we developed did not record images and was hence able to obtain the consent of the caregivers.

We also considered the ethical aspect of the experiment. We orally explained the purpose of the research to the PWD and their families. We promised to protect their privacy even if they later withdrew from the experiment. Finally, we documented their signatures to indicate their "agreement and cooperation" for the research.

3.2 Home Implementation

Although, many utilities are modified as in modern homes, the system in the group home was designed considering the image of a traditional Japanese home. Moreover, part of the group home did not eliminate barriers either. For instance, the doors clearly demarcate the outside and inside areas. In a common home, it is difficult to embed a large number of sensors. Therefore, while introducing the monitoring system in the real group home, all communication devices used were wireless. Hence, the environment appeared more ambient and no wires were needed to form a LAN. However, there was neither Internet connectivity nor any outlets for Internet inside the real home. Therefore, the necessary connections and outlets for equipments were added. Figure 3 depicts the arrangements of cameras in the group home. Before introducing the mimamori-care system, there were some blind spots for the caregivers while caregiving other PWD.

4 Results and Discussions

4.1 Evaluation Method

We decided to analyze the results based on an interview with the caregivers. We believed it more effective to investigate the efficiency of the system through open-ended responses, because the patient, caregiver, and situation changed according to the context [11]. We asked them about the aspects they considered effective and their impressions on using the system. In particular, we paid attention to the difference in the uses of the system during daytime and nighttime caregiving. We are strongly motivated to accept the demerits as a basis for further improvements.

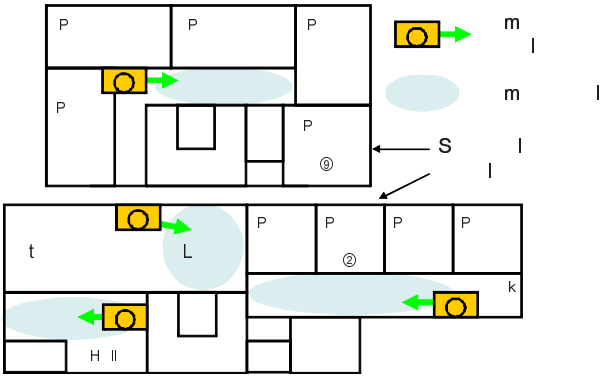


Fig. 3. Camera arrangements in the group home

4.2 Effects on Caregiving

Two months have passed since this system was introduced. As a result, several useful effects of this system in dementia care have already been observed. Our hypothesis was centered on the effect of the system during daytime care when many PWD actively engaged in their daily activities. However, of the positive comments received from the caregivers during nighttime caregiving, the most noteworthy points were as follows.

- Person A's private room is on the second floor. At night, Person A wandered frequently between the restroom and the private room. Before we introduced the system, the caregivers had observed this activity only based on the sounds of Person A's movements. Now, with the mimamori-care system, the caregivers are able to accurately track Person A's movements. The caregivers became aware of the fact that Person A used the restroom as many as 20 times within an hour. Moreover, they can confirm any possible danger with ease. At midnight, while returning from the restroom, Person A entered Person B's room by mistake. Through the monitor, the caregivers were able to verify this as soon as Person A entered the wrong room, and they were able to take the appropriate action. Figure 4 shows a caregiver using the system. The caregiver can carry the monitor to the desk, the bath, and the restroom.
- On the same day, Person C also used the restroom but returned back to his room as usual. Therefore, caregivers can easily confirm this difference with Person A. Hence, they assumed that Person A should have broken the door or that he was aware of the damage.
- Person D was discovered sleeping on the corridor of the second floor in the cold winter air. Apparently, he left his room and slept in the corridor to avoid the heat generated by the heating system in his room. Person D would have caught a cold if caregivers had not detected his position.
- During the daytime, the PWD typically spend their time in the living room. However, they were occasionally discovered napping in the corridor. With the mimamori-care system, the caregivers can easily monitor the activities of the PWD and take adequate action.



Fig. 4. A caregiver uses the system (left); inner view of the group home (right)

4.3 Discussion of the Real World-Oriented Approach

There are two important points that arose after commencing the operation of the system.

1. Effectiveness of caregiving during nighttime
   In our hypothesis, we assumed the effectiveness of caregiving during daytime. However, many effective episodes of caregiving were observed during the nighttime. Although nightshift care is important, it is difficult to have more than two caregivers on duty during the night. In this situation, we expect this system to be useful in most group homes.
2. Position and number of cameras
   We installed several cameras pointing in different directions. However, the caregivers discovered several blind spots after the installation and recommended that we install additional cameras in the inner and outer parts of doors, on the stairs, corridors etc. They also suggested that we change the direction of some cameras. The most unique opinion offered by the caregivers is their wish to see the appearance of the restroom when specific person uses it since some PWD tended to leave the restroom unclean. However, from the viewpoint of privacy, this has not yet been executed. For this, we need to examine other devices such as a distance detection system using ultrasonic devices.

5 Conclusion

This paper presents the mimamori-care system developed based on the real world-oriented approach. Our results show that the functions inherited by the system were effectively used in both daytime as well as nighttime caregiving. In this real world-oriented approach, the efforts of the engineers and the needs of the caregivers were effectively coordinated. In addition, the study revealed a few new findings and effects. One of the purposes of caregiving is the emphasis on the independence of PWD; this system encourages PWD to be more independent. Earlier, some caregivers experienced stress while caregiving PWD. However, this system makes it possible for the caregivers to monitor PWD in a relaxed manner without undermining their privacy. Although this system does not intend to be a substitute for caregivers, it provides them with considerable support. The reason for the positive vibes in group

homes is the comfortable, homelike environment. In these surroundings, PWD are not only caregiven but are also able to engage in other activities of daily life. Caregivers are enabled to peacefully make improvements in the quality of care. Consequently, the communication between the caregivers and PWD is activated. PWD sometimes help caregivers carry out chores such as cooking, washing, and cleaning. This leads to the independence of PWD and the improvement in QOL. In fact, the caregivers in the real group home often referred to mimamori-care as their eyes. We analyzed our observations and evaluated the information obtained from the caregivers and effectively implemented these to develop the present support system.

Acknowledgments. Our research is partly supported by the fund from Ministry of Education, Culture, Sports, Science and Technology, Japan, under the name of Cluster for Promotion of Science and Technology in Regional Areas.

References

[1] Wessener, C.: Public/Private Partnerships for Innovation, US National Academy of Sciences. In: OECD Workshop (2001)
[2] Branscomb, L.M., Auerswald, P.E.: Between Invention and Innovation An Analysis of the Funding for Early Stage Technology Development. Report to the Advanced Technology Program, NIST, US Department of Commerce, GCR-02-841 (2002)
[3] Takatsuka, R., Fujinami, T.: Aware Group Home: Person-Centered Care as Creative Problem Solving. In: Khosla, R., Howlett, R.J., Jain, L.C. (eds.) KES 2005. LNCS (LNAI), vol. 3681, pp. 451–457. Springer, Heidelberg (2005)
[4] ISO13407: Human-centred design processes for interactive systems (1999)
[5] Norman, D.A.: The design of everyday things (1988)
[6] Kanai, H., Nakada, H., Turuma, G., Kunifuji, S.: An Aware-Environment Enhanced Group Home. In: International Workshop on Smart Home (IWSH2006) in conjunction with International Conference on Hybrid Information Technology (2006)
[7] Nakagawa, K., Kato, N., Ueda, Y., Kunifuji, S.: A proposal and an implementation of a context-awareness system for web-based collaboration. IPSJ 47(7), 2081–2090 (2006)
[8] Yamada, I., Shiotsu, S., Itasaki, A., Inano, S., Yasaki, K., Takenaka, M.: Secure Active RFID Tag System. In: Proceeding of UbiComp 2005 Workshop on UbiComp Privacy PRIVACY IN CONTEXT (2005)
[9] Shimizu, K., Kawamura, K., Yamamoto, K.: Location System for Dementia Wandering. In: Proceedings of the 22nd Annual International Conference of the IEEE Engineering in Medicine and Biology Society, pp. 4928–92332. IEEE Computer Society Press, Los Alamitos (2000) pdf(CD-ROM)
[10] Kuwahara, N., Noma, H., Kogure, K., Hagita, N., Tetsutani, N., Iseki, H.: Wearable Auto-Event-Recording of Medical Nursing. In: Proceedings of INTERACT2003 (2003)
[11] Herbig, B., Bussing, A., Ewert, T.: The role of tacit knowledge in the work context of nursing. Journal of Advanced Nursing 34, 687–695 (2001)

Overview of Network Information by Using Anchored Maps

Kazuo Misue

Department of Computer Science, University of Tsukuba,
1-1-1 Tennoudai, Tsukuba, 305-8573 Japan
misue@cs.tsukuba.ac.jp

Abstract. Spring embedding is widely used to draw up networks. However, the readability of network diagrams drawn by using spring embedding is still not good enough. This paper proposes the use of "anchored maps," in which some of the nodes are fixed as anchors to improve the readability of network diagrams. First, the readability of network diagrams is discussed and the anchored maps are proposed. Then, an overview of the method the author developed for drawing anchored maps is explained. Next, examples of network diagrams that were created from a real sales database are used to prove the effectiveness of anchored maps.

Keywords: graph drawing, anchored map, visualization, knowledge mining.

1 Introduction

The purpose of this research is to develop visualization techniques to give overviews of networks that are highly readable for mining knowledge. The purpose of giving these overviews is to support the first movement to extract useful knowledge from network information. When we process large-scale numeric data, we usually draw graphs (e.g., scatter charts and histograms) to grasp the overall tendency of the data. When we process network information, we notice that we do not have enough techniques and tools to manipulate it. Techniques and tools for this type of manipulation are currently being developed.

The author proposes the introduction of dependable viewpoints from observations into network diagrams to improve the readability of the diagrams. It is possible to visualize network information as diagrams. However, it is often difficult to get any useful knowledge from these types of diagrams. Introducing viewpoints into network diagrams is expected to enable observers to adequately read useful information from them.

As an approach to introducing viewpoints on diagrams drawn by using spring embedding, which is often used to layout networks, we chose to fix some nodes at predetermined positions. The fixed nodes act like coordinate systems and perform functions as viewpoints. The author named this style of drawing "anchored map." Similar representation techniques are also used in some other tools, for example, Visual Who [2], which is a visualization tool for communities and SQWID [4], which is a WWW search tool.

B. Apolloni et al. (Eds.): KES 2007/ WIRN 2007, Part II, LNAI 4693, pp. 1269–1276, 2007.
© Springer-Verlag Berlin Heidelberg 2007

In the beginning, an anchored map has just three anchors. However, four or more viewpoints are often necessary, so general techniques for drawing anchored maps with four or more fixed nodes are necessary.

2 Readability of Network Diagrams

Drawing network diagrams (represent networks visually) and enabling to ability to read some of the meanings from the diagrams are not equal. If we expect the meanings of network diagrams to be readable, conventional techniques to draw networks (graph drawing techniques) cannot provide diagrams with enough readability.

2.1 Readability of the Spring Embedding

It has been a long time since we have seen the tools and applications that draw network diagrams by using spring embedding [3]. The author also has experience in developing the programs of spring embedding. The purpose of the programs is not just for drawing undirected graphs, but also for the representation of the relationships between concepts as network diagrams. The programs seem to work well at making the layouts of undirected graphs. However, when we tried to interpret the meaning in the application area of the networks, we often had trouble understanding what the diagrams expressed as a whole.

It is clear that the network diagrams represent the relationships between the nodes in each local segment. However, if we want to express the local relationships among the nodes, we do not need to draw network diagrams; tables and list expressions are also useful and are good enough for this purpose. One of the most important features of diagrams, especially network diagrams, is that we can look at the overview of the whole structure of the network. Drawing network diagrams provides overviews of the whole structure. However, most of us, including this author, are not satisfied with just these diagrams. The spring embedding enabled viewers to see an overview of the network information, but didn't enable them to read the meanings of the network information.

2.2 Introduction of Viewpoints into Network Diagrams

One of the factors of the low readability in network diagrams drawn using spring embedding is the lack of coordinate systems. Most layout techniques for directed graphs use the features of directed graphs; examples are upward/downward drawings [1] and hierarchical drawing [8]. These techniques orient the directions of the edges in one common direction, so they can visually express global direction. By using such layout techniques, we can grasp the overall structure of the networks and the position of each element in the global structure, even if we do not know the meanings of the nodes and edges.

Network diagrams drawn by using spring embedding do not have coordinate systems. This means that the diagrams do not give viewpoints to read the meanings of the network. To provide diagrams in which the meanings of the networks are highly readable it is critical to provide coordinate systems or viewpoints to the readers.

We developed a method to introduce the proxies of coordinate systems into undirected graphs; the method fixes three nodes chosen by the observers on the three vertices of a triangle [7]. The method fixes the three nodes (called "anchors"), which are the viewpoints, and arranges the other nodes (unfixed nodes, called "free nodes") by using spring embedding. Each unfixed node is placed in a stable position in the energy of springs through the push-pull processes by the anchors. When we consider the anchors as viewpoints, we can see the position of the free nodes as expressions of the relations with viewpoints.

2.3 Problem with Drawing Anchored Maps

There is a reason why only three anchors are used; it is to keep the anchors symmetrical. When three anchors are used, the network diagrams are independent to the arrangement of the anchors; the diagrams with any anchor arrangements are essentially the same. This means that we do not have to worry about the arrangement of the anchors.

However, the number of nodes to be used as viewpoints is not always three. When the number of anchors used is four or more, the arrangements of the anchors are not symmetrical. So we have to take the arrangement of the anchors into consideration.

We illustrate some of the problems that occurred when four or more anchors were used in Fig. 1. Assume that four anchors (a1, a2, a3, and a4) are arranged on four vertices of a square clockwise in this order. In addition, assume that two free nodes (f1 and f2) are related to two anchors with almost the same degrees. This means that the two free nodes are connected to the anchors by springs with almost the same lengths and strengths. If a free node has relations with a1 and a2, the node is arranged on the edge of the square (see f1 in Fig. 1 (a)). However, if it has relations with a1 and a3, the node is arranged near the center of the square (see f2). The difference in the two positions cannot be neglected. Additionally, when another free node has relations to a2 and a4, the free node is also arranged near the center of the square (see f3 in Fig. 1 (b)). In this case, although f2 and f3 are completely different from each other in relation to the anchors, they are arranged at close places.

It is necessary to choose the appropriate arrangement of anchors to avoid this type of visual misunderstanding.



Fig. 1. Problems with four anchors

By the way, anchors might have orders by their meanings. In such cases, the anchors can be arranged in the orders, even when four or more anchors are used. For example, when we choose the days of the week as anchors, we may arrange the anchors in the order of “Sunday,” “Monday,” …, “Saturday.”

3 Anchored Maps Drawing Method

When four or more anchors are used and the anchors do not have orders, we have to adequately decide the orders of the anchors. This problem cannot be solved by simple extending the spring embedding. We reorganized the drawing method of the anchored maps to solve the problem. A detailed description of the method is shown in [6].

3.1 Drawing Objects

We assume that the drawing objects are bipartite graphs. A bipartite graph is a graph where the set of nodes can be divided into two disjointed sets, such that no edge has both end-points in the same set. One of the two node sets is assumed to be a set of “anchors,” and the other is assumed to be a set of “free nodes.”

3.2 Drawing Conventions

The drawing convention of the anchored maps is written as follows:

(Arrangement convention for the nodes)
- Compound coordinate system: anchors are discretely arranged on a circumference, while the free nodes are freely arranged.

(Routing convention for the edges)
- Straight-line wiring: a straight line connects adjacent nodes.
- The edges are independent of the coordinate system.

3.3 Drawing Rules

We employed the following drawing rules for our anchored map drawing method.

(R1) The nodes are mutually separated more than the lowest distance.
(R2) The nodes connected to each other are laid out as closely as possible (minimizing the total length of the edges).
(R3) The number of crossings among the edges is minimized.
(R4) The anchors connected to common free nodes are laid out as closely as possible.
(R5) The free nodes connected to common anchors are laid out as closely as possible. (Free nodes belonging to the same cluster are laid out as closely as possible; free nodes that do not belong to the same cluster are not closely laid out).

The drawing rules (R1), (R2), and (R3) are also employed in many other drawing methods. We also judged them to be important in consideration of the adaptability to

the purpose of drawing anchored maps. Rules (R4) and (R5) are specific to anchored maps. A formal description of these rules is shown in [6].

3.4 Drawing Procedure of Anchored Maps

An anchored map is laid out using the following two steps:

(Step 1) Arrange anchors on the circumference at equal intervals. The size (i.e., radius) of the circumference is decided according to the size of the drawing area (i.e., window), and the order of the anchors on the circumference is decided.
(Step 2) Fix the anchors, and arrange the free nodes at the positions in which their relationships to the anchors are appropriately expressed.

The length of the circumference doesn't influence the quality of the layout; it influences only the size of the drawing. On the other hand, the order of the anchors largely influences the quality of the layout. We do not need to worry about the routing of the edges since the edges are drawn as straight-line segments.

4 Overview of Network Information

An example of the overview of the network information extracted from an actual database to explain the effectiveness of anchored maps is presented in this section.

4.1 Database

A sales database was exploited for experimentation purposes. One transaction is represented as one record, and each record has a purchase date, time, customer name, item name, item category, and price. We could not get a database from a real shop, so we used a sales database managed in our laboratory. It is not a database of a real shop, but the transactions are real. We started to use the database in June 2006, and 2652 purchases were recorded for half a year.

4.2 Overview Using Conventional Means

Various relations can be extracted from the database. In this paper, we focus on the relations between items and time zones (every hour; 00 ~ 23 hrs.) when the items were sold. We removed the items that were sold three times or less in any time zone to avoid noise. What remained was 37 kinds of items, which were sold four or more times in a given time zone.

Table 1 shows the relations between the items and time zones. Each row corresponds to an item, and each column corresponds to a time zone. Each cell represents the number of items sold in a time zone. Often the table is shown in a histogram. This type of representation style is widely used in the early stages of a statistical analysis. By using these representations, we might answer some pointed questions, such as what is the item sold most in a certain time zone. We can say that this representation is suitable for reading the relations between (items, time-zone) and the number of items sold.

Table 1. Overview by using a table

TE				3											3									
午後 紅 テ			3			3	3					3			3									3
城																3								
紫 (930g PRO)												3				3	3	3						
MONDA? グ ロ		3	3			3	3																	
マジ レ サ															3									3
サ 城																								
リ D		3										3				3					3	3		
		3																						
城				3																3			3	3
															3				3		3			
ナ c		3										3												
マ レ ュ																				3		3		
ネ												3				3								
藤																	3							
ド ド			3											3								3		
明 ュ レ													3				3			3				3
テリヤ																								
															3			3						
																		3						3
紫 100(?)																						3		
ポ レ 砲?															3					3				
FIRE???																								
															3			3						
KS??????																								
QOO????? グ ル																								
城			3													3								
日 U.F.O.															3									
リ			3																					
サ ル			3																3					
リ マ ム																								
n																								
京 原																								
o 郎			3																					
ホ ム																						3		
G u G																								
和 テ イ																								

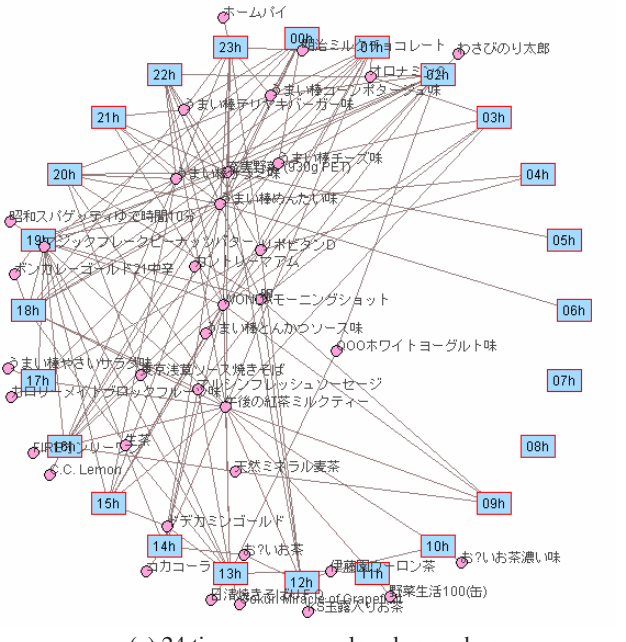
4.3 Overview by Using Anchored Maps

Figure 2 (a) shows the networks that represent the relations between the items and time zones when the items were bought. The time zones are placed as the 24 anchors. The time zones have a cyclic order, so they were placed in their proper order. From this map, we quickly find that the activity of the students is very low in the morning. Of course, if we carefully read Table 1, we can also find this information. This figure also intuitively informs us of this fact. Furthermore, we find that food such as the curries and spaghettis were bought in specific time zones, while canned coffee and sweets did not appear to depend on any specific time zone.

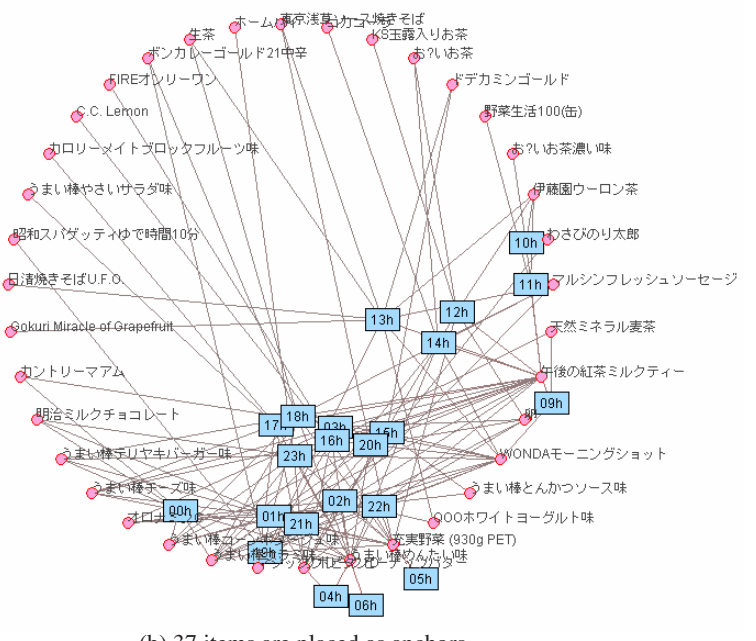
In Fig. 2 (b), the items are placed as the anchors. Since the items have no proper orders, we used the technique described in [6] to decide the order of the anchors. We see the following clusters in the map.

(1) Early morning 1: 04, 05, and 06 hrs.
(2) Early morning 2: 07 and 08 hrs. (no nodes = no activities)
(3) Morning: 09, 10, and 11 hrs.
(4) Lunch time: 12, 13, and 14 hrs.
(5) Evening ~ night: 15 ~ 03 hrs.

It might be possible to find these types of clusters without an overview of the networks using the clustering techniques that already exist. However, the advantage with the network information overview is that we can see the grounds and the background that compose the clusters as the connecting lines between the nodes. By using such maps, we can roughly observe the existence of the clusters from the beginning. Then, we better understand the reason why these clusters exist by touching on a virtual physical model of the relations.



(a) 24 time zones are placed as anchors



(b) 37 items are placed as anchors

Fig. 2. Overview by using anchored maps

5 Conclusion

This paper described the problem with the lack of coordinate systems in topological diagrams, which is the low readability of the diagrams. A drawing style called "anchored maps" was proposed to counteract this problem. In the anchored maps, some nodes, called "anchors", are fixed as viewpoints and thus effectively act like coordinate systems. The author developed a method to draw anchored maps with four or more anchors. The anchored maps are useful at the early analysis stage of network information.

There are several topics remaining as future works; one of them is the conceptual rearrangement of the "readability" of networks. Usually the readability of networks or graphs is described as the drawing conventions and drawing rules. The introduction of the viewpoints proposed in this paper might help to achieve a higher level of conceptual understanding than the drawing conventions and rules. These types of concepts should be organized as the aesthetic criteria for understandability.

References

1. Di Battista, G., Tamassia, R.: Algorithms for plane representations of acyclic digraphs. Theoretical Computer Science 61(2-3), 175–198 (1988)
2. Donath, J.S.: Visual Who: Animating the affinities and activities of an electronic community. In: ACM Multimedia 95 - Electronic Proceedings, ACM Press, New York (1995)
3. Eades, P.: A Heuristic for Graph Drawing. Congressus Numerantium 42, 149–160 (1984)
4. McCrickard, S.D., Kehoe, C.: Visualizing Search Results using SQWID. In: WWW6. Proceedings of the sixth International World Wide Web Conference (1997)
5. Misue, K.: Drawing Bipartite Graphs as Anchored Maps. In: APVIS 2006. Proceedings of Asia-Pacific Symposium on Information Visualization, pp. 169–177 (2006)
6. Misue, K.: Anchored Maps: Visualization Techniques for Drawing Bipartite Graphs. In: Proc. of HCI International (2007) (to appear)
7. Misue, K., Watanabe, I.: Visualization of Keyword Association for Text Mining. In: IPJS SIG Technical Report 1999-FI-55-8, Information Processing Society of Japan, pp. 65–72 (1999) (in Japanese)
8. Sugiyama, K., Tagawa, S., Toda, M.: Methods for Visual Understanding of Hierarchical Systems. IEEE Trans. SMC 11(2), 109–125 (1981)

Spreadsheet Interface Which Simplifies the Visualization of the Network Structures

Satoshi Koike, Kazuo Misue, and Jiro Tanaka

Department of Computer Sceince, University of Tsukuba
koike@iplab.cs.tsukuba.ac.jp, {misue, jiro}@cs.tsukuba.ac.jp

Abstract. We developed a system that extracts and visualizes the network structures from relational data on spreadsheets. Visualization through network diagrams obtained from relational data is indispensable to understanding their structure. However, time-consuming tasks are involved in visualizing these networks. The purpose of this research was to support the extraction and visualization of network structures from relational data. We developed an interface that coordinates spreadsheets with visualization tools. It simplifies the visualization of networks from relational data and enables their structures to be easily understood.

Keywords: Network Visualization, Spreadsheet, Graph Drawing, Wizard.

1 Introduction

There are numerous relational data throughout the world. Examples include data on personal connections and purchasing history. These relational data are stored in databases and are used in various scenarios. Retail stores, e.g., convenience stores, better understand customer attitudes from their purchasing history.

Relational data often compose network structures. We can find useful information by understanding the structures. However, it is difficult to understand the structure of networks from relational data when the network is represented as lists of characters and values. The visual representation of networks is essential to understanding the structure of networks from relational data.

Various methods of visualization and many visualization tools that enable network diagrams to be drawn have been developed. However, as databases and visualization tools are not sufficiently coordinated, we have to do time-consuming tasks such as extracting and translating data to visualize networks.

We took the approach of coordinating spreadsheets and visualization tools to solve this problem. Spreadsheets are tools for manipulating tabulated data, and they are also used to edit relational data. If it is easily to extract and visualize networks from relational data on spreadsheets, we can easily analyze these with network diagrams.

B. Apolloni et al. (Eds.): KES 2007/ WIRN 2007, Part II, LNAI 4693, pp. 1277–1284, 2007.
© Springer-Verlag Berlin Heidelberg 2007

2 Networks Extracted from Database

There is a connection between members written on the same record in a database. Various networks are composed of these relations. We focused on two kinds of networks in this research, “networks based on relations” and “networks based on similarities”.

2.1 Database

A database, D, is formulated as $D = \{r_1, r_2, ..., r_m\}$ in this paper. r_i is the ith record. The pth field in record r is formulated as $f_p(r)$.

2.2 Network Based on Relations

There are connections between fields’ members in the same record. A network is composed when the members of two fields in the database are identified as nodes. This network is based on the relation between two fields.

The network based on the relation between field f_p and field f_q can be formulated as weighted graph G_r (shown in (3)). Set V of nodes, set E of edges with weights, and function w_r that gives the weight of an edge comprise G_r (shown in (4), (5), and (6)).

$$G_r = (V, E, w_r) \tag{3}$$

$$V = F_p \bigcup F_q, F_p = \{f_p(r) \mid r \in D\} \tag{4}$$

$$E = \{(a,b) \mid f_p(r) = a \wedge f_q(r) = b \quad for^{\exists} r \in D\} \tag{5}$$

$$w_r : E \to R \tag{6}$$

Function w_r can be selected from various methods of calculation. There are some examples in the following.

Set $R_{p,q}(a,b)$ of records where a and b are placed in the same record, r, are shown in (7). The element count of $R_{p,q}(a,b)$ gives an indication of the strength of the relation between a and b. Considering this value as the weight of an edge, w_r is as shown in (8).

$$R_{p,q}(a,b) = \{r \in D \mid f_p(r) = a \wedge f_q(r) = b\} \tag{7}$$

$$w_r((a,b)) = \left| R_{p,q}(a,b) \right| \tag{8}$$

The value of numerical field $f_w(r)$ can be assumed to be the value of the strength of the relation between $f_p(r)$ and $f_q(r)$. The method of calculation to add the value of $f_w(r)$ is selected from various choices (e.g., sum or average). Considering this value as the weight of the edge, w_r is shown in (9) and (10).

$$w_r((a,b)) = \sum_{r \in R_{p,q}(a,b)} f_w(r) \tag{9}$$

$$w_r((a,b)) = \frac{\sum_{r \in R_{p,q}(a,b)} f_w(r)}{\left| R_{p,q}(a,b) \right|} \tag{10}$$

It is necessary to determine fields f_p, f_q as vertices and field f_w as the weight of an edge to extract the network based on the relation. In addition, it is necessary to determine function w_r.

2.3 Network Based on Similarities

The members in a database are connected with the other field's members written on the same record. The similarity of the members can be calculated by using its relations to the other field's members.

When focusing the relation between field f_p's members and field f_q's members, the similarities of f_p's members are calculated. Here, f_q is called a reference field for the similarities. Two or more fields can be used as reference fields at the same time.

A network is based on the similarity is composed when we consider that all the different members of a field as nodes. The network based on the similarities of all the different members of f_p can be formulated as a weighted complete graph, G_s (shown in (11)). Set V of nodes, set E of edges with weights, and function w_r that gives the weight i.e., similarity of an edge comprise G_s (shown in (12), (13), and (14)).

$$G_s = (V, E, w_s) \tag{11}$$

$$V = F_p, F_p = \{ f_p(r) \mid r \in D \} \tag{12}$$

$$E = F_p \times F_p \tag{13}$$

$$w_s : E \to R \tag{14}$$

Function w_s can also be selected from various methods of calculation. The following has some examples (shown in (15), (16), and (17)).

$$w_s((a_i, a_j)) = \frac{1}{\sum_{b \in F_q} \left| w_r((a_i,b)) - w_r((a_j,b)) \right|} \tag{15}$$

$$w_s((a_i, a_j)) = \frac{1}{\sqrt{\sum_{b \in F_q} (w_r((a_i,b)) - w_r((a_j,b)))^2}} \tag{16}$$

$$w_s((a_i,a_j)) = \frac{\sum_{b\in F_q}(w_r((a_i,b)) - v(a_i,b))(w_r((a_j,b) - v(a_j,b))}{\sqrt{\sum_{b\in F_q}(w_r((a_i,b)) - v(a_i,b))^2}\sqrt{\sum_{b\in F_q}(w_r((a_j,b)) - v(a_j,b))^2}}, \quad (17)$$

$$v(a,b) = \frac{\sum_{b\in F_q} w_r((a,b))}{|F_q|}$$

It is necessary to determine fields f_p as vertices, fields f_q as the target for calculating the similarities, and field f_w as the value of the strength of the relation to extract the network based on similarities. It is also necessary to determine function w_r. In addition, it is necessary to determine function w_s.

3 Extracting and Visualizing Networks

The steps involved in visualizing networks are as follows.

Step 1. Select the type of network
It is necessary to determine what kind of network is to be extracted. Two kinds of networks ("networks based on relation G_r" and "networks based on similarity G_s") were presented in Section 2.

Step 2. Select the fields
It is necessary to determine the extraction fields or method of calculation to extract the networks selected in Step 1. As described in Section 2, it is necessary to determine f_p, f_q, f_w, w_r to extract G_r and it is necessary to determine f_p, f_q, f_w, w_r, w_s to extract G_s.

Step 3. Select the style of visualization
It is necessary to determine the style of visualization of the network. As there are various methods of visualization, it is necessary to choose one according to the kind of the network and the purpose of the visualization. (AnchoredMap Layout [1] and Radial Layout [2] can be chosen for the system proposed in this paper.)

Step 4. Extract the network
The selected network is extracted after Step 3 and the network data are translated to data that correspond to the visualization tool. (The system proposed in this paper outputs GraphML [3] and CSV from the data on the spreadsheet. GraphML is extended XML to represent the graph. GraphML is used in many tools as an input data format. CSV is used in several Web applications as an input data format.)

Step 5. Visualize with tool
The network is visualized by inputting data to the tool.

There are numerous choices in extracting the network; therefore, it is necessary to select from these. Moreover, the extracted network needs to be translated to data that

correspond to the visualization tool. These tasks are time consuming and difficult for people who do not have skills to translate data.

4 Wizard for Extracting/Visualizing Network

A wizard is often used to support complex tasks. It is an interface that guides users through complex tasks, using step-by-step dialogs. The wizard is used for setting up applications.

It is further used for visualizing information. An interface called "Graph Wizard" is implemented on a spreadsheet to draft coordinate graphs such as bar and line graphs. Although it is necessary to select from various choices such as data fields and types of graphs, it is possible to easily draw these using the Graph Wizard.

Visualizing the network is similar to drawing these coordinate graphs. A spreadsheet is also used to present and edit relational data. It is also necessary to select from numerous choices such as data fields and methods of visualizing the network. The wizard can also be applied to visualizing the network for this reason.

We propose the wizard as an interface to support the extraction and visualization of networks from relational data. The user can extract and visualize networks from relational data with this wizard by answering its questions.

4.1 System Architecture

Figure 1 outlines the system architecture.

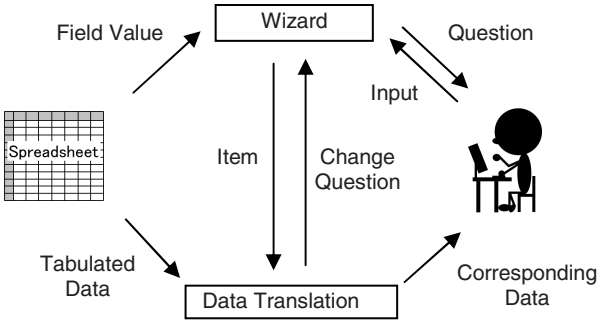


Fig. 1. System Architecture

The wizard component promotes input by users by asking them questions. If it needs the field value for the question, it obtains the value from the sheets. The data translation component changes the questions asked by the wizard. Based on the inputs by the wizard, it translates the tabulated data into the data that correspond to the visualization tool.

4.2 Implementation

The items that are necessary to extract and visualize networks can be classified into three steps.

Step 1: Type of network
Step 2: Field
Step 3: Style of visualization

The wizard presents a question to the user to set these items through dialog-input forms. He or she extracts and visualizes the network by inputting items into the dialog forms.

A dialog input form is divided into three parts (Figs. 2, 3, 4, and 5) according to the item that needs to be extracted to visualize the network.

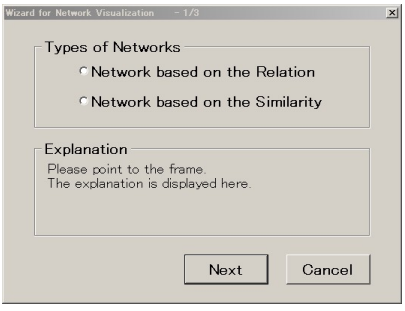


Fig. 2. Dialog-input form for Step 1

Figure 2 shows a dialog form for two types of networks. A text explanation is displayed in a frame in the middle of the form. The user selects a network ("relation" or "similarity") using the buttons on the form. The system changes to the next dialog form depending on the input selected.

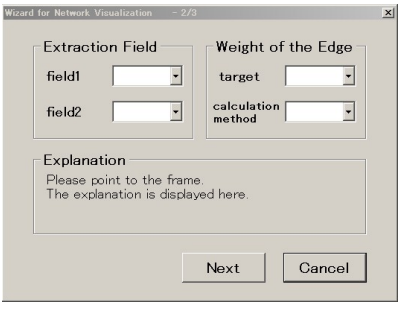


Fig. 3. Dialog-input form for Step 2 (relation)

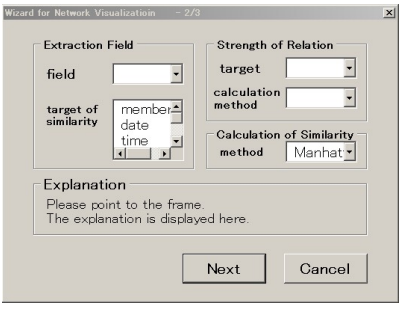


Fig. 4. Dialog-input form for Step 2 (similarity)

Figure 3 shows the dialog form to select a field to extract the network based on relations. The user selects a field extracted as nodes from the left frame. The value of

the box corresponds to the field on the sheet. He or she selects the field and method of calculating the edge weight from the right frame. The value of the "target" box corresponds to the numerical field on the sheet.

Figure 4 shows the dialog form to select a field to extract the network based on similarities. The user selects a field extracted as nodes and the target field of similarity from the left frame. The user selects the field and method of calculating the relational strength from the upper right frame. The user selects the method of calculating similarity from the lower right frame.

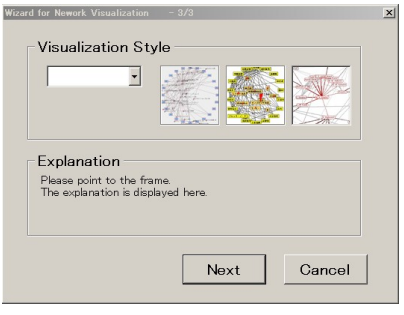


Fig. 5. Dialog-input form for Step 3

Figure 5 shows the dialog form for the style of visualization. The user selects a method from the frame.

The corresponding data are output after he or she has selected from the form.

5 Discussion

We compared the tasks of manually extracting and visualizing the network based on relations and similarities with those done by the wizard.

Table 1. Comparison of tasks

manual procedure	wizard
0. Prepare the translation program	
1. Extract the network by manual	1. Extract the network and output the corresponding data by answering the question from the wizard
2. Output the intermediary file	
3. Translate the intermediary file to the corresponding data	
4. Visualize by the tool	2. Visualize by the tool

It is necessary to output an intermediary file to output corresponding data in the manual procedure. In addition, users have to prepare the translation program that translates the intermediary file to corresponding data. With the wizard, on the other

hand, they can output the corresponding data simply by answering questions asked by the wizard, which enables networks to be visualized.

6 Related Works

There has been some research that has taken the approach of coordinating tabulated data and network diagrams.

Henry and Felete developed a tool called "MatrixExplorer" to assist in the understanding of network structures [4]. This tool displays tabulated data and network diagrams simultaneously and visually represents their correspondences.

Mackinlay developed A Presentation Tool (APT) that designs effective representations of the relational data [5]. This tool designs representations including bar and scatter charts according to the features of the relational data.

However, we took a new approach that uses a wizard to coordinate tabulated data and network diagrams.

7 Conclusions

We proposed a wizard as an interface to extract and visualize the structures of networks. We implemented it to extract and visualize networks from the relational data available on spreadsheets. It is much easier to visualize networks with this wizard and we intend to apply it to surveying other networks.

References

1. Misue, K.: Drawing Bipartite Graphis as Anchored Maps. In: Proceedings of Asia-Pacific Symposium on Information Visualization, pp. 169–177 (2006)
2. Heer, J., Card, S.K., Landay, J.A.: Prefuse: a toolkit for Interactive Information Visualization. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, pp. 421–430 (2005)
3. Brandes, U., Pich, C.: GraphML Transformation. In: Pach, J. (ed.) GD 2004. LNCS, vol. 3383, pp. 89–99. Springer, Heidelberg (2005)
4. Henry, N., Felete, J.D.: MatrixExplorer: a Dual-Representation System to Explorer Social Networks. IEEE Transaction on Visualization and Computer Graphics 12(5), 677–684 (2006)
5. Mackinlay, J.: Automating the Design of Graphical Presentations of Relational Information. ACM Transactions on Graphics 5(2), 110–141 (1986)

A Fuzzy Ranking Approach to Data Envelopment Analysis

Hsuan-Shih Lee[1], Pei-Di Shen[2,⋆], and Wen-Li Chyr[2]

[1] Department of Shipping and Transportation Management
National Taiwan Ocean University
Keelung 202, Taiwan
[2] Department of Information Management
Ming Chung University
Taipei 111, Taiwan
pdshen@mcu.edu.tw

Abstract. One problem with data envelopment analysis (DEA), a prominent evaluation method in social sciences, is its low discriminating power when evaluated units are insufficient or inputs and outputs are too many relative to the number of units. To deal with this problem, we incorporate fuzzy set theory with classical DEA so that a broader aspect of evaluation could be taken into account. We propose a method to encapsulate the efficiencies of an unit in different aspects as a fuzzy efficiency. A fuzzy efficiency is further compared with other fuzzy efficiencies to determine its strength and weakness based on extended fuzzy preference relation. With the strength and weakness of an unit, we aggregate them into a total performance index so that a complete ranking of units is obtained. The method proposed in this paper demonstrates a high discrimination in DEA applications compared to classical DEA.

Keywords: DEA; Fuzzy efficiency; Fuzzy ranking.

1 Introduction

DEA was initiated in 1978 when Charnes, Cooper and Rodes (CCR) demonstrated how to change a fractional linear measure of efficiency into a linear programming (LP) format [2]. As a result, decision-making units (DMUs) could be assessed on the basis of multiple inputs and outputs, even if the production function was unknown. This non-parametric approach solves an LP formulation per DMU and the weights assigned to each linear aggregation are the results of the corresponding LP. The weights are chosen so as to show the specific DMU in as positive a light as possible, under the restriction that no other DMU, given the same weights, is more than 100% efficient. Consequently, a Pareto frontier is attained, marked by specific DMUs on the boundary envelope of input-output variable space. The frontier is considered a sign of relative efficiency, which has been achieved by at least one DMU.

⋆ Corresponding author.

B. Apolloni et al. (Eds.): KES 2007/ WIRN 2007, Part II, LNAI 4693, pp. 1285–1292, 2007.
© Springer-Verlag Berlin Heidelberg 2007

Often decision-makers are interested in a complete ranking, beyond the dichotomized classification, in order to refine the evaluation of the units. Various ranking methods have been developed. Sexton et al. [15] proposed the cross-efficiency technique in which the DMUs are both self and peer evaluated. In [1], Andersen and Petersen proposed the super-efficiency technique in which DMUs are ranked through the exclusion of the unit being scored from the DEA models of envelopment form. Torgersen et al. [16] suggested an approach in which an efficient unit is highly ranked if it appears frequently in the reference sets of inefficient DMUs. Friedman and Sinuany-Stern [3] first suggested the application of multi-variate statistical tools, such as canonical correlation analysis and discriminant analysis, in order to reach a complete DMU rank.

Fuzzy set theory has been successfully applied in various fields of social sciences such as decision making [4,5,6,7,8,9,10,11]. In [12,13], we have incorporated fuzzy set theory into DEA to increase the discrimination in DEA applications. In this paper, instead of considering only the classical DEA efficiency which is the most favorable to the DMU being evaluated, we take other evaluations called peer evaluations into consideration and represent these evaluations by a fuzzy efficiency in the sense that the efficiency of a DMU is not a crisp value, instead, a fuzzy value designating efficiency lies in a range with different degree. A fuzzy efficiency is compared with other fuzzy efficiencies to identify its strength and weakness based on extend fuzzy preference relation [9]. By aggregating the strength and weakness of a fuzzy efficiency, we have the total performance of a DMU and thus obtain a complete rank of DMUs. The fuzzy method proposed in this paper demonstrates a high discrimination in DEA applications especially when DMUs are insufficient or inputs and outputs are too many relative to the number of DMUs.

2 Identifying Fuzzy Efficiencies of Decision Making Units

Originally established by Charnes et al. [2], the DEA model for evaluating the efficiency of a DMU (decision making unit), say DMU k, is as follows:

$$\begin{array}{ll} max\ h_k = \sum_{r=1}^{s} u_r y_{rk} & \\ s.t. \quad \sum_{i=1}^{m} v_i x_{ik} = 1, & \\ \quad \sum_{r=1}^{s} u_r y_{rj} - \sum_{i=1}^{m} v_i x_{ij} \le 0, j = 1, \ldots, n, & \\ \quad u_r, v_i \ge 0, r = 1, \ldots, s, i = 1, \ldots, m & \end{array} \tag{1}$$

where j is the DMU index, $j = 1, \ldots, n$, r is the output index, $r = 1, \ldots, s$, i is the input index, $i = 1, \ldots, m$, x_{ij} is the value of the ith input for the jth DMU, y_{rj} is the value of the rth output for the jth DMU, u_r is the weight given to the rth output, v_i is the weight given to the ith input, and h_k is the relative efficiency of DMU_k, the DMU under evaluation. In this model, DMU_k is efficient if and only if its optimal $h_k^* = 1$ and all corresponding optimal weights of input and output are greater than 0.

Basically, DEA model (1) identifies the most favorable weights to DMU_k in the sense that DMU_k achieves its highest efficiency under the constraint that all

other DMUs can not exceed efficiency 1 with same weights. The main drawback of DEA is its discriminating power. If the number of DMUs is not enough, it may result in that many DMUs to be identified as efficient. To enhance the discriminating power, we rewrite model (1):

$$\begin{array}{ll} max\ h_k = \sum_{r=1}^{s} u_r^{(k)} y_{rk} & \\ s.t. \quad \sum_{i=1}^{m} v_i^{(k)} x_{ik} = 1, & \\ \quad \sum_{r=1}^{s} u_r^{(k)} y_{rj} - \sum_{i=1}^{m} v_i^{(k)} x_{ij} \leq 0, j = 1, \ldots, n, & \\ \quad u_r^{(k)}, v_i^{(k)} \geq 0, r = 1, \ldots, s, i = 1, \ldots, m & \end{array} \tag{2}$$

Let $u_r^{(k)^*}$ and $v_i^{(k)^*}$ be the optimal solution of model (2). The efficiency of DMU_l in the shadow of DMU_k is given as follows:

$$E_{lk} = \frac{\sum_{r=1}^{s} u_r^{(k)^*} y_{rl}}{\sum_{i=1}^{m} v_i^{(k)^*} x_{il}} \tag{3}$$

E_{lk} is the peer evaluation of DMU_l in the shadow of DMU_k if $k \neq l$. E_{ll} is the self-evaluation for DMU_l, which is the same as h_l in model (1). Let $\cdot_{(k)}$ indicate that the indices have been permuted so that

$$E_{l(1)} \leq E_{l(2)} \ldots \leq E_{l(i)} \ldots \leq E_{l(n)}$$

For DMU_l, to encapsulate the sel-evaluation and peer evaluations across "all shadows", we represent these efficiencies by a fuzzy efficiency in terms of the following triangular fuzzy number:

$$\tilde{E}_l = (E_l^L, E_l^M, E_l^R), \tag{4}$$

where

$$E_l^L = \frac{\sum_{k=1}^{[n/2]} E_{l(k)}}{[n/2]}, E_l^M = \frac{\sum_{k=1}^{n} E_{l(k)}}{n}, \text{ and } E_l^R = \frac{\sum_{k=[n/2]}^{n} E_{l(k)}}{n - [n/2] + 1}.$$

3 Ranking of Fuzzy Efficiencies

As outlined in previous section, efficiencies of a decision making unit in sel-evaluation and peer evaluations can be encapsulated as a fuzzy efficiency represented by a triangular fuzzy number. In this section, we propose a ranking method for fuzzy efficiencies based on the concept of the strength and weakness proposed in [9]. For clarity, some results in [9] are reiterated here.

Definition 1. *The α-cut of fuzzy set $\tilde{A}$, $\tilde{A}^\alpha$, is the crisp set $\tilde{A}^\alpha = \{x \mid \mu_{\tilde{A}}(x) \geq \alpha\}$. The support of $\tilde{A}$ is the crisp set $Supp(\tilde{A}) = \{x \mid \mu_{\tilde{A}}(x) > 0\}$. $\tilde{A}$ is normal iff $sup_{x \in U} \mu_{\tilde{A}}(x) = 1$, where U is the universe set.*

Definition 2. *$\tilde{A}$ is a fuzzy number iff $\tilde{A}$ is a normal and convex fuzzy subset of real number.*

Definition 3. *A triangular fuzzy number $\tilde{A}$ is a fuzzy number with piecewise linear membership function $\mu_{\tilde{A}}$ defined by*

$$\mu_{\tilde{A}}(x) = \begin{cases} \frac{x-a_1}{a_2-a_1}, & a_1 \le x \le a_2, \\ \frac{a_3-x}{a_3-a_2}, & a_2 \le x \le a_3, \\ 0, & otherwise, \end{cases}$$

which can be denoted as a triplet (a_1, a_2, a_3).

Definition 4. *Let $\tilde{A}$ be a fuzzy number. Then $\tilde{A}_\alpha^L$ and $\tilde{A}_\alpha^U$ are defined as $\tilde{A}_\alpha^L = \inf_{\mu_{\tilde{A}}(z) \ge \alpha}(z)$ and $\tilde{A}_\alpha^U = \sup_{\mu_{\tilde{A}}(z) \ge \alpha}(z)$ respectively.*

Definition 5. *[9] An extended fuzzy preference relation R on fuzzy numbers is an extended fuzzy subset of the product of fuzzy numbers with membership function $-\infty \le \mu_R(\tilde{A}, \tilde{B}) \le \infty$ being the preference degree of fuzzy number $\tilde{A}$ to fuzzy number $\tilde{B}$.*

1. *R is reciprocal iff $\mu_R(\tilde{A}, \tilde{B}) = -\mu_R(\tilde{B}, \tilde{A})$ for all fuzzy numbers $\tilde{A}$ and $\tilde{B}$.*
2. *R is transitive iff $\mu_R(\tilde{A}, \tilde{B}) \ge 0$ and $\mu_R(\tilde{B}, \tilde{C}) \ge 0 \Rightarrow \mu_R(\tilde{A}, \tilde{C}) \ge 0$ for all fuzzy numbers $\tilde{A}$, $\tilde{B}$ and $\tilde{C}$.*
3. *R is additive iff $\mu_R(\tilde{A}, \tilde{C}) = \mu_R(\tilde{A}, \tilde{B}) + \mu_R(\tilde{B}, \tilde{C})$*
4. *R is a total ordering iff R is reciprocal, transitive and additive.*

In [9], we have defined an extended fuzzy preference relation on fuzzy numbers, which is reciprocal, transitive and additive. Some results in [9] are as follows.

Definition 6. *[9] For any fuzzy numbers $\tilde{A}$ and $\tilde{B}$, we define the extended fuzzy preference relation $F(\tilde{A}, \tilde{B})$ by the membership function*

$$\mu_F(\tilde{A}, \tilde{B}) = \int_0^1 ((\tilde{A} - \tilde{B})_\alpha^L + (\tilde{A} - \tilde{B})_\alpha^U) d\alpha \tag{5}$$

Lemma 1. *[9] F is reciprocal, i.e., $\mu_F(\tilde{B}, \tilde{A}) = -\mu_F(\tilde{A}, \tilde{B})$.*

Lemma 2. *[9] F is additive, i.e., $\mu_F(\tilde{A}, \tilde{B}) + \mu_F(\tilde{B}, \tilde{C}) = \mu_F(\tilde{A}, \tilde{C})$.*

Lemma 3. *[9] F is transitive, i.e., $\mu_F(\tilde{A}, \tilde{B}) \ge 0$ and $\mu_F(\tilde{B}, \tilde{C}) \ge 0 \Rightarrow \mu_F(\tilde{A}, \tilde{C}) \ge 0$.*

Lemma 4. *[9] Let $\tilde{A} = (a_1, a_2, a_3)$ and $\tilde{B} = (b_1, b_2, b_3)$ be two triangular fuzzy numbers. Then $\mu_F(\tilde{A}, \tilde{B}) = (a_1 + 2a_2 + a_3 - b_1 - 2b_2 - b_3)/2$.*

We define the preference intensity function of one fuzzy number $\tilde{A}$ over another number $\tilde{B}$ as follows:

$$Q(\tilde{A}, \tilde{B}) = \max\{\mu_F(\tilde{A}, \tilde{B}), 0\}. \tag{6}$$

The strength of the fuzzy efficiency $\tilde{E}_i$ can be defined by

$$S_i = \sum_{k \ne i} Q(\tilde{E}_i, \tilde{E}_k) \tag{7}$$

and the weakness of $\tilde{E}_i$ can be defined similarly as

$$W_i = \sum_{k \neq i} Q(\tilde{E}_k, \tilde{E}_i). \tag{8}$$

Total performance index of the ith DMU can be obtained by aggregating its strength and weakness as follows:

$$t_i = \frac{S_i}{S_i + W_i}. \tag{9}$$

4 Numeric Example

Example 1. (*Efficiency evaluation of six nursing homes.*) Sexton [15] considered a case of six nursing homes whose input and output data for a given year are described in Table 1, where the input and output variables are defined as follows:

StHr (x_1) staff hours per day, including nurses, physicians. etc.
Supp (x_2) supplies per day, measured in thousands of dollars
MCPD (y_1) total Medicare-plus Medicaid-reimbursed patient days (0,000)
PPPD (y_2) total privately paid patient days (0,000)

The results of classical DEA method, solved by LP, are given in Table 2. Using (2) and (3), we obtained the results shown in Table 3. As shown in Table 2,

Table 1. Data of Example 1 (Sexton, 1986, p.19)

DUM	Inputs		Outputs	
	StHr(x_1)	Supp(x_2)	MCPD(y_1)	PPPD(y_2)
A	1.50	0.2	1.40	0.35
B	4.00	0.7	1.40	2.10
C	3.20	1.2	4.20	1.05
D	5.20	2.0	2.80	4.20
E	3.50	1.2	1.90	2.50
F	3.20	0.7	1.40	1.50

Table 2. Classical DEA results of Example 1

DMU	Efficiency	Input weights		Output weights	
		v_1	v_2	u_1	u_2
A	1	0.517	1.121	0.714	0
B	1	0.138	0.642	0.144	0.380
C	1	0.313	0	0.170	0.274
D	1	0.069	0.321	0.072	0.190
E	0.977	0.110	0.513	0.115	0.304
F	0.867	0.155	0.722	0.162	0.427

Table 3. Cross-efficiencies of Example 1

	Cross-efficiencies with respect to					
DMU	A	B	C	D	E	F
A	1	1	0.711	1	1	1
B	0.350	1	0.650	1	1	1
C	1	0.831	1	0.831	0.831	0.831
D	0.411	1	1	1	1	1
E	0.430	0.977	0.921	0.977	0.977	0.977
F	0.411	0.867	0.648	0.867	0.867	0.867

Table 4. Fuzzy efficiencies of Example 1

DMU	Efficiency
A	(0.904,0.952,1.000)
B	(0.667,0.833,1.000)
C	(0.830,0.886,0.943)
D	(0.802,0.901,1.000)
E	(0.766,0.877,0.977)
F	(0.642,0.755,0.867)

Table 5. Extended fuzzy preference relation of Example 1

F(A,B)	A	B	C	D	E	F
A	0.000	0.237	0.131	0.102	0.150	0.394
B	-0.237	0.000	-0.106	-0.135	-0.087	0.158
C	-0.131	0.106	0.000	-0.029	0.019	0.263
D	-0.102	0.135	0.029	0.000	0.048	0.293
E	-0.150	0.087	-0.019	-0.048	0.000	0.244
F	-0.394	-0.158	-0.263	-0.293	-0.244	0.000

there are two-third of DMUs with efficiency equal to 1 which indicate the discriminating power of classical DEA is very low in this case. Each row of Table 3 shows the efficiencies of a DMU measured in different aspect, which is referred as peer evaluation. Note that the efficiencies in the diagonal of Table 3 are the results of classical DEA. To enhance the discriminating power, we regard each row of Table 3 as the overall performance of a DMU. Therefore, each row is further encapsulated into a fuzzy number shown in Table 4 by (4). Extended fuzzy preference relation in (5) is employed to identify the extended fuzzy preference relation of DMUs shown in Table 5 which is further converted into preference intensities shown in Table 6 by (6). The strength and weakness of a DMU are identified by (7) and (8) respectively as shown in Table 7. Total performance index of a DMU is obtained by aggregating the strength and weakness with (9), which is shown in Table 8 accompanied by the rank. As we can see from Table 8, the DMUs with efficiencies equal to 1 are differentiated with total performance index.

Table 6. Preference intensities of Example 1

Q(A,B)	A	B	C	D	E	F
A	0.000	0.237	0.131	0.102	0.150	0.394
B	0.000	0.000	0.000	0.000	0.000	0.158
C	0.000	0.106	0.000	0.000	0.019	0.263
D	0.000	0.135	0.029	0.000	0.048	0.293
E	0.000	0.087	0.000	0.000	0.000	0.244
F	0.000	0.000	0.000	0.000	0.000	0.000

Table 7. The strength and weakness of Example 1

DMU	Strength(S_i)	Weakness(W_i)
A	1.014	0.000
B	0.158	0.565
C	0.389	0.160
D	0.505	0.102
E	0.331	0.218
F	0.000	1.352

Table 8. Total performance indices and ranking of Example 1

DMU	Total performance index	Rank
A	1.000	1
B	0.218	5
C	0.708	3
D	0.832	2
E	0.603	4
F	0.000	6

5 Conclusions

Low discrimination of DEA results in that a complete ranking of DMUs with DEA has been one of the concerns of decision makers. In this paper we incorporated fuzzy set theory with DEA to enhance the discriminating power of DEA. We have proposed a method to encapsulate the efficiencies of an unit in different aspects as a fuzzy efficiency. A fuzzy efficiency is further compared with other fuzzy efficiencies to determine its strength and weakness based on extended fuzzy preference. With the strength and weakness of an unit, we aggregate them into a total performance index so that a complete ranking of units is obtained. Finally, we have provided a example to justify that our method provides a high discrimination in DEA applications compared to classical DEA.

Acknowledgement

This research work was partially supported by the National Science Council of the Republic of China under grant No. NSC95-2416-H-019-005-.

References

1. Andersen, P., Petersen, N.C.: A procedure for ranking efficient units in data envelopment analysis. Management Science 39(10), 1261–1294 (1993)
2. Charnes, A., Cooper, W.W., Rhodes, E.: Measuring the efficiency of decision-making units. European Journal of Operational Research 2, 429–444 (1978)
3. Friedman, L., Sinuany-Stern, Z.: Scaling units via the canonical correlation analysis and the data envelopment analysis. European Journal of Operational Research 100, 629–637 (1997)
4. Fodor, J., Roubens, M.: Fuzzy Preference Modelling and Multicriteria Decision Support. Kluwer, Dordrecht (1994)
5. Kacprzyk, J.: Group decision making with a fuzzy linguistic majority. Fuzzy Sets and Systems 18, 105–118 (1986)
6. Lee, H.-S.: A new fuzzy ranking method based on fuzzy preference relation. In: 2000 IEEE International Conference on Systems, Man And Cybernetics, pp. 3416–3420. IEEE Computer Society Press, Los Alamitos (2000)
7. Lee, H.-S.: Optimal consensus of fuzzy opinions under group decision making environment. Fuzzy Sets and Systems 132, 303–315 (2002)
8. Lee, H.-S.: On fuzzy preference relation in group decision making. International Journal of Computer Mathematics 82(2), 133–140 (2005)
9. Lee, H.-S.: A fuzzy multi-criteria decision making model for the selection of distribution center. In: Wang, L., Chen, K., Ong, Y.S. (eds.) ICNC 2005. LNCS, vol. 3612, pp. 1290–1299. Springer, Heidelberg (2005)
10. Lee, H.-S.: A fuzzy method for evaluating suppliers. In: Wang, L., Jiao, L., Shi, G., Li, X., Liu, J. (eds.) FSKD 2006. LNCS (LNAI), vol. 4223, pp. 1035–1043. Springer, Heidelberg (2006)
11. Lee, H.-S., Chou, M.-T.: A fuzzy multiple criteria decision making model for airline competitiveness evaluation. In: Gabrys, B., Howlett, R.J., Jain, L.C. (eds.) KES 2006. LNCS (LNAI), vol. 4252, pp. 902–909. Springer, Heidelberg (2006)
12. Lee, H.-S., Lin, K., Fang, H.-H: A fuzzy multiple objective DEA for the human development index. In: Gabrys, B., Howlett, R.J., Jain, L.C. (eds.) KES 2006. LNCS (LNAI), vol. 4252, pp. 922–928. Springer, Heidelberg (2006)
13. Lee, H.-S., Yeh, C.-H.: A multiple criteria decision making model based on fuzzy multiple objective DEA. In: Gabrys, B., Howlett, R.J., Jain, L.C. (eds.) KES 2006. LNCS (LNAI), vol. 4252, pp. 917–921. Springer, Heidelberg (2006)
14. Orlovski, S.A.: Decision-making with fuzzy preference relations. Fuzzy Sets and Systems 90, 155–167 (1978)
15. Sexton, T.R., Silkman, R.H., Hogan, A.J.: Data envelopment analysis: Critique and extensions. In: Silkman, R.H. (ed.) Measuring Efficiency: An Assessment of Data Envelopment Analysis, pp. 73–105. Jossey-Bass, San Francisco, CA (1986)
16. Torgersen, A.M., Forsund, F.R., Kittelsen, S.A.C.: Slack-adjusted efficiency measures and ranking of efficient units. The Journal of Productivity Analysis 7, 379–398 (1996)

A New Information Fusion Method for Fuzzy Information Retrieval

Hsuan-Shih Lee[1], Ming-Tao Chou[2,⋆], Wei-Kuo Tseng[3], Hsin-Hsiung Fang[1,3], and Chen-Huei Yeh[1,4]

[1] Department of Shipping and Transportation Management
National Taiwan Ocean University
Keelung 202, Taiwan
[2] Department of Aviation and Maritime Management
Chang Jung Christian University
Tainan 711, Taiwan
mtchou@mail.cjcu.edu.tw
[3] Department of Shipping and Business Management
China College of Marine Technology and Commerce
Taipei 111, Taiwan
[4] Yang Ming Marine Transport Corporation

Abstract. In this paper, we present a new prioritization method for information fusion under fuzzy environment. We propose a new measure to measure the similarity between generalized trapezoidal fuzzy numbers. Based on the proposed similarity measure, we develop an utility function of generalized trapezoidal fuzzy numbers. With the utility function of generalized trapezoidal fuzzy numbers, we propose a new method to handle the problems of fuzzy information retrieval where fuzzy numbers are used to represent the degrees of strength at which documents satisfy prioritized criteria.

Keywords: Information fusion; Prioritized criteria, Similarity.

1 Introduction

Fuzzy subsets provide an useful structure for the representation and subsequent manipulation of information. The information represented by the fuzzy subsets can be of many different types, for example it could be a preference function in a decision making problem [2] or the representation of uncertain information via a possibility distribution [7]. For a decision making problem in which we have to select a solution from a set of alternatives based on the satisfaction of some criteria, Bellman and Zadeh [1] suggested that we can represent these criteria as fuzzy subsets over the space of alternatives. Yager [6] considered the conjunction fuzzy subsets in situations in which a priority relationship exists between the sets being aggregated. Prioritization is manifested by the fact the influence of some the arguments is determined by its compatibility with higher priority argument.

⋆ Corresponding author.

B. Apolloni et al. (Eds.): KES 2007/ WIRN 2007, Part II, LNAI 4693, pp. 1293–1298, 2007.
© Springer-Verlag Berlin Heidelberg 2007

Yager [6] modified the intersection operator introduced in [4] and constructed prioritized aggregation operators using fuzzy modeling techniques [5].

Consider an information retrieval problem in which users can retrieve the document that has the highest satisfaction of prioritized criteria in which a priority relationship exits between the criteria being evaluated. Under fuzzy environment, the degrees of satisfaction of criteria may be represented by fuzzy numbers. Assume the fuzzy numbers under consideration are generalized trapezoidal fuzzy numbers. In this paper, we propose a new similarity measure to measure the similarity between generalized trapezoidal fuzzy numbers. We also propose a method to defuzzify the degree of the strength at which a document satisfies a specific criterion. We propose a new fusion method to aggregate all degrees of satisfaction of prioritized criteria as the degree of satisfaction of user's query.

2 Similarity of Generalized Trapezoidal Fuzzy Numbers

Definition 1. *A generalized trapezoidal fuzzy number* $A = (a_1, a_2, a_3, a_4; h_a)$ *is a fuzzy number with piecewise linear membership function* μ_A *defined by*

$$\mu_A(x) = \begin{cases} \frac{(x-a_1)h_a}{a_2-a_1}, & a_1 \le x \le a_2, \\ h_a, & a_2 \le x \le a_3, \\ \frac{(a_4-x)h_a}{a_4-a_3}, & a_3 \le x \le a_4, \\ 0, & otherwise, \end{cases}$$

where $0 \le h_a \le 1$.

If $h_a = 1$, A is normal and thus a trapezoidal fuzzy number we usually refer to. That is, the definition of trapezoidal fuzzy numbers is generalized by this definition.

To measure the similarity of generalized trapezoidal fuzzy numbers, we extend the similarity measure in [3]. The similarity between generalized trapezoidal fuzzy numbers $A = (a_1, a_2, a_3, a_4; h_a)$ and $B = (b_1, b_2, b_3, b_4; h_b)$ is given by

$$S_p(A, B) = ([1 - \frac{(\sum_{i=1}^{4} |a_i - b_i|^p)^{\frac{1}{p}}}{4u^p}] \times \frac{1}{d(h_a, h_b)})^{0.5} \tag{1}$$

where

$$d(h_a, h_b) = e^{|lnh_a - lnh_b|} = (\frac{h_a}{h_b})^{(\frac{ln\frac{h_a}{h_b}}{|lnh_a - lnh_b|})} \tag{2}$$

and u is the span of the universe of discourse.

The similarity measure $S_p(A, B)$ consists of two parts. The first part

$$1 - \frac{(\sum_{i=1}^{4} |a_i - b_i|^p)^{\frac{1}{p}}}{4u^p} \tag{3}$$

is the similarity measure in [3], which measures the similarity in supports, and the second part is

$$\frac{1}{d(h_a, h_b)} \tag{4}$$

which measures the similarity in heights of two fuzzy numbers. Since the height of a fuzzy number is a ratio scale between 0 and 1, we take the inverse of log distance between two heights as the similarity. Overall similarity is obtained by taking the geometric mean of the similarity in supports and the similarity in heights.

Example 1
Let $A = (1, 1, 1, 1; 1)$ and $B = (1, 1, 1, 1; 0.5)$ be two generalized trapezoidal fuzzy numbers. Suppose the support of fuzzy numbers lies in [0,1]. Then $u = 1$. Hence the similarity in supports of A and B is

$$1 - \frac{(\sum_{i=1}^{4} |a_i - b_i|^p)^{\frac{1}{p}}}{4u^p} = 1$$

and the similarity in heights is

$$\frac{1}{d(h_a, h_b)} = 1/2 = 0.5.$$

The height of A is two times of B. Therefore the similarity in heights is 0.5. By taking the geometric mean of these two similarities, we have the overall similarity 0.707107. The reason why we take the geometric mean is that both similarities are ratio scales.

3 The Prioritized Information Fusion Method

Assume there are n documents d_1, d_2, ..., d_n in an information system and m different prioritized criteria C_1, C_2, ..., C_m to be satisfied for a specific query. Let $C_j(d_i)$ denote the degree of strength at which document d_i satisfies criterion C_j. Assume the degree $C_j(d_i)$ is a generalized trapezoidal fuzzy number whose support is in [0,1]. That is, if $C_j(d_i) = (a, b, c, d; f)$ then $0 \le a, b, c, d, f \le 1$.

Definition 2. *The utility of $C_j(d_i)$ is defined to be*

$$U_{ji} = \frac{S_p((1, 1, 1, 1; 1), C_j(d_i))}{S_p((1, 1, 1, 1; 1), C_j(d_i)) + S_p((0, 0, 0, 0; 1), C_j(d_i))}. \tag{5}$$

Example 2
Assume the degree at which document d_1 satisfies criterion C_2 is $(0.3, 0.4, 0.5, 0.6; 0.9)$. That is, $C_2(d_1) = (0.3, 0.4, 0.5, 0.6; 0.9)$. If the Euclidean distance is assumed in (1), then $p = 2$. The similarity between the largest degree $(1, 1, 1, 1; 1)$ and $C_2(d_1)$ is

$$S_2((1, 1, 1, 1; 1), C_2(d_1)) = 0.804635.$$

Then similarity between the smallest degree $(0, 0, 0, 0; 1)$ and $C_2(d_1)$ is

$$S_2((0, 0, 0, 0; 1), C_2(d_1)) = 0.831471.$$

Hence the utility of $C_2(d_1)$ is

$$U_{21} = \frac{S_2((1,1,1,1;1), C_2(d_1))}{S_2((1,1,1,1;1), C_2(d_1)) + S_2((0,0,0,0;1), C_2(d_1))} = 0.491799.$$

Let

$$\overline{U_{ji}} = \sum_{i=1}^{n} U_{ji}/n \tag{6}$$

and

$$\overline{U_{ki}} = \sum_{i=1}^{n} U_{ki}/n. \tag{7}$$

The correlation between criterion C_j and criterion C_k is given by

$$\gamma(C_j, C_k) = \frac{\sum_{i=1}^{n}(U_{ji} - \overline{U_{ji}})(U_{ki} - \overline{U_{ki}})}{\sqrt{\sum_{i=1}^{n}(U_{ji} - \overline{U_{ji}})^2}\sqrt{\sum_{i=1}^{n}(U_{ki} - \overline{U_{ki}})^2}}. \tag{8}$$

Definition 3. *The compatibility between criteria C_j and C_k is defined to be*

$$\rho_{jk} = \frac{1 + \gamma(C_j, C_k)}{2}. \tag{9}$$

If the criteria are prioritized in the way that a criterion with smaller index has higher priority, the fusion weights of criteria are given by

$$W(C_j) = \begin{cases} \frac{1}{1+\rho_{12}}, & j = 1 \\ \frac{\rho_{12}}{1+\rho_{12}} \times \frac{\rho_{23}}{1+\rho_{23}} \times \ldots \times \frac{\rho_{(j-1)j}}{1+\rho_{(j-1)j}} \times \frac{1}{1+\rho_{j(j+1)}}, & 1 < j < m \\ \frac{\rho_{12}}{1+\rho_{12}} \times \frac{\rho_{23}}{1+\rho_{23}} \times \ldots \times \frac{\rho_{(m-1)m}}{1+\rho_{(m-1)m}}, & j = m \end{cases} \tag{10}$$

Moreover, it is very easy to verify the following property:

Property 1. *The fusion weight given by (10) satisfies*

$$\sum_{j=1}^{m} W(C_j) = 1 \tag{11}$$

and

$$W(C_j) > W(C_k) \text{ if } k > j. \tag{12}$$

The fused degree of satisfaction for document d_i is given by

$$F(d_i) = W(C_1) \times U_{1i} + W(C_2) \times U_{2i} + \ldots + W(C_m)U_{mi}. \tag{13}$$

Example 3

Assume there are there are 3 documents to be searched with respect to three prioritized criteria C_1, C_2 and C_3. The criterion with the highest priority is C_1 and the criterion with lowest priority is C_3. Assume the utilities of criteria with respect to each document are shown in Table 1. Following (8), we have the correlation coefficients of criteria shown in Table 2.

Table 1. The utilities of criteria with respect to different document

Utility	C_1	C_2	C_3
d_1	0.9	0.7	0.6
d_2	0.8	0.88	0.92
d_3	0.4	0.9	0.6

Table 2. The correlation between criteria

Correlation	C_1	C_2	C_3
C_1	1	-0.72058	0.327327
C_2	-0.72058	1	0.419314
C_3	0.327327	0.419314	1

Table 3. The compatibility between criteria

Compatibility	C_1	C_2	C_3
C_1	1	0.139712	0.663663
C_2	0.139712	1	0.709657
C_3	0.663663	0.709657	1

The compatibility between criteria is shown in Table 3. Substituting compatibility into (10), we have the following fusing weights of criteria:

$$W(C_1) = 0.877415,$$

$$W(C_2) = 0.071702$$

and

$$W(C_3) = 0.050884.$$

Finally, the fused degrees of satisfaction for documents are shown in Table 4.

Table 4. The fused degrees of satisfaction

	d_1	d_2	d_3
$F(d_i)$	0.8704	0.8118	0.4460

The prioritized fuzzy fusing method is summarized as follows:

Algorithm. Prioritized_information_fusion

1. Compute the utilities of criteria with respect to documents by (5).
2. Compute the correlation between criteria by (8).
3. Compute the compatibility between criteria by (9).
4. Compute fusion weights of the prioritized criteria by (10).
5. Aggregate all degrees of satisfaction by (13).

4 Conclusions

In this paper, we have proposed a prioritized fusing method for aggregation information under fuzzy environment. The fuzzy information is assumed to be in form of generalized trapezoidal fuzzy numbers. We have proposed a new similarity measure to measure the similarity between generalized trapezoidal fuzzy numbers. We further developed an utility function of generalized trapezoidal fuzzy numbers so that compatibility between criteria can be computed. Based on the compatibility relationship obtained, we have proposed a method to compute the fusion weights of multiple criteria in which a priority relationship exists between the criteria being evaluated.

Acknowledgment

This research work was partially supported by the National Science Council of the Republic of China under grant No. NSC95-2416-H-019-005-.

References

1. Bellman, R.E., Zadeh, L.A.: Decision-making in a fuzzy environment. Management Science 17(4), 141–164 (1970)
2. Dubois, D., Prade, H.: The three semantics of fuzzy sets. Fuzzy Sets and Systems 90(2), 141–150 (1997)
3. Lee, H.-S.: Optimal consensus of fuzzy opinions under group decision making environment. Fuzzy Sets and Systems 132, 303–315 (2002)
4. Yager, R.R.: Nonmonotonic set theoretic operations. Fuzzy Sets and Systems 42, 173–190 (1991)
5. Yager, R.R., Filev, D.P.: Essentials of Fuzzy Modeling and Control, New York (1994)
6. Yager, R.R.: Structures for prioritized fusion of fuzzy information. Information Sciences 108, 71–90 (1998)
7. Zadeh, L.A.: Fuzzy sets as a basis for a theory of possibility. Fuzzy Sets and Systems 1, 3–28 (1978)

An Analytic Approach for Synthesizing Comparison Matrices in AHP

Hsuan-Shih Lee[1], Cheng-Chi Chung[1], Su-Man Wang[2], Ming-Tao Chou[3], and Chen-Huei Yeh[1,4]

[1] Department of Shipping and Transportation Management
National Taiwan Ocean University, Keelung 202, Taiwan
[2] Department of International Trade
St. John's University, Tamsui, Taipei 251, Taiwan
[3] Department of Aviation and Maritime Management
Chang Jung Christian University, Tainan 711, Taiwan
[4] Yang Ming Marine Transport Corporation

Abstract. In AHP, judgements of decision makers are represented by positive reciprocal matrices (pairwise comparison matrices) with ratio judgements representing the preference of a decision maker. Each decision maker's opinion is characterized by a pairwise comparison matrix in the AHP. Under group decision environment, a group comparison matrix has to be drawn from the matrices of the members of the group as the consensus of the group. In 2001, Yeh, Kerng and Lin [11] proposed a goal programming method to synthesize the pairwise comparison matrices into a consensus matrix for the group. Yeh et al. proposed a generic algorithm to find a solution approximating the objective function proposed by them. In this paper, we are going to show that their formulation can be solved analytically and the optimal solution can be easily computed.

Keywords: AHP, Comparison matrices, Group decision.

1 Introduction

In the AHP context of a group decision, each matrix represents the opinion of a decision maker. It is vital to synthesize the set of matrices of decision makers in the group decision. It is an important issue regarding how to draw a group consensus under the group decision environment. Several methods have been proposed to draw a group consensus from a group of experts under the fuzzy environment [2,3,10].

Basically, there are three types of the methods appeared in the literatures to aggregate AHP matrices: the geometric mean method (GMM) [8], the weighted arithmetic mean method (WAMM) [6], and the goal programming [1,4,11]. With GMM, a entry of the group matrix is obtained by taking the geometric mean of corresponding entries of all matrices and the group priority vector is then derived from the group matrix. With WAMM, all matrices are prioritized first to obtain the priority vector of each matrix. Then priority vectors are synthesized into the

B. Apolloni et al. (Eds.): KES 2007/ WIRN 2007, Part II, LNAI 4693, pp. 1299–1306, 2007.
© Springer-Verlag Berlin Heidelberg 2007

group priority vector by WAMM. In the goal programming methods, the group priority vector is obtained by optimizing the objective functions that measure the fitness of the group priority vector. Among these methods, all matrices are equal footing per se.

Yeh *et al.* [11] formulated the problem of AHP matrices synthesis as a mathematical programming problem with the goal that minimizes the squared Euclidean distance among the consensus priority vector and the priority vectors of the matrices to be synthesized. Yeh et al. designed a genetic algorithm to find the solution that meets the objective function. However, the algorithm is limited to the solution matrix with 9-scale ratio values. The algorithm will not work if the solution space consists of general positive reciprocal matrices.

In this paper, we are going to show that the formulation of Yeh *et al.*[11] can be solved analytically. The solution can be easily computed. In the formulation of Yeh *et al.* [11], the scale is from 1 to 9 rather than any other of the possible scales so that solution can be encoded as a binary string. Since the problem can be solved analytically, our solution is applicable to the matrices with ratios of real numbers like the formulations of Bryson [1] and Mikhailov [4].

2 The Analytic Approach

Suppose there are R numbers of decision makers participating in a group decision activity. In the group decision activity, there are n criteria or factors to be considered. Each decision maker represents their preference of criteria with a pairwise comparison matrix. Let $A_r = (a_{kjr})_{n\times n}$ denote the comparison matrix of rth decision maker, where a_{kjr} is the preference ratio of the kth criterion over the jth criterion of the rth decision maker. Let w_{ri} be the weight of the ith criterion derived from comparison matrix A_r. For exact weights of criteria, the Eigenvector method may be employed. There are several practical ways to find the priorities (weights) instead of solving the eigenvalue problem. To be comparable with the result of the work of Yeh et al. [11], we follow their convention which calculates the weight by

$$w_{ri} = \frac{(\prod_{j=1}^{n} a_{ijr})^{1/n}}{\sum_{k=1}^{n}(\prod_{j=1}^{n} a_{kjr})^{1/n})}.$$

Let $A = (a_{ij})_{n\times n}$ be the consensus matrix derived from $A_1, A_2, \ldots, A_R$. The priorities derived from A is calculated in the same way:

$$w_i = \frac{(\prod_{j=1}^{n} a_{ij})^{1/n}}{\sum_{k=1}^{n}(\prod_{j=1}^{n} a_{kj})^{1/n}}.$$

Suppose there are n pairwise comparison matrices $(A_1, A_2, \ldots, A_n)$ to be aggregated. Yeh et al. [11] formulated the following optimization problem to find the consensus matrix $A = (a_{ij})_{n\times n}$:

$$\begin{array}{l} \min \sum_{i=1}^{n} \sqrt{\sum_{r=1}^{R}(w_i - w_{ri})^2} \\ s.t.\ \sum_{i=1}^{n} w_i = 1 \end{array} \tag{1}$$

where

$$w_i = \frac{(\prod_{j=1}^{n} a_{ij})^{1/n}}{\sum_{k=1}^{n}(\prod_{j=1}^{n} a_{kj})^{1/n}} \tag{2}$$

and

$$w_{ri} = \frac{(\prod_{j=1}^{n} a_{ijr})^{1/n}}{\sum_{k=1}^{n}(\prod_{j=1}^{n} a_{kjr})^{1/n}}. \tag{3}$$

Yeh *et al.* proposed a genetic algorithm to find consensus matrix A that would minimize (1).

We find that problem (1) has analytic solution. Before we proceed to derive analytic solution of (1), some results of mathematical programming are reviewed.

Definition 1. *[7] Solution* $\mathbf{x}$ *is a stationary point of a smooth function* f *if* $\nabla f(\mathbf{x}) = 0$.

Theorem 1. *[7] A stationary point of a smooth function* f *is an unconstrained strict local maximum if the Hessian matrix at the point is negative definite. A stationary point is an unconstrained strict local minimum if the Hessian matrix is positive definite.*

Theorem 2. *[5] Matrix* H *is positive definite if all the eigenvalues of* H *are greater than 0.*

Definition 2. *[7] The Lagrangian function associated with function* $f(\mathbf{x})$ *over constraints* $g_1(\mathbf{x}) = b_1, \ldots, g_m(\mathbf{x}) = b_m$ *is*

$$L(\mathbf{x}, \mathbf{v}) = f(\mathbf{x}) + \sum_{i=1}^{m} v_i(b_i - g_i(\mathbf{x}))$$

where v_i *is the Lagrange multiplier for the* i*th constraint.*

Theorem 3. *[7] If* $(\mathbf{x}^*, \mathbf{v}^*)$ *is a stationary point of Lagrangian function* $L(\mathbf{x}, \mathbf{v})$ *and* $\mathbf{x}^*$ *is an unconstrained (local) optimum of* $L(\mathbf{x}, \mathbf{v}^*)$*, then* $\mathbf{x}^*$ *is a (local) optimum of original constrained problem.*

The following theorem presents the necessary and sufficient condition for problem (1):

Theorem 4. *The objective function (1) will be minimized if and only if*

$$w_i = \frac{\sum_{r=1}^{R} w_{ri}}{R}. \tag{4}$$

Proof: Let $\mathbf{w} = (w_1, w_2, \ldots, w_n)$. The Lagrangian function associated with the objective function (1) subject to $\sum_{i=1} w_i = 1$ is

$$L(\mathbf{w}, \lambda) = \sum_{i=1}^{n} \sqrt{(\sum_{r=1}^{R}(w_i - w_{ri})^2)} + \lambda(\sum_{i=1}^{n} w_i - 1). \tag{5}$$

For (**w**) to be the minimum of (1), it must satisfy the following:

$$\frac{\partial L}{\partial w_i} = \frac{\sum_{r=1}^{R}(w_i - w_{ri})}{\sqrt{\sum_{r=1}^{R}(w_i - w_{ri})^2}} + \lambda = 0 \tag{6}$$

and

$$\frac{\partial L}{\partial \lambda} = \sum_{i=1}^{n} w_i - 1 = 0. \tag{7}$$

(7) yields

$$\sum_{i=1}^{n} w_i = 1 \tag{8}$$

and (6) yields

$$\frac{Rw_i - \sum_{r=1}^{R} w_{ri} + \lambda\sqrt{\sum_{r=1}^{R}(w_i - w_{ri})^2}}{\sqrt{\sum_{r=1}^{R}(w_i - w_{ri})^2}} = 0. \tag{9}$$

Since

$$\sqrt{\sum_{r=1}^{R}(w_i - w_{ri})^2} > 0,$$

we have

$$Rw_i = \sum_{r=1}^{R} w_{ri} + \lambda\sqrt{\sum_{r=1}^{R}(w_i - w_{ri})^2}. \tag{10}$$

By summing up both sides of (10), we obtain

$$\sum_{i=1}^{n} Rw_i = \sum_{i=1}^{n}\sum_{r=1}^{R} w_{ri} + \lambda\sum_{i=1}^{n}\sqrt{\sum_{r=1}^{R}(w_i - w_{ri})^2}.$$

Since

$$\sum_{i=1}^{n} w_i = 1$$

and

$$\sum_{i=1} w_{ri} = 1,$$

we have

$$R = R + \lambda\sum_{i=1}^{n}\sqrt{\sum_{r=1}^{R}(w_i - w_{ri})^2}.$$

Hence

$$\lambda = 0. \tag{11}$$

Substituting (11) into (6), we have

$$w_i = \frac{\sum_{r=1}^{R} w_{ri}}{R}. \tag{12}$$

The necessary condition for (1) is thus obtained.

In the reset of the proof, we shall show that the necessary condition is also a sufficient one. Let

$$\mathbf{w}^* = (w_1^*, w_2^*, \ldots, w_n^*)$$

and

$$\lambda^* = 0$$

where

$$w_i^* = \frac{\sum_{r=1}^{R} w_{ri}}{R}.$$

Then $(\mathbf{w}^*, \lambda^*)$ is a stationary point of $L(\mathbf{w}, \lambda)$. Let

$$L'(\mathbf{w}) = L(\mathbf{w}, \lambda^*).$$

Then $\mathbf{w}^*$ is a stationary point of $L'(\mathbf{w})$. The Hessian matrix of $L'(\mathbf{w})$ is

$$H = (\frac{\partial^2 L'}{\partial w_i \partial w_j})$$

where

$$\frac{\partial^2 L'}{\partial w_i \partial w_j} = 0 \text{ for } i \neq j$$

and

$$\frac{\partial^2 L'}{\partial w_i \partial w_i} = -(\sum_{r=1}^{R}(w_i - w_{ri})^2))^{-\frac{3}{2}}(\sum_{r=1}^{R}(w_i - w_{ri}))^2 + R\sqrt{\sum_{r=1}^{R}(w_i - w_{ri})^2}.$$

Eigenvalues of H are

$$-(\sum_{r=1}^{R}(w_i - w_{ri})^2))^{-\frac{3}{2}}(\sum_{r=1}^{R}(w_i - w_{ri}))^2 + R\sqrt{\sum_{r=1}^{R}(w_i - w_{ri})^2}. \tag{13}$$

When $w_i = w_i^*$, (13) becomes

$$R\sqrt{\sum_{r=1}^{R}(w_i - w_{ri})^2}. \tag{14}$$

Since

$$R\sqrt{\sum_{r=1}^{R}(w_i - w_{ri})^2} > 0,$$

by Theorem 2, H is positive at $(\mathbf{w}^*, \lambda^*)$. According to Theorem 1, $\mathbf{w}^*$ is a strict local minimum of (1) subject to $\sum_{i=1}^{n} w_i = 1$. Since $\mathbf{w}^*$ is the only stationary point, $\mathbf{w}^*$ is the minimum of (1) subject to $\sum_{i=1}^{n} w_i = 1$. □

Once the consensus priorities are obtained by (4), the optimal consensus comparison matrix A can be derived according to the following theorem.

Theorem 5. *The objective function (1) will be minimized when*

$$a_{ij} = \frac{\sum_{r=1}^{R} w_{ri}}{\sum_{r=1}^{R} w_{rj}}. \tag{15}$$

Proof: If $a_{ij} = \frac{\sum_{r=1}^{R} w_{ri}}{\sum_{r=1}^{R} w_{rj}}$, following (2), we have

$$w_i = \frac{(\prod_{j=1}^{n} a_{ij})^{1/n}}{\sum_{k=1}^{n}(\prod_{j=1}^{n} a_{kj})^{1/n}} = \frac{(\prod_{j=1}^{n}(\frac{\sum_{r=1}^{R} w_{ri}}{\sum_{r=1}^{n} w_{rj}}))^{1/n}}{\sum_{k=1}^{n}(\prod_{j=1}^{n}(\frac{\sum_{r=1}^{R} w_{rk}}{\sum_{r=1}^{R} w_{rj}}))^{1/n}}$$
$$= \frac{\sum_{r=1}^{R} w_{ri}}{\sum_{k=1}^{n}\sum_{r=1}^{R} w_{rk}} = \frac{\sum_{r=1}^{R} w_{ri}}{\sum_{r=1}^{R}\sum_{k=1}^{n} w_{rk}} = \frac{\sum_{r=1}^{R} w_{ri}}{R}$$

which will minimize (1) following Theorem 4. That is, (1) reaches its minimum when

$$a_{ij} = \frac{\sum_{r=1}^{R} w_{ri}}{\sum_{r=1}^{R} w_{rj}}.$$ □

3 Numeric Illustration

To be comparable with the result of Yeh et al [11], we follow the same example in [11], which is shown in Table 1, 2 and 3. The result of Yeh is shown in Table 4. Applying our analytic method to the same data, we have different result as shown in Table 5. To measure the quality of the solution, Yeh et al used the following fitness function:

$$\sum_{i=1}^{n} \sqrt{\sum_{r=1}^{R} (w_i - w_{ri})^2} \tag{16}$$

Table 1. The comparison matrix of the design manager

	cost	manufacturability	quality	tech. improvement	market share
cost	1	5	7	3	1/3
manufacturability	1/5	1	3	1/3	1/5
quality	1/7	1/3	1	1/7	1/9
tech. improvement	1/3	3	7	1	1/3
market share	3	5	9	3	1
weight	0.2812	0.0320	0.0320	0.1590	0.4582

Table 2. The comparison matrix of the manufacturing manager

	cost	manufacturability	quality	tech. improvement	market share
cost	1	1/3	1	1/2	3
manufacturability	3	1	3	1	5
quality	1	1/3	1	1/3	3
tech. improvement	2	1	3	1	5
market share	1/3	1/5	1/3	1/5	1
weight	0.1418	0.3498	0.1312	0.3217	0.0555

Table 3. The comparison matrix of the marketing manager

	cost	manufacturability	quality	tech. improvement	market share
cost	1	7	5	4	3
manufacturability	1/7	1	1/3	1/4	1/5
quality	1/5	3	1	1/3	1/4
tech. improvement	1/4	4	3	1	1
market share	1/3	5	4	1	1
weight	0.4877	0.0435	0.0809	0.1781	0.2098

Table 4. The consensus matrix obtained by the method of Yeh et al

	cost	manufacturability	quality	tech. improvement	market share
cost	1.0000	3.0000	3.2500	2.6000	1.6770
manufacturability	0.3333	1.0000	0.9000	0.7143	0.7500
quality	0.3080	1.1111	1.0000	0.2857	0.3333
tech. improvement	0.3846	1.4000	3.5000	1.0000	0.6923
market share	0.6000	1.3333	3.0000	1.4444	1.0000
weight	0.3743	0.1228	0.0893	0.1867	0.2270

Table 5. The consensus matrix obtained by the analytic method

	cost	manufacturability	quality	tech. improvement	market share
cost	1.0000	2.1413	3.7308	1.3824	1.2587
manufacturability	0.4670	1.0000	1.7423	0.6456	0.5878
quality	0.2680	0.5739	1.0000	0.3705	0.3374
tech. improvement	0.7234	1.5490	2.6989	1.0000	0.9106
market share	0.7944	1.7012	2.9639	1.0982	1.0000
weight	0.3074	0.1436	0.0824	0.2224	0.2442

Table 6. Quality of the solution

	fitness
Yeh's method	0.44718
Our analytic method	0.45028

As shown in Table 6, the fitness of the analytic method is 0.44718 while the fitness of the genetic method proposed by Yeh et al is 0.45028, which indicates that the analytic method outperforms Yeh's genetic method.

4 Conclusions

Based on the same optimization formulation of Yeh, we have proposed an analytic approach toward the formulation. We have shown that the analytic approach outperforms the genetic algorithm proposed by Yeh. The outperformance of the analytic method has been shown by rigid proof and illustration. We have proposed an analytic approach which is simpler and more accurate than the genetic method.

Acknowledgements

This research work was partially supported by the National Science Council of the Republic of China under grant No. NSC-95-2416-H-019-005.

References

1. Bryson, N.: Generating consensus priority point vectors: a logarithmic goal programming. Computers & Operations Research 26, 637–643 (1999)
2. Kacprzyk, J., Fedrizzi, M., Nurmi, H.: Group decision making and consensus under fuzzy preferences and fuzzy majority. Fuzzy Sets and Systems 49, 21–31 (1992)
3. Lee, H.-S.: Optimal consensus of fuzzy opinions under group decision making environment. Fuzzy Sets and Systems 132, 303–315 (2002)
4. Mikhailov, L.: Group prioritization in the AHP by fuzzy preference programming method. Computers & Operations Research 31, 293–301 (2004)
5. Nakos, G., Joyner, D.: Linear Algebra with Application. Brroks/Cole Publishing Company, Pacific Grove, CA (1998)
6. Ramanathan, R., Ganesh, L.S.: Group preference aggregation methods employed in AHP: an evaluation and an intrinsic process for deriving members' weightages. European Journal of Operational Research 79, 249–265 (1994)
7. Rockafellar, R.T.: Lagrange multipliers and optimality. SIAM Review 35, 183–238 (1993)
8. Saaty, T.L.: The Analytic Hierarchy Process. McGraw-Hill, New York (1980)
9. Sen, A.: Quasi-transitivity, rational choice and collective decisions. Review of Economic Studies 36, 381–393 (1969)
10. Tanino, T.: On group decision making under fuzzy preferences. In: Kacprzyk, J., Fedrizzi, M. (eds.) Multiperson Decision Making Using Fuzzy Sets and Possibility Theory, pp. 172–185. Kluwer Academic Publishers, Dordrecht (1990)
11. Yeh, J.M., Kreng, B., Lin, C.: A consensus approach for synthesizing the elements of comparison matrix in the Analytic Hierarchy Process. International Journal of System Science 32(11), 1353–1363 (2001)

Estimating Missing Values in Incomplete Additive Fuzzy Preference Relations

Hsuan-Shih Lee[1], Ming-Tao Chou[2], Hsin-Hsiung Fang[1,3,⋆], Wei-Kuo Tseng[3], and Chen-Huei Yeh[1,4]

[1] Department of Shipping and Transportation Management
National Taiwan Ocean University
Keelung 202, Taiwan
[2] Department of Aviation and Maritime Management
Chang Jung Christian University
Tainan 711, Taiwan
[3] Department of Shipping and Business Management
China College of Marine Technology and Commerce
Taipei 111, Taiwan
[4] Yang Ming Marine Transport Corporation
fhhbear@yahoo.com.tw

Abstract. In some occasions, decision makers may have to deal with the problems in which only partial information is available. As a result, decision makers embody their preferences as incomplete fuzzy preference relations. In this paper, we propose an iterative procedure to estimate the missing values of the incomplete fuzzy preference relations that are assumed to be additive. The procedure is based on additive transitivity property. Measures of consistency and completeness of an incomplete fuzzy preference are also developed to assist decision makers to identify the quality of their decisions.

Keywords: Estimation, Incomplete fuzzy preference relation, Transitivity.

1 Introduction

Decision processes involve tasks such as identifying possible alternatives, evaluating alternatives and selecting alternatives. In the evaluation of alternatives, decision makers usually compare alternatives pairwisely against the criterion under consideration. One of the most popular representations of preference is fuzzy preference relation [6], which provides preference degrees between all pairs of alternatives. Since decision makers focus on pairwise comparison, it would be easier for decision makers to specify their preference. However, preferences provided in such local-focus manner may result in global inconsistency. Due to the limited knowledge, decision makers may not be able to differentiate some pairs of alternatives to provide their preference such that some entries of the preference

⋆ Corresponding author.

B. Apolloni et al. (Eds.): KES 2007/ WIRN 2007, Part II, LNAI 4693, pp. 1307–1314, 2007.
© Springer-Verlag Berlin Heidelberg 2007

relation remains unknown. Decision makers' reluctance to provide preference that they are not sure of results in incomplete fuzzy preference relation [3,4,5].

One way to recover the missing values in an incomplete fuzzy preference relation is to obtain the value from other decision makers [5]. Alonso et al. [1] proposed a different way to estimate the missing values based on additive transitivity. In their paper, they suggested the preference of alternative i over alternative j can be estimated if the preference of alternative i over alternative k and the preference of alternative k over alternative j are known. The alternative k is an intermediator acting as the bridge between alternative i and j. However, if the preference of alternative k over i and the preference of alternative j over k are available but the preference of alternative i over k and the preference of alternative k over j are unknown, they could not estimate the preference of alternative i over j via alternative k. To remedy their drawbacks, we propose a four-way method to estimate missing values in place of their "one-way" estimation. Since more information is adopted in our method, the estimation power and accuracy of our method would be higher.

2 Additive Fuzzy Preference Relations

Let $N = \{1, 2, \ldots, n\}$. We rephrase the definition of fuzzy preference relations given in [2,6] as follows:

Definition 1. *Let $R = (r_{ij})_{n\times n}$ be a preference relation. R is called a fuzzy preference relation if $r_{ij} \in [0, 1]$ and $r_{ii} = 0.5$ for all $i, j \in N$.*

The interpretation of r_{ij} is as follows. $r_{ij} = 0.5$ indicates that the preference between alternative i and alternative j is indifferent. If $r_{ij} > 0.5$, then alternative i is preferred to alternative j and vice versa.

The property of additive transitivity for fuzzy preference relations is homomorphic to the transitivity property for pairwise comparison matrices in AHP advocated by Saaty [7]. The additive transitivity given by Tanino [8] is as follows:

$$r_{ij} - 0.5 + r_{jk} - 0.5 = r_{ik} - 0.5. \tag{1}$$

If all elements in a fuzzy preference relation satisfy equation (1), then there would be no contradiction in that relation. Furthermore, we will have the reciprocal property:

$$r_{ij} + r_{ji} = 1, \forall i, j \in N. \tag{2}$$

Hence, we have the following definition:

Definition 2. *Let $R = (r_{ij})_{n\times n}$ be a fuzzy preference relation. R is called an additive consistent fuzzy preference relation if it satisfies the following:*

$$r_{ij} + r_{jk} = r_{ik} + 0.5, \forall i, j, k \in N \tag{3}$$

If some of the values in a fuzzy preference relation R are unknown, R is said to be incomplete. If a fuzzy preference relation is assumed to be additively consistent, it must satisfy equation (3). Let j be the intermediator between i and k. r_{ik} can be estimated in four different ways with intermediator j:

Case 1. Since $r_{ij} + r_{jk} = r_{ik} + 0.5$, we can estimate r_{ik} by

$$r_{ik}^{j1} = r_{ij} + r_{jk} - 0.5. \tag{4}$$

Case 2. Since $r_{ji} + r_{ik} = r_{jk} + 0.5$, we can estimate r_{ik} by

$$r_{ik}^{j2} = r_{jk} - r_{ji} + 0.5. \tag{5}$$

Case 3. Since $r_{ik} + r_{kj} = r_{ij} + 0.5$, we can estimate r_{ik} by

$$r_{ik}^{j3} = r_{ij} - r_{kj} + 0.5. \tag{6}$$

Case 4. Since $(r_{ik} + r_{kj} - 0.5) + (r_{ji} - 0.5) = r_{ii} = 0.5$, we have

$$r_{ik} + r_{kj} = 1.5 - r_{ji}. \tag{7}$$

Hence we can estimate r_{ik} by

$$r_{ik}^{j4} = 1.5 - r_{ji} - r_{kj}. \tag{8}$$

The estimation of r_{ik} via intermediator j can be obtained by

$$\hat{r}_{ik}^{j} = (r_{ik}^{j1} + r_{ik}^{j2} + r_{ik}^{j3} + r_{ik}^{j4})/4 \tag{9}$$

Let the discrepancy between r_{ik} and $\hat{r}_{ik}^{j}$ be

$$d_{ik}^{j} = |\hat{r}_{ik}^{j} - r_{ik}| \tag{10}$$

The average discrepancy of r_{ik} can be obtained by

$$d_{ik} = \frac{\sum_{j \in N} d_{ik}^{j}}{n} \tag{11}$$

The consistence level of r_{ik} can be defined as

$$c_{ik} = 1 - d_{ik} \tag{12}$$

The consistency of fuzzy preference relation R can be defined as follows:

Definition 3. *The consistency of fuzzy preference relation R is defined to be*

$$c_R = \frac{\sum_{i \in N, k \in N, i \neq k} c_{ik}}{n^2 - n} \tag{13}$$

If R is an incomplete fuzzy preference relation, some of entries are unknown. For an intermediator j between i and k, equation (4), (5), (6) and (8) may not applicable. Therefore, equation (9) may not suitable for R. To develop the estimation function of r_{ik} via intermediator j for an incomplete fuzzy preference relation R, we define functional intermediator as follows.

Definition 4. *The intermediator j is said to be a functional intermediator of (i,k) if one of (4), (5), (6) and (8) is applicable. That is, one of $\{r_{ij}, r_{jk}\}$, $\{r_{jk}, r_{ji}\}$, $\{r_{ij}, r_{kj}\}$ and $\{r_{ji}, r_{kj}\}$ is known.*

Let

$$F_{ik}^{j} = \{l | r_{ik}^{jl} \text{ is computable}\}. \tag{14}$$

That is, if $F_{ik}^{j} \neq \phi$, j is a functional intermediator for (i,k). For an incomplete fuzzy preference relation R, the estimation of r_{ik} via functional intermediator j is obtained by

$$\hat{r}_{ik}^{j} = \frac{\sum_{l \in \{1,2,3,4\} \text{ and } r_{ik}^{jl} \text{ is computable}} r_{ik}^{jl}}{|F_{ik}^{j}|} \tag{15}$$

Let

$$F_{ik} = \{j | j \text{ is a functional intermediator of } (i,k)\}.$$

For an incomplete fuzzy preference relation R, the estimation of r_{ik} can be obtained by

$$\hat{r}_{ik} = \frac{\sum_{j \in F_{ik}} \hat{r}_{ik}^{j}}{|F_{ik}|} \tag{16}$$

For an incomplete fuzzy preference relation R, the discrepancy of r_{ik} with respect to j can be obtained by

$$d_{ik}^{j} = |\hat{r}_{ik}^{j} - r_{ik}| \tag{17}$$

The discrepancy of r_{ik} can be obtained by

$$d_{ik} = \frac{\sum_{j \in F_{ik}} d_{ik}^{j}}{|F_{ik}|} \tag{18}$$

The consistency level of r_{ik} can be defined as

$$c_{ik} = 1 - d_{ik} \tag{19}$$

Let

$$K = \{(i,k) | i \neq k, r_{ik} \text{ is known }, i,k \in N\}. \tag{20}$$

The consistency of an incomplete fuzzy preference relation R is defined as follows:

Definition 5. *The consistency of an incomplete fuzzy preference relation R is defined to be*

$$c_R = \frac{\sum_{(i,k) \in K} c_{ik}}{|K|}. \tag{21}$$

Since some values in an incomplete fuzzy preference relation are missing, the information provided for estimation may not be adequate. We propose the following measure to measure the completeness of estimation information.

Definition 6. *For an incomplete fuzzy preference relation R, the completeness of the estimation information for r_{ik} with respect to functional intermediator j is defined to be*

$$p_{ik}^{j} = \frac{|F_{ik}^{j}|}{4}. \tag{22}$$

The following definition follows immediately.

Definition 7. *For an incomplete fuzzy preference relation R, the completeness of the estimation information for r_{ik} is defined to be*

$$p_{ik} = \frac{\sum_{j \in F_{ik}} p_{ik}^{j}}{|F_{ik}|}. \tag{23}$$

The completeness of the estimation information provided by incomplete fuzzy preference relation R can be defined as follows:

Definition 8. *For an incomplete fuzzy preference relation R, the completeness of R is defined to be*

$$p_R = \frac{\sum_{i,k \in N, i \neq k} p_{ik}}{n^2 - n}. \tag{24}$$

Following the definition of consistency, a consistent fuzzy preference relation may not be complete whereas a complete fuzzy preference relation may not be consistent. A fuzzy preference relation to be informative should be as complete and consistent as possible. Therefore, we combine the two concepts "consistency" and "completeness" to propose the concept of "informativeness" in the following definition.

Definition 9. *The informativeness of a fuzzy preference relation R is defined to be*

$$I_R = \alpha c_R + (1 - \alpha) p_R \tag{25}$$

where $0 \leq \alpha \leq 1$.

3 Estimation Procedure

Let U denote the set of all unknown entries. Let l be the index of the iteration and every variable associated with a superscript (l) denote the sate of the variable in iteration l. Initially, the index is set to zero by letting $U^{(0)} = U$, $F_{ik}^{(0)} = F_{ik}$, and $R^{(0)} = R$. In the procedure, $r_{ik}^{(l)jt}$ is given by (4), (5), (6) and (8) by letting $R = R^{(l)}$. The procedure to estimate missing values is as follows:

Procedure of Estimation

$l = 0$

Do while $U^{(l)} \neq \phi$

$U^{(l+1)} = U^{(l)}$, $R^{(l+1)} = R^{(l)}$

For $(i,k) \in U^{(l)}$ and $F_{ik}^{(l)} \neq 0$ do

$$\hat{r}_{ik}^{(l)j} = \frac{\sum_{t\in\{1,2,3,4\}} \text{and } {}_{r_{ik}^{(l)jt}} \text{ is computable } r_{ik}^{(l)jt}}{|F_{ik}^{(l)j}|}$$
$$\hat{r}_{ik}^{(l)} = \frac{\sum_{j\in F_{ik}^{(l)}} \hat{r}_{ik}^{(l)j}}{|F_{ik}^{(l)}|}$$
$$r_{ik}^{(l+1)} = \hat{r}_{ik}^{(l)}$$
$$U^{(l+1)} = U^{(l)} - \{(i,k)\}$$
$$l = l + 1$$

End of Estimation

Consider the following fuzzy preference relation over a set of four alternatives:

$$R = \begin{pmatrix} 0.5 & x & 0.7 & x \\ 0.4 & 0.5 & x & 0.7 \\ 0.3 & x & 0.5 & x \\ x & 0.4 & x & 0.5 \end{pmatrix}$$

where x denotes the entry is unknown. Then set of unknown entries

$$U = \{(1,2),(1,4),(2,3),(3,2),(3,4),(4,1),4,3)\}.$$

Since $F_{12}^{(0)3} = F_{12}^{(0)4} = \phi$, there is no functional intermediator of $(1,2)$. Hence r_{12} is unestimable. Likewise, r_{34} and r_{43} are unestimable. The functional intermediator of (1,4) is 2. $F_{14}^{(0)2} = \{2,4\}$. Hence

$$\begin{aligned} &\hat{r}_{14}^{(0)2} \\ &= (\hat{r}_{14}^{(0)22} + \hat{r}_{14}^{(0)24})/2 \\ &= ((r_{24}^{(0)} - r_{21}^{(0)} + 0.5) + (1.5 - r_{21}^{(0)} - r_{42}^{(0)}))/2 \\ &= ((0.7 - 0.4 + 0.5) + (1.5 - 0.4 - 0.4))/2 \\ &= 0.75 \end{aligned}$$

Thus

$$\hat{r}_{14}^{(1)} = 0.75.$$

We can have the following estimation in the same way:

$$\hat{r}_{23}^{(1)} = 0.6, \hat{r}_{32}^{(1)} = 0.4, \hat{r}_{41}^{(1)} = 0.25$$

After iteration $l = 0$, the fuzzy preference relation becomes

$$R = \begin{pmatrix} 0.5 & x & 0.7 & 0.75 \\ 0.4 & 0.5 & 0.6 & 0.7 \\ 0.3 & 0.4 & 0.5 & x \\ 0.25 & 0.4 & x & 0.5 \end{pmatrix}$$

The unknown set becomes

$$U^{(1)} = \{(1,2),(3,4),(4,3)\}.$$

Since

$$\begin{aligned}
&\hat{r}_{12}^{(1)3} \\
&= (\hat{r}_{12}^{(1)31} + \hat{r}_{12}^{(1)32} + \hat{r}_{12}^{(1)33} + \hat{r}_{12}^{(1)34})/4 \\
&= ((r_{13}^{(1)} + r_{32}^{(0)} + 0.5) + (r_{32}^{(1)} - r_{31}^{(1)} + 0.5) \\
&\quad + (r_{13}^{(1)} - r_{23}^{(1)} + 0.5) + (1.5 - r_{31}^{(1)} - r_{23}^{(1)}))/4 \\
&= (0.6 + 0.6 + 0.6 + 0.6)/4 \\
&= 0.6
\end{aligned}$$

and

$$\begin{aligned}
&\hat{r}_{12}^{(1)4} \\
&= (\hat{r}_{12}^{(1)41} + \hat{r}_{12}^{(1)42} + \hat{r}_{12}^{(1)43} + \hat{r}_{12}^{(1)44})/4 \\
&= ((r_{14}^{(1)} + r_{42}^{(0)} + 0.5) + (r_{42}^{(1)} - r_{41}^{(1)} + 0.5) \\
&\quad + (r_{14}^{(1)} - r_{24}^{(1)} + 0.5) + (1.5 - r_{41}^{(1)} - r_{24}^{(1)}))/4 \\
&= (0.65 + 0.65 + 0.55 + 0.55)/4 \\
&= 0.6
\end{aligned}$$

we have

$$\begin{aligned}
\hat{r}_{12}^{(2)} &= (\hat{r}_{12}^{(1)3} + \hat{r}_{12}^{(1)4})/2 \\
&= 0.6.
\end{aligned}$$

Estimating the other unknown in the same way, we have

$$\hat{r}_{34}^{(2)} = 0.55$$

and

$$\hat{r}_{43}^{(2)} = 0.45.$$

After iteration $l = 1$, we have the following complete fuzzy preference relation:

$$R = \begin{pmatrix} 0.5 & 0.6 & 0.7 & 0.75 \\ 0.4 & 0.5 & 0.6 & 0.7 \\ 0.3 & 0.4 & 0.5 & 0.55 \\ 0.25 & 0.4 & 0.45 & 0.5 \end{pmatrix}.$$

4 Conclusions

Due to the lack of knowledge, the preference provided by decision makers may be incomplete. The problem of learning missing values from known values for incomplete fuzzy preference relations is addressed in this paper. Based on additive

consistency, we have proposed a four-way estimation method to learn missing values from known values. For each pair of preference, we estimate it based on four different transitivity via an intermediator. In the future research, we will focus the estimability of incomplete fuzzy preference relation based on additive transitivity. The necessary and sufficient condition for an incomplete fuzzy preference relation to be estimable shall be investigated in the future study.

Acknowledgment

This research work was partially supported by the National Science Council of the Republic of China under grant No. NSC95-2416-H-019-005-

References

1. Alonso, S., Chiclana, F., Herrera, F., Herrera-Viedma, E.: A learning procedure to estimate missing values in fuzzy preference relations. In: Torra, V., Narukawa, Y. (eds.) MDAI 2004. LNCS (LNAI), vol. 3131, pp. 227–238. Springer, Heidelberg (2004)
2. Kacprzyk, J.: Group decision making with a fuzzy linguistic majority. Fuzzy Sets and Systems 18, 105–118 (1986)
3. Kim, S.H., Ahn, B.S.: Group decision making procedure considering preference strength under incomplete information. Computers and Operations Research 24, 1101–1112 (1997)
4. Kim, S.H., Ahn, B.S.: Interactive group decision making procedure under incomplete information. European Journal of Operational Research 116, 498–507 (1999)
5. Kim, S.H., Choi, S.H., Kim, J.K.: An interactive procedure for multiple attribute group decision making with incomplete information: Range-based approach. European Journal of Operational Research 118, 139–152 (1999)
6. Orlovski, S.A.: Decision-making with fuzzy preference relations. Fuzzy Sets and Systems 1, 155–167 (1978)
7. Saaty, T.L.: The Analytic Hierarchy Process. McGraw-Hill, New York (1980)
8. Tanino, T.: Fuzzy preference orderings in group decision making. Fuzzy Sets and Systems 12, 117–131 (1984)

Fuzzy Logic Based Reputation System for Mobile Ad Hoc Networks

Jin-Long Wang and Shih-Ping Huang

Department of Information and Telecommunication Engineering
Ming Chuan University, Taipei 11120, Taiwan
jlwang@mcu.edu.tw

Abstract. Mobile ad hoc networks (MANET) consist of some mobile nodes, but do not have the requirement of infrastructure or centralized management. As a result, there may have a lot of critical issues occurred in mobile ad hoc networks, such as the security of data transmission, routing, and so on. The misbehavior of nodes which are due to selfish or malicious may lead to the losing of packets, rejecting of services, etc. In this paper, a new scheme based on reputation, bandwidth and hop count is proposed to solve these problems. This scheme employs the fuzzy logic to perform the routing path decision in order to choose the best and feasible routing path. Finally, the SMPL, a simulation tool, is utilized to estimate the efficiency of the proposed scheme. From the results, the proposed scheme is superior to other existing schemes.

Keywords: Mobile ad hoc networks, Reputation systems, Bandwidth, Hop count, Fuzzy logic routing path decision.

1 Introduction

The mobile ad hoc network (MANET) is composed of some nodes that communicate with other nodes through wireless network, which is a kind of peer-to-peer communication. MANET is a kind of self-organized network and uses multi-hop routes to forward data. Thus, without the cooperation of all nodes the MANET will not be able to transmit data among nodes. Namely, the efficacy of a MANET will be reduced, if some nodes cannot work properly, such as misbehavior or intentional behavior. Misbehavior of a node means that the node routes or forwards abnormally; the intentional behaviors are those nodes which selfishly mind its advantage or attack malevolently other nodes.

In the past, some of researches are based on the evaluation of reputation to determine the choice of steady routing. However, the reputation is not the only consideration factor for determining routing path. This is because the topology of MANET is changing all the time and the network status is not always steady due to the movement frequently and unsteady bandwidth.

This paper considers not only the reputation on the MAMET, but also two other critical factors, including the bandwidth and the hop count. This paper proposes the new evaluation mechanism of nodes reputation and the fuzzy logic decision scheme

B. Apolloni et al. (Eds.): KES 2007/ WIRN 2007, Part II, LNAI 4693, pp. 1315–1322, 2007.
© Springer-Verlag Berlin Heidelberg 2007

for determining the most feasible routing path. At last the SMPL, a simulation tool, is utilized to estimate the efficiency of the proposed scheme. From the results, the proposed scheme is superior to other existing schemes.

In this paper, some previous works are discussed in section two. The system model will be introduced in section three. The evaluation mechanism and the decision model are presented in section four. In section five, the simulations and results are illustrated. The efficacy of the propose scheme is compared with other methods. Finally, the conclusion remark is given in section six.

2 Previous Works

In this section, some previous works are discussed, including the routing protocol and the reputation systems in MANET.

2.1 Routing Protocol of MANET

Currently, there are many routing protocols in MANET. These routing protocols can be categorized into two classes, including the proactive routing protocol and the reactive routing protocol.

In the proactive routing protocol, each node will broadcast the routing information for each fixed period of time. When a node receives the routing information, it will change its routing table accordingly. Because the packet carrying the routing information is sent to destination immediately with little delay, it will consume the bandwidth of wireless network and the capacity of nodes.

In the reactive routing protocol, such as the Ad Hoc On-Demand Distance Vector (AODV) and the Dynamic Source Routing (DSR), one node will discover the routing information of network only when it has some packets to send. The benefit of this protocol is that the required bandwidth is less than the proactive routing protocol.

In this paper, the proposed scheme is based on routing protocol of AODV to discover the feasible and reliable path.

2.2 Reputation Systems

The reputation system of MANET is very important, because it is able to avoid misbehavior of nodes. Some reputation systems based on reactive routing protocol have been proposed for MANET, such as CONFIDANT[1], CORE[5], Self-policing[2], and SORI[3].

CONFIDANT proposed by Buchegger, S. and Le Boudec, J.Y.[1] is the abbreviation of Cooperation of Nodes, Fairness in Dynamic Ad Hoc Networks. CONFIDAT is based on the algorithm of Bayesian reputation and trust system which is used to distinguish the normal nodes from the misbehavior nodes. In this manner, the routing path will evade the misbehavior nodes, which will be isolated and cannot do anything.

CORE (collaborative reputation) proposed by Michiardi, P. and Molva, R.[5] can be monitored by using Watchdog. The mechanism of reputation system which is composed of subjective reputation, indirect reputation and functional reputation can employ the value of reputation to determine whether a node should be isolated or not.

Self-policing scheme is proposed by Buchegger, S. and Boudec, J. V.[2]. By using the self-policing scheme, one node is able to avoid the interference of misbehavior nodes and operates correctly. In this scheme, every node can discover the misbehavior nodes according to the information collected by themselves or from other nodes. Once a misbehavior node is located, that node will be isolated from the network.

In the previous studies, the source node always passes the packet to the neighbor node with the highest reputation. However, the neighbor node may not be on the most effective path. It may conduct the packet to an unstable area, as shown in Figure 1.

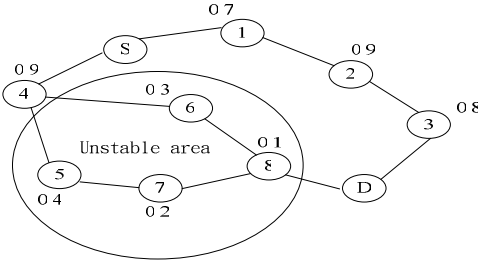


Fig. 1. The disadvantage of choosing the node with the highest reputation

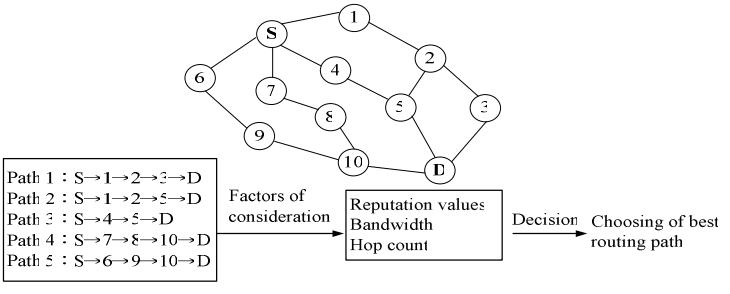


Fig. 2. The disadvantage of choosing the node with the highest reputation

3 System Model

In this paper, three important factors are taken into consideration on routing decision for MANET, including reputation, bandwidth and hop count. For example, in Figure 2, there are five paths between the source and the destination. For effectively routing packets, the proposed scheme is based on the fuzzy logic and three factors to discover the most feasible path.

In the following sectors, the evaluation mechanism of reputation and the method of path decision are introduced, in which the fuzzy logic decision is adopted to perform the path selection.

4 Proposed Schemes

In this section, the scheme for choosing the most feasible path in MANET is proposed, and also the evaluation mechanism of reputation and the fuzzy logic routing decision is introduced.

4.1 Evaluation Mechanism of Reputation

It is assumed that the reputation of every node is between 0 and 1 in an MENET. The lower reputation indicates that the node is low reliable, and vice versa. The reputation of a node is refreshed after forwarding packets. The reputation can be obtained by the equation (1).

$$\begin{aligned} g(n+1) &= \alpha f(n) + (1-\alpha) g(n) \\ &= \alpha f(n) + (1-\alpha)\left[\alpha f(n-1) + \alpha(1-\alpha) g(n-1)\right] \\ &= \alpha f(n) + (1-\alpha)\alpha f(n-1) + \alpha(1-\alpha)^2 g(n-1) \\ &= \alpha f(n) + (1-\alpha)\alpha f(n-1) + (1-\alpha)^2 \alpha f(n-2) + \ldots + (1-\alpha)^n \alpha f(n-n) \\ &= \sum_{i=0}^{n} \alpha(1-\alpha)^i f(n-i) \end{aligned} \tag{1}$$

$$SERR(n+1) = f(n) - g(n)$$

$$SDEV(n+1) = (1-h) \times SDEV(n) + h \times \left| SERR(n+1) \right|$$

$$REP(n+1) = g(n+1) + k \times SDEV(n+1)$$

$g(n)$ is the reputation REP_{old} predicted in the past, f(n) is the latest reputation REP_{latest}, $f(n+1)$ is the real value of reputation at time $n+1$, and $g(n+1)$ is the predicted reputation $REP_{forecast}$ at time $n+1$. Evidently, it will have a difference between $g(n+1)$ and $f(n+1)$. Hence, the *SERR* and the *SDEV* are estimated to refine the evaluation of reputation $REP(n+1)$.α is a weighted value for the latest reputation which is between 0 and 1 and can be adjusted according to the current status of MANET. When the traffic status of MANET is changed quickly, α can be set as a higher value, such as 0.8. On the other hand, when the traffic is steady, α can be set as a lower value, such as 0.3. In addition, every node has to record the reputation and the neighbor nodes in the routing table.

4.2 Collections of Bandwidth

The routing protocol of AODV is adopted by each node to evaluate the bandwidth of MANET, as shown in Figure 7. The source node S sends an RREQ message. Whenever the destination node receives an RREQ, it will respond with an RREP message. The bandwidth is normalized to the value between 0 and 1. For evaluating the bandwidth, a new field of BW_value is added to the RREP and RREQ messages. The BW_value will be the minimum bandwidth among all links on the path from the source to the destination. When the destination node receives an RREQ message, the minimum bandwidth along the path will be evaluated and put on the BW_value field of RREP message, which then will be sent back to the source node. When the source node receives the RREP packets, it will record BW_value on its routing table.

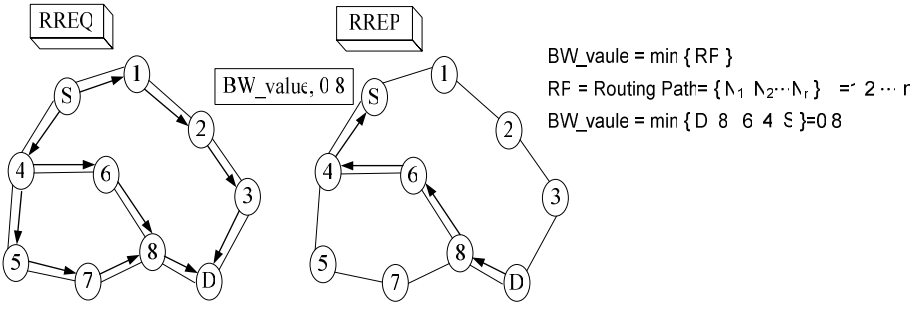


Fig. 3. Collections of bandwidth

4.3 Fuzzy Logic Routing Decision

The proposed scheme perform the routing decision is based on the fuzzy logic and three important factors, including the reputation, the bandwidth and the hop count. In this section, the fuzzy logic routing scheme is described, in which there are three main steps, including the fuzzification, the rule evaluation, and the defuzzification.

4.3.1 Fuzzification

The first step of the proposed routing scheme is fuzzification, in which the membership function should be defined. Firstly, three input variables and six membership functions are defined. Three input variables are REP(Reputation), BW(Bandwidth), and HC(hop count). Six membership functions are HREP, LREP, HBW, LBW, HHC, and LHC. The membership values can be obtained by applying the values of input variables to different membership functions. The membership value is located at the range from 0 to 1.

4.3.2 Rule Evaluation

The fuzzy rules are defined according to the membership value, in which there are a lot of IF-THEN rules for deciding which path should be chosen. Three membership values can be used to produce eight combinatorial rules. The fuzzy number (FN) of each rule is defined as the minimum value of the three membership values, as shown in function (2).

$$FN_i=\min\{HREP/LREP, HBW/LBW, HHC/LHC\} . \quad (2)$$

FN: the minimum value of each combination of membership values

i:1 to 8

If HREP=0.7, HBW=0.5, and HHC=0.2, the fuzzy number will be 0.2. Each rule may have four kinds of different results, including Yes(Y), Probably Yes (PY), Probably No (PN), and No (N). Thus, the routing path decision (RPD) can be defined as the largest fuzzy number of individual results, as shown in function (3).

$$RPD(p)=Max\{FN_1, FN_2,\ldots\ldots,FN_t\} . \quad (3)$$

RPD: p is one of the four possibilities, Y, PY and PN or N

4.3.3 Defuzzification

In the phase of defuzzification, the process of transforming the result of fuzzy inference into an exact number is called fuzzy decision (FD). A weight value is set for each possible result, including Y, PY, PN and N. The final fuzzy decision can be obtained by the equation (4).

$$FD=\frac{\sum M_m \times W_m}{\sum M_m} \tag{4}$$

M_m: RPD_m value
W_m: the weight value for the result m
m: Y, PY, PN or N

The value of FD is between 0 and 1. The path with the largest value of FD will be the most feasible choice.

5 Simulations and Results

In this section, the environment and result of our simulation are introduced. The most important environment parameters in this simulation are illustrated in table 1, including the simulation time, the simulation area, the number of nodes, and possible ratios of misbehavior.

Table 1. Simulation parameters

Simulation time	500(sec)
Simulation areas	1000m x 1000m
Number of nodes	20, 40, 60, 80, 100
Ratio of misbehavior	2%, 5%, 10%, 20%, 30%

5.1 Simulation Results

In this simulation, the performance of proposed scheme is compared with the schemes considering only reputation, bandwidth, and hop count, respectively.

From the results of Figure 4, it is evident that the packet delivery ratio will decrease as the ratio of misbehavior nodes increases. The scheme considering reputation is better than the schemes considering hop count and bandwidth in MANET. Furthermore, the packet delivery ratio of proposed scheme is superior to other existing schemes. This is due to fact that the proposed scheme considers three different factors, including the reputation, the bandwidth, and the hop count

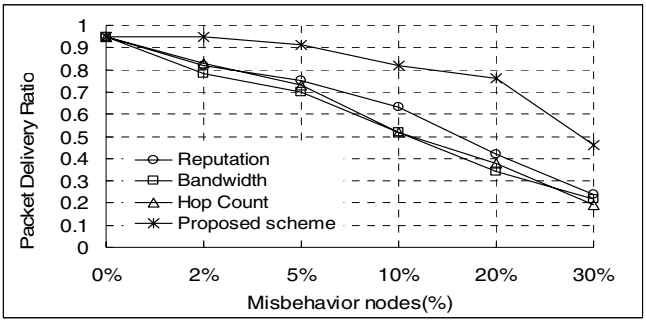


Fig. 4. Comparison of Reputation Systems, Bandwidth, Hop Count and Proposed scheme

In Figure 5, the packet delivery ratio of the proposed scheme for different number of nodes (20, 40, 60, 80 and 100) is compared with other existing schemes. The packet delivery ratio will decrease as the number of nodes increases. This is because the more the number of nodes, the more the uncertainty and the abnormal nodes in MANET. The increment of number of nodes also implies the higher ratio of lost packets. However, the proposed scheme overcomes this kind of problem by considering the reputation, the bandwidth, and the hop count.

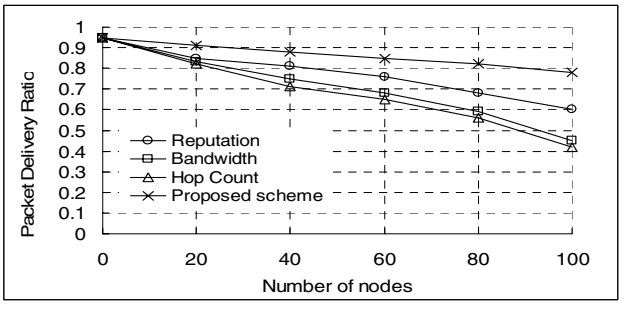


Fig. 5. Packet delivery ratio versus different number of nodes

From the results of simulation, the proposed scheme is evidently superior to other existing schemes, and able to offer the lower packet lost ratio and the higher efficiency in MANET.

6 Conclusions

In MANET, the network status is changed from time to time due to the movement and the unreliability of nodes, which will lead to the lower packet delivery ratio. In this paper, the proposed scheme considers not only the reputation, but also the bandwidth and the hop count. Besides, the fuzzy logic is used to determine the most feasible routing path. From the simulation results, it is shown that the proposed scheme can

enhance the efficiency of whole network and is superior to the existing schemes. Especially, the packet loss rate of proposed scheme performs better than other schemes even though the ratio of misbehavior nodes becomes larger.

References

1. Buchegger, S., Le Boudec, J.V.: Performance analysis of the CONFIDANT protocol: Cooperation of Nodes-Fairness in dynamic ad-hoc networks. In: Proceedings of The Third ACM International Symposium on Mobile Ad Hoc Networking and Computing, Lausanne, Switzerland, pp. 226–236. ACM Press, New York (2002)
2. Buchegger, S., Le Boudec, J.V.: Self-Policing Mobile Ad hoc Networks by Reputation Systems. In: IEEE Communications, pp. 101–107. IEEE Computer Society Press, Los Alamitos (2005)
3. He, Q., Wu, D., Khosla, P.: SORI:A Secure and Objective Reputation-based Incentive Scheme for Ad-hoc Networks. In: Proceedings of IEEE, pp. 825–830. IEEE Computer Society Press, Los Alamitos (2004)
4. Marti, S., Giuli, T.J., Lai, K., Baker, M.: Mitigating Routing Misbehavior in Mobile Ad Hoc Networks. In: Proc. MOBICOM, pp. 255–265 (2000)
5. Michiardi, P., Molva, R.: CORE: A Collaborative Reputation Mechanism to Enforce Node Cooperation in Mobile Ad Hoc Networks. In: 6th IFIP Conf. Sec. Commun. and Multimedia, Portoroz, Slovenia (2002)
6. Wang, J.L., Chen, C.W.: Fuzzy Logic Based Mobility Management for 4G Heterogeneous Networks, pp. 888–895. Springer, Heidelberg (2006)

Synthesizing Comparison Matrices of AHP Under Group Decision

Ming-Tao Chou[1], Hsuan-Shih Lee[2,⋆], Ching-Wu Chu[2], and Chao-Yuan Cheng[3]

[1] Department of Aviation and Maritime Management
Chang Jung Christian University
Tainan 711, Taiwan
[2] Department of Shipping and Transportation Management
National Taiwan Ocean University
Keelung 202, Taiwan
hslee@mail.ntou.edu.tw
[3] Department of Electrical Engineering
St. John's University
Tamsui, Taipei 251, Taiwan

Abstract. Tackling multi-criteria decision making problems with AHP, judgements of decision makers are represented by positive reciprocal matrices with ratio judgements representing the preference of a decision maker. In this paper, we are going to propose a new method for aggregating individual ratio judgments into a group consensus. With the aggregation we proposed, the objective weight of each opinion is determined automatically in such a way that it is larger if the opinion is closer to the consensus and lesser otherwise. An iterative procedure is proposed to reach the consensus.

Keywords: Aggregation, AHP, Group consensus.

1 Introduction

Ever since the advent of Analytic Hierarchy Process (AHP) [10], the researches and applications related to AHP were booming. Multi-Criteria Decision Making (MCDM) and Multi-Criteria Group Decision Making are prominent areas of research in normative decision theory. Group decision-making techniques, such as Nominal group technique, Delphi technique, dialectical approach and AHP, have been widely implemented as useful tools in the process of group decision making. Among them, AHP has been utilized successfully in group decision making to solve economic, political, social and technological problems [11].

AHP provides ratio-scale measurements of the priorities of elements in various levels of a hierarchy. These ratio-measurements are embodied in a matrix consisting of pairwise comparison, which is positive and reciprocal, representing the preference of a decision maker. Another way to model the preference of decision makers is the fuzzy preference relations, which represent the preference

⋆ Corresponding author.

B. Apolloni et al. (Eds.): KES 2007/ WIRN 2007, Part II, LNAI 4693, pp. 1323–1330, 2007.
© Springer-Verlag Berlin Heidelberg 2007

by a number between 0 and 1. The priority vector indicating the weights of the elements under consideration can be obtained by either the right eigenvector method [10] or goal programming methods [3,4,7] such as the logarithmic least squares [5].

Forman and Peniwati [6] noted that there are several possible ways to aggregate information when more than one individuals participate in a decision process, including: (1) aggregating the individual judgments for each set of pairwise comparisons into a consensus matrix (2) aggregating the individuals' derived priority vectors into a consensus priority vector. Forman et al. referred the first as aggregating individual judgments (AIJ) and the second as aggregating individual priorities (AIP). Aczel, Saaty and Roberts [1,2] have shown that when aggregating the judgements of individuals where the reciprocal property is assumed, only the geometric mean satisfies the Pareto principle and homogeneity condition. That is, for AIJ, the geometric mean must be used. As for AIP, either the arithmetic or geometric average may be assumed. The choice of AIJ or AIP under different circumstances was also addressed by Forman et al. [6]

Bryson [3] formulated the AHP matrices aggregation problem as a linear programming problem by logarithmic goal programming. Bryson suggested the process of generating group consensus priority vector consists of three stages: (1) individual specification and review (2) group synthesis (3) prioritization. Rather than explicitly focusing on each pairwise comparison, Bryson focused on the entire set of pairwise comparison values. Their method works even if the matrices are not reciprocal.

In this paper, we are going to propose a new aggregation procedure in which the objective weights are determined automatically. With the proposed procedure, the aggregation weight of a opinion is determined automatically in such a way that it is larger if the opinion is closer to the consensus and lesser otherwise. By varying the aggregation parameters, different aggregation result may be obtained. We show that the arithmetic mean also can be obtained by setting parameters to a specific value. The proposed aggregation process is an iterative procedure whose convergence is also guaranteed.

2 Problem Formulation

Definition 1. *An $n \times n$ square matrix $A = (a_{ij})$ is reciprocal if $a_{ij} = \frac{1}{a_{ij}}$ for $1 \leq i, j \leq n$.*

Definition 2. *An $n \times n$ positive reciprocal matrix A is consistent iff*

$$a_{ij}a_{jk} = a_{ik} \text{ for } 1 \leq i, j, k \leq n.$$

Assume there are m decision makers and n elements (criteria) to be compared. Let $A_k = (a_{ijk})$ be the comparison matrix given by k-th decision maker, where a_{ijk} representing the preference ratio of the i-th element over the j-th element by the judgement of the k-th decision maker. The comparison matrix A_k is positive and reciprocal. Let $A = (a_{ij})$ be the consensus matrix synthesized from

$A_1, A_2, \ldots, A_m$ and $\mathbf{w} = (w_1, w_2, \ldots, w_n)$ be its corresponding priority vector. In this paper, we are assuming the aggregation weights u_ks are unknown and want to determine the consensus vector and the aggregation weights simultaneously such that extreme opinions have lesser aggregation weights. We formalize our idea as follows. Aggregating the set of positive reciprocal matrices $A_1, A_2, \ldots, A_m$ amounts to solving the following optimization problem:

$$\text{(P):} \quad \min J_{r,c}(\mathbf{u}, \mathbf{w}) \sum_{k=1}^{m} (u_k)^r (\|\mathbf{w} - \mathbf{w}_k\|^2 + c)$$
$$\text{subject to} \sum_{i=1}^{m} u_i = 1, \sum_{j=1}^{n} w_j = 1,$$

where $r \in (1, \infty]$ and $c > 0$ are two constants to control the aggregation. How the aggregation is affected by r and c will be discussed in the subsequent of the paper. Let $(\mathbf{u}^*, \mathbf{w}^*)$ be the optimal solution of problem **(P)**. Then the consensus matrix $A = (a_{ij})$ can be obtained by letting

$$a_{ij} = \frac{w_i^*}{w_j^*}.$$

Problem **(P)** is an analytical problem which has the advantage that using differential calculus one can determine necessary conditions for local optima. In next section, the necessary conditions are derived. An iterative procedure based on the conditions is also proposed.

3 The Aggregation Algorithm

Before we proceed, some results of mathematical programming are reviewed.

Definition 3. *[9] Solution* $\mathbf{x}$ *is a stationary point of a smooth function* f *if* $\nabla f(\mathbf{x}) = 0$.

Theorem 1. *[9] A stationary point of a smooth function* f *is an unconstrained strict local maximum if the Hessian matrix at the point is negative definite. A stationary point is an unconstrained strict local minimum if the Hessian matrix is positive definite.*

Theorem 2. *[8] Matrix* H *is positive definite if all the eigenvalues of* H *are greater than 0.*

Definition 4. *[9] The Lagrangian function associated with function* $f(\mathbf{x})$ *over constraints* $g_1(\mathbf{x}) = b_1, \ldots, g_m(\mathbf{x}) = b_m$ *is*

$$L(\mathbf{x}, \mathbf{v}) = f(\mathbf{x}) + \sum_{i=1}^{m} v_i (b_i - g_i(\mathbf{x}))$$

where v_i *is the Lagrange multiplier for the* i*-th constraint.*

Theorem 3. *[9] If* $(\mathbf{x}^*, \mathbf{v}^*)$ *is a stationary point of Lagrangian function* $L(\mathbf{x}, \mathbf{v})$ *and* $\mathbf{x}^*$ *is an unconstrained (local) optimum of* $L(\mathbf{x}, \mathbf{v}^*)$*, then* $\mathbf{x}^*$ *is a (local) optimum of original constrained problem.*

Now we are ready to derive the necessary conditions for the local optima of **(P)**.

Theorem 4. *If* $(\mathbf{u}, \mathbf{w})$ *is a local minimum of* **(P)** *then*

$$u_k = \frac{(\frac{1}{\|\mathbf{w}-\mathbf{w}_k\|^2+c})^{1/(r-1)}}{\sum_{i=1}^{m}(\frac{1}{\|\mathbf{w}-\mathbf{w}_i\|^2+c})^{1/(r-1)}}, \tag{1}$$

$$\mathbf{w} = \frac{1}{\sum_{k=1}^{m}(u_k)^r}\sum_{k=1}^{m}(u_k)^r\mathbf{w}_k. \tag{2}$$

Let

$$S^* = \{(\mathbf{u}, \mathbf{w}) | (\mathbf{u}, \mathbf{w}) \text{ satisfies equations (1) and (2)}\}.$$

That is

$$S^* = \{(\mathbf{u}, \mathbf{w}) | \mathbf{w} = \frac{1}{\sum_{k=1}^{m}(u_k)^r}\sum_{k=1}^{m}(u_k)^r\mathbf{w}_k, u_k = \frac{(\frac{1}{\|\mathbf{w}-\mathbf{w}_k\|^2+c})^{1/(r-1)}}{\sum_{i=1}^{m}(\frac{1}{\|\mathbf{w}-\mathbf{w}_i\|^2+c})^{1/(r-1)}}\}. \tag{3}$$

Then, the optimal solution of (**P**), $(\mathbf{u}^*, \mathbf{w}^*) \in S^*$. We want to find a $(\mathbf{u}, \mathbf{w}) \in S^*$ as the approximation of $(\mathbf{u}^*, \mathbf{w}^*)$. The systems described by equations (1) and (2) cannot be solved analytically. However, there exists an algorithm approximating them iteratively. Actually, $(\mathbf{u}, \mathbf{w}) \in S^*$ can be approximated iteratively by

$$u_k^{(l+1)} = \frac{(\frac{1}{\|\mathbf{w}^{(l)}-\mathbf{w}_k\|^2+c})^{1/(r-1)}}{\sum_{i=1}^{m}(\frac{1}{\|\mathbf{w}^{(l)}-\mathbf{w}_i\|^2+c})^{1/(r-1)}} \tag{4}$$

and

$$\mathbf{w}^{(l+1)} = \frac{1}{\sum_{k=1}^{m}(u_k^{(l+1)})^r}\sum_{k=1}^{m}(u_k^{(l+1)})^r\mathbf{w}_k. \tag{5}$$

This can be characterized by the following theorem:

Theorem 5. $(\mathbf{u}^{(l)}, \mathbf{w}^{(l)})$ *satisfies (1) and (2) as* l *approaches infinity. That is*

$$(\mathbf{u}^{(l)}, \mathbf{w}^{(l)}) \in S^* \text{ as } l \to \infty.$$

Now, we are ready to present an iterative algorithm that aggregates pairwise comparison matrices into a consensus matrix satisfying equations (1) and (2). The procedure can be summarized in the following steps:

Algorithm A

1. Each DM (decision maker) constructs a positive reciprocal matrix representing his personal preference. Let it be A_i. Assume these matrices are with satisfactory degree of consistency.

2. Compute the priority vector $\mathbf{w}_k = (w_{1k}, w_{2k}, \ldots, w_{nk})$ of the matrix $A_k = (a_{ijk})$ for $k = 1, \ldots, m$.
3. Each iteration in this algorithm will be labeled l, where $l = 0, 1, 2, \ldots$. Let $\mathbf{u}^{(l)} = (u_1^{(l)}, u_2^{(l)}, \ldots, u_m^{(l)})$ and $\mathbf{w}^{(l)} = (w_1^{(l)}, w_2^{(l)}, \ldots, w_n^{(l)})$.
4. Set initial aggregation weights such that $0 \leq u_k^{(0)} (k = 1, 2, \ldots, m) \leq 1$ and $\sum_{k=1}^{m} u_k^{(0)} = 1$.
5. Calculate
$$\mathbf{w}^{(0)} = \frac{1}{\sum_{k=1}^{m} (u_k^{(0)})^r} \sum_{k=1}^{m} (u_k^{(0)})^r \mathbf{w}_k.$$
6. Set $l = 0$.
7. Calculate $\mathbf{u}^{(l+1)}$ as follows:
$$u_k^{(l+1)} = \frac{\left(\frac{1}{\|\mathbf{w}^{(l)} - \mathbf{w}_k\|^2 + c}\right)^{1/(r-1)}}{\sum_{k=1}^{m} \left(\frac{1}{\|\mathbf{w}^{(l)} - \mathbf{w}_k\|^2 + c}\right)^{1/(r-1)}}.$$
8. Calculate
$$\mathbf{w}^{(l+1)} = \frac{1}{\sum_{k=1}^{m} (u_k^{(l+1)})^r} \sum_{k=1}^{m} (u_k^{(l+1)})^r \mathbf{w}_k.$$
9. If $\|\mathbf{u}^{(l+1)} - \mathbf{u}^{(l)}\| > \varepsilon$, set $l = l + 1$ and go to step 7.
10. The consensus priority vector is $\mathbf{w}^{(l+1)}$. Let $\mathbf{w}^{(l+1)}$ be denoted as $\mathbf{w} = (w_1, w_2, \ldots, w_n)$. The consensus matrix $A = (a_{ij})$ is given by
$$a_{ij} = \frac{w_i}{w_j}.$$

4 Illustrative Example

Example 1

A company plans to launch a new product on the market, which is a vital move

Table 1. The comparison matrix of the design manager

	cost	manufacturability	quality	tech. improvement	market share
cost	1	5	7	3	1/3
manufacturability	1/5	1	3	1/3	1/5
quality	1/7	1/3	1	1/7	1/9
tech. improvement	1/3	3	7	1	1/3
market share	3	5	9	3	1

for the company to remain competitive and expand its market shares. There are three departments in this company. The company has to decide its business strategy of the new product based on the following factors: cost, manufacturability, quality, technological improvement, and market share. There are debates

Table 2. The comparison matrix of the manufacturing manager

	cost	manufacturability	quality	tech. improvement	market share
cost	1	1/3	1	1/2	3
manufacturability	3	1	3	1	5
quality	1	1/3	1	1/3	3
tech. improvement	2	1	3	1	5
market share	1/3	1/5	1/3	1/5	1

on the priorities of these factors. The company decides to elicit the priorities of the factors by AHP. A questionnaire survey is conducted to elicit the subject preference of the managers of the departments. The results are shown in Table 2, and Table 3. The company has to derive the priorities of the factors from these matrices and then sets up its strategic moves based on the priorities.

Following our algorithm, priority vectors from managers are derived first. These priority vectors are then aggregated into the group priority vector. The priority vector of the design manager derived by the eigenvector method is:

$$(0.281301, 0.069533, 0.032079, 0.159038, 0.458048),$$

which shows that the market share is the most important factor and the cost of the product comes second. The priority vector of the manufacturing manager derived by the eigenvector method is

$$(0.141833, 0.349706, 0.131204, 0.32169, 0.055567),$$

which shows that the manufacturing manager prefers the manufacturability first and technological improvement second. The priority vector of the marketing

Table 3. The comparison matrix of the marketing manager

	cost	manufacturability	quality	tech. improvement	market share
cost	1	7	5	4	3
manufacturability	1/7	1	1/3	1/4	1/5
quality	1/5	3	1	1/3	1/4
tech. improvement	1/4	4	3	1	1
market share	1/3	5	4	1	1

manager derived by the eigenvector method is

$$(0.487798, 0.04353, 0.080874, 0.178036, 0.209762),$$

which indicates that the cost is the most favorable factor and the market share is the second from the viewpoint of the marketing manager. From these priority vectors, we can see that the opinions of the design manager and the marketing manager are closer, and the opinion of the manufacturing manager is divergent.

In the aggregation of the their opinions, the opinions of the design manager and marketing manager should be reflected much more in the consensus than that of the manufacturing manager.

Let the aggregation parameters $r = 1.2$ and $c = 0.5$. The aggregation process of the managers' opinions is shown in Table 4. The process terminates with the aggregation weights:

$$(0.436836, 0.099758, 0.463406),$$

which coincides with our intuition. The weights of the manufacturing manager's opinion is 0.099758 which is much less compared to the other two since it is divergent from others. The consensus priority vector is

$$(0.369540, 0.0783174, 0.062936, 0.180451, 0.308754),$$

with which the consensus comparison matrix can be derived as shown in Table 5.

Table 4. Results of each iteration of the aggregation algorithm for example 1 with $r = 1.2$ and $c = 0.5$

l	$\mathbf{u}^{(l)}$	$\mathbf{w}^{(l)}$
0	(1.000000,0.000000,0.000000)	(0.281301,0.0695330,0.032079,0.159038,0.458048)
1	(0.678109,0.066219,0.255671)	(0.321771,0.0761784,0.047547,0.170607,0.383895)
2	(0.568862,0.083592,0.347546)	(0.341983,0.0777929,0.054416,0.175247,0.350560)
3	(0.511720,0.091375,0.396905)	(0.353514,0.0782670,0.058109,0.177594,0.332515)
4	(0.479625,0.095277,0.425098)	(0.360249,0.0783867,0.060187,0.178861,0.322315)
5	(0.461299,0.097319,0.441381)	(0.364179,0.0783958,0.061370,0.179559,0.316494)
6	(0.450811,0.098413,0.450776)	(0.366460,0.0783773,0.062044,0.179948,0.313170)
7	(0.444813,0.099009,0.456178)	(0.367775,0.0783573,0.062428,0.180167,0.311272)
8	(0.441387,0.099337,0.459276)	(0.368531,0.0783423,0.062647,0.180290,0.310189)
9	(0.439431,0.099521,0.461048)	(0.368964,0.0783324,0.062771,0.180359,0.309572)
10	(0.438316,0.099624,0.462061)	(0.369211,0.0783262,0.062842,0.180399,0.309220)
11	(0.437679,0.099682,0.462639)	(0.369352,0.0783225,0.062883,0.180421,0.309019)
12	(0.437317,0.099715,0.462969)	(0.369433,0.0783204,0.062906,0.180434,0.308905)
13	(0.437110,0.099733,0.463157)	(0.369479,0.0783191,0.062919,0.180441,0.308840)
14	(0.436992,0.099744,0.463264)	(0.369505,0.0783184,0.062927,0.180446,0.308803)
15	(0.436925,0.099750,0.463325)	(0.369520,0.0783180,0.062931,0.180448,0.308781)
16	(0.436886,0.099753,0.463360)	(0.369527,0.0783177,0.062933,0.180449,0.308769)
17	(0.436864,0.099755,0.463380)	(0.369534,0.0783176,0.062935,0.180450,0.308762)
18	(0.436852,0.099757,0.463391)	(0.369537,0.0783175,0.062936,0.180450,0.308758)
19	(0.436845,0.099757,0.463398)	(0.369538,0.0783175,0.062936,0.180451,0.308756)
20	(0.436841,0.099758,0.463402)	(0.369539,0.0783174,0.062936,0.180451,0.308755)
21	(0.436839,0.099758,0.463404)	(0.369540,0.0783174,0.062936,0.180451,0.308754)
22	(0.436837,0.099758,0.463405)	(0.369540,0.0783174,0.062936,0.180451,0.308754)
23	(0.436836,0.099758,0.463406)	(0.369540,0.0783174,0.062936,0.180451,0.308754)
24	(0.436836,0.099758,0.463406)	(0.369540,0.0783174,0.062936,0.180451,0.308754)

Table 5. The consensus comparison matrix of example 1 with $r = 1.2$ and $c = 0.5$

	cost	manufacturability	quality	tech. improvement	market share
cost	1	4.718	5.872	2.049	1.197
manufacturability	0.212	1	1.244	0.434009199	0.254
quality	0.170	0.804	1	0.349	0.204
tech. improvement	0.488	2.304	2.867	1	0.584
market share	0.836	3.942	4.906	1.711	1

5 Conclusions

In this paper, we have proposed an iterative procedure to aggregate comparison matrices into consensus matrix such that the consensus matrix is close to the majority and less biased by the extreme opinions. Illustrative example was provided to demonstrate the aggregation procedure.

Acknowledgements

This research work was partially supported by the National Science Council of the Republic of China under grant No. NSC-95-2416-H-019-005.

References

1. Aczel, J., Saaty, T.L.: Procedures for synthesizing ratio judgments. Journal of Mathematical Psychology 27, 93–102 (1983)
2. Aczel, J., Roberts, F.S.: On the possible merging functions. Mathematical Social Sciences 17, 205–243 (1989)
3. Bryson, N.: Generating consensus priority point vectors: a logarithmic goal programming. Computers & Operations Research 26, 637–643 (1999)
4. Cook, W.D., Kress, M.: Deriving weights from pairwise comparisons. Social Economic Planning Sciences 20(6), 341–345 (1986)
5. Fichtner, J.: On deriving priority vectors from matrices of pairwise comparisons. Social Economic Planning Sciences 20(6), 399–405 (1986)
6. Forman, E., Peniwati, K.: Aggregating individual judgments and priorities with the Analytic Hierarchy Process. European Journal of Operational Research 108, 165–169 (1998)
7. Lee, H.-S., Tseng, W.-K.: Goal programming methods for constructing additive consistency fuzzy preference relations. In: Gabrys, B., Howlett, R.J., Jain, L.C. (eds.) KES 2006. LNCS (LNAI), vol. 4252, pp. 910–916. Springer, Heidelberg (2006)
8. Nakos, G., Joyner, D.: Linear Algebra with Application. Brroks/Cole Publishing Company, Pacific Grove, CA (1998)
9. Rockafellar, R.T.: Lagrange multipliers and optimality. SIAM Review 35, 183–238 (1993)
10. Saaty, T.L.: The Analytic Hierarchy Process. McGraw-Hill, New York (1980)
11. Wedley, W.C.: Combining qualitative and quantitative factors-An analytic hierarchy approach. Socio Economic Planning Sciences 24, 57–64 (1990)

A Proposal of Adequate and Efficient Designing of UML Documents for Beginners

Masakazu Takahashi[1,2], Satoru Takahashi[3], and Yoshikatsu Fujita[4]

[1] Shimane University, 1060 Nishikawatsu, Matsue, Shimane, Japan
masakazu@cis.shimane-u.ac.jp
[2] Galaxy Express Corporation, 18-16-1 Hamamatsucho, Minato-ku, Tokyo, Japan
masakazu_takahashi@galaxy-express.co.jp
[3] Mitsui Asset Trust Banking Co., Ltd, 2-1jyonan-mishima, Tokyo, Japan
Satoru_1_Takahashi@mitsuitrust.co.jp
[4] Matsushita Electric Industrial Co., Ltd, , Tokyo, Japan
fujita.yoshikatsu@ji.panasonic.com

Abstract. In this paper, we present automatic Unified Modeling Language (UML) documents generation method using Natural Language Processing (NLP). This method consists of following five steps: 1) apply NLP to requirement specifications written in Japanese, pick up all nouns and predications, and develop noun and predication lists, 2) clarify all objects and classes from the noun list, add some operations to them, and develop the Object and Class diagrams, 3) define all message sequences and directions between classes, and develop the Sequence diagrams, 4) give a name to each Sequence diagram as a Use-Case, gather all Use-Cases, and develop the Use-Case diagrams, 5) define the states and transition of the classes, and develop the State Machine diagrams. We also developed the support tools for processing this method and could develop UML documents adequately and efficiently. Average correct answer rate of UML documents improved 20[%] in comparison with no method.

Keywords: Unified Modeling Language, Automatic Generation, Designing Education, Design Procedure, Support Tools.

1 Introduction

This paper proposes a development method of Unified Modelling Language (hereafter, UML) [7] documents (such as the Object diagrams, Class diagrams, Sequence diagram, Use-Case diagrams and State Machine diagrams) from the requirement specifications written in Japanese for designers who have little developmental experience using UML (hereafter, inexperienced designer).

Recently, the object-oriented methods become popular in software developments [3]. When we develop the software using the object-oriented methods, development documents are generally written in UML. So the companies and universities prepare and provide the UML education for inexperienced designers ardently. Otherwise, it is

B. Apolloni et al. (Eds.): KES 2007/ WIRN 2007, Part II, LNAI 4693, pp. 1331–1338, 2007.
© Springer-Verlag Berlin Heidelberg 2007

difficult to teach UML for inexperienced designers, because of following reasons: 1) information (contents) in requirement specifications as the input is not enough to develop them, 2) UML has various diagrams.

In this paper, we propose the development method that consists of following steps:

1) Conduct natural language processing to requirement specifications, and develop edited text that contains enough information for the development of UML documents.
2) Select candidates of objects (instance specifications) from edited text and clarify slots and links. Then develop the Object diagrams. Compare with the Object diagrams, and group similar objects into one Class. Then decide operations and develop the Class diagrams.
3) Align the operations in the classes according to time-lines that are written in requirement specifications. Then develop the Sequence diagrams.
4) Decide a name that abridges contents of the Sequence diagram, gather all names as Use-Cases, and develop the Use-Case diagram.
5) Identify states and transitions of each class appearing in the Sequence diagrams, and develop the State Machine diagrams using them.

2 Evaluation of Existing Methods

We explain some advanced researches and discuss issues when we apply following methods to UML developments for inexperienced designers.

There are many development processes, such as Rational Unified Process (RUP) [4], development of technical documents using UML [5] and educational courses for inexperienced designers [8]. These processes are suitable for top-down developments. But it is hard for inexperienced designers, because skills required to realize this process are very severe.

There are developed many tools, such as tools supporting repeat-design-process of UML [9], the training system, ASSET [10] and integrated development environment UML [1]. These tools are suitable to actual software developments, but it is hard to apply to education of UML development.

And many supporting methods are investigated, such as automatic UML generation method by applying inductive reasoning [2] and verification method for developed UML documents [6]. These methods are effective, but they are not suitable to education for inexperienced designers.

As the result, we need adequate development method suitable for inexperienced designers.

3 Proposed Method for Designing of UML Documents

In this chapter, we will explain the proposed method that derives UML documents from requirement specifications. Section 3.1 explains an outline of the proposed method, section 3.2 explains the support tools that we have developed.

3.1 Outline of the Proposed Method

Figure 1 explains an outline of the proposed method. The proposed method consists of five steps: 1) preparation of input data, 2) development of the Object and Class diagrams, 3) development of the Sequence diagrams, 4) development of the Use-Case Diagrams and 5) development of the State Machine Diagrams.

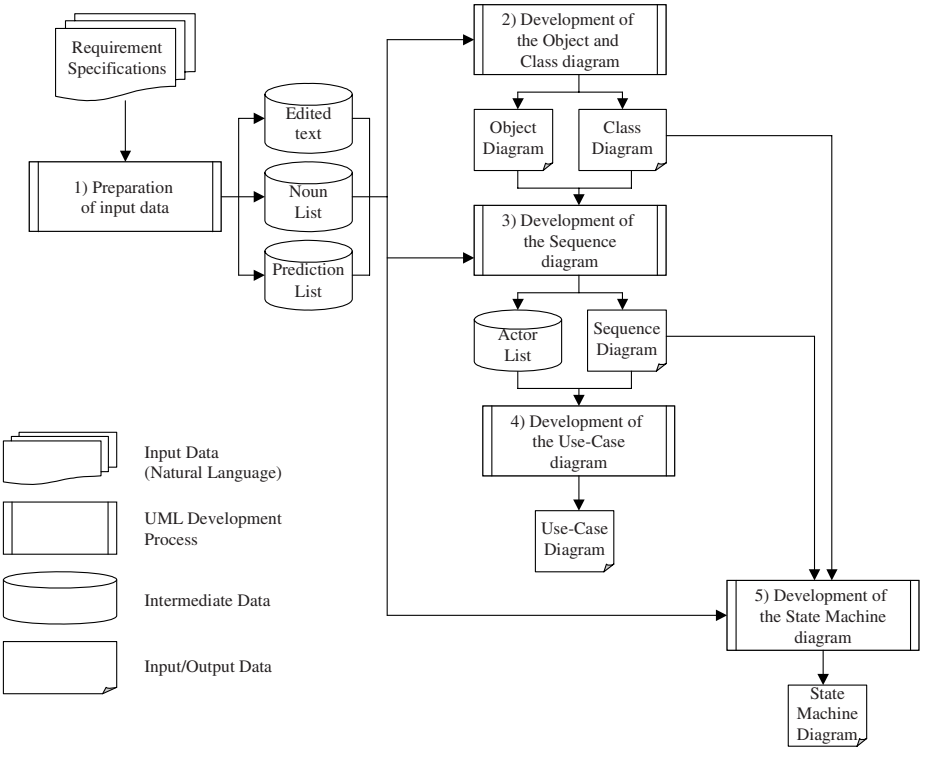


Fig. 1. Outline of the Proposed Method

Step 1), preparation of input data, consists of four sub-steps, such as developing various lists, substituting pronouns, adding subjects and objects, and dividing complex sentences. In step 1), requirement specifications are analyzed using natural language processing techniques, noun and predication list is developed by extracting nouns and predications, and edited text that has enough information to develop UML documents is developed by adding omitted information, such as subjects and objects.

Step 2), development of the Object and Class diagrams, consists of eight sub-steps, such as selecting instance specifications of objects, selecting slots of objects, setting the links between objects (above three sub-steps are for the development of the Object diagrams), defining names of classes, setting attributes of classes, setting associations between classes, setting operations of each class, and setting other information of classes (above five sub-steps are for the development of the Class diagrams). In step 2), the Object diagrams are developed by clarifying candidates of objects. Based

on the design results of the Object diagrams, Class diagrams are developed by grouping and comparing developed objects.

Step 3), development of the Sequence diagrams, consists of six sub-steps, such as developing actor list, selecting actors, deciding actors, deciding lifelines, defining all messages, setting execution specifications, and setting other information. In step 3), sctor list is developed by selecting candidates of actors from noun list. Using the list, the Sequence diagrams are developed by adding massages between lifelines (actors).

Step 4), development of the Use-Case diagrams, consists of four sub-steps, such as adoption of actors, adoption of Use-Cases, setting of Associations and setting of Subjects. In step 4), Actors and use-cases are derived from developed sequence diagrams by consolidating information. The Use-Case diagrams are developed by setting associations between actors and use-cases and giving adequate names.

Step E), development of the State Machine diagrams consists of five sub-steps, such as selecting classes that have state transitions, deciding states of each class, setting triggers and guards, setting effects and consolidating classes. In step E), classes including state transitions and states are clarified by analysing massages between actors (lifelines). And triggers and guards when one state changes to other state are clarified by analyzing the Sequence diagrams and the edited text.

3.2 Supporting Environment of Proposed Method

An outline of support tools is shown in Figure 2. The support tools consist of six sub-tools, such as the Requirement Specification Pre-Processing Tool (RSPPT), Object Diagram Development Support Tool (ODDST), Class Diagram Development Support Tool (CDDT), Sequence Diagram Development Support Tool (SDDST), Use Case Diagram Development Support Tool (UCDST) and the State Machine Diagram Development Support Tool (SMDST).

RSPPT consists of three functions, such as the instance selection, slot setting, and link setting function. CDDST consists of five functions, such as the class selection, attribute setting, method setting, association setting, and additional information setting function. SDDST consists of four functions, such as the actor selection, lifeline setting, message setting, and implementation specification setting function. UCDST consists of three functions, such as the actor setting, use case setting, and association setting function. SMDST consists of three functions, such as the state setting, state transition condition setting, and action setting function.

4 Application and Evaluation of the Proposed Method

Section 4.1 explains the experimental results applying the proposed method, and section 4.2 describes the evaluations of the results applying the proposed method.

4.1 Experimental Results of the Proposed Method

We conducted some experiments to evaluate the proposed method, and we prepared fifty testees who had university graduate level skills and some questions that

corresponded to the L2 level (L2 level means the level that the engineer can be in charge of the development for the part of UML modeling) examination conducted by the Consortium for UML based Modeling Technology Promotion. We made twenty-five testees answer with no supports (hereafter, conventional method) and other twenty-five testees answer with the proposed method (hereafter, proposed method). An abridgement from the examination is shown in Figure 3. Table 1 shows the result of correct answer rates of each UML documents.

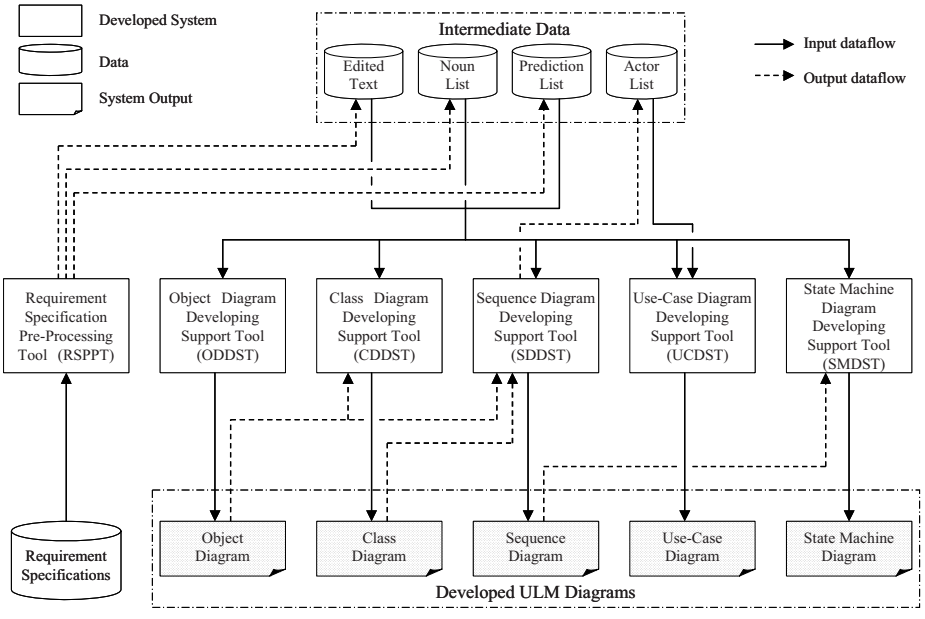


Fig. 2. Outline of Support Tools

Table 1. Results of correct answer rates

	Object and Class diagrams	Sequence diagrams	Use-Case diagrams	State Machine diagrams
Conventional method	68[%]	76[%]	80[%]	76[%]
Proposed method	84[%]	88[%]	92[%]	88[%]
Proposed/Conventional	131[%]	115[%]	116[%]	117[%]

4.2 Evaluation of the Proposed Method

This section evaluates the results of the proposed method. At first, we evaluate supporting functions of developing each UML document. Next, we evaluate whole of the proposed method.

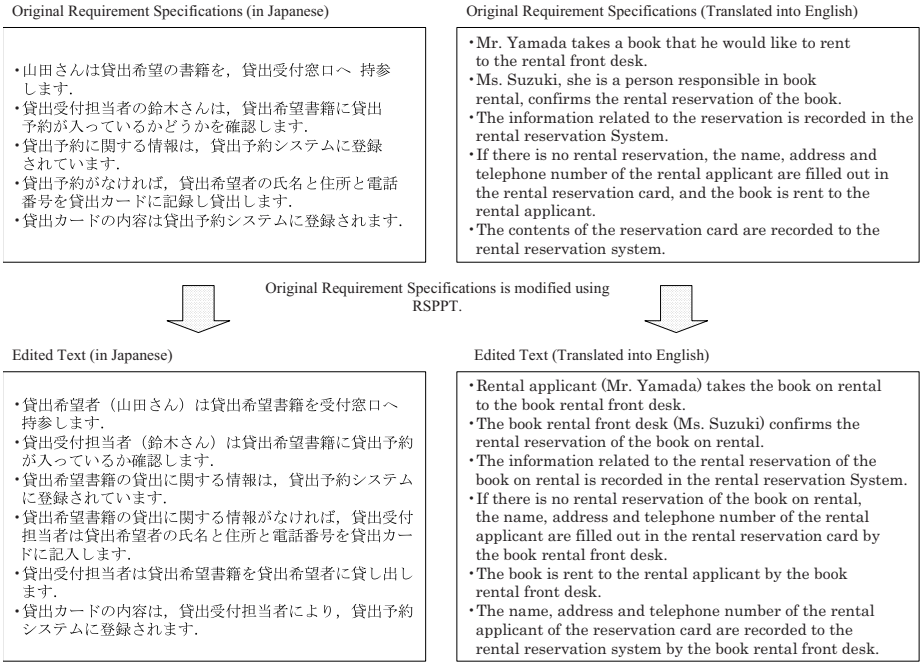


Fig. 3. Abridgement from the questions

A correct answer rate of the examinations about the Object and Class diagrams became 92[%] from 80[%], and it improved 31[%] in average. A correct answer rate of the examinations about the Object and Class diagrams are improved, because testees can develop the Object diagrams automatically only by selecting corresponding nouns in the list, and they can develop the Class diagrams by referring and comparing with the Object diagrams.

A correct answer rate of the examinations about the Sequence diagrams became 84[%] from 76[%], and it improved 17[%] in average. A correct answer rate of the examinations about the Sequence diagrams are improved, because lifelines (such as the actors, classes and objects) are clarified, and testee can develop the Sequence diagrams only by defining message transfers automatically.

A correct answer rate of the examinations about the Use-Case diagrams improved 80[%] to 92[%], and it improved 15[%] in average. A correct answer rate of the examinations about the Use-Case diagrams are improved, because testees can develop the Use-Case diagrams only by selecting corresponding actors in actor list.

A correct answer rate of the examinations about the State Machine diagrams improved 76[%] to 84[%] and it improved 17[%] in average. A correct answer rate of the examinations about the State Machine diagrams are improved, because testees can develop the State Machine diagrams only by selecting necessary information from the Sequence and Class diagrams.

Additionally, we send out questionnaires about the proposed method to the testees. Typical answers related to the proposed method are as followings.

[Merits of the proposed method]

- Because edited text includes in all necessary information, designers can concentrate on developments of the UML documents.
- Because the development process has already been defined uniquely, there is no need to consider the next operations.
- Because the UML documents are basically developed only by selecting information from prepared candidates, the UML documents are developed easily and efficiently.
- Because information transfers between diagrams is clear, it is easy to check consistency.

[Demerits of the proposed method]

- The preparation of input for developing edited text is complicated.
- Because there are some differences between method that designers would like to adopt and the proposed method, designers fell some cumbersomeness.

To investigate the results of the questionnaires, inexperienced designers tend to point out merits of the proposed method, and experienced designers tend to point out demerits of it. The proposed method cause inexperienced designers the reduction of the trial and error about the modeling. We consider this is favorable for inexperienced designers. Otherwise, it is considered that experienced designers hesitate to accept because usability of development process is limited by applying the proposed method. The proposed method provides adequate supports for inexperienced designers, but the proposed method provides too much supports for experienced designers.

5 Conclusion

This paper has proposed the development method of adequate UML documents for inexperienced designers. As the result of applying this method to some L2 level examination, we confirm 20[%] improvement of correct answer rate of UML documents in average. Consequently, the proposed method is useful for inexperienced designers to develop the UML documents. Otherwise, we found some botherations for experienced designers, because the proposed method forces development process strongly. We will refine the proposed method and support tools to be adapted to all kinds of designers in future.

Acknowledgments

This research is conducted as a part of "Basic Technology for Next Generation Transportation System Design", which is a delegation from New Energy and Industrial Technology Development Organization (NEDO).

References

[1] Ge, K., Zang, L., Yang, S., Zhou, B.: A Development and Management Environment Based on UML. Technology of Object Oriented Languages and Systems, 214–221 (1998)
[2] Harada, M., Tamura, H., Nomura, Y.: Object Oriented Analysis System, CAMEO/A, and Rule Leaning by Inductive Reasoning. In: Proceedings of OO'96 Symposium, Information Processing Society of Japan, pp. 57–64 (1996) (written in Japanese)
[3] Jaaksi, A., Aalto, J., Aalto, A., Vatto, K.: Tried & True Object Development – Industry-Proven Approaches with UML. Cambridge University Press, Cambridge (1999)
[4] Kruchten, P.: The Rational Unified Process – An introduction, 3rd edn. Addison Wesley, Reading (2003)
[5] Mackinnon, N., Murphy, S.: Designing UML diagrams for technical documentation: continuing the collaborative approach to publishing class diagram. In: Proceedings of the 22nd annual international conference on Design of communication: The engineering of quality documentation, pp. 120–127 (2004)
[6] Ohnishi, A.: A Supporting System for verification among Models of the UML. IEIEC Transaction D-I J84-D-I(6), 671–681 (2001) (written in Japanese)
[7] Rumbaugh, J., Jacobson, I., Booch, G.: The Unified Modeling Language Reference Manual. Addison-Wesley Object Technology Series. Addison-Wesley, Reading (2004)
[8] Tabrizi, M., Collins, C., Ozan, E., Li, K.: Implementation of object-orientation using UML in entry level software development courses. In: Proceedings of the 5th conference on Information technology education, pp. 128–131 (2004)
[9] Tanzer, J.: Improving UML design tools by formal games. In: Proceedings of the 26th International Conference on Software Engineering (2004) (no page number)
[10] Virvou, M., Tourtoglou, K.: An Adaptive Training Environment for UML. In: Proceedings of the 6th International Conference on Advanced Learning technology (2006) (no page number)

Analysis of the Relation Between Stock Price Returns and Headline News Using Text Categorization

Satoru Takahashi[1,2], Masakazu Takahashi[3], Hiroshi Takahashi[4], and Kazuhiko Tsuda[2]

[1] Mitsui Asset Trust and Banking Co., Ltd., 3-23-1 Shiba, Minato-ku, Tokyo, Japan
Satoru_1_Takahashi@mitsuitrust-fg.co.jp
[2] Graduate School of Systems Management, The University of Tsukuba, 3-29-1 Otsuka, Bunkyo-ku, Tokyo, Japan
{satoru, tsuda}@gssm.otsuka.tsukuba.ac.jp
[3] Shimane University, 1060 Nishikawatsu-cho, Matsue-shi, Shimane, Japan
smasakazu@cis.shimane-u.ac.jp
[4] Okayama University, 1-1, Tsushima-Naka, 1-Chome, Okayama, Japan
htaka@e.okayama-u.ac.jp

Abstract. In this paper, we analyze about the relation between stock price returns and Headline News. Headline News is very important sources of information in asset management, and is sent in large quantities every day. We study the effect of more than 13,000 Headline News sent from JIJI PRESS. We classify Headline News using Text Categorization and analyze the reaction of a stock price return for every type of News. From our research, we figure out following issues; 1) we make the Text Categorization System that has about 80% of classification accuracy, 2) this system can extract effective information to stock price returns from Headline News.

Keywords: Headline News, Stock Price Return, Text Mining, Naive Bayes.

1 Introduction

In order to trade in various financial assets in the financial market, it is important to evaluate the financial asset accurately. Therefore, the fund manager - an institutional investor – has been evaluating financial assets using a lot of type of information, such as information on macroeconomics, and micro information about an individual firm. Many researches about valuation of financial assets have been conducted in the academic field and the field of business. However, these researches were the analysis only using numerical information, and there are few researches that used data except numerical information. In particular, fund managers and traders, as institutional investors, use a lot of text information other than number information. Furthermore, it has been reported that text information has different value from numerical information [1][2][5].

In this paper, we analyze the influence which text information has on a stock price. We focus on Headline News as text information, which is transmitted immediately and is very important information for fund manager.

B. Apolloni et al. (Eds.): KES 2007/ WIRN 2007, Part II, LNAI 4693, pp. 1339–1345, 2007.
© Springer-Verlag Berlin Heidelberg 2007

The paper is organized as follows. We explain the data we used in section 2, and analysis method in section 3. We illustrate the result in section 4, and then we conclude our discussion in Section 5.

2 About Headline News

In this paper, we analyze the reaction of the stock price to Headline News. Headline News is information that is transmitted timely by the exclusive computer of Bloomberg or Reuters. Headline News is one of the sources of information transmitted earliest for a fund manager (We express Headline News as News after this). We use the News data offered from JIJI Press. Content of News includes many different fields, such as a business trend, macroeconomic indicators, a trend of a foreign market, information of an individual stock.

As those contents, the following things are included.

Domestic news of mass media, international news such as Reuters, weather news or news that watched on television.
A statement of Bank of Japan or Government.
Market information, such as interest rates or exchange rate.
The analyst comment or rating information of brokerage firm of rating company.
A message or statement of CEO, or estimation of the statement by the third person.

First we extract specific News that has high possibility to affect the stock prices from these News data, such as “scandalous affair of the company”, “mergers and acquisitions”, "evaluation of an analyst" or "the statement of CEO". And then, we analyze how the stock price of the company that was reported in the News changes before and after publishing the News. Figure 1 is the sample of the News we use for analysis.

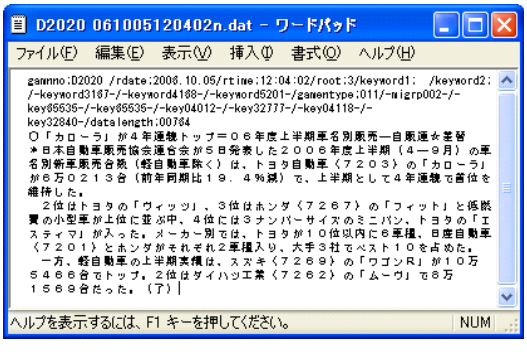


Fig. 1. Example of Headline News

As shown in figure 1, News contains (1) type of News, (2) delivered date, (3) delivered time, (4) system code, (5) length of this News and (6) body of News. Furthermore, if it is the News for a specific company, the code of the target company is describing in the body of News, like the form of <7262>.

We collected the News updated between 2006/8/10 and 2006/11/24, and defined the period as a period of analysis. The total number of News published all over the period by JIJI Press amounts to 346,975 cases and is an enormous number. We choose the company that listed on the first section of Tokyo stock exchange market (TSE1) during analysis period for analysis. The number of companies for analysis is 1,719 (22 companies were delisted and 18 companies were new listed on TSE1). Table 1 shows the basic statistic of the News that we analyze. As shown in table 1, the number of News about the company of TSE1 is 13,402, and accounts for about 4% of all type of News. The number of averages of the News about TSE1 published on one day is 144.

Table 1. Basic statistics about News

	All e e	e m	③ e m ( )		③
l m er e	46		4	66 %	6 %
A er ge m e e (er e)	6		44	%	4 %
x m e e (er e)		4	4		
m m e e (er e)					

We analyze about the effect and influence of News by measuring stock price change before and after publishing the News.

3 Analysis Method

In this paper, although the purpose is to analyze the relation between information of News and a stock price, we assume that the influence which News gives the stock price changes with contents of News. Therefore, at first we classify News into three types, and then we analyze the effect of News for each type.

3.1 Naïve Bayesian Classifier Algorithm

We use Naïve Bayesian classifier method to interpret News, which is frequently used for text mining [1][4][6]. It is a well-known and well-established algorithm such as Decision tree, K-Nearest Neighbors or Neural network. Naïve Bayes can be understood most intuitively and has high classification performance as well as other methods.

Let n be the number of keywords that characterize a text data. By using these keywords, a text data is described by the n-tuple of attribute value $(x_1, x_2, \cdots, x_n)$. x_i takes the value of 0 or 1 that expresses whether the i-th keyword exists in text data. For finding the class that the text data belongs to, we try to assign the most probable target value of the following equation.

To expand equation, we consider Bayes's rule and the fact that $P(\mathbf{x})$ takes the same value in all classes. Furthermore, we assume that occurrences of key words are

independent of each other under the condition that class c has been given. By using a set of labeled training data, $P(c)$ can be defined from the relative frequency of each class, and $P(x_i \mid c)$ can be defined from the relative frequency of keyword in each class. And then, we verify classification accuracy of the algorithm that we made using test data.

$$\begin{aligned}\hat{c} &= \arg\left\{\max_c (P(c \mid (x_1, x_2, \cdots, x_n))\right\} \\ &= \arg\left\{\max_c \left(\frac{P((x_1, x_2, \cdots, x_n) \mid c)P(c)}{P((x_1, x_2, \cdots, x_n))}\right)\right\} \\ &= \arg\left\{\max_c (P(c)P((x_1, x_2, \cdots, x_n) \mid c))\right\} \\ &= \arg\left\{\max_c (P(c)\prod_{i=1}^{n} P(x_i \mid c))\right\}\end{aligned}$$

In this paper, we use three types of class, “Good News”, “Bad News” and “Neutral News”. We chose each 400 News data at random as training data and test data.

3.2 Stock Price Return Analysis

We measure stock price change before and after publishing the News. We already showed in the section 2 that delivered time is recorded in the News. Therefore, if the delivered date is a business day and delivered time is during the morning, a stock price return is measured on the basis of the closing price of delivered day. And, if the delivered date is a holiday or delivered time is the afternoon of a business day, a stock price return is measured on the basis of the closing price of following business day. We use a daily stock price return. In order to eliminate the effect of the market, we analyze the excess stock price return from the arithmetic average of the stock price return of the company listed on TSE1, we express this return as excess return after this.

We total an excess return for each type of News (Good News, Bad News or Neutral News), and we perform statistical test to examine whether a significant difference between the excess returns among the News of each type exists or not. We employ Welch test to measure the difference.

4 Empirical Analysis

We explain an analysis result in this section. First, we show the result of the text classification by Naïve Bayes, and then we show the result about the effect of News.

4.1 Result of Naïve Bayes Classification Algorithm

First, we performed morphological analysis and pattern matching to extract keywords. And then, we gave a category of class, such as Good News, Bad News or Neutral News, from the contents of News in order to give a teacher label. We show the classification result in table 2. We find that Good News accounts for about 49% of all type of News.

Table 2. News Classifications of Training Data

w			
w	97	97	
	492	242	2

Although the number of keywords first extracted from learning data was 1,360 pieces, we selected 205 keywords by the ratio to appear, in section 3.3 and the classification result of training data [1][6]. We show in-sample classification result of training data in Table 3.

Table 3. In-sample Classification Result by Training Data

			B	
y			2	2
f	B	2		
Naïve Bayes				3

			B	
y		8 %	2 %	22 %
f	B	2%	8 %	8 %
Naïve Bayes		82%	22%	88 %

Next, we verified the accuracy of the text classification algorithm by Naïve Bayes using test data. We also gave a teacher label of Good News, Bad News and Neutral News to test data. We show the classification result in table 4. The percentage of Good News is 46% and is the highest of three type of News in test data.

Table 4. News Classifications of Test Data

yp f N	G		N
N f N			

We classified test data into three classes using Naïve Bayes algorithm that we made from learning data. Table 5. shows the classification result. The algorithm which we made has about 78% of classification accuracy and performs reasonably well.

Table 5. Out-of-sample Classification Result by Training Data

		e e e		
			B	e
e y e		1		5
e	B		5	5
Naïve Bayes	e	1	5	5

		e e e		
			B	e
e y e		1		
e	B	1		
Naïve Bayes	e	5	55	

We classified whole data (13,402 News) using this algorithm. We show the classification result in table 6.

Table 6. News Classifications of Whole Data

yp f N	G		N
N f N			
	7		

Table 6 shows that Good News accounts for about 59% of all type of News, and the percentage of good News in whole data is slightly higher than the percentage of good News in learning data or test data. Table 5 shows that although the algorithm that we made has about 80% classification precision for whole data, it tends to classify Bad News as Good News a little. Therefore, there is bias that increases the number of good News in the classification result of whole data. To make algorithm that has higher classification accuracy is a future aim.

4.2 Analysis of the Effect of the News by Welch Test

We measured excess return before and after publishing each type of News. Then we calculated the average value of an excess return of each type of News, and we performed Welch Test to examine whether a significant difference between the average excess returns among the News of each type exists or not. We found that the level of an excess return is the order of Good News, Neutral News and Bad News, and there were significant differences in excess return of each type of News. From here, we explain the result of Welch Test between Good News and Bad News, since the difference of return of these classes is large. Table 7 shows the result of Welch test between excess return of Good News and that of Bad News

Table 7. Result of Welch Test

		av	av	av	av	av	av
A	Goo N w		5			5	
	a N w			5			
W	t				6		
t	t ft t	GOO	GOO	GOO	GOO	GOO	GOO

As shown in Table 7, the excess return of the company which Good News is published is higher than that of the company which Bad News is published. This result shows that it is possible to obtain an excess return by using News. Furthermore, Table 7 shows that the effect is the same before and after publishing News. As we shown in section 2.1, since the average number of News published per day is 144, the News of the same contents is often reported continuously and the influence of News is overlapped. In the analysis for Analyst's report by using text mining, we found same result [3].

Next, in order to investigate change of an excess return in detail, we calculated daily cumulative excess return in each type of News. We show the chart in Figure 2.

Figure 2 shows that the level of cumulative excess return is the order of Good News, Neutral News and Bad News, it's the same as Table 7. Moreover, a change of excess return in -1 business day in each type of News is extremely large. In this paper, the basic time when we measure an excess return is defined as follows.

> If the delivered time is before noon of a business day, a stock price return is measured on the basis of delivered day.
>
> If the delivered time is after noon of a business day, or if the delivered date is a holiday, a stock price return is measured on the basis of the following business day.

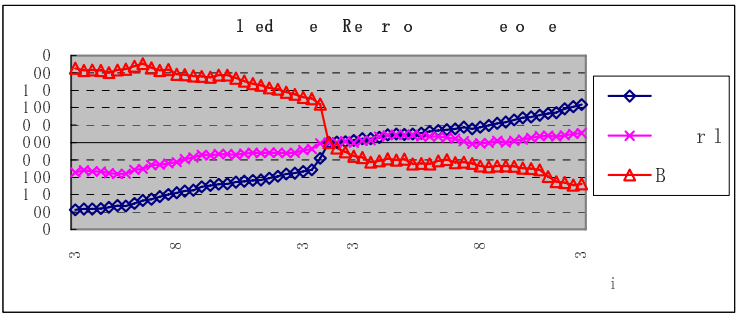


Fig. 2. Accumulated Excess Return of Each Type of News

From a definition, when we get News within the trading hours of a business day, we calculate the excess return in the business day as excess return in -1 business day. Therefore, Figure 2. shows that the reaction to the News is the largest in the business day to which the News delivered. We have to analyze the relationship between News and intraday stock price return in order to understand the effect of News in detail.

5 Conclusion

In this paper, we analyze the reaction of the stock price to Headline News. We classify Headline News using Naïve Bayes Classification Algorithm and analyze the reaction of a stock price return for each type of News. From our research, we figure out following issues; 1) we make the Text Categorization System that has about 80% of classification accuracy, 2) this system can extract effective information to stock price returns from Headline News. These results are useful for solving the problem about how information is reflected in the asset price.

References

1. Antweiler, W., Frank, M.Z.: Is All That Talk Just Noise? The Information Content of Internet Stock Message Boards. The Journal of Finance 59, 1259–1294
2. Dewally, M.: Internet Investment Advice: Investing with a Rock of Salt. Financial Analysts Journal 59, 65–77 (2003)
3. Takahashi, S., Takahashi, H., Tsuda, K., Terano, T.: Analyzing Asset Management Knowledge from Analyst's Reports through Text Mining. In: International IPSI-2004 (September 2004)
4. Tumarkin, R., Robert, F.W.: News or noise? Internet message board activity and stock prices. Financial Analyst Journal 57, 41–51 (2001)
5. Wuthrich, B., Cho, V., Leung, S., Permunetilleke, D., Sankaran, K., Zhang, J., Lam, W.: Daily Prediction of Major Stock Indices from Textual www Data. In: KDDM'98 Conference, NY, pp. 364–368. AAAI Press, USA (1998)
6. Lewis, D.D.: Naive (Bayes) at Forty: The Independence Assumption in Information Retrieval. In: Nédellec, C., Rouveirol, C. (eds.) Machine Learning: ECML-98. LNCS, vol. 1398, Springer, Heidelberg (1998)

Evaluation of a Hierarchical Shaper as a Policy Execution Point

Takeshi Aimoto[1], Takeki Yazaki[1], Takashi Isobe[2], Yoshihiko Sakata[1], and Kenichi Yoshida[3]

[1] Alaxala Networks, Corporation, 890 Kashimada, Kawasaki 212-0054, Japan
[2] Central Resaerch Lab., Hitachi Ltd., Tokyo 185-8601, Japan
[3] Graguate School of Business Sciences, University of Tsukuba, Tokyo 112-0012, Japan
{t-aimoto,yazaki,yoshihiko.sakata}@alaxala.com,
tisobe@crl.hitachi.co.jp, yoshida@gssm.otsuka.tsukuba.ac.jp

Abstract. The Next Generation Network (NGN) is expected to become the most important infrastructure for various services used for e-Business such as telephone services, Internet services and new generation broadcasting services. It is supposed to be integrated into all IP network. To satisfy various Quality of Service (QoS) requirements of these services not only for each service but also for every user, we have developed a hierarchical shaper. The shaper has been implemented in a QoS aware gigabit router environment. Although this shaper has to be evaluated to manage actual traffic, the influence of various network conditions on bandwidth control has not been studied thoroughly. This paper reports some experimental results and clarifies the characteristics of the shaper.

Keywords: QoS, Hierarchical shaper, Policy execution point.

1 Introduction

The Next Generation Network (NGN) is expected to become the most important infrastructure for various services used in e-Business such as telephone services, Internet services, and new generation broadcasting. It is supposed to be integrated into the all IP network over which End-to-End QoS functions will be deployed to provide stable and high quality communication services.

To satisfy QoS requirements of various network services, knowledge-based systems that consist of a policy-based network management system and a QoS aware gigabit router have been developed. Policy decision point (PDP) is a rule-based scheme within a policy-based network management system. It translates various business requirements into router commands after which a QoS aware gigabit router manages the actual QoS requirements as a policy execution point (PEP).

The requirements of a QoS aware gigabit router are summarized as follows:

1) It has to manage different QoS classes into which each type of traffic is to be categorized. A priority class is created to provide better QoS for voice communication and video streaming data traffic. These priority classes require low packet loss and low latency. A non-priority class is applied to control

B. Apolloni et al. (Eds.): KES 2007/ WIRN 2007, Part II, LNAI 4693, pp. 1346–1353, 2007.
© Springer-Verlag Berlin Heidelberg 2007

E-mail and Web browsing traffic. These types of non-priority class traffic permit some deterioration of communication quality.

2) It has to provide scalable End-to-End QoS-enabled service while assuring a fair and stable connection to every user. Bandwidth management based on the contract between the network provider and their users is also important.

The influence of various network conditions including heavy-user-traffic and attack-traffic on bandwidth control for each user and class has yet to be evaluated and will require thorough study in the future.

This paper evaluates the characteristics of our hierarchical shaper which is implemented in a QoS aware gigabit router, and shows that the proposed techniques can be applied to an actual network [1]. The organization of this paper is as follows: First Section 2 briefly reviews related studies. Section 3 describes the architecture of our hierarchical shaper. Section 4 reports on the experimental results of its performance, and Section 5 summaries our findings.

2 Related Studies

A shaper is used to control communication quality at a congestion point. Some studies undertaken discuss the characteristics of conventional shaping functions, such as Priority Queuing (PQ) [2] and Weighted Fair Queuing (WFQ) [3]. PQ is designed to select one class based on order of priority, and WFQ is a flow-based queuing algorithm that schedules low-volume traffic first, while letting high-volume traffic share the remaining bandwidth. Unfortunately, both PQ and WFQ cannot assure the bandwidth of each user's data transfer while maintaining the priority control of bandwidth for the user. For instance, WFQ assures the bandwidth of each class's data transfer by controlling the transfer with queues for each class. But it has difficulties in allocating bandwidth between users within a specific class. In WFQ, all transmitted data are accumulated in a single queue of the class. Thus, per user bandwidth control is not possible.

Another problem occurs with the interference between TCP congestion control and shaper function. Packet loss due to the shaper function is known to cause a decrease in TCP session bandwidth. The Random Early Detection (RED) algorithm has been developed to recover decreased bandwidth [4]. Hasegawa et al. [5] reports simulation results which show the effectiveness of the RED algorithm by analyzing the relationship between packet loss and the decrease in TCP session bandwidth. However, the effectiveness of a complex hierarchical shaper in an actual network has not been studied. The complexity of a hierarchical shaper makes the simulation studies difficult. This complexity becomes an obstacle in the modeling of the phenomena which is essential in the simulation studies. Thus in this paper, we conduct experiments which use actual network equipment.

3 IPv6 Hierarchical Shaper

To protect user's bandwidth while assuring QoS specifications, we have developed a hierarchical shaper [1]. This shaper secures the bandwidth of each user's IPv6 and

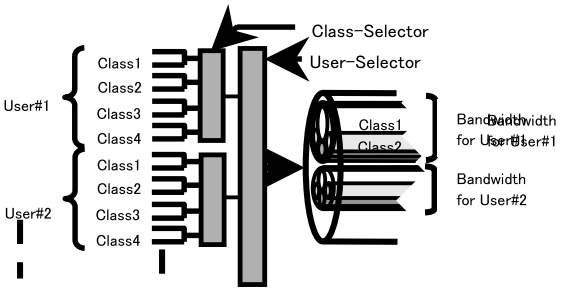


Fig. 1. Schematic diagram of the hierarchical shaper

IPv4 data transfers. It divides each user's bandwidth into its classes according to the importance of each class. Fig.1 shows its schematic diagram. It is supposed to control 1000 users' bandwidth while dividing allocated users' bandwidth into 4 classes. Class-selectors control priority of each user's traffic by using 4 queues for each user. The user-selector selects the class-selector based on the allocated bandwidth of each user, Packets in the top priority queue of the selected class-selector will be sent out. Thus a total of 4000 queues are used to control overall traffic.

In real business environments, network users can buy bandwidth from network providers. The bandwidth of the user-selector will be set up based on the contract between the network user and network provider. In addition to this, the network users can specify the priority of their traffic. For example, they can specify high priority class for VoIP traffic, and low priority class for E-mail and Web traffic. The network provider can set up a class-selector through PDP to reflect such priorities.

The function of each selector is realized using Field Programmable Gate Array (FPGA) circuits to achieve gigabit class performance. Within the forwarding time of a single packet, FPGA techniques enable the user-selector to select a class-selector for the next user based on the guaranteed bandwidth of each user. Next, the selected class-selector selects the class, i.e. the queue for sending packets, based on priority. The class-selector also uses either the PQ or WFQ scheduling mechanism. The network provider can choose the mechanism of the class-selector based on the contract with the user. In this paper, the results using WFQ are reported.

4 Experimental Results

4.1 System Configuration

To evaluate the actual performance of the hierarchical shaper and the interference between the shaper mechanism and TCP congestion control mechanism, we have conducted some experiments. This section reports on results of those experiments.

Fig.2 shows the block diagram of the system configuration used in the experiment. The system is composed of two switches (switch-1 and switch-2), two analyzers (analyzer-1 and analyzer-2), and two personal computers (PC1 and PC2). A hierarchical shaper is implemented in switch-1. In this configuration, the hierarchical

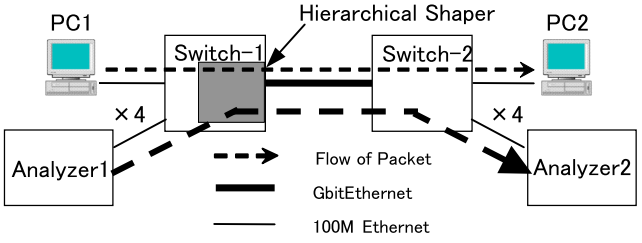


Fig. 2. Block diagram of the system configuration used in the experiment

shaper is arranged at the interface of switch-1 into the Gigabit Ether circuit, which is assumed to be the exit (bottleneck point) of NGN, while many users (up to 1000 users) are assumed to connect at switch-2 through a 100Mbps Ether circuit. Additionally, the Round Trip Time (RTT) of priority class is assumed to be relatively short, i.e., 1ms at the longest. All evaluation is done using IPv6 packets.

The hierarchical shaper is designed to protect the bandwidth assigned to each user. And user traffic is classified into two classes: one priority class for QoS guaranteed data and the other for non-priority-class data. The bandwidth allocated to each user is divided between the 2 classes according to a WFQ policy weighting of 2:1 for each class. If the allocated bandwidth for the user is 100Mbps, the allocated bandwidths for each class are 66.7Mbps and 33.3Mbps, respectively. If inbound traffic of the priority class for QoS guaranteed data is less than the guaranteed bandwidth of 66.7Mbps, all the traffic is supposed to be secured without influencing other traffic.

Two experiments are conducted as follows.

1) In the first experiment, the performance of the user-selector and the class-selector is examined. Two users (user#1 and user#2) are assumed to use UDP applications. As for the traffic pattern of priority-class, constant traffic (for example, high-quality streaming video) is assumed, while both constant traffic and burst traffic are assumed for the traffic pattern of non-priority-class. Two analyzers are used to control and measure the packets of user#1 and user#2, and up to 1000 users. Both users send priority and non-priority class traffic.
2) In the second experiment, the interference between the TCP congestion control mechanism and the shaper mechanism is analyzed by observing the forwarding performance of a critical file between two PCs. Precisely speaking, PC1 forwards a 100MByte file to PC2 as an FTP priority class of user#1. Two analyzers are used to control and measure the packets of the other users contained within the background traffic.

4.2 Performance of the User-Selector and the Class-Selector

4.2.1 Interactions between Two Classes of a Single User

In this experiment, we analyzed the characteristics of two classes of traffic from a single user. The influence of non-priority class traffic to the bandwidth of priority class traffic is examined to clarify the performance of the class-selector. The average bandwidth of non-priority class from user#1 is varied while the bandwidth of priority class is fixed. Both constant and burst traffic are examined for non-priority traffic.

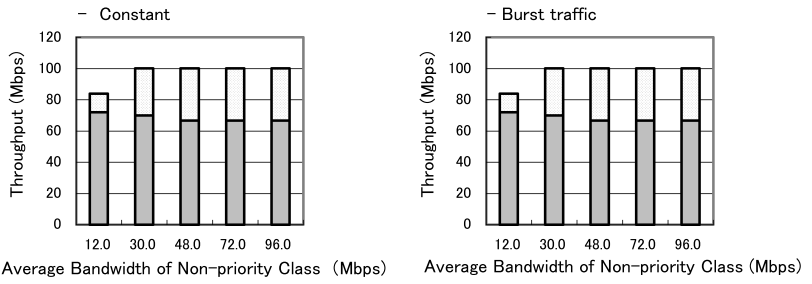


Fig. 3. Throughput related to average bandwidth of non-priority-class. Grayed bars:User#1 priority-class; Dotted bars:User#2 non-priority-class.

Burst traffic means that an intermittent transfer is repeated periodically, but bandwidth on average within a period is the same as the average bandwidth. Accordingly, average bandwidth is adjusted by changing the number of packets of continuous transfer, while the duration of stopping transfer is fixed at 1 msec.

Fig.3 shows the change of throughput related to the average bandwidth of non-priority class when inbound bandwidth of priority class is set to 72Mbps (over guaranteed bandwidth) and the traffic pattern of non-priority class is constant and burst. As is designed, more than 66.7 Mbps of priority class traffic is secured in all cases. However, over guaranteed bandwidth (i.e., 72 – 66.7Mbps) for the priority class traffic is not secured, and remaining bandwidth is used for the non-priority class traffic.

When inbound bandwidth of the priority class is 48Mbps (below guaranteed bandwidth), all priority class traffic is secured without the influence of average bandwidth of non-priority-class. And furthermore, no priority class packet is abandoned. Thus, inbound bandwidth for the priority class traffic is secured completely.

4.2.2 Interactions Between Two Users

In this experiment, the influence of user#2 traffic on user#1 bandwidth is examined. This experiment is designed to analyze the characteristics of the class-selector and the user-selector combination. The average bandwidth of user#2 is varied while the bandwidth of priority class and non-priority class of user#1 is fixed. Both constant and burst traffic were used.

Fig.4 shows the change of throughput related to the average bandwidth of user#2. The inbound bandwidth of priority class is set to 72Mbps. Both the constant and burst traffic pattern of the non-priority class is examined. We can observe that bandwidth of user#1 is always secured in any case.

Further experiments evaluate the effect of many users (up to 1000 users). Here allocated bandwidth of user#1 and user#101 #1000 is fixed at 1Mbps. Allocated bandwidth of user#2 #100 is varied. As shown in Fig.5, the bandwidth of user#1 is always secured for priority class and non-priority class regardless of the influence of the number of users. Precisely speaking, 99.7% of the total bandwidth for user#1 is always secured, and 99.7% of priority class traffic of user#1 is also always secured.

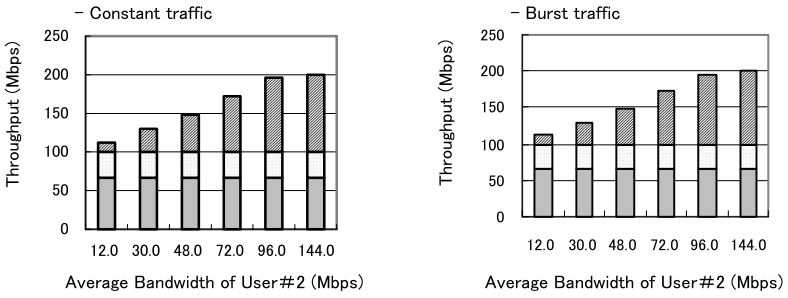


Fig. 4. Throughput related to average bandwidth of user#2. Grayed bars:User#1 priority-class; Dotted bars:User#1 non-priority-class; Slant-lined bars:User#2.

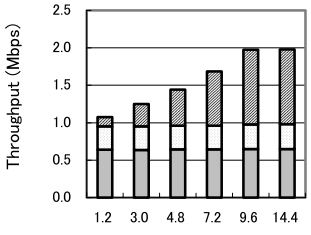


Fig. 5. Throughput related to average bandwidth of user#2 under the influence of many users. Grayed bars: User#1 priority-class; Dotted bars: User#1 non-priority-class; Slant-lined bars: User#2.

4.3 Performance of Forwarding a Critical File as Priority Class Traffic

In this experiment, the interference between the TCP congestion control mechanism and the shaper mechanism is analyzed by observing the performance of forwarding a critical file between two PCs. Large files are forwarded as priority class traffic by using TCP protocol between two PCs. The File size is 100 MByte. The total guaranteed traffic of user#1 is set to 10Mbps. The guaranteed bandwidth of priority class and non-priority class are set to 6.67Mbps and 3.33Mbps, respectively. The actual bandwidth transferred is estimated by using FTP performance data collected at each PC. TCP Reno [6], implemented in Windows XP, is used with the SACK option setting ON [7].

Primarily, the influence of non-priority class traffic is examined. As shown in Fig.6 (left), guaranteed bandwidth (10Mbps) and that of priority class (6.67Mbps) are always secured. Secondarily, the influence of user#2 traffic on user#1 throughput is examined by changing user#2's average bandwidth while user#1's non-priority class bandwidth is constant. Also, as shown in Fig.6 (right), user#1's throughput and that of priority class are always secured without influencing user#2's traffic.

In this experiment, we also measured the TCP packet loss ratio, which was about 0.05%. This packet loss was caused by the queue overflow at the shaper. TCP sent out "congestion window size/1500" packets without waiting for ACK packets from the

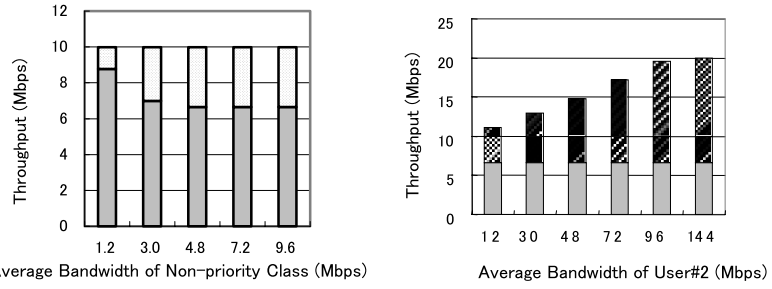


Fig. 6. Throughput related to the average bandwidth of user#1 non-priority class (left) and the average bandwidth of user#2 (right) when priority class is transferred through TCP. Grayed bars: User#1 priority-class; Dotted bars: User#1 non-priority-class; Slant-lined bars: User#2.

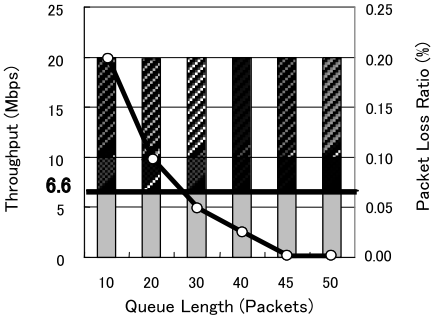


Fig. 7. Throughput and packet loss ratio related to queue length. Grayed bars: User#1 priority-class; Dotted bars: User#2 non-priority-class; Slant-lined bars: User#2; Open circles indicate packet loss ratio.

receiver. If all the packets under transfer are accumulated in the queue of the hierarchical shaper, the packets overflowing at the queue cause packet losses. This phenomenon is known to cause application level performance degradation.

To measure the effect of queue length on the packet loss ratio and application level performance, we have conducted several different experiments. In the experiments, the queue length was decreased until packet loss occurred, and then the packets loss ratio and the application level throughput were measured. Fig.7 shows this loss ratio and throughput in relation to the length of the queue. No packet loss is observed in a length of queue of over 45 packets. 45 packets can be deduced from the maximum value of congestion window size, i.e. 64kByte. When the congestion window size is set to 64kBytes, the maximum number of packets between PCs will be 45 under TCP congestion control. Thus, packet loss never occurs with a queue of length 45 packets.

Fig.7 shows user#1's throughput of priority class. Even if the queue length is set to a small number (i.e., less than 45 packets), the priority class traffic is always secured. However, as queue length decreases, packet loss increases slightly. So it is reasonable to assume that a slow start state has not occurred even with the shorter queue length. Accordingly, the packet loss ratio is small enough to avoid TCP slow starting. This supposition is consistent with the simulation result reported in [5].

Upon reviewing the results above it can be determined that large size files forwarded as priority class traffic can achieve, 98.03% or more throughput of the guaranteed bandwidth configured in a hierarchical shaper when it is transmitted without the influence of the traffic of non-priority class or other users. Although packet loss occurs depending on queue length, it does not affect packet throughput.

5 Conclusion

We have evaluated the performance of an IPv6 hierarchical shaper using actual network equipment. This shaper has both a user-selector and a class-selector to secure the bandwidth of both per user and per class traffic simultaneously. Experimental results under various network conditions show the following:

1. The combination of user-selector and class-selector from our hierarchical shaper can guarantee user bandwidth while protecting priority class transfer.
2. The actual interference between the TCP congestion control mechanism and a complex shaper's queuing function is consistent with that which is predicted by previous studies and which uses the simple simulation model [5]. In the actual situation, a 0.05% packet loss ratio does not cause any performance degradation. Additionally, a queue length of 30 packets is long enough to prevent packet loss of greater than 0.05%. Since a queue length of 30 packets is not difficult to be implemented and our hierarchical shaper actually has a longer queue length, we can use our shaper to protect TCP traffic bandwidth.

It can be concluded, that these experimental results support the fact that our hierarchical shaper has sufficient performance of bandwidth control in the next generation network.

Acknowledgments. This research was supported in part by contract research titled the "Development study of high reliable communications infrastructure based on IPv6 gigabit router" from the Telecommunications Advancement Organization of Japan.

References

1. Yazaki, T., Kanetake, T., Akahane, S., Sakata, Y., Sugai, K., Yano, H.: High-Speed IPv6 Router/Switch Architecture. In: Proceeding of 2004 Symposium on Applications and the Internet-Workshops (2004)
2. http://www.nwfusion.com/links/Encyclopedia/P/524.html
3. Golestani, S.: A Self-Clocked Fair Queueing Scheme for Broadband Applications. In: Proceedings of INFOCOM '94, pp. 636–646 (1994)
4. Floyd, S., Jacobson, V.: Random Early Detection gateways for Congestion Avoidance. IEEE/ACM Trans. on Networking 1, 397–413 (1993)
5. Hasegawa, G., Matsuo, T., Murata, M., Miyahara, H.: Comparisons of Packet Scheduling Algorithms for Fair Service among Connections on the Internet. In: IEEE INFOCOM, IEEE Computer Society Press, Los Alamitos (2000)
6. Steavens, W.R.: TCP/IP Illustrated, vol. I. Addison-Wesley Publishing Company, Reading, Massachusetts (1994)
7. Mathis, et al.: TCP Selective Acknowledgment Options. In: IETF, RFC 2018 (1996)

Query Message Delivery over Community-Based Overlay Network

Yoshikatsu Fujita[1,2], Yasufumi Saruwatari[2], Jun Yoshida[1], and Kazuhiko Tsuda[2]

[1] Matsushita Electric Industrial Co., Ltd., Corporate *e*Net Business Division
2-13-10, Kyobashi, Chuo-ku, Tokyo 104-0031, Japan
{fujita.yoshikatsu, yoshida.jun}@jp.panasonic.com
http://panasonic.co.jp/enet/
[2] Graduate School of Business Scineces, University of Tsukuba, Tokyo
3-29-1, Otsuka, Bunkyo-ku, Tokyo 112-0012, Japan
{saru, tsuda}@gssm.otsuka.tsukuba.ac.jp
http://www.gssm.otsuka.tsukuba.ac.jp

Abstract. We propose an efficient query message delivery mechanism over community-based overlay network. By applying the nature of power law distribution, our simulation shows this delivery mechanism will cover more than 80% of all the nodes, with relaying reverse-query message under probability as low as 10%. This platform is built on percolation theory to propagate the message and contents that each node in the community just relays reverse-query message to randomly selected their peers. This can be applied for electric flyer delivery targeting community member's common interests.

Keywords: Overlay Network, Percolation Theory, CDN, Power-law distribution.

1 Introduction

In Japan, with rapidly expanding demand for broadband contents over the Internet, five leading TV set manufacturers has established new company called "TV Portal Service (TVPS)" in July 2006, aiming for presenting TV-friendly contents which is easier to view, operate, pay and other transactions only on TV display. This clearly shows that not only PCs but TV sets become more and more important devices as a broadband contents receiver over the Internet. Sooner ore later, millions of TV sets will access to the Internet searching for their favorite broadband contents, which causes unprecedented huge amount of traffic to the backbone network.

However, when it comes to the matter of how we should deliver broadband contents over the Internet, the total throughput will be determined by any bottleneck somewhere between contents provider to consumers. For example, even a user purchases 100Mbps optical fiber service, one's requested contents comes from distant server only 100Kbps because of narrow path somewhere over the Internet. Most of proxy servers distributed over the Internet are aimed for static contents like homepage objects, and are not tuned for broadband stream contents.

B. Apolloni et al. (Eds.): KES 2007/ WIRN 2007, Part II, LNAI 4693, pp. 1354–1361, 2007.
© Springer-Verlag Berlin Heidelberg 2007

For instance, once many users try to pull large video streams at the same time, it is clear that backbone network is easily falls into overflow. This requires a new contents delivery technology which supports a huge simultaneous access transaction for broadband objects.

This study aimed for overcoming this problem by building our proposed community-based overlay network over the Internet, and to manage generated traffic within that community, to be new message propagation method to cover the whole network. This technique is based on the fact that any link status on the Internet follows power law distribution [1]. For example, one of the most popular pure P2P network Gnutella has been analyzed that its nodes' outgoing degree can be expressed as $P(k) \sim \mathrm{k}^{-\tau}(\tau \geq 0)$ [2]. However, such an unstructured network is not manageable in nature, and makes it difficult to apply for contents delivery for its fundamental network architecture. In this paper, we employ percolation theory [3] that is mainly studied in physics, to model the "percolating information" for query message delivery over pure P2P network. In other words, the query message released by client seeking for requested contents is not merely used for this purpose, but we define a new "reverse-query" message to find any client who needs a certain contents and try to apply the percolation theory to manage the generated traffic. This will lead to reduce the explosive P2P query traffic while maintaining fairly high clients cover rate over our proposed overlay network. An outlook of our model is shown in Fig.1.

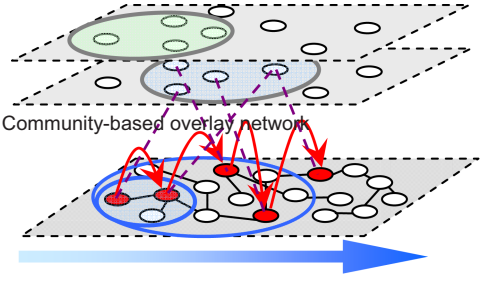


Fig. 1. Query Message Propagation on the Overlay Network

2 Related Work

2.1 Contents Delivery Network

In order to deliver broadband contents over the Internet, CDN(Contents Delivery Network) architecture [4], [5] and contents distribution algorithm for replication [6], [7] are actively studied. But such CDN solutions for large scale contents delivery faces difficulty because the number of acceptable simultaneous access is almost determined by hardware specification of cache servers, and this falls into optimal cache server distribution problem with considering dynamic request load balance under the exact forecast of contents popularity and hardware availability. This is also regarded as a big issue for realizing broadcast type traffic over the Internet. Existing technique to handle telephone call over the telephone network is specific for point to point

traffic, and it is not applicable for clearing simultaneous access to a contents sever, that makes it difficult to deliver broadband contents to many clients.

Recently, file sharing application over P2P network has been pervasive and it plays an important role as contents distribution infrastructure. When we apply P2P network for contents delivery, under the pure P2P architecture which has no center server, it is major concern how to find available resources and required contents from all over the network. In this field of study, such projects as CAN[8] Chord [9] Pastry [10] Tapestry [11] are trying to employ Distributed Hash Table (DHT)[12].

However, in order to apply for commercial services, there exist more problems in this DHT solution.

(1) When peers join/leave the network, they have to takeover management table to some other peers, results in heavy overhead.
(2) Network structure becomes complicated.

In addition, traffic generated by those P2P nodes has been increasing more and more, whose amount of load gives serious impact to today's ISP backbone network.

2.2 Power-Law Nature

In this study, we propose a new network architecture for query message delivery which can solve problems shown above and applicable for scattering electric flyer over the internet.

To estimate the amount of total traffic for retrieval query sent by each client, there is a report from [13]. And for modeling method of network with a power law link distribution, there is an algorithm by [1] with preferential attachment model. However, this method focuses mainly on the growth of network itself and analyzing dynamic traffic behavior has been still left for further study. Also, validation of those theoretical network structure and real P2P network on the Internet is not fully analyzed yet.

3 Reverse Query Mechanism

We have already introduced a fundamental idea to propagate a query message over the unstructured peer-to-peer network called “Reverse Query Mechanism” [14]

In that paper, we have defined new “do-you-need” query and shown that contents can be delivered just relaying this query from server to peer to peer. Details are explained in [14].

3.2 Analysis

3.1.1 Percolation on Generalized Random Graph

The percolation behavior on generalized random graph can be led as follows [15].

Suppose the degree distribution of each node to be p(k), then the general function of this distribution can be defined as:

$$G_0(x) = \sum_{k=0}^{\infty} p(k)x^k \tag{1}$$

Suppose the vertex distribution of connected component on generalized random graph holds general function $H_0(x)$, and general function for the size of connected component from a certain branch to be $H_1(x)$. Then the average size of connected component $<\!|\, c_0 \,|\!>$ to be

$$<\!|\, C_0 \,|\!> = H_0^{'}(1) = 1 + G_0^{'}(1) H_1^{'}(1)$$

$$= 1 + \frac{G_0^{'}(1)}{1 - G_1^{'}(1)} \tag{2}$$

The state transition will take place when right-hand side becomes 0,

$$G_1^{'}(1) = 1 \Leftrightarrow \sum_k k(k-2)\, p(k) = 0$$

$$\Leftrightarrow \frac{< k^2 >}{< k >} = 2 \tag{3}$$

Here, the percolation threshold can be

$$q_c = \frac{1}{\frac{< k^2 >}{k} - 1} \tag{4}$$

If we assume our delivery platform to be generalized random graph, we can say that almost all node can receive the transferred messages when each node relays received messages more than q_c.

3.1.2 Percolation on Power Law Overlay Network

The percolation behavior on the power law overlay network can be led as follows.

(1) From Origin to the First Neighbor

We will consider the number of vertex that is 1 hop away from the origin.

Let us employ the expression k_m^n, which shows the degree of vertex to be *m* hops away from origin (lower right) and *n*th vertex out of the set of *m* vertexes (upper right). Fig.2 shows an example.

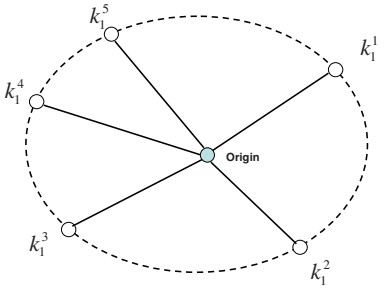


Fig. 2. First neighbors

Then, remember the assumption that every node on the network has degree according to power law. This assumption is observed from real P2P network [2].

When we assume the origin holds connection with k_0 vertexes, then the number of vertex that can be reachable within 1 hop from origin is

$$k_1^1 - 1 \tag{5}$$

As there are k_0 vertexes as a whole, then the total number of reachable vertexes is

$$\sum_{i=1}^{k_0} (k_1^i - 1) \tag{6}$$

(2) From Origin to the Second Neighbor

Then, we will consider the number of vertex that is 2 hops away from the origin.

Let the degree of the first hop vertex to be *l*,

$$\frac{k}{n-1} \bullet \frac{l}{n-2} \tag{7}$$

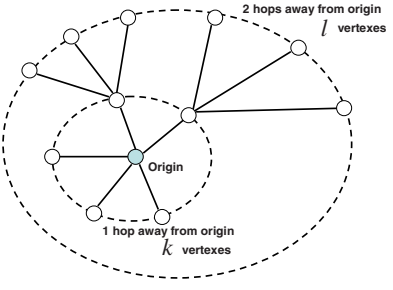


Fig. 3. Second neighbors

The expected value for the number of vertex, that k_i is 2 hops away from the origin is

$$E(2) = \sum_k \sum_l \frac{k(l-1)}{(n-1)^2} P_k P_l \tag{8}$$

(3) Distant Neighbor

The expected value of vertex V_m that *M* hops away from the origin, by letting $P_k = P_l = P_m$ for large *m*, can be expressed as

$$E(M) = \sum_k \sum_l \frac{k^{M-1}}{(m-1)^M} (l-1) P_k^{M-1} \tag{9}$$

When we want to deliver a certain message to more than 80 % of nodes on the overlay network by applying (4), it is enough to deliver the message *m* hops away by setting α as:

$$E(M) \geq 0.8 \tag{10}$$

It is shown by Sarshar[16] that a message delivered on power law overlay network can be reached any nodes within a certain number of iterating relay operation. In the next chapter, we have investigated the dynamics of α upon mathematical simulation.

4 Evaluation

In order to evaluate our proposed model, we prepared a random network with a power law link distribution generated by Pajek [17], and implemented our algorithm on R environment. This overlay network is generated based on generalized BA model, which presumes that every vertex has at least some baseline probability of gaining an edge, to generate edges by mixture of preferential attachment and uniform attachment [18]. For generating condition, we set the total node number $N = 1000$, $M_0 = 3$, $TTL = 5 \quad 25$ and average degree = 2.7. In order to evaluate the results, we compared cover rate (= how much of nodes receives the reverse-query) upon relay probability implanted in the reverse query message.

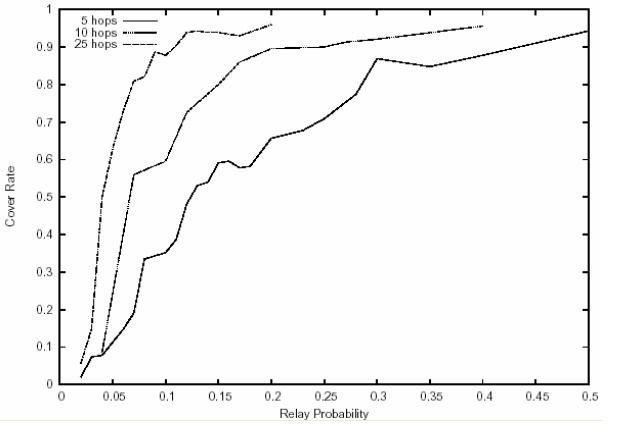


Fig. 4. Simulation Results

Fig.4 shows a typical example of message propagation over this generated overlay network. The propagating message will cover almost 80% of nodes upon relay probability 0.08 when TTL=25. After several hops of relaying messages, the copy of message body will just go out from the P2P network by using up the TTL, and this will lead to reduce the explosive P2P query traffic while maintaining fairly high clients cover rate over our proposed mechanism.

5 Conclusion

In this paper, we have applied percolation theory to model the "message delivery" over community-based overlay network. We defined a new "reverse-query" message to find any client who needs certain contents, and try to apply the percolation theory to prevail the message over the network. We analyzed validity of the model through dynamics of mathematical simulation. This concludes that our proposal is effective for message delivery to cover nodes that joins the same community-based overlay network.

For commercial application example, we can attach video clip in the reverse-query message, which is possible to use this mechanism as electric flyer distribution platform.

References

[1] Albert, R., Barabasi, A.L.: Statistical Mechanics of Complex Networks. Reviews of Modern Physics 74, 47–97 (2002)
[2] Faloutsos, M., Faloutsos, P., Faloutsos, C.: On Power-law Relationships of the Internet Topology. In: ACM SIGCOMM, pp. 251–262 (1999)
[3] Stauffer, D., Aharony, A.: Introduction to Percolation Theory. Taylor and Froncis, London (1994)
[4] Karger, D., Lehman, E., Leighton, T., Pnigrahy, R., Levine, M., Lewin, D.: Consistent Hashing and Random Trees: Distributed Caching Protocols for Relieving Hot Spots on the World Wide Web. In: Proceedings of the 29th annual ACM Symposium on Theory of Computing, pp. 654–663. ACM Press, New York (1997)
[5] Abrams, M., Standridge, C.R., Abdulla, G., Williams, S., Fox, E.A.: Caching Proxies: Limitations and Potentials. In: Proceedings of 4th International World Wide Web Conference, pp. 119–133 (1995)
[6] Chen, Y., Katz, R.H., Kubiatowicz, J.D.: Dynamic Replica Placement for Scalable Content Delivery. In: Druschel, P., Kaashoek, M.F., Rowstron, A. (eds.) IPTPS 2002. LNCS, vol. 2429, pp. 306–318. Springer, Heidelberg (2002)
[7] Li, Y., Liu, M.T.: Optimization of Performance Gain in Content Distribution Networks with Server Replicas. In: SAINT2003 Proceedings, pp. 182–189 (2003)
[8] Ratnasamy, S., Francis, P., Handley, M., Karp, R., Schenker, S.: A Scalable Content-addressable Network. In: Proceedings of the 2001 Conference on Applications, Technologies, Architectures, and Protocols for Computer Communications, pp. 161–172. ACM Press, New York (2001)
[9] Stoica, I., Morris, R., Karger, D., Kaashoek, M.F., Balakrishnan, H.: Chord: A Scalable Peer-to-Peer Lookup Service for Internet Applications. In: Proceedings of ACM SIGCOMM, pp. 149–160. ACM Press, New York (2001)
[10] Rowstron, A., Druschel, P.: Pastry: Scalable, Decentralized Object Location, and Routing for Large-Scale Peer-to-Peer Systems. In: Guerraoui, R. (ed.) Middleware 2001. LNCS, vol. 2218, pp. 329–350. Springer, Heidelberg (2001)
[11] Hildrum, K., Kubiatowicz, J.D., Rao, S., Zhao, B.Y.: Distributed Object Location in a Dynamic Network. In: Proceedings of the 14th ACM Symposium on Parallel Algorithms and Architectures, pp. 41–52. ACM Press, New York (2002)

[12] Sunaga, H., Hoshiai, T., Kamei, S., Kimura, S.: Technical Trends in P2P-Based Communications. IEICE Trans. Commun E87-B(10), 2831–2846 (2004)
[13] Adamic, L.A., Lukose, R.M., Puniyani, R., Huberman, B.A.: Search in power-law networks. Physical Review E 64(046135) (2001)
[14] Fujita, Y., Yoshida, J., Tsuda, K.: Reverse-Query Mechanism for Contents Delivery Management in Distributed Agent Network. In: Khosla, R., Howlett, R.J., Jain, L.C. (eds.) KES 2005. LNCS (LNAI), vol. 3684, pp. 758–764. Springer, Heidelberg (2005)
[15] Masuda, N., Konno, N.: Science of Complex Network. Sangyo Tosho (2005) (in Japanese)
[16] Sarshar, N., Boykin, P.O., Roychowdhury, V.P.: Percolation Search in Power Law Networks: Making Unstructured Peer-to-Peer Networks Scalable. In: IEEE P2P2004, IEEE Computer Society Press, Los Alamitos (2004)
[17] Pajek: http://vlado.fmf.uni-lj.si/pub/networks/pajek/
[18] Pennock, D.M., Flake, G.W., Lawrence, S., Glover, E.J., Giles, C.L.: Winners Don't Take All: Characterizing the Competition for Links on the Web. Proceedings of the National Academy of Sciences 99(8), 5207–5211 (2002)

Relation Analysis on Information System Life Cycle Processes by KeyGraph Algorithms

Tadashi Tsukahara[1], Atsuo Hazeyama[2], and Kazuhiko Tsuda[3]

[1,3] Graduate School of Business Sciences, University of Tsukuba
3-29-1 Otsuka, Bunkyo-ku, Tokyo 112-0012, Japan,
[2] Graduate School of Education, Tokyo Gakugei University,
4-1-1 Nukuikita-machi, Koganei-shi, Tokyo 184-8501, Japan
{Tadashi Tsukahara, t-tsuka}@poem.ocn.ne.jp,
{Atsuo Hazeyama, hazeyama}@u-gakugei.ac.jp,
{Kazuhiko Tsuda, tsuda}@gssm.otsuka.tsukuba.ac.jp}

Abstract. It is indispensable to provide a method for evaluating validity of information systems management. This paper proposes a method that analyzes the contents of information systems management by visualization with KeyGraph algorithms. Then we consider the relationships among the contents of each process in life cycle processes. We also analyze inter-relationships among the contents of several processes in stage. The results show effectiveness of our proposed method.

Keywords: Information System, Life Cycle Processes, KeyGraph Algorithms.

1 Introduction

Management of information systems should define life cycle of those systems and should be conducted according to the definition. Two types of management method are generally well-known: 1) a management method for information systems which constructs a software life cycle model and manages execution of activities in the model. The software life cycle model is consisted of requirement definition, requirement analysis, design, implementation, and verification that are technical processes for system development. 2) a management method which constructs a model based on time such as a framework of systematization, system development, and so on, and manages the business process or activities.

In recent years, it is indispensable the organizations in corporations possess and maintain information systems for their business management. It is said that it is also important to validate management of information systems under such management environments.

The goal of this study is to visualize processes and/or activities in the early stage of system life cycle that are subject to management and to investigate their relationships. More concretely speaking, we analyze a) relationships among activities of both information technology (IT) strategy and information system (IS) planning processes that composed of early stage, and b) inter-relationships between the processes of IT

B. Apolloni et al. (Eds.): KES 2007/ WIRN 2007, Part II, LNAI 4693, pp. 1362–1368, 2007.
© Springer-Verlag Berlin Heidelberg 2007

strategy and IS planning. We investigate success factors of life cycle management organizations perform by such analysis.

2 System Life Cycle Processes

There is little research on management of system life cycle processes because reports with respect to management are generally dealt with internal ones for companies. We start summarizing system life cycle processes and their management according to the related standards.

Risk management is based on a similar way of thinking to system life cycle management. Therefore we refer to literature of risk management.

2.1 Standards of Life Cycle Processes

Life cycle processes for system development has been already standardized as ISO/IEC 15288:2002 [1]. Fig. 1 shows correlation of processes among standards for system life cycle processes. System life cycle processes of this standard are composed of the following four process groups: 1) technical processes, 2) project processes, 3) enterprise processes, 4) agreement processes. The management processes of systems have been included in this standard.

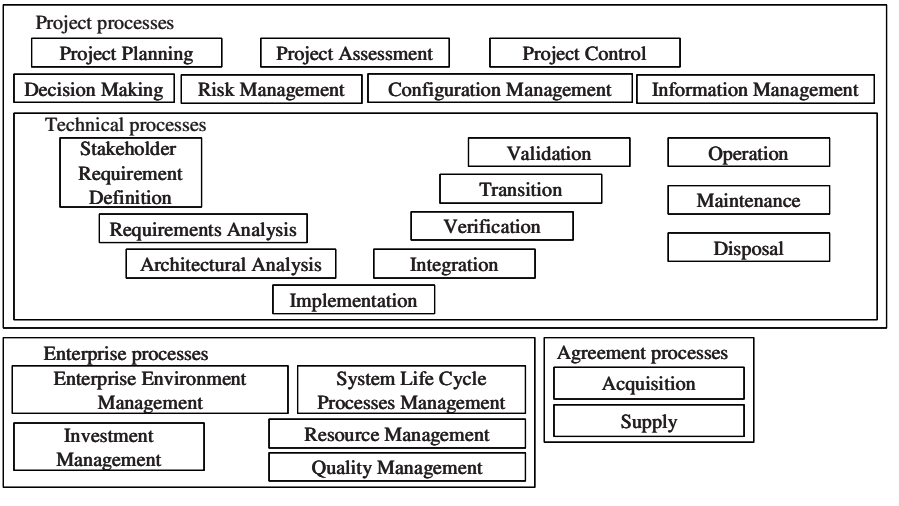


Fig. 1. Processes in ISO/IEC15288: 2002 System life cycle processes

2.2 Risk Management of Life Cycle Processes

McManus proposed a risk management method for technical process of system life cycle processes by using checklists [2].

Tsukahara proposed a risk assessment method that set a process model of system life cycle by the combination of several standards such as PMBOK Guide (ANSI/PMI 99-001 2004) [3] and created checklists [4].

These studies propose system life cycle management in the form of checklists.

3 Life Cycle Model

Many of traditional life cycle management for information systems focus on the development process we introduced in section 2. They utilized scenario of management that was created by some method like brain-storming in the form of checklists. However there are few discussions on the relationships among scenarios and the contexts. This study analyzes the relationships among scenarios that describe the objects and contents of management. Concretely speaking, 1) we adopt the stage model which manages system life cycle from the viewpoint of time. 2) We use the KeyGraph algorithms for the analysis of the stage model.

3.1 Life Cycle Stage Model

Table 1 shows the stage model of the system life cycle this study analyzes. We cite the stages Annex B and D of ISO/IEC 15288:2002. We use the System audit and management standards [5] as the contents of the stage. The number of sentences in the standard is shown in this table.

Table 1. Life Cycle Stage Model and System Management Standard

LIFE CYCLE STAGES (ISO/IEC15288:2002 Annex B and D)		**SYSTEM MANAGEMENT STANDARD**
CONCEPT	Identify Stakeholders' needs / Explore concepts / Propose viable solutions	Information Technology strategy (47)
		Planning Business (23)
DEVELOPMENT	Refine system requirements / Create solution description / Build system / Verify and validate system	Development Business (49)
PRODUCTION	Produce systems / Inspect and test	
UTILIZATION	Operate system to satisfy users' needs	Operation Business (73)
SUPPORT	Provide sustained system capability	Maintenance Business (19)
RETIREMENT	Store, archive or dispose of the system	

The System Management Standard is a standard for which a system auditor uses as criteria in audit when organizations conduct system audit. It is a practical norm for which organizations run appropriate investment for information systematization based on IT strategies and IS planning, development, operation, and maintenance of

system life cycle. It defines management items for each business process and describes the details. It represents system life cycle processes from the viewpoint of system audit.

3.2 Analysis by Using KeyGraph Algorithms

We analyze the contents of processes in the life cycle and the relationships between the processes by using KeyGraph algorithms [6, 7]. KeyGraph is a visualization method which combines inter-relationships of words that are included in several sentences written in a natural language by their frequency and/or co-occurrence. We think KeyGraph algorithms can be used in this study as follows:

1) we can check the inter-relationships between words and/or sentences that describe the processes in life cycle regardless of order by procedures
2) by processing descriptions of the several stages or processes with KeyGraph algorithms, we can ascertain their inter-relationships in a visualized form.

As the stages or processes are concepts that represent progress of time or business units, relationships among stages or processes KeyGraph algorithms represent are meaningful in a chronological way and from the viewpoint of business analysis. Such a feature of KeyGraph algorithms is useful for scenario analysis of management.

4 Analysis of Life Cycle Processes

We drew the contents in stage of IT strategy and IS planning by KeyGraph. We describe the contents of analysis as follows:

4.1 Analysis of IT Strategy Process

Fig. 2 shows a diagram for IT strategy process.

We find this diagram shows this process is composed of the following three groups:

1) head who executes management of the organization, and the management items (the overall optimization plan, business continuation plan, and compliance)
2) committee of information systems and department of information systems
3) information systems that optimize strategy for management

A group shows strong relationships between words with higher term frequency. Hereafter we call such a group “island”.

These results draw the following two conclusions and they are reasonable:

a) the head of executive is strongly connected with committee of information systems by the relationships which represent that (s)he obtains support information and that (s)he makes decisions on systematization.
b) committee of information systems and department of information systems divide roles by clarifying their authority, responsibility, and missions

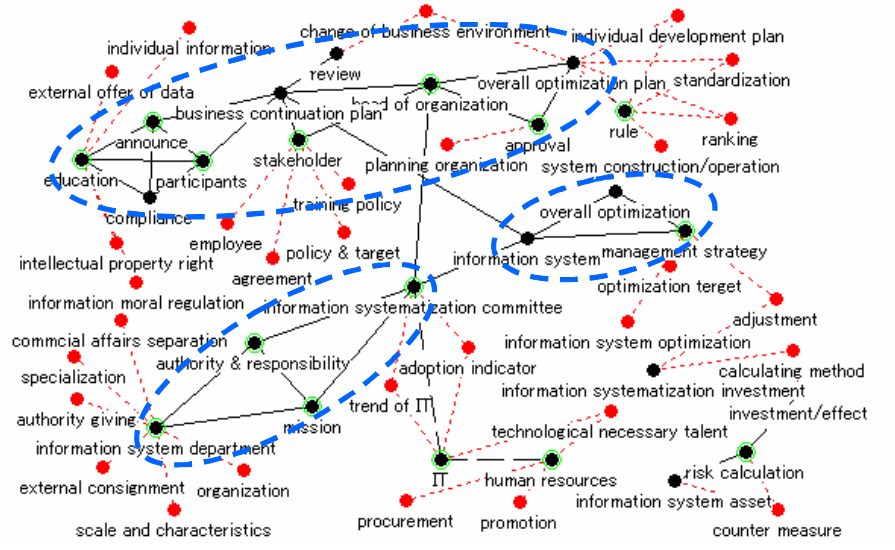


Fig. 2. KeyGraph for IT strategy process

4.2 Analysis of Planning Business Process

Fig. 3 shows a KeyGraph diagram for IS planning business process.

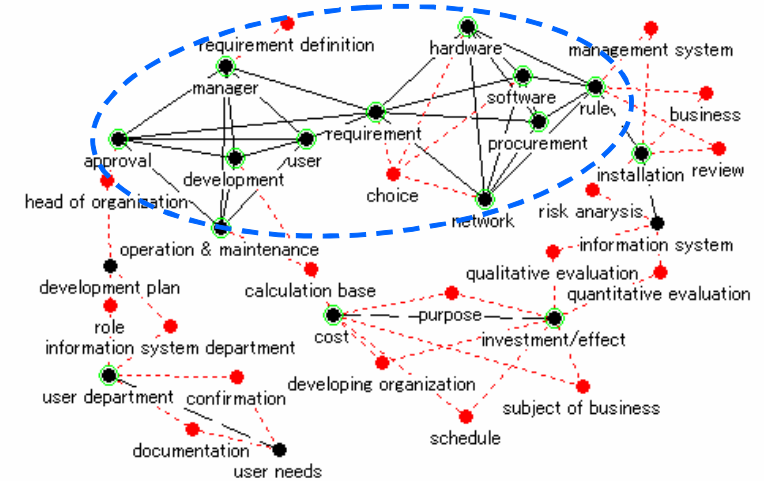


Fig. 3. KeyGraph for Planning Business Process

We find this diagram shows planning business process in building information systems has the following two important points:

1) not only department of development but also department of operation and maintenance participate in the task. Furthermore, department of user also participate. The participants determine requirements for the system.
2) Procurement of hardware, software, and network necessary for system building should be considered.

Both constitute one island. Various items that compose of this island are strongly inter-connected.

These results by analysis draw the following two conclusions and they are reasonable:

a) Procurement necessary for system building and stakeholders constitute one island around requirement.
b) The head of the planning department approves requirements and makes a development plan.

4.3 Unified Analysis of IT Strategy and Planning Business Processes

We visualized the contents that combined process of IT strategy with that of planning business and analyzed them. The contents are shown in Fig.4.

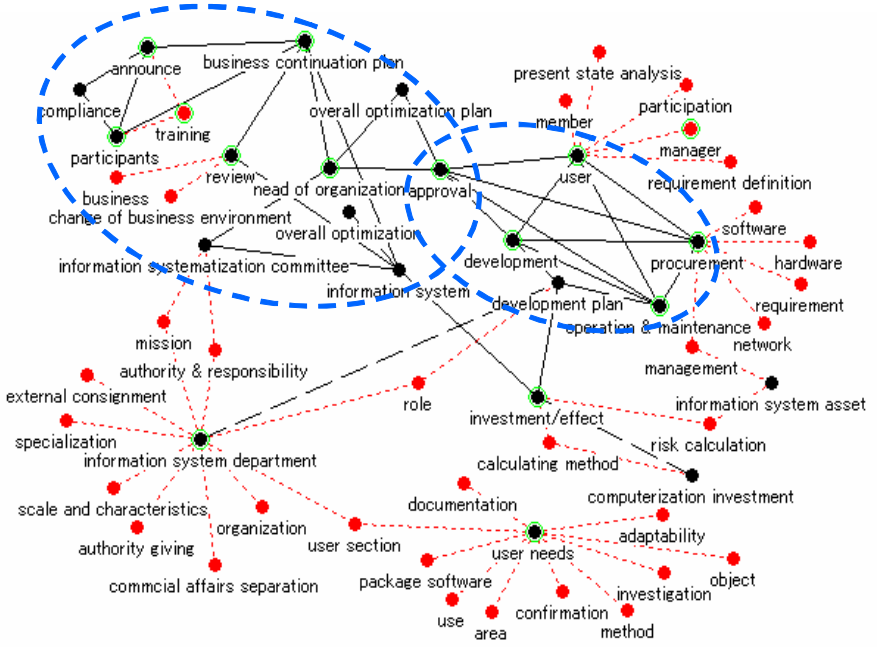


Fig. 4. KeyGraph for the processes that integrate IT strategy with IS planning

The management items (the overall optimization plan, business continuation plan, and compliance) that head who executes management of the organization treats still form an island. Similarly, creation of system requirements specification by the

department of development, operation, maintenance, and user also still forms an island. These two islands are connected by approval for the requirements by the head of the organization.

These results draw the following two conclusions and they are reasonable:

1) head who executes management of IT strategy has approved the development plan and procurement of the planning business process
2) information system for overall optimization is located in the IT strategy process. On the other hand, development plan is located in the planning business process

5 Conclusions

This study analyzed upstream stage in life cycle management of information systems by using KeyGraph. The result clarified the relationship between the head who executes management of the organization and the management items (the overall optimization plan, business continuation plan, and compliance) in the process of information technology strategy. We could find complicated situation with respect to requirements specification for systematization in the information system planning process. We also confirmed the method enables to analyze cases that range several processes.

KeyGraph enables to ascertain relationships among affairs in the life cycle processes. By making use of this feature, we will analyze all stages in life cycle management as well as the relationships among processes. Based on these results, we will be able to extract success factors and failure factors for life cycle management of information systems.

References

1. International Standardization for Organization: Systems Engineering - System life cycle processes, ISO/IEC 15288:2002 /JIS X 0170 (2004)
2. McManus, J.: Risk Management in Software Development Projects. Elsevier, Amsterdam (2004)
3. Project Management Institute, Inc.: A Guide to the Project Management Body of Knowledge Third Edition (PMBOK Guide) (2004)
4. Tsukahara, T.: Practice of Two-Step Risk Management in the Development of Information Systems. In: Proceedings of the 3rd International Conference on Project Management (2006)
5. System audit and management standards, Japan Information Processing Development Corporation (in Japanese) (2005)
6. Osawa, Y.: Data Analysis of Chance Discovery. Tokyo Denki University Press (2006) (in Japanese)
7. Kita, K., et al.: Information Retrieval Algorithms. Kyoritsu Publishing (2002) (in Japanese)

ERRATUM

Scalable 2-Pass Data Mining Technique for Large Scale Spatio-temporal Datasets

Tahar Kechadi and Michela Bertolotto

University College Dublin, Belfield, Dublin 4, Ireland.
{tahar.kechadi, michela.bertolotto}@ucd.ie
Sergio Di Martino, Filomena Ferrucci
Dip. di Matematica e Informatica, Università degli Studi di Salerno, Italy
{sdimartino,fferrucci}@unisa.it

B. Apolloni et al. (Eds.): KES 2007/ WIRN 2007, Part II, LNAI 4693, pp. 785–792, 2007.
© Springer-Verlag Berlin Heidelberg 2007

DOI 10.1007/978-3-540-74827-4_171

The names of the third and fourth authors have been incorrectly positioned between the affiliations. Consequently, these names are missing from the table of contents, author index, running heads, and metadata. The correct positioning is given below:

Tahar Kechadi[1], Michela Bertolotto[1], Sergio Di Martino[2], Filomena Ferrucci[2]

[1] University College Dublin, Belfield, Dublin 4, Ireland
{tahar.kechadi, michela.bertolotto}@ucd.ie
[2] Dip. di Matematica e Informatica, Università degli Studi di Salerno, Italy
{sdimartino,fferrucci}@unisa.it

The original online version of this chapter can be found at
http://dx.doi.org/10.1007/978-3-540-74827-4_99

Author Index

Abbattista, Fabio III-728
Abe, Akinori II-909
Abe, Hidenao II-934
Abe, Jair Minoro II-861, II-869, II-893
Abe, Norihiro III-1065
Abu Bakar, Rohani Binti III-308, III-370
Adachi, Yoshinori II-1013, II-1037, II-1045
Adli, Alexander II-287, II-295
Ahn, Tae-Chon III-49
Aimoto, Takeshi II-1346
Aiordachioaie, Dorel III-284
Akama, Seiki II-861, II-869, II-893
Akashi, Takuya III-91
Akdemir, Bayram I-125
Al-Hashel, Ebrahim III-909
Alba Castro, José Luis II-809
Alepis, E. II-1130
Almkvist, Esben III-259
Álvarez, Juan Antonio II-10, II-26
Alves, Ana I-346
Alves, Roberto T. III-830
Amarume, Hiraku III-761
Amitani, Shigeki II-901
Angulo, Cecilio II-10
Angulo, Jesús III-17
Antonelli, Michela I-303
Antoniotti, M. III-892
Antonucci, Alessandro II-604
Aoyama, Hirokazu III-453
Aoyama, Kouji II-1253
Apolloni, Bruno III-131, III-184, III-812
Aracil, Rafael II-1204
Araki, Kenji II-361
Arca, Stefano I-455
Argotte-Ramos, L. II-1138
Arima, Jun II-1253
Ariton, Doinita I-52
Ariton, Viorel I-52
Aritsugi, Masayoshi II-752, II-761
Arroyo-Figueroa, G. II-1138
Atis, Selcuk I-773
Aufaure, Marie-Aude III-695
Augello, Agnese I-42, I-156
Aulí-Llinàs, Francesc I-478
Avola, Danilo II-369
Azzini, Antonia II-801

Baba, Norio II-642, III-82
Baca, José II-1204
Bafna, Vibha I-485
Bajaj, Preeti I-471, I-740
Bajo, Javier II-453, II-461
Balachandran, Bala M. III-909, III-941
Ballesta, Mónica II-1170
Balvig, Jens J. III-1041, III-1049
Balzarotti, Stefania II-395
Bannore, Vivek I-510
Baptiste-Jessel, Nadine III-1000
Baruque, Bruno II-435
Basile, Pierpaolo III-711
Basile, Teresa M.A. I-132
Bassis, Simone III-184, III-812
Batres, Rafael II-712, II-720
Battistelli, Viviana I-463
Bawane, Narendra I-77
Becerra, J.A. II-1187
Bedini, Luigi III-9
Bellas, F. II-1187
Bellocchio, Francesco I-60
Benkert, Beatrice II-485
Bento, Carlos I-346
Bertolotto, Michela II-785
Bertoluzzo, Francesco II-619
Bertoni, Alberto III-886
Bertozzi, Massimo I-704
Bevilacqua, Vitoantonio III-830
Bi, Jun I-855
Biba, Marenglen I-132
Bilgin, Osman I-116
Blanco, Ángela I-140
Bo, Jiang Shen III-1041
Bolat, Emine Doğru I-804
Bombini, L. I-704
Bono, Salvatore I-647
Bonomi, Andrea III-686
Borghese, N. Alberto I-60, II-344

Borzemski, Leszek I-311, II-271
Bouchachia, Abdelhamid II-419
Brdiczka, Oliver I-363
Brigljević, Vuko I-319
Broggi, Alberto I-704
Brogi, Antonio III-703
Brown, David J. II-18, II-34, I-405
Brunelli, Matteo II-612
Brzostowski, Krzysztof II-247
Bueno, David III-1024
Bui, Viet Phuong III-431

Caamaño, P. II-1187
Cacciola, Matteo III-243
Cai, Xin II-175
Caiafa, Cesar F. III-1
Calefato, Fabio III-728
Camastra, Francesco III-267, III-300
Campolo, Maurizio III-243
Campos, Cristina II-1230
Cao, Cungen I-580
Car, Zeljka I-623
Caraffi, C. I-704
Carmagnola, Francesca III-745
Carmona, Cristina III-1024
Caruso, Costantina II-109
Casel, Nicolas II-58
Casiraghi, Elena I-421
Castrogiovanni, Mario III-804
Cattani, S. I-704
Cattinelli, Isabella II-344
Ceccarelli, Michele III-878
Cen, Zhongdi II-175
Cena, Federica III-745
Cerri, Pietro I-704
Cha, ByungRae II-141
Cha, Chang-Il I-397
Chadebec, Olivier III-431
Chae, Oksam I-501, I-526
Chai, Bencheng II-565
Chalmeta, Ricardo II-1230
Chang, Ji-Woong I-397
Chang, Pai-Yu II-696
Chang, Won-Du II-387
Chen, An-Pin II-90
Chen, Chun-Hsien I-262
Chen, Heng-Sheng III-354
Chen, Huifang III-477
Chen, Mu-Yen II-50
Chen, Wen-Yuan II-664
Chen, Yen-Wei II-303, II-311, II-318, II-326
Chen, Zaiping I-757
Chen, Zhigang II-194
Cheng, Chao-Yuan II-1323
Cheng, Jingde II-444, II-840
Chiarappa, Patrizia III-830
Chiho, Yoshida II-1089
Chiu, Hsiao-Ya II-90
Cho, Dongsub I-781
Cho, Seonghwi I-354
Choi, Byung-Uk II-99
Choi, Claudia II-485
Choi, Dong-Eun III-342
Choi, Gyoung Soon II-926
Choi, Hun-Young II-1146
Choi, Insung II-728
Choi, Seungok II-728, III-251
Choi, Woo-Kyung III-645
Chou, Ming-Tao II-1293, II-1299, II-1307, II-1323
Chu, Ching-Wu II-1323
Chuang, Chen-Chia II-689
Chung, Cheng-Chi II-1299
Chung, Hyunsook I-558
Chung, Jihoon II-99
Chung, Sheng-Luen II-664, II-877
Chung, Sheng Chieh II-90
Chung, Ta-Hsiang II-704
Chung, Yongwha II-816
Chung, Younky II-555
Chung, Yun-Koo II-533
Chyr, Wen-Li II-1285
Ciceri, Maria Rita II-395
Cimpian, Emilia III-736
Cioffi, Marco III-439
Cirrincione, Giansalvo I-688
Cirrincione, Maurizio I-688
Clark, Lillian II-66
Cococcioni, Marco I-303
Coghill, George M. II-18
Cogitore, Bruno III-422
Conejo, Ricardo III-1024
Consoli, Angela III-934
Conti, Vincenzo I-34, I-108
Corazza, Marco II-619
Corchado, Emilio II-133, II-435
Corchado, Juan M. II-453, II-461
Corfini, Sara III-703
Corona, Davide F.V. III-862

Cottington, David II-926
Coulomb, Jean-Louis III-431
Crippa, Paolo I-463
Crowley, James L. I-363
Cruz, M.D. II-26
Çunkaş, Mehmet I-116
Cuppens, Frédéric III-1107
Cuppens-Boulahia, Nora III-1107
Cutrupi, Vincenzo III-439
Cuzzocrea, Alfredo I-729

d'Amato, Claudia I-148, III-653
d'Anjou, Alicia II-1179
da Costa Pereira, Célia I-164
Dalbelo-Bašić, Bojana I-319
Dalke, Hilary II-926
Damiani, Ernesto II-832, III-1098
Danese, G. II-117
David, Amos A. III-1024
de Antonio, Angélica I-765
Debenham, John I-279
Degemmis, Marco III-711
Deguchi, Akiko III-753
Deguchi, Toshinori II-1029
Delgado, Myriam III-830
de Luis, Ana II-461
de Paz, Juan F. II-461
Deshmukh, Amol I-740
Dewan, M. Ali Akber I-501, I-526
Dharmendra Sharma III-917
Di Gesù, Vito III-862
Di Iorio, Ernesto I-9
Di Martino, Ferdinando I-101
Di Mauro, Nicola I-132
Di Noia, Tommaso III-711
Di Sciascio, Eugenio III-711
Ding, Ying III-736
Doh, Kyung-Goo III-626
Dong, Chen II-194, II-209
Donoso-Bach, Lluís I-478
Drapała, Jarosław II-279
Duret, Denis III-422
Duro, R.J. II-1187

Ebisawa, Yoko III-1058
Echegoyen, Zelmar II-1179
Eckert, Claudia III-1090
Edmonds, Ernest II-901
El Alami, Marc II-58
Encelle, Benoît III-1000
Encheva, Sylvia II-125, I-551, II-885
Enkhsaikhan, Majigsuren III-941
Erkan, Kadir I-804
Erkmen, Burcu I-17
Escalera, Juan A. II-1204
Esposito, Anna III-275
Esposito, Antonietta M. III-275

Fahimi, Farshad II-34
Fan, Liya III-661
Fan, Pei-Shu II-649
Fang, Hsin-Hsiung II-1293, II-1307
Fanizzi, Nicola I-148, III-653
Fansi, Majirus III-1098
Farooq, Nazni III-950
Fascioli, Alessandra I-704
Favaretto, Daniela II-627
Fedeli, Livio III-25
Fedriga, R.I. I-704
Fedrizzi, Michele II-612
Felisa, M. I-704
Feng, Jun II-1081, II-1097
Ferilli, Stefano I-132
Fernandez, Marcel III-992
Fernández, Miguel A. I-447, I-518
Fernández-Caballero, Antonio I-447, I-518
Fernández-Leal, Ángel I-327
Fernández-Montes, A. II-26
Ferraioli, Fabrizio III-439
Ferraresi, Simone II-149
Ferrari, Stefano I-60
Ferre, Manuel II-1204
Ferri, Fernando II-369, II-403
Fiasché, Maurizio III-414
Figuerola, Carlos G. II-427
Filippone, Maurizio III-300, III-901
Flórez-Revuelta, Francisco II-469
Formisano, Alessandro III-439
Francocci, Emanuele II-149
Frąś, Mariusz I-311
Freitas, Alex A. III-830
Freno, Antonino I-69
Frosini, Graziano I-303
Fuchino, Tetsuo II-712, II-720, II-744
Fujii, Arisa III-975
Fujii, Satoru III-453
Fujita, Yoshikatsu II-1331, II-1354
Fukui, Ken-ichi III-202
Fukui, Shinji II-1013, II-1021

Fukumi, Minoru III-91
Fukumura, Teruo II-777
Fukumura, Yoshimi III-761
Furukawa, Shuichi III-583
Furumura, Takashi III-469, III-522
Furutani, Michiko II-909
Furutani, Yoshiyuki II-909
Fyfe, Colin I-572

Gabbar, Hossam A. II-736
Gabillon, Alban III-1098
Gabrys, Bogdan II-419
Gaglio, Salvatore I-34, I-42, I-156
Gaito, Sabrina III-812
Gandhe, Sanjay T. I-797
Gao, Fei II-167
Gao, Kun II-194
Gao, Xueshan II-541, II-571
García-Chamizo, Juan Manuel II-469
García Mateo, Carmen II-809
García-Rodríguez, José II-469
Gastaldo, Paolo II-133
Gavrilova, Tatiana II-1154
Geem, Zong Woo I-371
Gena, Cristina III-745
Gendarmi, Domenico III-728
Gennari, Rosella III-678
Gerace, Ivan III-25
Gerbaud, Laurent III-422
Gervás, Pablo I-346
Ghidoni, S. I-704
Giachero, M. II-117
Gianfelici, Francesco I-463
Gianini, Gabriele II-832
Gigliotta, Onofrio II-74
Gil, Arturo II-1170
Giripunje, Shubhangi I-77
Giudicepietro, Flora III-275
Gómez, Francisco J. I-518
Gómez-Pérez, Asunción I-534
González Agulla, Elisardo II-809
González, Javier II-435
González, L. II-26
Goto, Kazuto I-631
Goto, Masato II-1073
Goto, Yuichi II-444, II-840
Gotoda, Naka III-608
Gotoh, Minori III-194
Graña, Manuel II-1179, II-1196
Grangel, Reyes II-1230
Grifoni, Patrizia II-369, II-403
Grisleri, Paolo I-704
Güneş, Salih I-125
Gunn, Steve III-148

Hagita, Norihiro II-909
Haindl, Michal III-33
Håkansson, Anne I-180, I-270
Hamaguchi, Takashi II-736, II-744
Han, Byeong-jun I-596
Han, Ki-Joon II-1
Hanaoka, Goichiro III-967, III-975, III-984
Hao, Jie II-209
Hara, Akira II-411
Harada, Kouji III-838
Haraguchi, Makoto II-1214
Harak, Hanane II-311
Hartung, Ronald I-180, I-270
Hashimoto, Yoshihiro II-736
Hatanaka, Toshiharu I-664
Hattori, Akira II-1073
Hatzilygeroudis, Ioannis II-477
Hayakawa, Jun-ichi II-987
Hayakawa, Kazuki III-1049
Hayakawa, Satoshi II-517
Hayashi, Yuki III-769
Hazeyama, Atsuo II-1362
Heh, Tzeng-Yuan III-406, II-696
Heo, Gunhaeng III-251, II-728
Herrero, Álvaro II-133
Hervás, Raquel I-346
Hien, Thai Duy II-311
Hilas, Constantinos S. II-1112
Hirano, Kazuhiko III-1065
Hirano, Shoji II-978
Hochin, Teruhisa II-793
Hong, Dong-Suk II-1
Hong, Kwang-Seok I-246, I-696
Hong, Seok-ju I-493
Hong, Sungeon I-354
Hong, Zhi-Yu I-747
Hori, Satoshi III-1058, III-1065
Hosoda, Takuma II-959
Hossain, M. Julius I-501, I-526
Howlett, R.J. II-671
Howson, P.A. II-671
Hsu, Chien-Chang I-747
Huang, Ding-Jie II-681
Huang, Houkuang III-166

Huang, Shih-Ping II-1315
Huang, Xinyin II-318
Huang, Xu III-917, III-926, III-950
Huang, Yi-Sheng II-656, II-704
Huneiti, Ammar II-1162
Hussain, Aini I-839
Hwang, Eenjun I-596
Hwang, Euy-Kyung I-222
Hwang, Sun-Myung II-533

Ichikawa, Teruhisa III-499, III-539
Ichimura, Hiroshi III-460
Ichimura, Takumi II-411
Ikeda, Mitsuru III-590, III-786
Ikeno, Hidetoshi III-123
Ikezaki, Masakazu I-712, II-769
Imai, Hideki III-967, III-984
Imai, Sayaka II-752
Inagaki, Kosuke III-210
Inagaki, Shigenori III-753
Inoue, Masaki II-736
Inuso, Giuseppina III-414, I-680
Inuzuka, Nobuhiro II-987
Iorio, F. III-870
Ishibuchi, Hisao II-509
Ishida, Yoshiteru III-796, III-821, III-838, III-846, III-854
Ishii, Naohiro II-995, II-1029, II-1037, II-1045, II-1051, II-1059
Ishikawa, Kohei II-1246
Ishino, Masahiko III-539
Isobe, Takashi II-1346
Isokawa, Teijiro III-115, III-123
Ito, Hideaki II-777
Ito, Masahiro II-326, II-335
Ito, Takaichi II-970
Itoh, Hidenori III-194
Itou, Junko III-484, III-547
Itou, Masahiro III-514
Iwahori, Yuji II-1005, II-1013, II-1021, II-1045
Iwakura, Tomoya II-1238
Iwata, Tomoharu III-600

Jahn, Dieter II-485
Jain, Lakhmi C. I-206, I-230, I-567, I-572, III-934
Jang, In-Hun III-73
Jędrzejowicz, Joanna I-172
Jędrzejowicz, Piotr I-172
Jen, Chan-Yung III-406
Jeng, Don Jyh-Fu III-370
Jeng, LiDer II-877
Jeng, MuDer II-649, II-877
Jeon, Hong-Tae III-645
Jeon, Moongu I-85
Jeulin, Dominique III-17
Jevtic, Dragan I-623
Jezic, Gordan I-214
Jimbo, Takashi II-1059
Jimenez, J.L. II-1196
Jin, Min II-217
Jin, Ran II-194, II-209
Jo, Geun-Sik II-255
Jrad, Zeina III-695
Jung, Jason J. II-255
Jung, Sangjoon II-555
Jurasovic, Kresimir I-214

Kabassi, K. II-1130
Kageyama, Hisayoshi III-453
Kakehi, Masahide III-600
Kamiura, Naotake III-123
Kanamori, Yoshinari II-752
Kanda, Taki III-393
Kanda, Tomoki II-318
Kando, Noriko II-377
Kanellopoulos, D. II-493
Kanenishi, Kazuhide III-608
Kaneyama, Takashi III-469, III-522
Kang, Euisun III-616
Kang, Sanggil I-588
Kang, Tae-Gu III-645
Kang, Tae-Koo III-65
Kanoh, Masayoshi III-194
Kanzaki, Ryohei III-123
Karungaru, Stephen III-91
Kataoka, Nobuhiro III-491
Katarzyniak, Radosław Piotr I-230
Kato, Hirofumi II-1005
Kato, Naotaka II-1261
Kato, Nariaki II-1051
Kato, Shohei III-194
Kato, Tsuneaki II-377
Kawaguchi, Masashi II-1059
Kawai, Yukiko I-612
Kawanaka, Haruki II-1005, II-1013
Kawano, Koji II-736
Kawaoka, Tsukasa I-604, I-631, I-720, I-813, I-847

Kawasaki, Yuichi II-517
Kayano, Akifumi III-446
Kazienko, Przemysław III-1016
Kechadi, Tahar II-785
Kei, Ikeda II-311
Keradec, Jean-Pierre III-422
Keskar, Avinash G. I-740, I-797
Kikuchi, Koki II-571
Kim, Dong-Oh II-1
Kim, Dongwon III-65
Kim, Haeng-Kon II-581
Kim, Hang-Kon II-533
Kim, Hanil I-379
Kim, Hea-Suk I-222, II-1146
Kim, Hong-Gee II-255
Kim, Hyung Kwan I-429
Kim, Jin-Il I-93
Kim, Jinho I-222, II-1146
Kim, Jong-Dae I-379
Kim, Joo-Hyung III-57
Kim, Jung-Hyun I-246, I-696
Kim, Jungmin I-558
Kim, Jungyeop I-354
Kim, Minkoo I-596
Kim, Sang-Wook I-379, I-397
Kim, Seong-Joo III-645
Kim, Sunghwan II-99
Kim, Ung Mo I-672
Kim, Young-Joong III-41
Kim, Yukyong III-626
Kimble, Chris II-66
Kimura, Kakeru III-469
Kimura, Manabu III-227
Kimura, Masahiro III-202, III-235
Kisada, Wataru III-1049
Kitajima, Teiji II-744
Kitakoshi, Daisuke III-218
Kitazima, Teiji II-736
Klawonn, Frank II-485
Kobara, Kazukuni III-984
Kocabey, Sureyya I-773
Koh, Takahiro II-840
Koike, Satoshi II-1277
Kojima, Masanori III-469
Kojiri, Tomoko III-600, III-769
Kokogawa, Tomohiro III-477
Kokubo, Mikinori III-499
Koleszynska, Joanna I-789
Korus, Lukasz I-335
Kosaka, Hiroaki III-399
Koshiba, Hitoshi II-1261
Kotsianti, S.B. II-493
Koutsojannis, Constantinos II-477
Kozierkiewicz, Adrianna I-238
Król, Dariusz III-1008
Kubota, Naoyuki III-98
Kucuk, Serdar I-655
Kudenko, Daniel II-66
Kudou, Tsukasa III-491
Kumamoto, Tadahiko I-612
Kumar, Harshit I-588
Kunifuji, Susumu III-777, II-1214, II-1222, II-1261
Kuriyama, Ken III-567
Kuroda, Ipei III-1041
Kurosawa, Yoshiaki II-411
Kusek, Mario I-214
Kusunoki, Fusako III-753
Kuwajima, Isao II-509
Kwak, Jin I-672

Lagos, Nikolaos II-1162
La Foresta, Fabio I-680, III-243, III-414
Lanubile, Filippo III-728
Lanzarotti, Raffaella I-455
Lasota, Tadeusz III-1008
Lazzerini, Beatrice I-303
Le, Son III-917
Lee, Byong I-781
Lee, Chil Woo I-429, I-493
Lee, DongSeob II-141
Lee, Eun-Ser II-581
Lee, Ho-Dong III-57
Lee, Hsuan-Shih II-1285, II-1293, II-1299, II-1307, II-1323
Lee, Huey-Ming III-325, III-354, III-363, III-386
Lee, Hyun-Gu III-57
Lee, Jeehyong III-251
Lee, Junghoon I-379
Lee, Jusang I-85
Lee, Kang-Jun II-1
Lee, Sang Ho II-581
Lee, Sang-Gu I-93
Lee, Sang-Yun II-99
Lee, Sangwook I-85
Lee, Sungju II-816
Lee, Tsang-Yean III-354
Lee, Wooram III-251
Lee, Yang Weon I-429

Lee, Youngsook I-672
Lee, Zne-Jung II-689
Leng, Jinsong I-572
Leporati, F. II-117
Leray, David II-1122
Li, Jifang II-209
Li, Pi-Hsiang II-42
Li, Qian II-225
Li, Shijin II-1066
Li, Zhonghua I-831
Lim, Myo-Taeg III-41
Lim, Younghwan III-616
Lin, Chii-Ruey II-649
Lin, Li-Hsueh III-363
Lin, Lily III-363, III-386
Lin, Po-Hung III-406, II-696
Lipori, Giuseppe I-455
Liu, Hanzhong II-167
Liu, Honghai I-405, II-18, II-34
Liu, James II-240
Liu, Qiang I-855
Liu, Xuejun II-167
Liu, Ying I-542
Liu, Zhen II-202
Lo Bosco, Giosuè III-862
Loh, Han Tong I-542
Loia, Vincenzo I-101
Lombardi, Gabriele I-421
López, María T. I-447, I-518
Lou, Rui I-757
Lovrek, Ignac I-198
Lu, Jiamin II-1097
Luccarini, Luca I-639

Ma, Wanli II-547, II-589
Maeno, Yoshiharu II-970
Magnani, Lorenzo II-917
Malchiodi, Dario III-131, III-184
Malerba, Donato II-109
Mammone, Nadia III-414, I-680
Manzano, Juan M. II-453
Mao, Ching-Hao III-325
Marasović, Kristina I-319
Maratea, Antonio III-878
Marcelloni, Francesco I-303
Marinaro, Maria III-275
Marrara, Stefania II-801, III-1098
Martín-Merino, Manuel I-140
Martinelli, Francesca III-25
Martone, Raffaele III-439
Mascherini, Massimiliano I-413
Mastrogiovanni, Fulvio III-670
Mastronardi, Giuseppe III-830
Masuda, Yasutaka I-813
Masulli, Francesco III-901
Mateos, Montserrat II-427
Matrone, G. II-117
Matsubara, Fuminori II-1005
Matsuda, Noriyuki III-1065, III-1073
Matsuda, Takahiro III-967
Matsui, Nobuyuki III-115, III-123
Matsuoka, Rumiko II-909
Matsushita, Mitsunori II-377
Matsuura, Kanta III-967, III-984
Matsuura, Kenji III-608
Mauri, G. III-892
Medici, P. I-704
Mello, Paola I-639
Menolascina, Filippo III-830
Mera, Kazuya II-411
Merico, D. III-892
Mich, Ornella III-678
Miele, G. III-870
Miglino, Orazio II-74
Miholca, Constantin III-284
Mikeš, Stanislav III-33
Milici, Giovanni I-108
Min, Suhong I-781
Mineno, Hiroshi III-469, III-477, III-522, III-531
Mirza, Hanane H. II-295
Misue, Kazuo II-1269, II-1277
Mitsukura, Yasue III-91
Miura, Hirokazu III-1065
Miura, Motoki II-1222, III-777
Miyachi, Taizo III-1041, III-1049
Miyadera, Youzou III-554
Miyaji, Isao III-460
Miyamaru, Takuya III-531
Miyamoto, Eiichi III-514
Miyamoto, Sadaaki II-517
Miyazaki, Michio I-93
Mizoguchi, Hiroshi III-753
Mizuno, Tadanori III-453, III-469, III-477, III-491, III-522, III-531, III-539
Mizuta, Masahiro II-501
Mogami, Yoshio III-82
Mohan, Sriram III-1081
Moniaci, Walter III-259

Montero, Calkin S. II-361
Montuori, Alfonso III-259
Moon, Daesung II-816
Moon, Il-Young I-390
Moon, Yang-Sae I-222, II-1146
Morabito, Francesco Carlo III-243, III-414, I-680
Moraga, Claudio I-26
Moret-Bonillo, Vicente I-327
Moretti, Elena II-627
Morgan, Ian I-405
Morihiro, Koichiro III-115
Mosca, Alessandro III-686
Motto Ros, Paolo III-292
Mukai, Naoto II-1081, II-1097, II-1105
Münch, Richard II-485
Munemori, Jun III-446, III-484, III-547
Musa, Zalili Binti III-378
Muselli, Marco III-174
Mushrif, M.M. I-485
Mutluer, Mümtaz I-116

Nagata, Yasunori II-287, II-295, II-311
Nakagawa, Kenichi II-1261
Nakagawa, Takashi II-1013
Nakamatsu, Kazumi II-869, II-893
Nakamtsu, Kazumi II-861
Nakamura, Masaaki III-499
Nakamura, Shoichi III-554
Nakano, Miyoko III-91
Nakano, Ryohei III-210, III-218, III-235
Nakao, Zensho II-287, II-295, II-303, II-311
Nam, Junghyun I-672
Napolitano, F. III-870
Nara, Shinsuke II-444, II-840
Nara, Yumiko II-942
Natali, Luca III-131
Nazzicari, N. II-117
Nguyen, Ngoc Thanh I-206, I-230, I-238
Nicolau, Viorel III-284
Nicosia, Giuseppe III-804
Niennattrakul, Vit II-824
Nin, Kou II-642
Nishikawa, Ikuko II-326, II-335
Nishimura, Haruhiko III-115
Nishitani, Hirokazu III-399
Niu, Zhendong I-580
Nogawa, Yasutsugu III-98
Nojima, Yusuke II-509
Nolfi, Stefano II-74
Nouno, Ikue II-335
Nowé, Ann III-107
Noyel, Guillaume III-17
Nozaki, Shinya II-303
Numao, Masayuki III-202

Obashi, Yoshitsugu III-477
Odagiri, Kazuya II-995
Oehlmann, Ruediger II-926
Ogawa, Kazuto III-975, III-984
Ogura, Kanayo II-1246
Ohsaki, Miho II-934
Ohsawa, Yukio II-959, II-970
Ohtake, Go III-975
Okada, Masashi II-1051
Okamoto, Seishi II-1238
Okamoto, Takeshi III-821
Okita, Makoto III-514
Okubo, Yoshiaki II-1214
Okuda, Kazuo II-1059
Okumura, Noriyuki I-604, I-631, I-720
Ortega, Juan Antonio II-10, II-26
Otero Muras, Enrique II-809
Ozaki, Masahiro II-1037, II-1045
Özbay, Yüksel I-1

Paik, Juryon I-672
Pak, Jung-Min III-41
Palade, Vasile III-284
Palmonari, Matteo III-686
Palsodkar, Prasanna I-740
Pan, Fuping I-821
Paolozzi, Stefano II-403
Paradiso, Angelo III-830
Park, Chang-Hyun III-73
Park, Daehyuck III-616
Park, Gwi-Tae III-57, III-65
Park, Kiheon II-728
Park, Min-Chul III-57
Park, Soohong I-354
Park, Sungjoon I-588
Parsai, Rashmi I-471
Pasero, Eros III-259, III-292
Paterlini, M. I-704
Payá, Luis II-1170
Pedrero, Jose M. II-1170
Pellegrini, Paola II-627
Perconti, Davide I-34
Pérez, Carlos II-10

Pérez-Jiménez, Rafael I-447
Pereira, Francisco C. I-346
Petrosino, Alfredo III-140
Pham, Tuan D. I-439
Phillips-Wren, Gloria I-567
Piatti, Alberto II-604
Piattini, Mario II-353
Picasso, Francesco II-133, II-149
Pilato, Giovanni I-42, I-156, III-720
Pinello, Luca III-862
Pirrone, Roberto III-720
Piuri, Vincenzo I-60
Polat, Kemal I-125
Ponticorvo, Michela II-74
Porta, P.P. I-704
Portillo-Rodríguez, Javier II-353
Posada, Jorge I-295
Postalcıoğlu, Seda I-804
Posterli, M. I-704
Pratissoli, Stella I-421
Proto, Araceli N. III-1
Purevjii, Bat-Odon II-752

Quaglini, Alessio II-149

Raiconi, G. III-870
Raimondo, Giovanni III-259
Rakowski, Maciej I-206
Ramírez, Jaime I-534, I-765
Ramli, Dzati Athiar I-839
Rao, ZhiMing I-262
Rascunà, Rosario III-804
Ratanamahatana, Chotirat Ann II-824
Reignier, Patrick I-363
Reinoso, Oscar II-1170
Resta, Marina II-635
Rhee, Sang-Min I-222
Rho, Seungmin I-596
Ribino, Patrizia I-108
Rizzi, Simone I-421
Rizzo, Riccardo III-720
Röder, Patrick III-1090
Rodríguez, Sara II-453, II-461
Rodríguez-Ortíz, G. II-1138
Rogers, Paul II-419
Roh, Seok-Beom III-49
Romano, Salvatore I-34
Rouve, Laure-Line III-431
Rovetta, Stefano III-901
Rovira, Jordi II-435
Rúbies-Feijoo, Carles I-478
Ruffino, Francesca III-174
Rusin, Michal I-206
Russo, Giuseppe III-720
Rybakov, V. II-950

Sahalos, John N. II-1112
Sahel, Zoheir II-419
Saito, Kazumi III-202, III-227, III-235
Sáiz, José Manuel II-133
Sakai, Kazuo III-567
Sakakibara, Kazutoshi II-335
Sakamoto, Hirotaka II-335
Sakamoto, Yasuyuki III-777
Sakata, Yoshihiko II-1346
Salerno, Emanuele III-1, III-9
Samad, Salina Abdul I-839
Sanín, Cesar I-295
Sans, Thierry III-1107
Sansonnet, Jean-Paul II-1122
Saruwatari, Yasufumi II-1354
Sassi, Roberto II-801
Satake, Hiroki III-590
Sato, Kenji II-1214
Sato-Ilic, Mika II-525
Satomi, Masashi III-98
Scarpa, Giuseppe III-33
Scarpetta, Silvia III-275
Scherer, Stefan III-156
Schmidt, Rainer I-287
Schobert, Max II-485
Schomburg, Dietmar II-485
Schwenker, Friedhelm III-156
Scotti, Fabio II-801
Scriminaci, Mario I-156
Seki, Yoich III-123
Semeraro, Giovanni III-711
Sengupta, Arijit III-1081
Serra-Sagristà, Joan I-478
Sessa, Salvatore I-101
Seta, Kazuhisa III-590, III-786
Setchi, Rossitza II-1162
Setiawan, Nurul Arif I-493
Sgorbissa, Antonio III-670
Shao, Gang I-757
Sharma, Dharmendra II-541, II-547, II-589, III-909, III-917, III-950
Shen, Pei-Di II-1285
Shen, Yan I-855
Shiah, Sheau-Wen II-696, III-406

Shiau, Yea-Jou II-649
Shigeki, Yokoi II-1089
Shigematsu, Takayuki III-514
Shim, D. I-85
Shimada, Yukiyasu II-712, II-736, II-744
Shimizu, Nobuo II-501
Shin, Jungpil II-387
Shin, Miyoung I-379, I-397
Shiratori, Norio III-547
Shyr, Wen-Jye II-82
Siemiński, Andrzej II-263
Sim, Kwee-Bo III-73
Sinkovic, Vjekoslav I-214
Sobecki, Janusz I-206, III-1033
Sobue, Shouta II-318
Soga, Masato III-1073
Solesvik, Maryna Z. I-551
Song, JooSeok III-959
Sorbello, Filippo I-34, I-108, I-647
Soriano, Miguel III-992
Soto, Juan Pablo II-353
Sottara, Davide I-639
Staiano, Antonino III-140
Stathopoulou, I.-O. II-1130
Stefanini, Federico M. I-413
Su, Pin-June II-656
Sugihara, Taro II-1261
Sugimura, Ai II-1037
Sugiyama, Kozo II-1246
Sulaiman, Rossilawati II-547
Suzuki, Keisuke III-554
Suzuki, Takeshi III-1049
Świątek, Jerzy II-247, II-279
Swierkowski, Leszek I-510
Szczerbicki, Edward I-295

Tadauchi, Masaharu II-995
Tafreschi, Omid III-1090
Tagliaferri, R. III-870
Tai, David Wen-Shung II-42
Takagi, Saeko III-1073
Takahagi, Eiichiro III-316
Takahashi, Hiroshi II-1339
Takahashi, Masakazu II-1331, II-1339
Takahashi, Osamu III-514
Takahashi, Satoru II-1331, II-1339
Takahashi, Yusuke III-453
Takami, Yasuda II-1089
Takamizawa, Hidehisa II-761
Takashima, Hikaru II-720
Takatsuka, Ryozo II-1261
Takeda, Kazuhiro II-736, II-744
Takeoka, Saori II-1037
Taki, Hirokazu III-1065, III-1073
Talele, K.T. I-797
Tamura, Yasuhisa III-583
Tan, Hong-Zhou I-831
Tanahashi, Yusuke III-218
Tanaka, Hirofumi II-959
Tanaka, Jiro II-1277
Tanaka, Katsumi I-612
Tanaka, Takushi II-852
Tanaka-Yamawaki, Mieko II-597
Taniguchi, Kazuhiko III-98
Tapia, Dante I. II-453, II-461
Tektas, Mehmet I-773
Teng, Wei-Chung II-681
Terashima, Yoshiaki III-531
Terziyan, Vagan III-634
Tettamanzi, Andrea G.B. I-164
Tezel, Gülay I-1
Thai, Hien D. II-295
Thiel, Christian III-156
Thielen, Bernhard II-485
Tian, Fengzhan III-166
Tinelli, Eufemia III-711
Ting, I-Hsien II-66
Tokunaga, Yuichi III-531
Tokuoka, Seiji II-597
Tomàs-Buliart, Joan III-992
Tommasi, Stefania III-830
Tona, G. I-34
Topuz, Vedat I-773, I-864
Toro, Carlos I-295
Toyoda, Sadamitsu II-1222
Tran, Dat II-547, II-589, III-926
Trawiński, Bogdan III-1008
Trawiński, Krzysztof III-1008
Trentin, Edmondo I-9
Trzec, Krunoslav I-198
Tseng, Wei-Kuo II-1293, II-1307
Tsihrintzis, G.A. II-1130
Tsuchiya, Seiji I-813, I-847
Tsuda, Kazuhiko II-1339, II-1354, II-1362
Tsuge, Satoru III-91
Tsukahara, Tadashi II-1362
Tsumoto, Shusaku II-934, II-978
Tumin, Sharil II-125, I-551, II-885

Turchetti, Claudio I-463
Turnbull, George I-405
Tweedale, Jeff III-934

Uchino, Kanji II-1238
Ueda, Misato II-959
Ueda, Naonori III-227
Uesugi, Kentaro II-1105
Ugai, Takanori II-1253
Ui, Takeshi II-959
Umano, Motohide III-590
Uosaki, Katsuji I-664
Urata, Hiroki III-123
Ushiama, Taketoshi III-575, I-712, II-769

Valentini, Giorgio III-886
Vaquero, Javier I-295
Vassallo, Giorgio I-42, I-156
Velasco, F. II-26
Verbeeck, Katja III-107
Versaci, Mario III-243, III-414
Villaverde, I. II-1196
Villazón-Terrazas, Boris I-534
Virvou, M. II-1130
Vitabile, Salvatore I-34, I-108, I-647
Vizcaíno, Aurora II-353
Vizzari, Giuseppe III-686
Vrancx, Peter III-107
Vukovic, Marin I-623

Waldenström, Christofer I-254
Waligora, Tina I-287
Walters, S.D. II-671
Wang, Chen-Shu II-50
Wang, Jin-Long II-1315
Wang, Qin II-217
Wang, Su-Man II-1299
Wang, Taowei II-233
Wang, Zhihai III-166
Wang, Zhudeng II-189
Wanichsan, Dachawut II-824
Watabe, Hirokazu I-604, I-631, I-720, I-813, I-847
Watada, Junzo III-308, III-333, III-370, III-378
Watanabe, Toyohide I-712, II-769, II-1081, II-1097, II-1105, III-575, III-600, III-769
Watanabe, Yuji III-846
Wen, YuanLin II-877
Won, Chan-Hee III-41
Won, Dongho I-672
Won, Jung-Im I-379
Woodham, Robert J. II-1005, II-1013, II-1021
Wright, Peter II-66
Wu, De II-1066
Wu, Hui-Ju II-42
Wu, Jianping I-855
Wu, Xiaobing II-541, II-571
Wu, Yuqing III-1081
Wurtz, Frederic III-422

Xi, Lifeng II-159, II-182, II-194, II-202, II-209, II-217, II-240
Xiao, Qili II-182
Xiao, Tianyuan III-661
Xu, Bing III-333
Xue, Wen II-793

Yabutani, Joji III-1058
Yaegashi, Rihito II-995
Yamada, Kunihiro III-460, III-469, III-522
Yamada, Takeshi III-600
Yamaguchi, Norio III-514
Yamaguchi, Takahira II-934
Yamakami, Toshihiko III-507
Yamaki, Naokazu III-539
Yamamoto, Nozomi II-1021
Yamano, Katsuhisa II-1059
Yamasaki, Hiroya II-744
Yamazaki, Atsuko K. III-1058
Yan, Yonghong I-821
Yan, Zhixian III-736
Yang, Aimin II-233
Yang, Fu-Cheng III-406, II-696
Yang, Heng-Li II-50
Yang, Jianqiang III-148
Yang, Nan I-757
Yang, Xuemei II-326
Yano, Yoneo III-608
Yasuda, Takami II-1073
Yasumura, Shoko II-769
Yazaki, Takeki II-1346
Yeh, Chen-Huei II-1293, II-1299, II-1307
Yıldırım, Tülay I-17

Yin, Hujun II-435
Yin, Xunlei I-757
Ying, Kuo-Ching II-689
Yokoi, Shigeki II-1073
Yokoyama, Setsuo III-554
Yoshida, Hajime III-547
Yoshida, Jun II-1354
Yoshida, Kenichi II-1346
Yoshida, Kouji III-460, III-469
Yoshida, Takamasa II-736
Yoshimoto, Fujiichi III-1073
Yoshimura, Eriko I-604, I-720, I-847
You, Kwanho III-251, II-728
Youn, Jeh-Seon I-246
Young-Joong Kim III-41
Yu, Donggang I-439
Yu, Jian II-565
Yu, SuJung III-959
Yuan, Guo-Cheng III-862
Yuizono, Takaya III-446
Yukawa, Takashi III-761
Zaccaria, Renato III-670
Zaffalon, Marco II-604
Zampunieris, Denis II-58
Zani, P. I-704
Zatwarnicka, Anna II-271
Zatwarnicki, Krzysztof II-271
Zgrzywa, Michał I-189
Zhang, Chunxia I-580
Zhang, Liangbin II-159, II-233
Zhang, Yanfeng III-166
Zhang, Yanghong II-194
Zhang, Yanhong II-240
Zhang, Yunong I-831
Zhao, Chen II-565
Zhao, Qingwei I-821
Zheng, Yi III-477
Zhou, Xiaobo I-439
Zhu, Yuelong II-1066, II-1097
Zoppis, I. III-892
Zunino, Rodolfo II-133