Jörg H. Siekmann
Carsten Ullrich (Eds.)

LNAI 4429

Cognitive Systems

Joint Chinese-German Workshop
Shanghai, China, March 2005
Revised Selected Papers

Springer

Lecture Notes in Artificial Intelligence 4429

Edited by J. G. Carbonell and J. Siekmann

Subseries of Lecture Notes in Computer Science

Ruqian Lu
Jörg H. Siekmann
Carsten Ullrich (Eds.)

Cognitive Systems

Joint Chinese-German Workshop
Shanghai, China, March 7-11, 2005
Revised Selected Papers

Springer

Series Editors

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

Volume Editors

Ruqian Lu
Chinese Academy of Sciences
Institute of Mathematics
Beijing, China
E-mail: rqlu@fudan.edu.cn

Jörg H. Siekmann
Universität des Saarlandes
and German Research Center for Artificial Intelligence (DFKI)
66123 Saarbrücken, Germany
E-mail: siekmann@dfki.de

Carsten Ullrich
German Research Center for Artificial Intelligence (DFKI)
Stuhlsatzenhausweg 3, 66123 Saarbücken, Germany
E-mail: cullrich@dfki.de

Library of Congress Control Number: 2007921307

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 3-540-70933-9 Springer Berlin Heidelberg New York
ISBN-13 978-3-540-70933-6 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: 12021215 06/3142 5 4 3 2 1 0

Preface

This special issue collects a subset of the papers presented at the Joint Chinese-German Workshop on Cognitive Systems, held March 7-11, 2005, at Fudan University, in Shanghai, the city that never sleeps and changes daily. Just as it is not easy to keep track of Shanghai's growth and modernisation, it is hard to keep up with research on the new transdiscipline of cognitive systems, which is emerging from computer science, the neurosciences, computational linguistics, neurological networks and the new philosophy of mind. The workshop served to present the current state of the art in these fields and brought together researchers from Fudan University and Jiao Tong University, both in Shanghai, China, and from Saarland University, Germany.

The Workshop on Cognitive Systems was the last in a series of events to mark the longstanding collaboration between the three universities, which includes numerous joint projects, exchange of researchers and research visits, as well as formal joint cooperation agreements, treatises and joint Ph.D. and student exchange programmes in the fields of computer science, artificial intelligence and in computational linguistics.

Well before 1995 there were several exchange visits between Ruqian Lu and Jörg Siekmann to Shanghai, Beijing and Saarbrücken/Kaiserslautern. In 1995, formal cooperation started with a visit to Shanghai by Jörg Siekmann, Xiaorong Huang and Hans Uszkoreit. The visit resulted in the joint project Applied Chinese Natural Language Generation on the automatic generation of Chinese, English and German languages for weather reports and for stock exchange news. It was jointly funded by the Volkswagen Foundation of Germany, the National Natural Science Foundation of China and the Shanghai Science and Technology Committee of China. This project, which successfully concluded in 1999, was the start of a long-term cooperation and additionally triggered much research and a number of student exchanges. The cooperation culminated just recently in the founding of the Joint Research Lab for Language Technology at Shanghai Jiao Tong University. It opened in March 2005 with a public ceremony that was widely presented in the press and television media of China and Germany.

The Workshop on Cognitive Systems, financed by the Deutsche Forschungsgemeinschaft (German National Science Foundation) and Fudan University, was a prerequisite for an international joint Chinese-German research training group to support about thirty Ph.D. students and postgraduate researchers at the three universities involved: Fudan University and Shanghai Jiao Tong University in China and University of Saarland in Germany.

This workshop marked the starting point of a long-term collaboration between the three universities in the broadly conceived area of cognitive systems. Cognitive systems are computational agents that perceive, understand, learn and develop through individual or social interaction with the environment and other agents.

Cognitive technologies model cognitive faculties such as perception, communication, planning, reasoning, autonomous action, social interaction and adaptive behaviour. The long-term research goal of our collaboration is to further develop these technologies and to use them in the context of the next generation of the Internet, which will be accessible *anytime* and *anywhere* using handheld devices such as PDAs. The *mobile access* to this worldwide information (i.e., multi-lingual text-processing) and its multimedia content (i.e., video, TV clips, pictures and diagrammatical information) will be *multimodal* using speech input and output, gestures, pointing and body features as well as the usual keyboard, mouse and screen.

Learning and education as well as the provision of daily life information for people in the Information Society will be facilitated by the next generation of the Internet as a backbone. Hence the focus of the planned graduate school will be on mobile access to the *Multimodal, Multimedia and Multi-lingual Semantic Web*.

The number of Internet users in China passed the 100 Million mark in the year 2005, and the current growth rate is 43%. Provided this trend is not disturbed by external events it is predicted that the number of registered Internet users in China will have surpassed the number of American (USA) users by 2006, and based on current growth rates it is also predicted that in the near future, i.e., in less than ten years' time, China alone will have more Internet users than the rest of the world: this is the fastest growing market in the world today.

The papers in these proceedings are organized as follows: There are four papers related to e-learning using the Internet to some extent:

- Christoph Benzmüller, Helmut Horacek, Ivana Kruijff-Korbayová, Manfred Pinkal, Jörg Siekmann, Magdalena Wolska: "Natural Language Dialog with a Tutor System for Mathematical Proofs"
- Rudolf Fleischer, Gerhard Trippen: "On the Effectiveness of Visualizations in a Theory of Computing Course"
- Ruqian Lu, Hongge Liu, Songmao Zhang, Zhi Jin, Zichu Wei: "Some Cognitive Aspects of a Turing Test for Children"
- Erica Melis, Ruimin Shen, Jörg Siekmann, Carsten Ullrich, Fan Yang, Peng Han: "Challenges in Search and Usage of Multi-media Learning Objects"

The second set of papers is concerned with natural language understanding and its use as a query language to access the Web:

- Fang Li, Xuanjing Huang: "An Intelligent Platform for Information Retrieval"
- Weining Qian, Feibo Chen, Bei Du, Aoying Zhou: "P-Terse: A Peer-to-Peer Based Text Retrieval and Search System"
- Tianfang Yao, Hans Uszkoreit: "Identifying Semantic Relations between Named Entities from Chinese Texts"
- Yueijie Zhang, Tao Zhang: "Research on English-Chinese Bi-directional Cross-Language Information Retrieval"

The final set of papers addresses multimodal access as well as multimodal content of picture (video) and textual information:

- Yi Yi Huang, Cun Lu Xu, Yan Qiu Chen: "Analyzing Image Texture from Blobs Perspective"
- Dietrich Klakow: "Access to Content"
- Hong Lu, Xiangyang Xue, Yap-Peng Tan: "Content-Based Image and Video Indexing and Retrieval"
- Huixuan Tang, Hui Wei: "Shape Recognition with Coarse-to-Fine Point Correspondence under Image Deformations"
- Keping Zhao, Shuigeng Zhou, Aoying Zhou: "Towards Efficient Ranked Query Processing in Peer-to-Peer Networks"

Each submission was reviewed by an internal reviewer, by an external reviewer and finally by the editors of this issue, who made the final decision with regard to acceptability and possible revision. We wish to thank Ann de Veire for the fine job she did in preparing these proceedings.

December 2006

Ruqian Lu
Jörg Siekmann
Carsten Ullrich

Table of Contents

Cognitive Systems

Natural Language Dialog with a Tutor System for Mathematical Proofs*

Christoph Benzmüller[1], Helmut Horacek[1], Ivana Kruijff-Korbayová[2], Manfred Pinkal[2], Jörg Siekmann[1], and Magdalena Wolska[2]

[1]Computer Science, Saarland University, Saarbrücken, Germany
[2]Computational Linguistics, Saarland University, Saarbrücken, Germany

Abstract. Natural language interaction between a student and a tutoring or an assistance system for mathematics is a new multi-disciplinary challenge that requires the interaction of (i) advanced natural language processing, (ii) flexible tutorial dialog strategies including hints, and (iii) mathematical domain reasoning. This paper provides an overview on the current research in the multi-disciplinary research project DIALOG, whose goal is to build a prototype dialog-enabled system for teaching to do mathematical proofs. We present the crucial sub-systems in our architecture: the input understanding component and the domain reasoner. We present an interpretation method for mixed-language input consisting of informal and imprecise verbalization of mathematical content, and a proof manager that supports assertion-level automated theorem proving that is a crucial part of our domain reasoning module. Finally, we briefly report on an implementation of a demo system.

1 Introduction

The goal of the DIALOG project is to develop a conversational tutoring system helping students to construct proofs of mathematical theorems. Empirical evidence shows that collaborative problem solving, question answering, and error correction are among the most prominent features of naturalistic one-to-one tutoring and that efficient tutoring exhibits certain dialog patterns characteristic of these collaborative processes [16]. In our project, we aim at a flexible tutorial dialog in which students interact with the system by proposing proof steps using an unconstrained mixture of natural language and mathematical symbols, and the system responds with pedagogically plausible and effective feedback and guidance toward the solution.

Since little is known about the use of natural language in student–tutor dialogs about proofs, we conducted two data–collection experiments. Students with varying educational backgrounds and little to fair prior mathematical knowledge

* This work has been funded by the Deutsche Forschungsgemeinschaft (DFG) Collaborative Research Center on *Resource-Adaptive Cognitive Processes*, SFB 378 (`http://www.coli.uni-saarland.de/projects/sfb378/`).

R. Lu et al. (Eds.): Chinese-German Ws on Cognitive Systems 2005, LNAI 4429, pp. 1–14, 2007.
© Springer-Verlag Berlin Heidelberg 2007

solved proofs in naive set theory[1] (e.g., $K((A \cup B) \cap (C \cup D)) = (K(A) \cap K(B)) \cup (K(C) \cap K(D))$) and binary relations (e.g., $(R \circ S)^{-1} = S^{-1} \circ R^{-1}$) with the help of a system simulated in a Wizard-of-Oz setup. Mathematics teachers were hired as tutors. Both students and tutors were allowed to formulate turns using natural language (German) typed on the keyboard and/or mathematical symbols available on a graphical interface. The students were instructed to enter proof steps, rather than complete proofs, to encourage a dialogue with the system.

More details on the setup of the first experiment can be found in [5] and the second in [8]. [32] and [9, 33] present the first and the second corpus respectively. Fig. 1 shows an excerpt from a typical session.[2]

The experience gained in the Wizard-of-Oz experiments and the analysis of the collected data let us identify three major research topics:

A. Interpretation of informal input mixing natural language and formal expressions. The mathematical content in the dialog utterances is (i) verbalized to varying degree, resulting in *a mixture of natural language and mathematical expressions*, and (ii) presented *informally* and *imprecisely*. These characteristics affect input analysis at the sentence-level, the discourse-level as well as at the level of domain interpretation. The language phenomena are by themselves not new, but the genre of an informal mathematical dialog adds new twists to them [6, 18]. The mixed language and the imprecision call for deep syntactic and semantic analysis to ensure a correct mapping of the surface input to the formal representation of the proof step.

B. Evaluation of proof steps. In order to evaluate a proof contribution, a domain reasoner that supports the tutoring process must be capable of judging contextual appropriateness of proof-steps. We identified the following appropriateness criteria pertinent to tutoring:

Soundness: Can the proof step be reconstructed by a formal inference system and logically and pedagogically verified?
Granularity: Is the 'argumentative complexity' or 'size' of the proof step logically and pedagogically acceptable?
Relevance: Is the proof step logically and pedagogically useful given the goal?

Existing Mathematical Assistant systems, such as ΩMEGA [28], are capable of interactive proof search and correctness verification. Soundness is a fairly tractable criterion for which different techniques are readily available [34]. However, proof step evaluation with respect to *granularity* and *relevance* is a novel and interesting application challenge for theorem proving systems.

C. Handling underspecification. A typical characteristic of proofs constructed by students is content underspecification and ambiguity. Interestingly, underspecificationalso occurs in shaped-up textbook proofs [34]. To illustrate

[1] K denotes the set complement.
[2] Here and further **T** and **S** denote tutor and student turns respectively.

T1: Bitte zeigen Sie: $A \cap B \in P((A \cup C) \cap (B \cup C))$!
Please show $A \cap B \in P((A \cup C) \cap (B \cup C))$ *!*
S1: $(A \cup C) \cap (B \cup C)) = C$
T2: Das ist falsch. Nach der Distributivität gilt: $(A \cup C) \cap (B \cup C) = (A \cap B) \cup C$. Somit gilt auch: $P((A \cup C) \cap (B \cup C)) = P((A \cap B) \cup C)$. Wie geht es nun weiter?
[This is wrong. By distributivity we have: $(A \cup C) \cap (B \cup C) = (A \cap B) \cup C$. *Therefore, we also have* $P((A \cup C) \cap (B \cup C)) = P((A \cap B) \cup C)$. *How to continue?]*
S2: $(A \cap B)$ muss in $P((A \cup C) \cap (B \cup C))$ sein, da $(A \cap B) \in (A \cap B) \cup C$
[$(A \cap B)$ *must be in* $P((A \cup C) \cap (B \cup C))$ *since* $(A \cap B) \in (A \cap B) \cup C$*]*
T3: Was meinen Sie mit: $(A \cap B)$ muss in $P((A \cup C) \cap (B \cup C))$ sein?
[What do you mean by: $(A \cap B)$ *must be in* $P((A \cup C) \cap (B \cup C))$*?]*
S3: da $P((A \cup C) \cap (B \cup C))$ nach Distributivgesetz gleich $P((A \cap B) \cup C)$ ist muss $(A \cap B)$ in $P((A \cup C) \cap (B \cup C))$ enthalten sein.
[since $P((A \cup C) \cap (B \cup C))$ *equals by distibutivity* $P((A \cap B) \cup C)$ $(A \cap B)$ *must be contained in* $P((A \cup C) \cap (B \cup C))$ *]*
T4: Was meinen Sie mit: muss $(A \cap B)$ in $P((A \cup C) \cap (B \cup C))$ enthalten sein. Meinen Sie Element oder Teilmenge?
[What do you mean by: $(A \cap B)$ *must be in* $P((A \cup C) \cap (B \cup C))$*. Do you mean as element or subset?]*
S4: $(A \cap B) \in P((A \cap B) \cup C)$
T5: Meinten Sie vorhin wirklich: $(A \cap B) \in (A \cap B) \cup C$?
[Do you really mean $(A \cap B) \in (A \cap B) \cup C$*?]*
S5: ja
[yes]
T6: Das ist falsch, denn $(A \cap B) \subseteq (A \cap B) \cup C$. $(A \cap B)$ und $(A \cap B) \cup C$ sind Mengen und keine Elemente voneinander. Was folgt als nächstes? ...
[This is wrong because $(A \cap B) \subseteq (A \cap B) \cup C$. $(A \cap B)$ *and* $(A \cap B) \cup C$ *are sets and not elements of each other. What comes next? ...]*

Fig. 1. An example dialog. P denotes the powerset.

proof-step underspecification let us consider the following excerpt from the first corpus:

T: Please show : $K((A \cup B) \cap (C \cup D)) = (K(A) \cap K(B)) \cup (K(C) \cap K(D))$
S: by the deMorgan rule $K((A \cup B) \cap (C \cup D)) = (K(A \cup B) \cup K(C \cup D))$ holds.

From the point of view of linguistic analysis, **S** is unambiguous. However, the proof-step that the utterance expresses is highly underspecified from a proof construction viewpoint: it is neither mentioned how the assertion is related to the target formula, nor how and which deMorgan rule was used. **S** can be obtained directly from the second deMorgan rule $\forall X, Y. K(X \cap Y) = K(X) \cup K(Y)$ by instantiating X with $(A \cup B)$ and Y with $(C \cup D)$. Alternatively, it could be inferred from **T** by applying the first deMorgan rule $\forall X, Y. K(X \cup Y) = K(X) \cap K(Y)$

from right to left to the subterms $K(A) \cap K(B)$ and $K(C) \cap K(D)$. Proof assistant systems, typically require such detailed specification to execute the user's proof-step directive. Differentiating between proof construction alternatives can be important from the tutoring perspective.

Based on the empirical findings, we implemented a prototype system that can handle variations on the dialog in Fig. 1 and several other dialogs from our corpora. The demo system consists of a graphical user interface, an input analyzer, a proof manager, a tutorial manager, and a natural language generator. The modules are connected and controlled by an Information State Update-based dialogue manager [29]. Our research and, in particular, the implementation focus has been mainly on the *input analyzer*, whose task is to interpret and formally represent the linguistic content of the student's dialog contributions, and the *proof manager*, whose task is to evaluate the student proof step proposals with the help of a domain reasoner: the automated theorem prover ΩMEGA. For the other modules we provided baseline functionality required to carry out the dialogs. More details on the DIALOG demo system can be found in [10].

The remainder of this paper is organized as follows: In Sections 2 and 3 we describe our approach to mixed language interpretation and proof step evaluation, respectively. In Section 4, we overview the related work. In Section 5, we summarize and present the conclusions.

2 Interpreting Informal Mathematical Discourse

For student utterances that contain proof-relevant parts, such as **S2** in Fig. 1, the task of input interpretation is to identify these and represent them in a format interpretable by a domain reasoner. To ensure correct mapping to this representation, deep analysis is needed. It is further justified by the varying degrees of mathematical content verbalization and imprecision, common in the informal mathematical discourse in our corpus, as well as the need for consistency of interpretation required for proof-step evaluation by the domain reasoner.

In this section, we present an overview of our input interpretation procedure; we omit obvious pre-processing such as sentence- and word-tokenization. We first describe three basic components that provide minimal functionality required to analyze simple cases of mixed language. Then we discuss extensions for some of the more complex phenomena. For a more detailed discussion of language phenomena and interpretation procedure see [30, 31, 18, 19]

2.1 Baseline Processing

A simple utterance consisting of a mixture of mathematical expressions and natural language is the utterance **S** presented in Section 1: "by the deMorgan rule $K((A \cup B) \cap (C \cup D)) = (K(A \cup B) \cup K(C \cup D))$ holds.". We shall use it to illustrate the step-wise analysis process that proceeds as follows: we first identify mathematical expressions in order to encapsulate them before syntactic parsing.

During syntactic parsing, a domain-independent semantic representation is constructed. This we then refine to obtain a domain-specific interpretation suitable as input to the domain reasoner.

Mathematical expression parsing. To recognize and parse mathematical expressions we use knowledge about operators and identifiers the domain and relevant for the given problem, e.g., $\cup, \cap, K$. For the purpose of subsequent syntactic and semantic parsing, each mathematical expression is assigned a symbolic token corresponding to its type.

In our example, the expression $K((A \cup B) \cap (C \cup D)) = (K(A \cup B) \cup K(C \cup D))$ is assigned the type `FORMULA`. The token representing the expression type is substituted for the original expression resulting in the following input to the parser: "by the deMorgan rule `FORMULA` holds."

Syntactic parsing and semantic interpretation. The parser processes sentences and syntactically well-formed fragments, and constructs a representation of their *linguistic meaning* (LM). The LM is represented as a relational dependency structure closely corresponding to the *tectogrammatical level* in [26].

To obtain the LM, we use the OpenCCG parser (`openccg.sourceforge.net`) for which we develop a lexically-based Combinatory Categorial Grammar for German [13]. Our motivation for using this grammar framework is two-fold: first, it is well known for its account of coordination phenomena, widely present in mathematical discourse. Second, mathematical expressions, represented by their types, lend themselves to a perspicuous categorial treatment as follows: In the course of parsing, we treat symbolic tokens representing mathematical expressions on a par with lexical units. The parser's lexicon encodes entries for each mathematical expression type represented by its token (*e.g.* `TERM`, `FORMULA`) together with the syntactic categories the expression may take (*e.g.* the category of a noun phrase, `np`, for `TERM`, the category of a sentence, `s`, for `FORMULA`). By designing one grammar that allows a uniform treatment of the linguistic content and the mathematical content, we aim at a consistent analysis of different degrees of mathematical content verbalization.

Domain interpretation. The LM representations built by the parser are domain-independent. To obtain domain-specific interpretations, we implemented a stepwise procedure that gradually assigns domain-specific semantics to predicates and relations in the LM.

As a first step, we use a *s*emantic lexicon to map (parts of) the LM representations to domain-independent conceptual frames. The input structures are described in terms of tectogrammatical valency frames of lexemes that evoke a given concept. The output structures are the evoked concepts with roles indexed by tectogrammatical frame elements. Where relevant, sortal information for role fillers is given. For example, the `Norm` tectogrammatical relation (TR) evokes the concept of a *Rule*. The dependent in the `Norm`-relation identifies the rule according to which the main proposition holds.

As a second step, semantic lexicon concepts are mapped to domain-specific interpretations using a *domain ontology*. The ontology is an intermediate

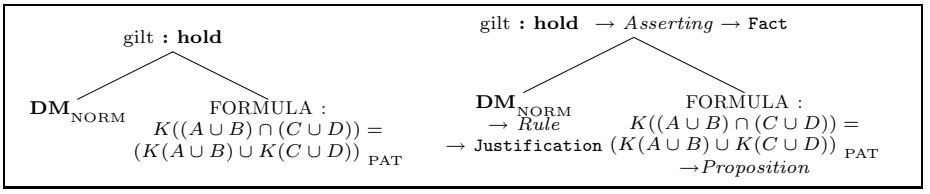


Fig. 2. Interpretation of the utterance "by the deMorgan rule $K((A \cup B) \cap (C \cup D)) = (K(A \cup B) \cup K(C \cup D))$ holds."; DM stands for DeMorgan

representation that mediates between the discrepant views of linguistic analysis and deduction systems' representation [17]; it thus has a potential of providing a direct link to logical definitions in a *mathematical knowledge base*, such as MBase [21]. The motivation for using an intermediate representation instead of directly accessing a mathematical knowledge base will become clear when we discuss ambiguity in Section 2.2.

Let us return to the example utterance: In Fig. 2 on the left, we show the representation of its linguistic meaning built by the parser. The structure consists of the German verb, "gilt", with the symbolic meaning **hold**, as the head, and two dependents in the TRs: Norm and Patient. The right part of Fig. 2 shows the assignment of domain-specific meaning: First, based on the semantic lexicon, the concept *Assertion* is assigned to **hold**, with Patient and Norm dependents as the *Proposition* and *Rule* respectively. Next, *Assertion* is interpreted as the **Fact** and the *Rule* as **Justification** in a proof-step. Applying re-writing transformations, we obtain the following underspecified representation used by the domain reasoner [4]: "**Fact** $K((A \cup B) \cap (C \cup D)) = (K(A \cup B) \cup K(C \cup D))$ **By** DeMorgan-1 **From .**".

The baseline processing described so far covers simple cases of the mixed language: it suffices to interpret utterances where terms or complete formulas are embedded within natural language parts. However, our corpus contains more complex cases of interleaved mathematical expressions and natural language. We turn to their processing in the next section.

2.2 Domain- and Context-Specific Processing

Our corpus contains a range of more complex phenomena typical for informal mathematical discourse, which the baseline processing described above cannot handle. In this section, we describe several extensions that we have implemented so far: parsing extensions to handle tightly interleaving mathematical expressions and natural language, and domain model extensions to handle ambiguous and imprecise formulations.

Parsing extensions. Input utterances often contain incomplete mathematical formulas interleaved with natural language expressions, where these two modes interact, *e.g.* (1) and (2) below. To handle these cases we made the following extensions: (i) the mathematical expression parser recovers information about

incomplete formulas using domain-knowledge of syntax and semantics of formal expressions, and (ii) the syntactic parser's lexicon contains the corresponding categories and their semantics.

For example, in (1), the parser recognizes the operator $\in$ as requiring two arguments: one of type *inhabitant* and the other *set*.

(1) $A \cap B$ ist $\in$ von $C \cup (A \cap B)$
$A \cap B$ *is* $\in$ *of* $C \cup (A \cap B)$

Accordingly, $\in$ is assigned a symbolic type `0_FORMULA_0`, where `0` indicates the arguments missing in the left and the right context. We have included a lexical entry `0_FORMULA_0` of syntactic category $\texttt{s/pp}_{lex:von}\texttt{\textbackslash np}$ in the lexicon of the syntactic parser.

Example (2) illustrates a case of tight interaction between mathematical expressions and the surrounding natural language:

(2) B enthaelt kein $x \in A$
B *contains no* $x \in A$

Here, the negation word "kein" has x, i.e. part of the expression $x \in A$, in its scope. The intended meaning of (2) can be paraphrased as B *contains no* x *such that* $x \in A$.

To account for this interaction we identify substructures of mathematical expressions that may lie in the scope of a natural language expression. In (2), the expression $x \in A$ is of type `FORMULA`. The relevant substructures are obtained by splitting the formula at the top node. As a result, we obtain two readings of the expression: $\texttt{TERM}_1$ $\texttt{0\_FORMULA}_1$ and $\texttt{FORMULA\_0}_2$ $\texttt{TERM}_2$ and parse the utterance with them. The lexical entry for `0_FORMULA` (formula with a missing left argument) is of syntactic category `s\np` (and semantics corresponding to `such that TERM has property FORMULA`), while the entry for `FORMULA_0` (formula with a missing right argument) is of category `s/np`. This allows us to obtain the intended reading of (2).

Domain modeling extensions. Another common phenomenon in informal mathematical discourse is ambiguity and/or imprecision. For example, the verb "enthalten" in the previously mentioned example (2), can in principle mean a subset, a membership, or a substring relation. To handle such cases, we extended the domain model. We added *semantic relations* that represent general concepts that subsume the specific mathematical relations. For example, the verb "enthalten" (*contain*) evokes either the concept of CONTAINMENT or that of STRUCTURAL COMPOSITION. CONTAINMENT in its most common interpretation specializes to the domain relations of (strict) SUBSET or ELEMENT with two roles: CONTAINER and CONTENTS. These are filled by the fillers of the TRs *Actor* (*act*) and *Patient* (*pat*) in the tectogrammatical valency frame of "enthalten", respectively. The STRUCTURAL COMPOSITION relation holds between a STRUCTURED OBJECT and its structural sub-component in the SUBSTRUCTURE role. Similarly, these roles are filled by the *Actor* and the *Patient* dependent, respectively. The semantic

$(\textbf{contain}_{pred}, \mathrm{x}_{act}, \mathrm{y}_{pat}) \rightarrow (\textsc{containment}_{pred}, \mathrm{container}_{act}, \mathrm{contents}_{pat})$ (a)
$(\textbf{contain}_{pred}, \mathrm{x}_{act:formula}, \mathrm{y}_{pat:formula}) \rightarrow$
$(\textsc{structural composition}_{pred}, \mathrm{structured\ object}_{act}, \mathrm{substructure}_{pat})$ (b)

Fig. 3. Example entries from the semantic lexicon

lexicon entry of the verb "enthalten" (*contain*) with the mappings of the concept roles described above is shown in Fig. 3. The domain ontology and semantic lexicon are presented in more detail in [31, 19].

We have so far concentrated on a proof-of-concept implementation of the input interpretation components. Our interpretation module is capable of producing analyzes of utterances similar to the ones in Fig. 1 as well as several variants thereof. We have implemented an OpenCCG grammar that covers variants of the syntactic structures of the utterances. We have also manually encoded the relevant part of the domain ontology required to account for the domain-specific interpretation. Our focus so far has not been on robust coverage, but rather on a systematic consistent representation of the most frequent constructions in the format readable by the domain reasoner.

3 Mathematical Domain Reasoning

Determining whether a proof step is appropriate requires that a mathematical domain reasoner can represent, reconstruct, and validate the proof step (including all the justifications used by the student) within its representation of the current proof.

Consider, for instance, utterance (a) in Fig. 4: Verification of the soundness of this utterance boils down to adding D as a new assertion to the proof state and to proving that: **(P1)** $(A \wedge B), (A \Rightarrow C), (C \Rightarrow D), (F \Rightarrow B) \vdash D$. Solving this proof task confirms the logical soundness of utterance (a). If further explicit justifications are provided in the student's utterance (e.g. a proof rule) then we have to take them into consideration and, for example, prove **(P1)** modulo these additional constraints.

Soundness evaluation can be supported by different methods — including some that avoid dynamic theorem proving system. On one extreme "gold-standard" proofs could be selected and the proposed partial proofs could be matched against them. The other extreme would be to interpret the problem as a challenge to proof theory and try to develop and implement a proper proof theoretic approach to differentiate between 'pedagogically good' proofs and proof steps and pedagogically 'less acceptable' proofs and proof steps in the space of all proofs for a given problem. A notion of a 'good proof' is, for instance, presented in [11].

Soundness is, however, only one of the criteria along which a proof step should be evaluated in a tutorial context. For instance, a proof step may be formally relevant in purely logical terms, but considered irrelevant when additional tutorial aspects are taken into account. This is, for instance, the case when the goal

Proof State	**Example Student Utterances**
(A1) $A \wedge B$.	(a) From the assertions follows D.
(A2) $A \Rightarrow C$.	(b) B holds.
(A3) $C \Rightarrow D$.	(c) It is sufficient to show D.
(A4) $F \Rightarrow B$.	(d) We show E.
(G) $D \vee E$.	

Fig. 4. Proof step evaluation scenario: (A1)-(A4) are assertions that have been introduced in the discourse and that are available to prove the goal (G). (a)-(d) are typical examples of proof step proposed by students.

of the session is to teach a particular proof technique. Well known examples are proof by induction on the naturals, proof by structural induction, and proof by diagonalization. Often different proof techniques are applicable to one and the same problem and this causes pedagogically different, but formally correct and relevant, proof directions.

On the other hand, a step that is sufficiently close to a valid proof step may be considered pedagogically relevant while being logically irrelevant. Proof step evaluation should therefore support dynamic step-by-step analysis of the proof constructed by the student using the criteria of soundness, granularity and relevance not only with respect to a purely logical dimension, but also a tutorial dimension.

So far we have mainly focused on the logical dimension; the hypothesis is that the solution in the logical dimension is a prerequisite for solving the proof step evaluation problem involving also the tutorial dimension. Much further research in this direction is clearly needed.

In the following sections, we discuss some of the issues related to evaluating *granularity* and *relevance*. We illustrate the challenges using a constructed example in Fig. 4. See also [7].

3.1 Granularity

Granularity judgment refers to analysis of the 'complexity' or 'size' of proofs instead of the mere existence of proofs. For instance, evaluation of (a) boils down to judging the complexity of the generated proof task **(P1)**.

Let us, for example, use Gentzen's natural deduction (ND) calculus [12] as the proof system $\vdash$. As a first and naive logical granularity measure, we may determine the number of $\vdash$-steps in the smallest $\vdash$-proof of the proof task for the proof step utterance in question; this number is taken as the argumentative complexity of the uttered proof step. For example, the smallest ND proof for utterance (a) requires three proof steps: we need one 'Conjunction-Elimination' step to extract A from $A \wedge B$, one 'Modus Ponens' step to obtain C from A and $A \Rightarrow C$, and another 'Modus Ponens' step to obtain D from C and $C \Rightarrow D$. On the other hand, the smallest ND proof for utterance (b) requires only '1' step: B follows from assertion $A \wedge B$ by 'Conjunction-Elimination'. If we now

fix a threshold that tries to capture, in this sense, the 'maximally acceptable size' of a single argument, then we can distinguish between proof steps whose granularity is acceptable and those which are not. This threshold may be treated as a parameter determined by the tutorial setting.

However, the ND calculus together with naive proof step counting does not always provide a cognitively adequate basis for granularity analysis. The reason is that two intuitively very similar student proof steps (such as **(i)** from $A = B$ and $B = C$ infer $A = C$ and **(ii)** from $A \Leftrightarrow B$ and $B \Leftrightarrow C$ infer $A \Leftrightarrow C$) may actually expand into base-level ND proofs of completely different size. Related research has shown that the standard ND calculus does not adequately reflect human-reasoning in this respect [24]. Two important and cognitively interesting questions thus concern the appropriate choice of a proof system $\vdash$ and ways to measure the 'argumentative complexity' of an admissible proof step.

3.2 Relevance

Relevance is about usefulness and importance of a proof step with respect to the given proof task. For instance, in utterance (c) the proof goal $D \vee E$ is refined to a new goal D using backward reasoning, i.e., the previously open goal $D \vee E$ is closed and justified by a new goal.

Solving logical relevance problem requires in this case checking whether a proof can still be generated in the new proof situation. In this case, the task is thus identical to **(P1)**. An irrelevant backward proof step, according to this criterion, is (d) since it reduces to the proof task: **(P2)** $(A \wedge B), (A \Rightarrow C), (C \Rightarrow D), (F \Rightarrow B) \vdash E$ for which no proof can be generated. Thus, (d) is a sound refinement step that is however irrelevant. This simple approach appears plausible, but needs to be refined. The challenge is to exclude detours and to take tutorial aspects into account (in a tutorial setting we are often interested in teaching particular styles of proofs, particular proof methods, etc.). This also applies to the more challenging forward reasoning case where for instance, utterance (b) should be identified as an irrelevant step.

4 Related Work

Input analysis in dialog systems is commonly done with shallow syntactic analysis combined with keyword spotting where short answers may be elicited by asking closed-questions [14]. Slot-filling templates, however, are not suitable representations of the content in our domain. Moreover, the interleaving of natural and symbolic language makes keyword spotting difficult because of the variety of possible verbalizations.

Statistical methods are often employed in tutorial systems to compare student responses with pre-constructed gold-standard answers [15]. In our context, such a static modeling solution is impossible because of the wide quantitative and qualitative range of acceptable proofs, i.e., generally, our set of gold-standard answers is infinite. In this respect our approach is closely related to the Why2-Atlas tutoring system [22]. This system presents students with qualitative physics

questions and encourages them to explain their answers with natural language. Different to our approach is that the students first present a complete essay of their answer to the system. The system then employs propositional logic representations and propositional abductive reasoning to analyze the answer with respect to a set of anticipated solutions. The analysis results are then used to create a dialog in which misconceptions in the students essay are addressed. In our scenario propositional logic appears insufficient and we employ first-order and higher-order representations and reasoning techniques. Similar to our scenario, however, is the problem of multiple proof alternatives that have to considered in the analysis tasks.

To analyze human-oriented mathematical proofs, shaped-up textbook proofs have been investigated in the deduction systems community (see [34, 27]). Claus Zinn [34], for instance, addresses complete, carefully structured textbook proofs, and relies on given text-structure, typesetting and additional information that identifies mathematical symbols, formulas, and proof steps. The DIALOG corpus provides an important alternative view, since textbook proofs neither reveal the dynamics of proof construction nor do they show the actual student's utterances, i.e., the student's proof step directives. Our corpus also illustrates the style and logical granularity of human-constructed proofs. The style is mainly declarative, for example, the students declaratively described the conclusions and some (or none) of the premises of their inferences. By contrast, many proof assistants employ a procedural style in which proof steps are invoked by calling rules, tactics, or methods, i.e., some proof refinement procedures.

Recent research into dialog modeling has delivered a variety of approaches more or less suitable for the tutorial dialog setting. For instance, scripting is employed in AutoTutor [23] and a knowledge-based approach similar to ours is implemented in the Geometry Tutor [1, 2]. Outside the tutorial domain, the framework of Information State Update (ISU) has been developed in the EU projects TRINDI (`http://www.ling.gu.se/research/projects/trindi/`) and SIRIDUS (`http://www.ling.gu.se/projekt/siridus/`) [29], and applied in various projects targeting flexible dialog. An ISU-based approach with several layers of planning is used in the tutorial dialog system BEETLE [35].

5 Conclusion

We presented our approach to interpreting informal mathematical discourse in the context of tutorial dialogue and to evaluating the proof steps proposed by the student by a back-end domain reasoning component. We employ a stratified approach to interpreting the mixed natural- and mathematical language; we first developed methods for basic cases, and enhanced them by techniques for handling additional levels of complexity. Our interpretation method has the potential of putting a tutorial system in a good position to apply strategies for enhancing higher-level problem-solving skills of the student.

We have identified two previously unconsidered aspects of proof-step evaluation: relevance and granularity of proof-steps, that are important from the

tutoring point of view. To address these criteria, it is not sufficient to merely establish the existence of proofs. The system has to construct proofs with particular properties. It may be the case that evaluating different criteria requires different theorem provers. Moreover, the system also needs to closely mirror and reflect reasoning steps as they are typically preferred by humans. Generally, the system will need to adapt to the capabilities of individual students and the requirements of varying tutorial settings.

We have implemented an input interpretation component capable of representing student utterances consisting of a mixture of natural language and mathematical expressions, in a format that is interpretable by an automated reasoning engine. The application of our approach to mathematical domains that are more challenging than naive set theory, and its evaluation therein is ongoing work. The hypothesis that assertion level reasoning [20] plays an essential role in evaluating appropriateness of underspecified partial proofs has been confirmed. The fact that assertion level reasoning may be highly underspecified in human-constructed proofs is a novel finding of our project (cf. [4]).

The implemented proof manager demonstrator helps to resolve underspecification and to evaluate proof steps based on heuristically guided abstract-level domain reasoning realized of the ΩMEGA-CORE framework [3]. The PM has been integrated also into the overall DIALOG system to communicate with the other components of the system.

The evaluation of the system so far concentrates mainly on individual analysis of specific aspects of single modules. One example presented in [25] is the evaluation of mechanized granularity judgments of proof steps, using deductive techniques based on natural deduction calculus and the PSYCOP approach [24].

References

1. V. Aleven, O. Popescu, and K. Koedinger. Towards tutorial dialog to support self-explanation: Adding natural language understanding to a cognitive tutor. In *Proceedings of the 10th International Conference on Artificial Intelligence and Education (AIED-01)*, pages 246–255, 2001.
2. V. Aleven, O. Popescu, and K. Koedinger. Pilot-testing a tutorial dialogue system that supports self-explanation. In *Proceedings of 6th International Conference on Intelligent Tutoring Systems (ITS-02)*, pages 344–354. Springer Verlag, 2002.
3. S. Autexier. *Hierarchical Contextual Reasoning.* PhD thesis, Saarland University, Germany, 2003.
4. S. Autexier, C. Benzmüller, A. Fiedler, H. Horacek, and B.Q. Vo. Assertion-level proof representation with under-specification. *Electronic Notes in Theoretical Computer Science*, 93:5–23, 2003.
5. C. Benzmüller, A. Fiedler, M. Gabsdil, H. Horacek, I. Kruijff-Korbayová, M. Pinkal, J. Siekmann, D. Tsovaltzi, B.Q. Vo, and M. Wolska. A Wizard-of-Oz experiment for tutorial dialogues in mathematics. In *AIED2003 — Supplementary Proceedings of the 11th International Conference on Artificial Intelligence in Education*, pages 471–481, Sydney, Australia, 2003.

6. C. Benzmüller, A. Fiedler, M. Gabsdil, H. Horacek, I. Kruijff-Korbayová, D. Tsovaltzi, B.Q. Vo, and M. Wolska. Language phenomena in tutorial dialogs on mathematical proofs. In *Proceedings of DiaBruck, the 7th Workshop on the Semantics and Pragmatics of Dialogue*, pages 165–166, Saarbrücken, Germany, 2003.
7. C.E. Benzmüller and Q.B. Vo. Mathematical domain reasoning tasks in natural language tutorial dialog on proofs. In M. Veloso and S. Kambhampati, editors, *Proceedings of the 20th National Conference on Artificial Intelligence (AAAI-05)*, pages 516–522, Pittsburgh, PA, 2005.
8. C. Benzmüller, H. Horacek, I. Kruijff-Korbayová, H. Lesourd, M. Schiller, and M. Wolska. DiaWozII – A Tool for Wizard-of-Oz Experiments in Mathematics. In *Proceedings of the 29th Annual German Conference on Artificial Intelligence (KI-06), Lecture Notes in Computer Science*, Bremen, Germany, 2006. Springer-Verlag. To Appear.
9. C. Benzmüller, H. Horacek, H. Lesourd, I. Kruijff-Korbayová, M. Schiller, and M. Wolska. A corpus of tutorial dialogs on theorem proving; the influence of the presentation of the study-material. In *Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC-06)*, pages 1766–1769, Genoa, Italy, 2006. ELDA.
10. M. Buckley and C. Benzmüller. System description: A dialog manager supporting tutorial natural language dialogue on proofs. In *Proceedings of the ETAPS Satellite Workshop on User Interfaces for Theorem Provers*, Edinburgh, Scotland, 2005.
11. N. Dershowitz and C. Kirchner. Abstract saturation-based inference. In Phokion Kolaitis, editor, *Proceedings of the 18th Annual IEEE Symposium on Logic in Computer Science (LICS-03)*, Ottawa, Ontario, June 2003. IEEE.
12. G. Gentzen. Untersuchungen über das logische Schließen I & II. *Mathematische Zeitschrift*, 39:176–210, 572–595, 1935.
13. C. Gerstenberger and M. Wolska. Introducing Topological Field Information into CCG. In *Proceedings of the 17th European Summer School in Logic, Language and Information (ESSLLI-05) Student Session*, pages 62–74, Edinburgh, Scotland, 2005.
14. M. Glass. Processing language input in the CIRCSIM-tutor intelligent tutoring system. In *Proceedings of the 10th International Conference on Artificial Intelligence and Education in Education (AIED-01), San Antonio, Texas*, pages 210–221, 2001.
15. A. Graesser, P. Wiemer-Hastings, K. Wiemer-Hastings, D. Harter, and N. Person. Using Latent Semantic Analysis to evaluate the contributions of students in AutoTutor. *Interactive Learning Environments*, 8, 2000.
16. A. C. Graesser, N. K. Person, and J. P. Magliano. Collaborative dialogue patterns in naturalistic one-to-one tutoring. *Applied Cognitive Psychology*, 9:1–28, 1995.
17. H. Horacek, A. Fiedler, A. Franke, M. Moschner, M. Pollet, and V. Sorge. Representation of mathematical objects for inferencing and for presentation purposes. In *Proceedings of the 17th European Meetings on Cybernetics and Systems Research (EMCSR-04), Vienna, Austria*, pages 683–688, 2004.
18. H. Horacek and M. Wolska. Interpretation of mixed language input in a mathematics tutoring system. In *Proceedings of the AIED-05 Workshop on Mixed Language Explanations in Learning Environments*, pages 27–34, Amsterdam, the Netherlands, 2005.
19. H. Horacek and M. Wolska. Interpreting semi-formal utterances in dialogs about mathematical proofs. *Data and Knowledge Engineering Journal*, 58(1):90–106, 2006.

20. X. Huang. Reconstructing Proofs at the Assertion Level. In A. Bundy, editor, *Proceedings of the 12th Conference on Automated Deduction*, number 814 in LNAI, pages 738–752. Springer, 1994.
21. M. Kohlhase and A. Franke. MBase: Representing knowledge and context for the integration of mathematical software systems. *Journal of Symbolic Computation*, 32(4), 2001.
22. M. Makatchev, P. W. Jordan, and K. VanLehn. Abductive theorem proving for analyzing student explanations to guide feedback in intelligent tutoring systems. *Journal of Automated Reasoning*, 32(3):187–226, 2004.
23. N. K. Person, A. C. Graesser, D. Harter, E. Mathews, and the Tutoring Research Group. Dialog move generation and conversation management in AutoTutor. In Carolyn Penstein Rosé and Reva Freedman, editors, *Building Dialog Systems for Tutorial Applications—Papers from the AAAI Fall Symposium*, pages 45–51, North Falmouth, MA, 2000. AAAI press.
24. L. J. Rips. *The psychology of proof*. MIT Press, Cambridge, Mass., 1994.
25. M. Schiller. Mechanizing proof step evaluation for mathematics tutoring – the case of granularity. Master's thesis, Saarland University, 2005.
26. P. Sgall, E. Hajičová, and J. Panevová. *The meaning of the sentence in its semantic and pragmatic aspects*. Reidel Publishing Company, Dordrecht, The Netherlands, 1986.
27. J. Siekmann. Proof presentation. In *Proof Presentation*. Elsevier, To Appear.
28. J. Siekmann, C. Benzmüller, A. Fiedler, A. Meier, I. Normann, and M. Pollet. *Proof Development in OMEGA: The Irrationality of Square Root of 2*, pages 271–314. Kluwer Applied Logic series (28). Kluwer Academic Publishers, 2003.
29. D. R. Traum and S. Larsson. The information state approach to dialogue management. In Jan van Kuppevelt and Ronnie Smith, editors, *Current and New Directions in Discourse and Dialogue*. Kluwer, 2003.
30. M. Wolska and I. Kruijff-Korbayová. Analysis of mixed natural and symbolic language input in mathematical dialogs. In *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics (ACL-04)*, Barcelona, Spain, 2004.
31. M. Wolska, I. Kruijff-Korbayová, and H. Horacek. Lexical-semantic interpretation of language input in mathematical dialogs. In *Proceedings of the ACL 2nd Workshop on Text Meaning and Interpretation*, pages 81–88, Barcelona, Spain, 2004.
32. M. Wolska, B.Q. Vo, D. Tsovaltzi, I. Kruijff-Korbayová, E. Karagjosova, H. Horacek, M. Gabsdil, A. Fiedler, and C. Benzmüller. An annotated corpus of tutorial dialogs on mathematical theorem proving. In *Proceedings of the 4th International Conference on Language Resources and Evaluation (LREC-04)*, 2004.
33. M. Wolska and I. Kruijff-Korbayová. Factors influencing input styles in tutoring systems: the case of the study-material presentation format. In *Proceedings of the ECAI-06 Workshop on Language-enabled Educational Technology*, To Appear, Riva del Garda, Italy, 2006.
34. C. Zinn. *Understanding Informal Mathematical Discourse*. PhD thesis, University of Erlangen-Nuremberg, 2004.
35. C. Zinn, J.D. Moore, M.G. Core, S. Varges, and K. Porayska-Pomsta. The BE&E tutorial learning environment (BEETLE). In *Proceedings of DiaBruck, the 7th Workshop on the Semantics and Pragmatics of Dialogue*, pages 209-210, Saarbrücken, Germany, 2003.

On the Effectiveness of Visualizations in a Theory of Computing Course

Rudolf Fleischer[1,⋆] and Gerhard Trippen[2]

[1] Fudan University, Shanghai Key Laboratory of Intelligent Information Processing, Department of Computer Science and Engineering, Shanghai, China
rudolf@fudan.edu.cn
[2] The Hong Kong University of Science and Technology, Department of Computer Science, Hong Kong
trippen@cs.ust.hk

Abstract. We report on two tests we performed in Hong Kong and Shanghai to verify the hypothesis that one can learn better when being given access to visualizations beyond the standard verbal explanations in a classroom. The outcome of the first test at HKUST was inconclusive, while the second test at Fudan University showed a clear advantage for those students who had access to visualizations.

1 Introduction

Visualizations of algorithmic concepts are widely believed to enhance learning [27, 29, 36], and quite some effort is put into creating nice visualizations [1, 2, 5, 13, 14, 16, 26] or (semi-)automatic visualization systems [3, 4, 6, 7, 9–11, 17, 25, 31, 32, 35]. However, there is still not much conclusive scientific evidence supporting (or disproving) this hypothesis [20, 28, 34]. In particular, Hundhausen's meta-analysis of 21 experimental evaluations [18] had the disheartening outcome that only about one half of the studies actually showed some positive benefit of using visualization techniques.

To shed more light on this phenomenon, there have been recently several more case studies trying to evaluate certain aspects of the effectiveness of visualizations in teaching. Cooper *et al.* [8] reported about a study using program visualization for introducing objects in an object-oriented programming course. Koldehofe *et al.* [22] used a simulation-visualization environment in a distributed systems course. Korhonen *et al.* [23] studied the effects of immediate feedback in a virtual learning environment. Kuitinen and Sajaniemi [24] evaluated the use of role-based visualization in teaching introductory programming courses. Grimson *et al.* [15] compared different learner engagement levels with visualization (i.e., how

⋆ This research was partially supported by a HKUST Teaching Development Grant CLI (Continuous Learning and Improvement Through Teaching Innovation), `Study on the Learning Effectiveness of Visualizations`, and by a grant from the National Natural Science Fund China (grant no. 60573025).

R. Lu et al. (Eds.): Chinese-German Ws on Cognitive Systems 2005, LNAI 4429, pp. 15–24, 2007.
© Springer-Verlag Berlin Heidelberg 2007

active the learner can interact with the system, only view, or answer questions, or play around with different parameter sets or own algorithms, etc.).

The first author participated in two recent ITiCSE Workshops, on "Exploring the Role of Visualization and Engagement in Computer Science Education" in 2002 [29], and on "Evaluating the Educational Impact of Visualization" in 2003 [28]. While the former one focused on the needs of good visualizations and presented a framework for experimental studies of their effectiveness, the latter one focused on the problem of disseminating good visualization tools and on how to evaluate learner outcomes in courses using visualization techniques.

Following the guidelines on how to evaluate the effectiveness of visualizations set up in these two workshops, we designed a study on the effectiveness of visualizations in a Theory of Computing course that the first author taught for several years in Hong Kong and Shanghai. In Hong Kong, we did the study in Spring 2004 at HKUST (The Hong Kong University of Science and Technology) in the course COMP 272, which is a second-year course in a three-year curriculum. In Shanghai, we did the study in Spring 2005 at Fudan University in the course *Theory of Computing*, which is an optional third-year course for computer science students and a mandatory final-year course for software engineering students. The result of the study is inconclusive. While the students at HKUST did not seem to benefit from the visualizations, there was a considerable improvement in the learning of the students at Fudan.

In Section 2 we explain the details of our study. In Section 3 we present the data collected in the two studies and give some statistical interpretations. We close with some remarks in Section 4.

2 The Study

In this section we will describe the course, the visualizations presented to the students, and how we evaluated their usefulness.

2.1 Theory of Computing: Course Description

At HKUST, COMP 272 is a second-year undergraduate course on automata theory and computability. The syllabus spans finite automata, context-free grammars, Turing machines, and non-computability. NP-completeness is taught in a different algorithms course, COMP 271. Our course followed closely the first six chapters of the textbook by Kinber and Smith [21]. In 2004 we had 99 students in the class. There were three one-hour classes a week, on Monday, Wednesday, and Friday, for a duration of 15 weeks.

We were teaching Comp 272 using the framework of *Just-in-Time Teaching (JiTT)* [19, 30]. The main feature of JiTT is that students are supposed to come to class well prepared, i.e., they are supposed to read all the material beforehand in a textbook or lecture notes. This gives the instructor the freedom to skip basic definitions and easy lemmas and instead focus on the important and difficult parts of the material. The students worked in teams of size three or four.

At Fudan, *Theory of Computing* in Spring 2005 was a combined course for computer science and software engineering students. For the former the course was an optional third-year course and only five students signed up, for the latter the course was a mandatory final-year course and we had 39 future software engineers in the class. Because the software engineering students had already learned about finite automata and context-free grammars in a previous course, they joined our class after the first five weeks in which the computer science students learned these topics. There was one three-hour class per week, for a duration of 15 weeks. The course was taught as a traditional course using the whiteboard (no PowerPoint slides), the textbook was the book by Sipser [33] but we covered the same material as in the course at HKUST.

2.2 Setup of the Study

As teachers, we hope that seeing visualizations of (abstract) definitions and playing around with algorithms will help the students to better understand the course material. In [28] we argued that the usefulness of visualizations in teaching strongly depends on the engagement level of the students. In this study, we used visualizations supporting the engagement levels of *viewing* (seeing definitions or step-by-step explanations of the algorithms) and *responding* (doing step exercises), but not *changing* (running algorithms on own input data).

Equal treatment of all students prohibited us to split the students into a test group and a control group without access to visualizations. Instead, for each single test the students where randomly selected for access to the visualization, and after finishing the test every student could see the visualizations.

The Tests. We prepared four tests by adapting publicly available visualization applets. The tests can be seen at
`ihome.ust.hk/~trippen/TA/COMP272_04/test[1,2,3,4].html`.
The tests only covered the first part of the course about finite automata and a little bit about context-free grammars. The first test was about transforming a nondeterministic finite automaton into an equivalent deterministic one. The second test was about transforming a finite automaton into an equivalent regular expression. The third test dealt with state minimization of deterministic finite automata, and the fourth test with transforming a finite automaton into an equivalent context-free grammar.

Formative Evaluations. Each test consisted of twelve questions. The first six questions were usually of the form "What would the algorithm do in the next step" or "Which states of the finite automaton have a certain property". They served as a pre-test. Afterwards, the students got verbal explanations related to the questions. Randomly selected, half of the students were also treated to nice visualizations of the verbal explanations. Fig. 1 shows screenshots of the verbal explanations and visualizations for the first test (making an NFA deterministic). Then followed the post-test in form of another six questions identical to the questions of the pre-test, but on different examples.

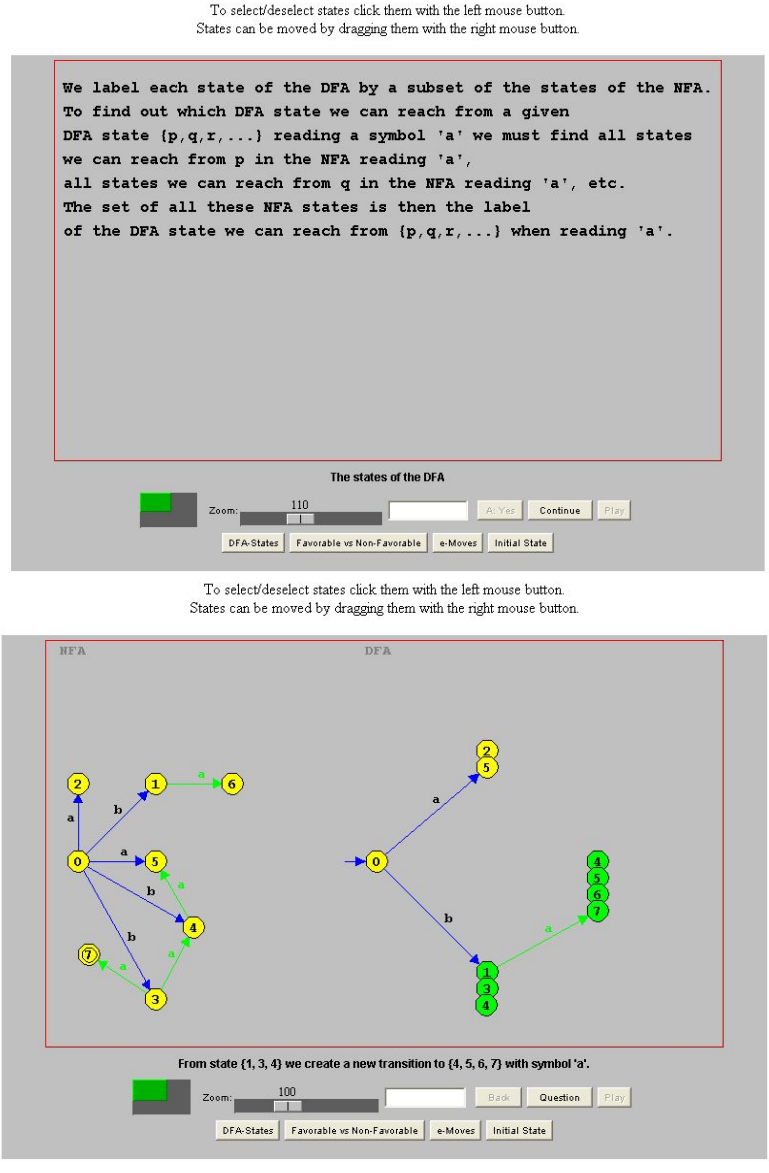


Fig. 1. Verbal explanations and visualization for transforming an NFA into an equivalent DFA

For each participant, access logs recorded all key strokes with time stamps to measure the time-on-task. This allowed us, for example, to identify students who justed clicked through the questions and explanations/visualizations without spending any time for thinking; these data are not included in our statistics in Section 3.

To get immediate student feedback after the tests, at the end each student also had to answer three questions about the test.

- Question 13: "Have the explanations been helpful for you?"
- Question 14: "Do you think animations could help you to better understand?"
- Question 15: "Was the second round of questions easier to answer for you?"

Execution. At HKUST, the students were asked to do the tests alone at home after they had learned the material in class. At Fudan, the five computer science students did the first two tests in class right after learning the material, and the other two tests alone at home after learning the material. The software engineer students, who had learned the material in the previous term, were given all four tests as home work assignment without any preparation. As expected, being unprepared they did rather poor on the pre-test questions, but after having gone through the verbal/visual explanations they did much better on the post-test questions.

Summative Evaluations. Felder and Silverman [12] distinguished four categories of student behaviour: sensory vs. intuitive learners, visual vs. verbal learners, active vs. reflexive learners, and sequential vs. global learners. In particular the second categoy, visual vs. verbal learners, is at the core of this study: we would expect the visual learners to profit more from the visualizations than the non-visual learners. Each student was asked to fill an online questionnaire determining its learning type. At HKUST, we used the questionnaire at
`www.metamath.com/multiple/multiple_choice_questions.cgi`,
which is unfortunately no longer available. At Fudan, we therefore used the questionnaire at `www.ldpride.net/learning_style.html`.

3 Statistics

Table 1 shows the data of the tests at HKUST in 2004. Test 2 had a low participation because the web server had crashed after a third of the students had finished the test, so all the other answers were not recorded. Overall, there is no difference in improvement between students who had seen the visualizations and the other students. Averaged over all four tests, the former gave 32% fewer wrong answers, while the latter gave 34% fewer wrong answers. Also, there is no difference between visual learners and non-visual learners. Unexpectedly, the visual learners performed worse with visualizations (32% improvement) than without (36% improvement), while the non-visual learners seemed even to be confused by the additional visualizations (29% improvement with visualizations versus 38% improvement without them). These results do not reflect the students' positive impression of the test. In general, the students who had seen the visualizations were more happy about the test when answering the feedback questions.

Table 1. The tests in 2004 at HKUST. Column *LS* distinguishes the learning styles: *VL* is visual learner, *L0* is unknown style, and *NVL* is non-visual learner. Column *St* gives the number of students in each category. Columns *Q1-6* and *Q7-12* list the number of wrong answers, and column *I%* gives the improvement in percent. Columns *Q13*-*Q15* give the percentage of students answering the feedback questions positively.

Test	LS	Wrong answers														Total			
		without visualization							with visualization										
		St	Q1-6	Q7-12	I%	Q13	Q14	Q15	St	Q1-6	Q7-12	I%	Q13	Q14	Q15	I%	Q13	Q14	Q15
1	VL	22	63	58	7				26	80	62	22				16			
	L0	9	24	18	25				2	7	5	28				25			
	NVL	10	34	25	26				9	37	30	18				22			
	All	41	121	101	16				37	124	97	21				19			
2	VL	7	21	7	66	71	71	57	5	15	7	53	80	60	80	61	75	66	66
	L0	3	6	7	-16	33	66	66	3	11	4	63	100	100	100	35	66	83	83
	NVL	4	9	5	44	0	25	0	2	6	0	100	100	100	100	66	33	50	33
	All	14	36	19	47	42	57	42	10	32	11	65	90	80	90	55	62	66	62
3	VL	23	40	23	42	73	78	73	24	43	33	23	75	83	75	32	74	80	74
	L0	5	11	11	0	60	100	80	10	29	19	34	70	100	80	25	66	100	80
	NVL	9	16	11	31	66	77	66	8	18	16	11	75	75	50	20	70	76	58
	All	37	67	45	32	70	81	72	42	90	68	24	73	85	71	28	72	69	47
4	VL	23	37	15	59	73	86	82	23	25	8	68	78	86	86	62	76	86	84
	L0	4	16	7	56	75	100	100	10	14	10	28	90	100	100	43	85	100	100
	NVL	9	14	4	71	55	55	77	6	13	6	53	50	66	66	62	53	60	73
	All	36	67	26	61	69	80	83	39	52	24	53	76	87	87	57	73	84	85
All	VL	75	161	103	36	81	85	82	78	163	110	32	84	88	87	34	83	86	84
	L0	21	57	43	24	76	95	90	25	61	38	37	84	100	92	31	80	97	91
	NVL	32	73	45	38	65	71	71	25	74	52	29	80	84	76	34	71	77	73
	All	128	291	191	34	76	84	81	128	298	200	32	83	89	85	33	80	87	83

Table 2 shows the data of the tests at Fudan in 2005. Overall, the students with access to visualizations fared much better in the second round of questions: 47% improvement versus 29% improvement for the students without visualizations (42% versus 26% for the visual learners). For the software engineering students, the performance gap is even bigger: 53% versus 21% (53% versus 8% for the visual learners). Remember that these students did the tests without any preparation and therefore really depended on the explanations/visualizations to learn or remember the material, so their performance data should be the most credible ones. Again, the students who had seen the visualizations were much more happy about the test when answering the feedback questions.

In the pre-test questions, the students at HKUST gave 2.3 wrong answers on the average, versus 1.8 wrong answers by the computer science students and 3.1 wrong answers by the software engineering students at Fudan. In the post-test questions, the students at HKUST improved to 1.5 wrong answers on the average, while the students at Fudan gave 1.1 and 1.8 wrong answers, respectively. This shows that the software engineering students at Fudan benefitted a lot more from the online explanations than the students at HKUST. The numbers for the computer science students at Fudan should not be misinterpreted, there were only five students, and they were the top five students in the class (so their group was not a good random sample).

Table 2. The tests in 2005 at Fudan. *CS* denotes the computer science students, *SE* the software engineering students, and *All* both together. Missing rows had no entries.

Test		LS	Wrong answers														Total			
			without visualization							with visualization										
			St	Q1-6	Q7-12	I%	Q13	Q14	Q15	St	Q1-6	Q7-12	I%	Q13	Q14	Q15	I%	Q13	Q14	Q15
1	CS	VL	1	1	2	-100	0	0	0	3	3	3	0	66	66	66	-25	100	100	100
		L0	1	3	1	66	100	100	100	0							66	100	100	100
		All	2	4	3	25	50	50	50	3	3	3	0	100	100	100	14	80	80	80
	SE	VL	2	7	10	-42	100	100	100	3	11	4	63	66	100	100	22	80	100	100
		L0	2	3	3	0	50	50	50	6	15	9	40	66	100	83	33	62	87	75
		NVL	2	7	3	57	50	50	50	0							57	50	50	50
		All	6	17	16	5	66	66	66	9	26	13	50	66	100	88	32	66	86	80
	All	VL	3	8	12	-50	66	66	66	6	14	7	50	83	100	100	13	77	88	88
		L0	3	6	4	33	66	66	66	6	15	9	40	66	100	83	38	66	88	77
		NVL	2	7	3	57	50	50	50	0							57	50	50	50
		All	8	21	19	9	62	62	62	12	29	16	44	75	100	91	30	70	85	80
2	CS	VL	4	9	5	44	100	100	75	2	4	3	25	100	100	100	38	100	100	83
		L0	1	1	1	0	100	100	100	0							66	100	100	85
		All	5	10	6	40	100	100	80	2	4	3	25	100	100	100	35	100	100	85
	SE	VL	2	5	4	20	100	100	50	2	9	1	88	50	50	100	64	75	75	75
		L0	3	12	8	33	100	100	33	2	4	0	100	100	100	100	50	100	100	60
		NVL	2	4	4	0	50	50	50	0							16	50	50	50
		All	7	21	16	23	85	85	42	4	13	1	92	75	75	100	50	81	81	63
	All	VL	6	14	9	35	100	100	66	4	13	4	69	75	75	100	51	90	90	80
		L0	4	13	9	30	100	100	50	2	4	0	100	100	100	100	47	100	100	66
		NVL	2	4	4	0	50	50	50	0							0	50	50	50
		All	12	31	22	29	91	91	58	6	17	4	76	83	83	100	45	88	88	72
3	CS	VL	3	10	3	70	66	100	100	2	2	2	0	100	100	50	58	80	100	80
		All	3	10	3	70	66	100	100	2	2	2	0	100	100	50	58	80	100	80
	SE	VL	1	6	6	0	100	100	100	3	13	10	23	66	100	100	15	75	100	100
		L0	1	2	0	100	100	100	100	2	12	5	58	100	50	50	64	100	66	66
		NVL	1	5	6	-20	0	0	0	1	5	5	0	100	100	100	16	50	50	50
		All	3	13	12	7	66	66	66	6	30	20	33	83	83	83	25	77	77	77
	All	VL	4	16	9	43	75	100	100	5	15	12	20	80	100	80	32	88	100	88
		L0	1	2	0	100	100	100	100	2	12	5	58	100	50	50	64	100	66	66
		NVL	1	5	6	-20	0	0	0	1	5	5	0	100	100	100	0	50	50	50
		All	6	23	15	34	66	83	83	8	32	22	31	87	87	75	32	78	85	78
4	CS	VL	0							3	2	3	-50	100	100	66	-50	100	100	66
		All	0							3	2	3	-50	100	100	66	-50	100	100	66
	SE	VL	4	7	3	57	75	75	100	1	6	3	50	0	100	100	53	60	80	100
		L0	1	1	0	100	100	100	100	4	9	2	77	100	100	100	80	100	100	100
		NVL	2	1	0	100	100	100	100	0							100	100	100	100
		All	7	9	3	66	85	85	100	5	15	5	66	80	100	100	66	83	91	100
	All	VL	4	7	3	57	75	75	100	4	8	6	25	75	100	75	40	75	87	87
		L0	1	1	0	100	100	100	100	4	9	2	77	100	100	100	80	100	100	100
		NVL	2	1	0	100	100	100	100	0							100	100	100	100
		All	7	9	3	66	85	85	100	8	17	8	52	87	100	87	57	86	93	93
All	CS	VL	8	20	10	50	75	87	75	10	11	11	0	100	100	80	32	88	94	77
		L0	2	4	2	50	100	100	100	0							50	100	100	100
		All	10	24	12	50	80	90	80	10	11	11	0	100	100	80	34	90	95	80
	SE	VL	9	25	23	8	88	88	88	9	39	18	53	55	88	100	35	72	88	94
		L0	7	18	11	38	85	85	57	14	40	16	60	85	92	85	53	85	90	76
		NVL	7	17	13	23	57	57	57	1	5	5	0	100	100	100	18	62	62	62
		All	23	60	47	21	78	78	69	24	84	39	53	78	95	95	40	78	86	82
	All	VL	17	45	33	26	82	88	82	19	50	29	42	78	94	89	34	80	91	86
		L0	9	22	13	40	88	88	66	14	40	16	60	85	92	85	53	86	91	95
		NVL	7	17	13	23	57	57	57	1	5	5	0	100	100	100	18	62	62	62
		All	33	84	59	29	78	81	72	34	95	50	47	82	94	88	39	80	88	80

4 Conclusions

We performed a series of tests in two courses at two different universities fo find out whether algorithm visualizations are more helpful for students than just verbal explanations. While the results of the first test at HKUST did not show

any difference in the learning output of the students, the results of the second test at Fudan (which we believe to be more credible than the test at HKUST, because of a different setup of the tests) showed a distinctive advantage of having access to visualizations. This is very encouraging for those instructors who spend much time and effort in creating good visualizations for their courses.

References

1. R. Baecker. Sorting out Sorting: A case study of software visualization for teaching computer science. In J. T. Stasko, J. Domingue, M. H. Brown, and B. A. Price, editors, *Software Visualization: Programming as a Multimedia Experience*, chapter 24, pages 369–381. The MIT Press, Cambridge, MA, and London, England, 1997.
2. R. M. Baecker. Sorting out sorting, 1983. Narrated colour videotape, 30 minutes, presented at ACM SIGGRAPH '81 and excerpted in ACM SIGGRAPH Video Review No. 7, 1983.
3. M. H. Brown. Exploring algorithms using Balsa-II. *Computer*, 21(5):14–36, 1988.
4. M. H. Brown. Zeus: A a system for algorithm animation and multi-view editing. In *Proceedings of the 7th IEEE Workshop on Visual Languages*, pages 4–9, 1991.
5. M. H. Brown and J. Hershberger. Color and sound in algorithm animation. *Computer*, 25:52–63, 1992.
6. M. H. Brown and R. Sedgewick. A system for algorithm animation. *Computer Graphics*, 18(3):177–186, 1984.
7. G. Cattaneo, U. Ferraro, G. F. Italiano, and V. Scarano. Cooperative algorithm and data types animation over the net. In *Proceedings of the IFIP 15th World Computer Congress on Information Processing (IFIP'98)*, pages 63–80, 1998. System home page: `http://isi.dia.unisa.it/catai`.
8. S. Cooper, W. Dann, and R. Pausch. Introduction to OO: Teaching objects first in Introductory Computer Science. In *Proceedings of the 34th Technical Symposium on Computer Science Education (SIGCSE'03)*, pages 191–195, 2003.
9. P. Crescenzi, C. Demetrescu, I. Finocchi, and R. Petreschi. Reversible execution and visualization of programs with LEONARDO. *Journal of Visual Languages and Computing*, 11(2):125–150, 2000. System home page: `http://www.dis.uniroma1.it/~demetres/Leonardo`.
10. C. Demetrescu, I. Finocchi, G. F. Italiano, and S. Näher. Visualization in algorithm engineering: Tools and techniques. In *Experimental Algorithics — The State of the Art*, pages 24–50. Springer-Verlag, Heidelberg, 2002.
11. C. Demetrescu, I. Finocchi, and G. Liotta. Visualizing algorithms over the Web with the publication-driven approach. In *Proceedings of the 4th Workshop of Algorithms and Engineering (WAE'00)*, 2000.
12. R. M. Felder and L. K. Silverman. Learning styles and teaching styles in engineering education. *Engineering Education*, 78(7):674–681, 1988.
13. V. Fix and P. Sriram. Empirical studies of algorithm animation for the selection sort. In W. Gray and D. Boehm-Davis, editors, *Empirical Studies of Programmers: 6th Workshop*, pages 271–282. Ablex Publishing Corporation, Norwood, NJ, 1996.
14. R. Fleischer and L. Kučera. Algorithm animation for teaching. In S. Diehl, editor, *Software Visualization, State-of-the-Art Survey*, Springer Lecture Notes in Computer Science 2269, pages 113–128. Springer-Verlag, Heidelberg, 2002.

15. S. Grimson, M. McNally, and T. Naps. Algorithm visualization in computer science education: Comparing levels of student engagement. In *Proceedings of the 1st ACM Symposium on Software Visualization (SOFTVIS'03)*, pages 87–94, 2003.
16. R. R. Henry, K. M. Whaley, and B. Forstall. The University of Washington Program Illustrator. In *Proceedings of the 1990 ACM SIGPLAN Conference on Programming Language Design and Implementation (PLDI'90)*, pages 223–233, 1990.
17. C. A. Hipke and S. Schuierer. VEGA: A user centered approach to the distributed visualization of geometric algorithms. In *Proceedings of the 7th International Conference in Central Europe on Computer Graphics, Visualization and Interactive Digital Media (WSCG'99)*, pages 110–117, 1999.
18. C. D. Hundhausen, S. A. Douglas, and J. T. Stasko. A meta-study of algorithm visualization effectiveness. *Journal of Visual Languages and Computing*, 13(3):259–290, 2002.
19. Just-in-Time Teaching Home Page. `http://webphysics.iupui.edu/jitt/jitt.html#`.
20. C. Kehoe, J. Stasko, and A. Taylor. Rethinking the evaluation of algorithm animations as learning aids: An observational study. *International Journal of Human-Computer Studies*, 54(2):265–284, 2001.
21. E. Kinber and C. Smith. *Theory of Computing*. Prentice Hall, Englewood Cliffs, NJ, 2001.
22. B. Koldehofe, M. Papatriantafilou, and P. Tsigas. Integrating a simulation-visualization environment in a basic distributed systems course: A case study using LYDIAN. In *Proceedings of the 8th Annual Conference on Innovation and Technology in Computer Science Education (ITiCSE'03)*, 2003.
23. A. Korhonen, L. Malmi, P. Myllyselkä, and P. Scheinin. Does it make a difference if students exercise on the Web or in the classroom? In *Proceedings of the 7th Annual Conference on Innovation and Technology in Computer Science Education (ITiCSE'02)*, pages 121–124, 2002.
24. M. Kuittinen and J. Sajaniemi. First results of an experiment on using roles of variables in teaching. In *Proceedings of the 15th Annual Workshop of the Psychology of Programming Interest Group (PPIG'03)*, pages 347–357, 2003.
25. S. P. Lahtinen, E. Sutinen, and J. Tarhio. Automated animation of algorithms with Eliot. *Journal of Visual Languages and Computing*, 9:337–349, 1998.
26. B. P. Miller. What to draw? When to draw? An essay on parallel program visualization. *Journal of Parallel and Distributed Computing*, 18:265–269, 1993.
27. P. Mulholland and M. Eisenstadt. Using software to teach computer programming: Past, present and future. In J. T. Stasko, J. Domingue, M. H. Brown, and B. A. Price, editors, *Software Visualization: Programming as a Multimedia Experience*, chapter 26, pages 399–408. The MIT Press, Cambridge, MA, and London, England, 1997.
28. T. L. Naps (co-chair), G. Rößling (co-chair), V. Almstrum, W. Dann, R. Fleischer, C. Hundhausen, A. Korhonen, L. Malmi, M. McNally, S. Rodger, and J. Á. Velázquez-Iturbide. Exploring the role of visualization and engagement in computer science education. Report of the ITiCSE 2002 Working Group on "Improving the Educational Impact of Algorithm Visualization". *ACM SIGCSE Bulletin*, 35(2):131–152, 2003.
29. T. L. Naps (co-chair), G. Rößling (co-chair), J. Anderson, S. Cooper, W. Dann, R. Fleischer, B. Koldehofe, A. Korhonen, M. Kuittinen, L. Malmi, C. Leska, M. McNally, J. Rantakokko, and R. J. Ross. Evaluating the educational impact of visualization. Report of the ITiCSE 2003 Working Group on "Evaluating the Educational Impact of Algorithm Visualization". *ACM SIGCSE Bulletin*, 35(4):124–136, 2003.

30. G. M. Novak, E. T. Patterson, A. D. Gavrin, and W. Christian. *Just-in-Time Teaching: Blending Active Learning with Web Technology*. Prentice Hall, Englewood Cliffs, NJ, 1999.
31. W. C. Pierson and S. H. Rodger. Web-based animation of data structures using JAWAA. In *29th SIGCSE Technical Symposium on Computer Science Education*, 1998. System home page: `http://www.cs.duke.edu/csed/jawaa/JAWAA.html`.
32. G. C. Roman, K. C. Cox, C. D. Wilcox, and J. Y. Plun. PAVANE: A system for declarative visualization of concurrent computations. *Journal of Visual Languages and Computing*, 3:161–193, 1992.
33. M. Sipser. *Introduction to the Theory of Computation*. China Machine Press, 2 (english). edition, 2002.
34. J. Stasko and A. Lawrence. Empirically assessing algorithm animations as learning aids. In J. T. Stasko, J. Domingue, M. H. Brown, and B. A. Price, editors, *Software Visualization: Programming as a Multimedia Experience*, chapter 28, pages 419–438. The MIT Press, Cambridge, MA, and London, England, 1997.
35. J. T. Stasko. Tango: A framework and system for algorithm animation. *Computer*, 23(9):27–39, 1990.
36. J. T. Stasko, J. Domingue, M. H. Brown, and B. A. Price. *Software Visualization: Programming as a Multimedia Experience*. The MIT Press, Cambridge, MA, and London, England, 1997.

Some Cognitive Aspects of a Turing Test for Children

Ruqian Lu[1,2,3], Hongge Liu[2,3], Songmao Zhang[2,3], Zhi Jin[2,3], and Zichu Wei[3]

[1] Shanghai Key Lab for IIP, Dept. of Computer Science and Engineering, Fudan University
[2] Institute of Computing Technology, Key Lab of IIP, Academia Sinica
[3] Institute of Mathematics, Key Lab MADIS, Academia Sinica

Abstract. Knowledge, cognition and intelligence are three tightly connected concepts. The Turing test is widely accepted as a test stone for machine intelligence. This paper analyzes experiences obtained in a research project on a Turing test for children and discusses its meaning with respect to some knowledge and cognition issues.

Keywords: Turing test, Children Turing test, intelligence, knowledge, cognitive system.

1 Controversies Around Turing Test

Since the Turing test (TT) was proposed half a century ago [20], it has been always a topic of controversial discussions. Three of the frequently discussed problems are listed below.

First problem: Is the Turing test a meaningful test stone for machine intelligence? We list here only the negative answers to this question. Apart from the classical argument raised by Searle on the Chinese room problem [19], the strongest contra among them may be stated by Hayes and Ford: "The Turing test had a historical role in getting AI started, but it is now a burden to the field, damaging its public reputation and its own intellectual coherence. We must explicitly reject the Turing test in order to find a more mature description of our goals; it is time to move it from the textbooks to the history books" [8].

Second problem: To which degree is a Turing test a suitable test stone for machine intelligence? People were wondering if the Turing test is a necessary or sufficient condition for confirming machine intelligence. Donald Michie thought [16] that the Turing test should not be understood as a means for testing the equivalence of machine and human intelligence. It should only be used to test whether a machine can have intelligence. He also said that it is not enough to test the conversation capability of a machine. Many other things have to be tested and TT itself should be generalized. Stevan Harnad proposed to include the tests of non-linguistic functions such as pattern recognition [7]. David Bell went a step further and proposed the concept of Turing robots [1].On summarizing, James H Fetzer pointed out that the total TT (TTT) proposed by Harnad should not be limited to a short time interval like the usual exams. A meaningful TT should extend over a long time period like the life time of a person [5].

R. Lu et al. (Eds.): Chinese-German Ws on Cognitive Systems 2005, LNAI 4429, pp. 25 – 35, 2007.
© Springer-Verlag Berlin Heidelberg 2007

There are also other opinions. Robert M. French for example said: “it is unnecessary to propose even harder versions of the Test in which all physical and behavioral aspects of the two candidates had to be indistinguishable before allowing the machine to pass the test”. Moreover, he claimed that “even in its original form, the Turing test is too hard”, “is not a reasonable test for general machine intelligence”[6]. This claim was soon refuted by Peter D. Turney [21].

Third problem: Is the Turing Test the most suitable form of testing machine intelligence? It is certainly not the only possible one. Actually, many different versions of such tests have been proposed. For example, Bringsjord thought that creativity is the key issue of human intelligence (according to Lady Lovelace) and proposed the Lovelace test as a replacement for Turing test [2]. Erion mentioned Rene Descartes, a great philosopher and scientist of 18th century, who had proposed a test for differentiating human from machine, which contained an idea of conversation test similar to TT but yet something more, which he called action test (as opposed to language test) [3]. Feigenbaum suggested a new scheme of intelligence test, which tests only professional knowledge, but not commonsense knowledge. He called it Feigenbaum test [4]. He thought it should be easier to pass the Feigenbaum test than to pass a general Turing test and thus the former can present the first step of completing the latter.

Our research has been focused on the third problem above. In a previous paper [14] we have reported our experiments of children Turing test (CTT), which is another variant of Turing test. While in that paper we focused on the implementation aspects of CTT, including the overall test design, the knowledge representation, the knowledge base, the language understanding, the engineering platform, etc., this paper has another focus: the cognitive aspects of CTT.

Since 1991, an international competition called Loebner test has been held each year to encourage the researchers of Turing test [11]. Though a bronze award has always been handed out to the best program of each year, the award of a gold medal with 100,000 Dollars is still pending. The discussion in this paper is based on both our and Losbner’s experiences.

2 CTT and Children Cognition Development

Although we have reported the first results of CTT in a previous paper [14], it is still worthwhile to give a short summary in this section to benefit the reader. We have performed CTT in three different types.

Type 1: The judges and person confederates are all children. Each session of conversation has 10 rounds. Each round consists of a sentence of both sides each. After 10 rounds the judge should decide who of the two confederates was a child and who was a computer. This type was again divided in two sub-types (children’s age between 5-7 years old or between 7-11 years old.)

Type 2: Only confederates are children. Judges are now adult teachers of primary schools or kindergartens. The number of rounds in each conversation session is

unlimited. The judge may perform the conversation continuously until she thought she had collected enough information to make the decision.

Type 3: All possible combinations of the two confederates are allowed: child and computer, two children or two computers.

In all these types, conversation was done in written texts (no voice recognition or generation). Working language is daily children Chinese. No information exchange between the two confederates was allowed.

In all experiments, the judges had about 10-20 percents of failure rate. For detailed data see [14].

Piaget has divided the development of children cognition in four stages: the stages of perception (age 0-2), preoperation (age 2-7), concrete operation (age 7-11) and formal operation (age 11-16) [18]. The children taken part at our CTT were mainly in the Piaget's stage of concrete operation. The characteristics of this stage are the acquisition of capability of basic mental operations, including some logical operations. In our observation, the attitude of Children towards CTT can be classified in three age intervals: Children of 5-7 years old consider CTT as an interesting game. Children of age 7-10 consider it as a knowledge competition. They believe they will win the CTT if the conversation partner can not answer their questions correctly. Children of age 10-11 expose a more reasonable attitude: they are already partially aware of the real purpose of CTT. An interesting question can be the following: do these different behaviors towards CTT correspond to the essential difference between Piaget's last three cognition stages?

Considering the inability of little children in understanding the real meaning of CTT, we have designed a qualification test for the eligibility of a child as a CTT participant:

Qualification Test for Child Judge (Imitation Game for deciding the Parents):

The imitation game is played by child C as judge with father (A) and mother (B) as confederates. C tries to determine through conversation who of (A, B) is the father and who is the mother.

We believe that if C has passed the qualification test, then

1. C should have understood the meaning of the concept "identification" that is neither a random guess nor a joke at free will. Rather, it should mean an effort to find out the real capacity of the contestants.

2. C should have understood that in order to identify the real capacity of the contestants one should have enough reasons that are based on differences between A and B. In order to make a right identification, one should first find out their most striking differences.

3. C should have understood that during the conversation A and B do not necessary cooperate with C in form of providing correct information. On the contrary, they could provide wrong information intentionally to mislead C (The father may try to introduce himself as mother, or vice versa).

We believe that a child who has passed this qualification test will understand the meaning of CTT, even if s/he has not yet played with a computer. The qualification problem appears also in other related tests, such as the Feigenbaum test [4], where it is

requested that all human participants should be members of the American National Academy in the particular domain of test. These persons must have passed another "qualification test", namely the election procedure of the National Academy of USA.

3 Cognitive Troubles in Detecting Commonsense Knowledge Failures

Roger Penrose pointed out that the major difficulty of programming Turing Test is to let the computer answer questions about commonsense knowledge [17]. This opinion is also shared by other experts. For example, the argument of French that the Turing Test is too hard for computers (see section 1) is based on so-called subcognitive questions, like "does fresh baked bread smell nicer than freshly mowed lawn?"[6], which are also of this commonsense knowledge character.

Therefore, any conversation system, which claims to be knowledge based, should try to avoid any commonsense knowledge failures. There are two kinds of such failures: explicit or implicit. The answer "yes" to the question "does an ox have a wing?" is a knowledge mistake; the answer "don't know" to the same question is knowledge ignorance. Both failures are explicit. But what would a program say to the following question:

Does every ox have a wing? (1)

Possibly, it would say: 'no'. More exactly, the answer would be 'no, not every ox has a wing', which is logically correct. But this formulation would raise the false impression that there may be some oxen that do have wings. A normal human being would not answer the question in this way. The right answer should be: 'no, no ox has a wing'. Let us now check why the program failed to produce a correct answer. Logically, the question can be expressed as follows:

For all x, (x is an ox →x has a wing) ?

Formally, in order to refute this statement, the program poses a negation symbol before the questioned proposition:

~ for all x, (x is an ox → x has a wing)

which equals to

there is an x, (x is an ox & x does not have wing)

which does not exclude the following possibility:

there is an x, (x is an ox & x has a wing)

Therefore the program would produce the impression about the possible existence of an ox with wing if it just answered: "no" to the question (1). This would be a big difference between a program's answer and a human's answer. We call this phenomenon the negation paradox.

Question (1) is not the only example of questions with commonsense errors (traps). Many such examples could be built, such as:

Does a green ox have a wing?

Does every green ox put on red jackets when it attends an international academic conference?

Often, a judge may prepare a commonsense trap for the confederates. If a program is not able to detect the commonsense errors contained in the judge's input, it may fall in the trap and fail to pass the test.

One of the principles of the Loebner contest was to keep the conversation as natural as possible. "Judges were ……also told to refrain from "trickery or guile"…; obscure or unexpected questions and gibberish designed to expose nonhumans would not be allowed……judges should respond naturally, as they would in a conversation with another person" [11]

However, the ability of detecting commonsense errors in conversation is also a part of human intelligence. It should not be an extravagant requirement to include some implied commonsense test in a TT. Let alone according to the reports the judges of Loebner contests did not quite follow these instructions: "But transcripts of interactions from restricted Loebner contests clearly reveal tensions among participants. Judges did not blindly follow the instruction to "respond naturally, as they would in a conversation with another person," simply because the very thing assumed by that instruction –namely, the personhood or humanness of each contestant—was in serious doubt" [11]

4 Cognitive Capability vs. Knowledge Capacity

Now it is appropriate to ask the question: what is the relation between cognition and knowledge in the context of the Turing test? Is the amount of knowledge equal to the capability of cognition? Or is the former positively proportional to the latter? Or is there is no obvious relation between intelligence and knowledge?

In order to answer this question, we have designed two lists of exam questions for school children. The first list was designed for testing their intelligent ability in solving puzzle like problems. It contains 20 problems, where 9 problems were of observation type, other 11 ones were of imagination type. The second list was designed for testing

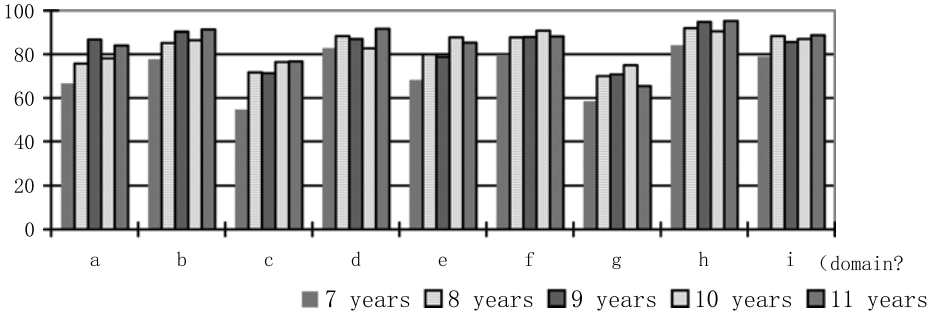


Fig. 1. Differences in Knowledge level among children of same age. a = astronomy, b = zoology, c = botanic, d = human body, e = history, f = culture, g = science and technology, h = daily life, i = social relations.

the children's mastering of knowledge, mostly commonsense knowledge. There were in total 269 questions and were divided in four groups: nature and environment, mankind and society, culture and arts, science and technology. About 200 children took part at this experiment. Statistics of the scores are made according to their ages (from 7 to 11 years old) separately. Figure 1 illustrates the impact of age on children's knowledge in different domains. It can be seen that the increase of children's knowledge is more rapid in domains such as life, culture, social relation and humanity, but rather slow in domains of science and technology.

Furthermore, they know more about animals than about plants. This result reflects the characteristics of children living in cities. Figure 2 shows the difference of children knowledge levels in general. It can be seen easily, that the difference in knowledge levels decreases with the increase of ages of tested children. This reflects the fact that children (of pre-school age) receive different quality of education in their families, but roughly the same kind of education in schools. It provides also a partially positive evidence for the validity of our conjecture 2. Figure 3 displays the increase curve of children's intelligence. It conforms roughly to the increase of knowledge. But since the data collected are not rich and complete enough, a statistical conclusion can not yet be drawn [9].

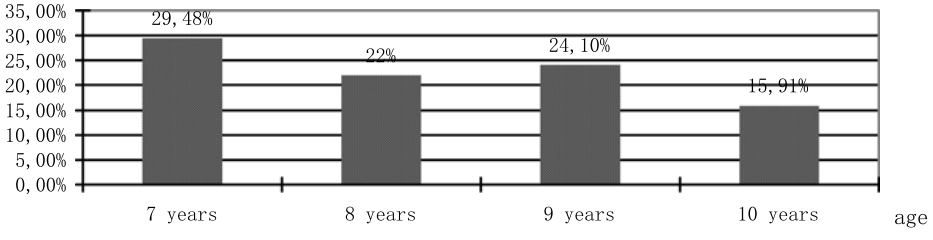


Fig. 2. Decrease of knowledge differences with increase of age

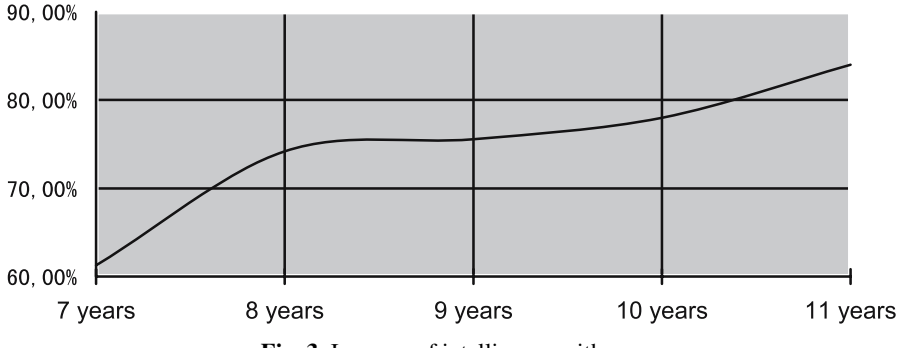


Fig. 3. Increase of intelligence with age

5 Emotion Quotient vs. Intelligence Quotient

In the Loebner test, judges are requested to be polite when interrogating the contestants, in order to conform to the rule of the social relationships: "In "normal" conversations,

we expect non-intimates to adhere to politeness rules, with each party striving to avoid conflict and to maintain social harmony" [11].

Therefore, the organizers of the Loebner test "Suggest a view of Loebner discourse in which polite and cooperative non-imitates avoid face-threatening acts" [11].

However, while suggesting a conversation strategy like this, they could not stick to it thoroughly in the whole contest, since the real situation was: "when judges are selected on the basis of their expertise in AI, cognitive science and computer journalism, the notion that Loebner participants are polite strangers – or that Loebner discourse is primarily interactional—is undermined" [11].

The reason rooted in the conflict between two different styles of judges' conversation with the contestants, as it was pointed out by Sean Zdenek: "Loebner test is defined 1) as a friendly, casual conversation between two strangers of equal power, and 2) as a one-way transaction in which judges control the conversational floor in an attempt to expose contestants that are not human" [23].

During the appraisal meeting of our CTT project, a member of the appraisal committee has tried to irritate a contestant, which was a program, by insulting it. Then the following conversation happened:

Appraisal committee member: You are stupid!

Program: No, it is you yourself who is stupid!

Interestingly, this expert appreciated the program very much because he thought that if a computer would show some degree of emotion and reaction to outside provocation, then this computer exposed a behavior close to that of a human.

In three of the four sessions of a CTT, the judges used some principles, which had nothing to do with knowledge. We list some of there justifications in the following:

Session 2: ……..

Judge (to A): Do you like your mother, and why?

A: Of course.

Judge (to A): Does your father take good care of you?

A: I am not telling you that.

……

Judge's conclusion: A is a program since a child should not be so indifferent towards his father. Besides, most of the children are honest. If A were a child, s/he should have responded with "I don't know" to any question, of which the answer is not known to him/her.

Session 3: ……

Judge (to A) do you know how a rabbit walks.

A: No, I will ask my mother.

……..

Judge's (mistaken) conclusion: A is a child since the sentence "I will ask my mother" sounds very much like a child's pet speech in the kindergarten where she is working. A child of this age has a strong tendency of parent dependency, in particular

mother dependency. When a child meets difficulty or even danger, the first person s/he thinks could be relied upon is the mother. Psychologically, this answer reminds of the first reaction of a child when this child does not know how to answer a question.

From these examples we see a principle frequently used by judges of our CTT. They identify the contestants not only by the contestant's intelligence quotient, but also by the contestant's behavior characteristics and emotion quotient. In section 2 we have been talking about qualification of judges. We have compared the advantage and disadvantage of taking adult teachers or children as judges of our CTT in [14] and have reached to the conclusion that teachers know the children better than the children know themselves. In fact, the teacher judges have made all conclusions based on their experiences of taking care of children. The conclusions of the sessions cited above were based on something, which people now count as emotion quotient of a child. This is a phenomenon that should also interest the psychologists and experts in developmental psychology and children education. This shows that the Turing test does not only test the intelligence of a computer, but also the emotion of a computer. It may be absurd to talk about the emotion of a computer. But is it not the same absurd to talk about the intelligence of a computer for some people?

According to our experience, it is easier to increase the intelligence quotient of a computer than to increase its emotion quotient. In order to do that, we need to study the emotional behavior of children and must cooperate with experts of developmental psychology. We are sure that emotion test applies not only to CTT, but also to other groups of people. For example, the patients usually worry about their health, and the criminals are usually in a standing fear of being caught by policemen. All these can be used in tests designed specifically for these particular people groups.

6 The Reverse Turing Test

Historically, the Turing Test has been used only to determine machine intelligence. Actually, it should have been used also in the reverse direction, namely to determine human intelligence. Image the following

Speculation: Assume human and machine intelligence can both be measured with a positive real number. Assume a judge has y per cent chance of making the right identification after a standard time of questioning then the ratio of human/machine intelligence is roughly equal to $y/(100 - y)$.

However, there is a problem with this speculation: we have neglected the impact of the human confederate. If the judge C makes a wrong identification in a TT, then this is not necessary the contribution of the computer confederate B. An inappropriate statement of the human confederate A may make the judge C think A were the computer B (see [14] for examples). Thus the meaning of the formula $y/(100 - y)$ becomes fuzzy and questionable.

In order to remove the influence of the human confederate, we introduce the concept of Mixed Turing Test (MTT). A MTT involves only one (human) judge and one subject, which is inaccessible for the judge except by text communication. The judge has to decide if the subject is a human or a computer solely by text conversation.

Mixed Turing Test: The judge has to go over a suite of N test sessions: { $t_1, t_2, \ldots, t_N$ }, where each t_j is either a J-H session or a J-C session. A J-H session is conversation between judge and human subject, while a J-C session is conversation between judge and computer subject. For each test, the number N_H of J-H sessions and the number N_C of J-C sessions with $N_H + N_C = N$ are randomly determined. These sessions are randomly ordered in a mixed way such that the judge is not aware with whom (human or computer) s/he is talking. We say that the computer has (partially) passed MTT if in all (part of) J-C sessions the judge has misidentified the computer.

Note that the test results of all J-H sessions are not used and will be thrown away. They will not affect the final result. The function of J-H sessions is only to "shuffle the cards". Note also that the judge will not be told whether his decision is correct after each session.

If the computer at least partially passed MTT, then we say that it has some degree of "machine intelligence". We give the name "conversational intelligence" to such kind of machine intelligence. It does not include image and voice recognition, for example. The following hypothesis is meaningful.

Hypothesis 1: MTT is a zero-sum game of conversational intelligence between judge and computer, which can be measured with a positive real number. If the judge successfully identifies the computer within a standard time of questioning [see 14] in y of the total N_c J-C sessions, then judge wins y points while the computer wins $N_c - y$ points. Assume $y < N_c$, $y/(N_c - y)$ is the (instantaneous) ratio of human/computer intelligence.

We say that the judge shows some level of "human intelligence" if $y \neq 0$. Therefore the number y provides some information about human intelligence with respect to computer.

Hypothesis 2: Assume J(x) be a group of judges of age x. Let each of them participate in MTT with the same subjects. Denote the average value of $y/(N_c - y)$ with $R_{J(x)}$. If the members of J are uniformly distributed among all people of the same age x, and if the values of $R_{J(x)}$ converge and $\lim_{|J(x)|\to\infty} R_{J(x)} = R(x)$, then we can assume that the average ratio of human/computer (conversational) intelligence in age x is R(x).

Note that we would get different values of R(x) with different computers. The computer serves here only as a reference system. Therefore we will use the notation R(x, c) instead of R(x) below, where c denotes the computer c used as the reference system. Now we are ready to introduce Reverse Turing Test (RTT) to measure human intelligence.

Reverse Turing Test: A RTT is a MTT with respect to some particular age x and computer subject c, where the value R(x, c) is known. If the value $y/(N_c - y)$ of a judge of age x is less, equal or larger than R(x, c), then we say that the conversational intelligence of this judge is below, equal or above the average level (among equal aged people).

However, we must be assured that RTT is meaningful before we accept it. The RTT would be meaningless if there are two reference systems (i.e. computers) c1 and c2, such that for the same judge of age x the two inequalities $y1/(N_{c1} - y1) < R(x, c1)$ and $y2/(N_{c2} - y2) > R(x, c)$ both hold, because it would mean that the judge were considered as below and also above the average intelligence level by different reference systems at the same time. For that we need some additional assumptions.

Definition: An implemented computer program for RTT is considered reasonable if for any two judges A and B, where A is less intelligent than B, it is always $y/(N_c - y)|_A \leq y/(N_c - y)|_B$.

But "A is less intelligent than B" may be a pure subjective opinion. Do we have a scientific measure for testing this statement? For this purpose, we propose a new hypothesis:

Hypothesis 3: The average intelligence level of people of age x is equal to or less than that of people of age x +1

We think most of the readers will accept this hypothesis, in particular if these people are children. Of course we should exclude some extreme cases such as people very old. Then we have the further

Hypothesis 4: An implemented computer program for RTT is reasonable if for any $x > 0$, it is always $R(x, c) \leq R(x+1, c)$.

These hypotheses correspond to our intuition. Thus, RTT can be considered as an exam for human judges if the computer program is reasonable. The values $y/(N_c - y)$ are the notes they obtain.

7 Some Concluding Remarks

It seems that many critics of TT are mainly based on the fact that it is too difficult to reach the goal set by Turing half a century ago. In fact, passing Turing test is not only difficult, but is also highly implausible in the near future, say during another fifty years. But this should not be a reason to deny its meaning for further research. Our motivation of studying the children Turing test is not to chase its final and complete success. CTT has raised many interesting research problems in AI and cognitive science. These problems deserve their own value of a deep study.

Acknowledgements. We owe many thanks to the anonymous referees of this paper, who have helped to improve the manuscript a lot. We are also very grateful to Joerg Siekmann and Carsten Ulrich, who have helped to turn this manuscript in its current form. **O**ur experiments in Turing test are partially supported by NSFC Major Research Program 60496324, CAS Project of Brain and Mind science, 863 project 2001AA113130, 973 project 2002CB312004, the innovation foundation of IOM, AMSS and ICT.

Other participants of these projects include Prof. Chunyi Shi, Prof. Shixian Li, with their research groups; Ping Yang, Lu Fan, Fan Yang, Xiaolong Jin and Hong Zheng.

The pupil exams on knowledge/intelligence have been made by Cheng Shu, Hongge Liu, Sikang Hu, Nan Yi, Xiaoling Zhao of AMSS. We thank all of them for their contribution.

References

[1] Bell, D, Turing robots, private communication
[2] Bringsjord, S. et.al., Creativity, the Turing Test, and the (Better) Lovelace Test, Minds and Machines 11:3-27, 2001.
[3] Erion, G..J.,The Cartesian Test for Automatism, Minds and Machines 11:29-39, 2001.
[4] Feigenbaum E. A., Some Challenges and Grand Challenges for Computational Intelligence, JACM 50:1, 32-40, 2003.
[5] Fetzer J. H., Constructions of the Mind, SEHR 4(2), 1995.
[6] French, R.M, Peeking behind the screen: the unsuspected power of the standard Turing Test, JETAI, 12(2000) 331-340.
[7] Harnad S., The Turing Test is not a Trick: Turing Indistinguishability is a Scientific Creterion, SIGART Bulletin 3(4), P: 9-10, 1992.
[8] Hayes, P.J., Ford, K.M., Turing Test considered harmful, Proc. of the Fourteenth IJCAI, 972-977. 1995.
[9] Liu, Hongge. Shu, Cheng et. al., Statistic Report on Testing Children Knowledge and Intelligence Levels, 2000.
[10] Liu, Lengning, A Study on children Turing test, master thesis, Institute of Mathematics, Academia Sinica, 2000.
[11] 1997 Loebner Prize Contest Results, http://www.loebner.net/Prizef/loebner-prize-1997.html, 2000.
[12] Lu, Ruqian et al., Agent oriented commonsense knowledge base, Acta Scientia, V.43, N.6, pp.641-652, 2000.
[13] Lu, Ruqian & Songmao Zhang, PANGU—An Agent-Oriented Knowledge Base, World Computer Congress 2000.
[14] Lu, Ruqian & Songmao Zhang, A retrospective view on children Turing test, Journal of software, V. 15, N. 12, pp.1751-1763, 2004.
[15] McCarthy J., Formalizing Commonsense: Papers by John McCarthy, Ablex Publishing Cooperation, 1990.
[16] Michie D., Turing's Test and Conscious Thought, Artificial Intelligence 60, P. 1-22, 1993.
[17] Penrose R, The Emperor's New Mind, Oxford University Press, 1989.
[18] Piaget, J., Inhelder, B., Memory and Intelligence, New York: Basic Books, 1973.
[19] Searle, J.R., Minds, Brains, and programs, Behavioral and Brain Sciences, Vol. 3, 1980.
[20] Turing, A, Computing Machinery and Intelligence, Mind, 1950.
[21] Turney, P.D., Answering subcognitive Turing Test questions: a reply to French, JETAI, 13 (2001) 409-419.
[22] Weizenbaum, J. ELIZA—A computer program for the study of natural language communication between man and machine, CACM 9:36-45, 1966.
[23] Zdenek, Sean, Passing Loebner's Turing Test: A Case of Conflicting Discourse Functions, Minds and Machines, 11: 53-76, 2001.

Challenges in Search and Usage of Multi-media Learning Objects

Erica Melis[1], Ruimin Shen[2], Jörg Siekmann[3],
Carsten Ullrich[1], Fan Yang[2], and Peng Han[2]

[1] DFKI, German Research Center for Artificial Intelligence, Stuhlsatzenhausweg 3, 66123 Saarbrücken, Germany
[2] Shanghai Jiao Tong University, 200030 Shanghai, China
[3] Saarland University, Department of Computer Science, 66041 Saarbrücken, Germany
melis@dfki.de, rmshen@sjtu.edu.cn, siekmann@ags.uni-sb.de, cullrich@dfki.de, fyang@sjtu.edu.cn, phan@sjtu.edu.cn

Abstract. The definition, assembly and manipulation of learning objects is becoming more and more popular in learning environments. But despite standardization efforts their appropriate markup and practical usage still faces many difficulties, such as retrieval, true interoperability and cognitively adequate selection and presentation. This paper describes current work of the authors tackling some of these challenges.

Keywords: Computers and Education, Information Search and Retrieval, Intelligent Web Services and Semantic Web, Metadata, Modeling structured, textual and multimedia data, CBR.

1 Introduction

The definition, assembly and manipulation of learning objects (LOs) is becoming more and more popular in learning environments. But despite standardization efforts (e.g., [1], [2], [3]), their appropriate markup and practical usage still faces many difficulties.

Challenges include the retrieval of appropriate multi-media learning objects (MMLOs) from distributed web-repositories, true interoperability and cognitively adequate selection and presentation. This paper describes current work of the authors tackling some of these challenges. The German co-authors focus on the semantic representation, search, presentation and (re-)use of LOs. The Chinese partners investigated solutions to authoring and usage of multi-media learning objects (MMLOs) in large-scale learning environments and automatic question answering. In this context, large-scale equals to several thousands of users from all over China and more than hundreds of students per class. This leads to high requirements regarding support (question answering) and the difficulty of controlling the teaching and learning effect.

In this paper, we present some solutions for the automatic usage of MMLO in large-scale distributed learning environments. The paper has two parts: Section 2 describes challenges and solutions of searching for LOs; Section 3 describes the usage of LOs.

R. Lu et al. (Eds.): Chinese-German Ws on Cognitive Systems 2005, LNAI 4429, pp. 36 – 44, 2007.
© Springer-Verlag Berlin Heidelberg 2007

2 Search of Learning Objects

This section focuses on the semantic representation of LOs and their retrieval. Section 2.1 discusses how can we annotate content in such a way that the semantics of the domain being learned is represented, and how to annotate the LOs in order to achieve a representation expressive enough for automatic processing. What are the additional techniques and annotations required by MMLOs such as video lectures or animated beamer presentations? The annotated content needs to be accessed with tools that allow full but easy access of the representation. One such tool, our semantic search facility, will be presented in Section 2.2. New solutions are also required for the search in Web-based learning environments where the LOs are distributed among several repositories. These will be discussed in Section 2.3.

2.1 Metadata Required for Search

Domain knowledge is a basis for learning. More often than not, the links to a formal description of the objects in the domain are realized by links from the LOs to some formal description. However, especially in domains of a formal nature, it is possible to have a more sophisticated internal representation of LOs that also provides the semantics of the LO.

Such an approach was realized in the learning environment ActiveMath, which uses the semantic XML-markup language OMDoc ([4], [5]) for mathematical documents. OMDoc has evolved as an extension of the OpenMath European standard for mathematical expressions (www.openmath.org) and provides a rather fine-grained breakdown into LOs. One objective of using this generic semantic markup language is to keep the encoded content reusable and interoperable with other, even non-educational, mathematical applications.

ActiveMath's content is represented as a collection of typed items in OMDoc annotated with metadata. The types indicate a structural characterization of the items which are either *concept* or *satellite* items: an OpenMath symbol defines a mathematical concept abstractly, i.e., an element of a formal ontology. Concepts (e.g., definitions or algorithms) are the main items of mathematical content, whereas satellites (e.g, exercises or examples) are additional items of the content which are related to one or several concepts. All items are accessible via a unique identifier.

A more hierarchical ontology in mathematics can be reached by grouping concepts into theories. Relations exist between such collections of items and the corresponding areas of mathematics. The element *theory* assembles knowledge items into mathematical theories, which can be assembled into larger theories via the import mechanisms of OMDoc. Some theories are relatively small and just concerned with particular concepts, e.g., the theory of differentiation. Large theories, e.g., Calculus, have substructures.

OMDoc has been extended for educational purposes. For this, there are metadata, which characterize not only organizational and mathematical but also educational properties of the OMDoc items. These metadata include relations. Overall, this establishes a mathematical and educational ontology.

The LOM- and DC-compatible metadata describe intellectual property rights as well as properties of learning objects that help to adapt to the learner, e.g., *difficulty* and *field*. The metadata also support the adaptation to the context of learning such as language of the learner or his educational level.

Currently, relations are expressed by metadata in OMDoc which can be translated to relations in RDF. For instance, the *prerequisite relation* expresses a mathematical dependency between concepts c_1 and c_2, i.e., it describes which concepts c_2 are mathematically necessary in order to define c_1. The *for* relation links satellite LOs to concepts and definitions to symbols. For example, a symbol can have several, typically equivalent, definitions. Additional relation include *against* and *is-a*.

To make the content more animative, the e-learning environment of the Chinese co-authors provides multimedia courseware which gives the illusion to the students as if they study in the traditional face-to-face classroom. That is to say, this courseware includes all the didactical data, such as teacher's video, audio, tutorials, computer screen, blackboard, mouse trace, etc. During the teaching process, the system compresses all teaching scenario data synchronously and automatically and edits this to a new courseware. Furthermore, a MarkerEdit Tool inserts indexing marker automatically (e.g., the title of a slide).

2.2 Semantic Search

Making the semantics of the LOs accessible to users offers new possibilities. For instance, in the LeActiveMath project, we developed a semantic search facility that takes advantage of the semantic content encoding, i.e., of its structural elements, its metadata, and the OpenMath semantics of mathematical expressions. It offers the following advantages over traditional, text-based search: to search for mathematical expressions that are formal, i.e., not just text; to search for types and other meta-information of content-items, and finally to search for other relevant sources located anywhere in the Web.

Traditionally, information retrieval focuses on the retrieval of text documents and on textual search based on two essential ingredients: first, an analysis (tokenization) process converts the text documents that will be searched into a sequence of tokens (typically words). Second, an index is built up, which stores the occurrence of tokens in documents and which can then be efficiently searched.

In order to make use of the semantic content provided in OMDoc we process the following information as well in order to build the index: the titles, the metadata information, the textual content, and the mathematical formulae.

2.2.1 Tokenization of Mathematical Formulae

Information retrieval from textual corpora processes a linear sequence of tokens. Mathematical formulae in OpenMath, however, are represented in a tree-structure. This tree structure offers valuable information for searching. Therefore, we use the sibling order of a tree-walk to produce a special sequence of tokens. More specifically, we tokenize the application with a marker of the depth, the symbols, strings, floats, and integers of OpenMath. For example, the formula $\sin x^2$ is tokenized into:

```
_(_1
_OMS_mbase://openmath-cds/transc1/sin
_(_2
_OMS_mbase://openmath-cds/arith1/power
_OMI_2
_OMV_x
_)_2
_)_1
```

Using this tokenization, we can query math expressions by an exact phrase match, that is, a match for a sequence of tokens. As a given expression can occur at any depth of a mathematical expression, exact phrase queries have to be expanded into a disjunction of queries for each depth.

Formulae with jokers can also be queried: the example above is matched by the query *sin(*)* that translates into the following sequence of tokens:

```
_(_1
_OMS_mbase://openmath-cds/transc1/sin
*
_)_1
```

where the * indicates a joker in the phrase query which matches anything as long as the remaining part is matched.

2.2.2 Integration of Relevant Web-Sources

In order to enable search for mathematical Web sources and to compare the results of this search, we automatically add links such that the same query can be submitted to search engines and content collections such as Google, Wikipedia, and MathWorld.

2.2.3 User-Adaptivity of the Search-Tool

The results of this kind of search may be overwhelming for a learner. Therefore, LeActiveMath's search tool is adaptive in the following sense: Per default, it searches only for concepts in the current course. Searching in the complete content is available via a link. Search in relevant Web-sources is only activated if the user model of the learner supports the fact the learner is able to process the information adequately. This information is represented as the *autonomy* value in the situational learner model.

For the multimedia courseware, we believe that it is even more important to design an interface for the students to decide whether the knowledge they are searching for is inside the courseware and how to locate it. For example, if a student wants to review “Probability”, he/she can input the phrase through a textbox or microphone, and then the computer can locate the relevant material in the courseware automatically.

Fig. 1. illustrates the workflow of our content-based indexing and retrieval system. Based on the marked courseware, we proposed the *Content based information retrieval system*, which indexes and locates the courseware based on an *on-demand keyword*

input. For each retrieval process, the system will give several possible choices in ascending order using *matching weight*:

$$Weight_i = Freq_i / \sum_{j=1}^{N} Freq(j) \quad (1)$$

Here $Freq_i$ is the choosen frequency relative to an on-demand knowledge concept which is incremented by 1 if it is chosen by the student. The sum is the total frequency value of the whole knowledge point in this course. This method helps to increase the indexing accuracy and efficiency.

2.3 Integration of Distributed Knowledge

In a Web-based environment, LOs may be distributed over several repositories and annotated with different metadata schemas. Hence, it is important that components processing the LOs can abstract from the actual knowledge sources. They should not need to locate the relevant sources nor should they interact with each source separately.

ActiveMath achieves this separation by using the mediator approach to information integration [6]. In this approach, a special service acts as a mediator between the knowledge processing and knowledge storing components, thus providing a uniform query interface to a multitude of autonomous data sources.

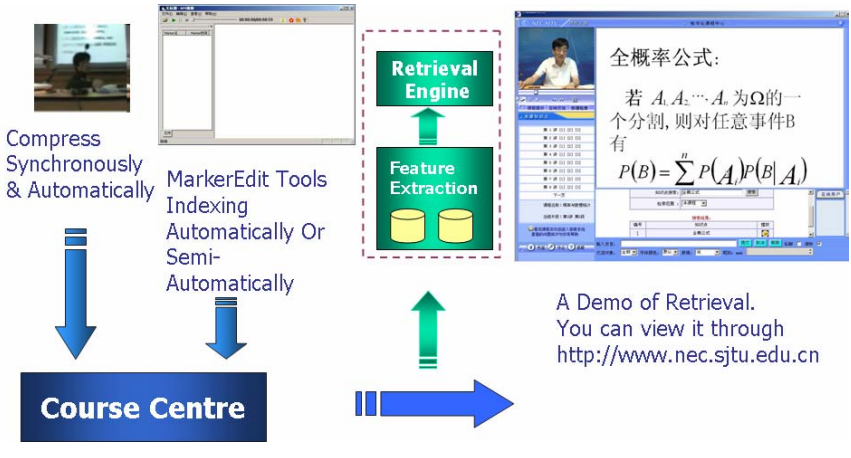


Fig. 1. Workflow of the indexing and retrieval system

Queries in ActiveMath are processed as follows: first, a client component passes a query formulated in its own schema to the mediator. The mediator translates the query into the mediated schema and then for each knowledge source from the mediated schema it translates the query into its respective schema. The queries are then sent to each knowledge base, and the answers (lists of identifiers) are merged. Finally, the mediator sends the answers back to the client.

The mediated schema is based on an ontology of instructional objects. Because existing metadata standards such as IEEE LOM can not represent sufficient information about the sources for a completely automatic search, we developed an ontology (see Fig. 2.) that describes different types of learning objects from an instructional point of view [7].

Central to the ontology is the distinction between *fundamentals* and *auxiliaries*. The class *fundamental* subsumes instructional objects that describe the central pieces of knowledge. *Auxiliary* elements include instructional objects which provide additional information about the concepts.

3 Usage of Multi-media Learning Objects

Once the LOs have been semantically annotated, *efficient* usage must be targeted as in large-scale learning environments individual support is nearly impossible. For instance, the e-learning lab of the Chinese co-authors counts 18,000 registered students. Therefore, as much support as possible must be provided automatically and the following section shows how to select LOs automatically. Section 3.2 describes how to answer students questions automatically. Section 3.3 sketches a solution for a multi-lingual environment, where special care needs to be taken with MMLOs in order to keep them re-usable.

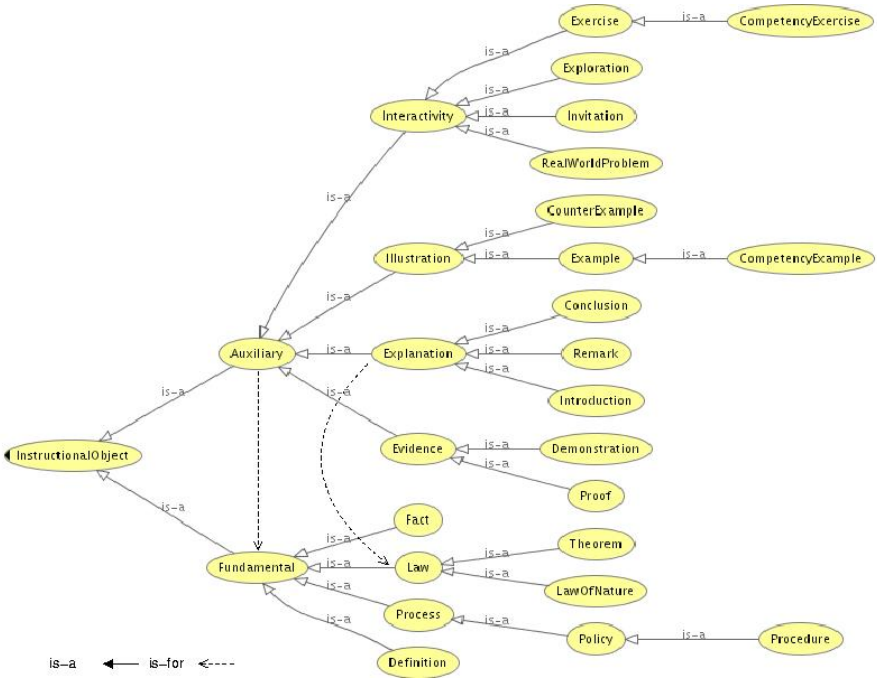


Fig. 2. Overview on the ontology of instructional objects

3.1 Intelligent Selection of Learning Objects

Students need to work through sequences of LOs to achieve their learning goals and these sequences are often manually assembled. Yet, it is impossible to foresee all potential paths towards a learning goal and to compose the corresponding courses in advance.

That is where course generation comes into play. A course generator automatically assembles learning objects into larger units, which support the learner to reach a given learning goal. The pedagogical decisions involved in such a task are complex and require an elaborate representation of the involved pedagogical knowledge.

ActiveMath uses a hierarchical task network (HTN) planner for course generation, which is an efficient planning technique that offers a relatively straight-forward way to represent human expert knowledge [8]. The HTN-planner has also heuristic knowledge in the form of decomposition rules: A planning problem is represented by sets of tasks; methods decompose non-primitive tasks into sub-tasks until a level of primitive tasks is reached, which can be solved by the given operators.

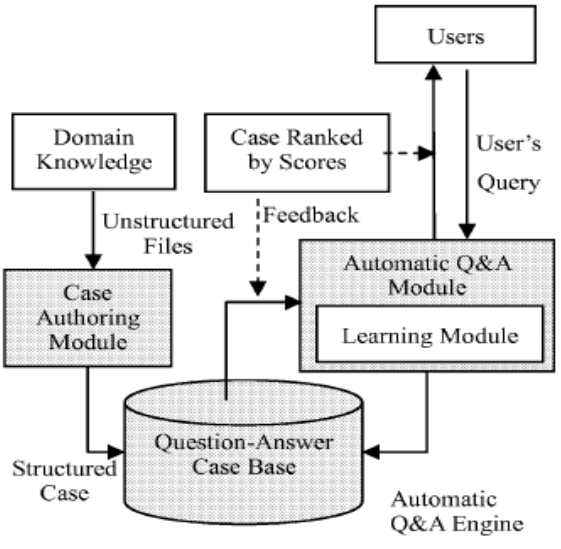


Fig. 3. Architecture of the Q&A System

A pedagogical task is defined as a tuple $t = (l, c)$, where l is a tutorial objective and c a unique identifier of a learning object (content goal). While c specifies the concept the course will primarily target, the tutorial objective determines the kinds of learning objects selected for c. A top-level task that serves as the starting point of course generation is *teachConcept id*. The goal of this task is to assemble a structured sequence of learning objects that help the learner to understand the content goal c. Using different collections of tasks and methods (i.e., different tutorial strategies), this task can be planned differently. Hence, the task-based approach can serve to represent a variety of pedagogical strategies.

Because tasks serve to represent a vast range of pedagogical goals, the size of the generated courses can range from a single element to a complete curriculum. For instance, while the task *teachConcept* results in sequences of several learning objects, other tasks may be achieved by a single element. Frequently occurring examples are tasks for exercise and example selection.

Tasks are also used to provide short term support to the learner by pedagogical agents. These agents monitor the learner's behaviour and, if they diagnose a potential learning problem, offer suggestions what content the learner should read to overcome the problems [9].

3.2 Intelligent and Automatic Question Answering

For the large-scale e-learning environment in China, it may take a teacher several hours to answer all the submitted questions. From our experience however, many questions, though put differently, usually have the same or a similar meaning. A solution is to share the answers among the students and let a computer recognize similar questions and then answer them automatically.

We have developed an interactive Q&A engine based on CBR as shown in Figure 3. This engine uses keywords (with weights) in the question to trigger a special case, which then has a standard answer. The weights of the keywords can be modified dynamically depending on feedback from the user.

If the computer cannot find an answer, it transfers the question to a teacher. After the teacher answers the question, the answer is added to the Q&A database and can now be shared among the students: as the Q&A database accumulates questions and answers, the hit rate grows over time.

3.3 Multi-lingual MMLOs

Applets and figures are an attractive and important part of multi-media content for Web-based learning environments. These LOs are sometimes interactive and they constitute an important ingredient to show the advantage of multi-media based learning versus traditional books.

One of the challenges is to make the representation of multi-media LOs multi-lingual as well and also to render the content according to the preferences of the learner. In ActiveMath textual LOs can have different texts for different languages. For applets we realized this 'internationalization' by separating the textual part from the applet. The different text for different languages can be stored in files attached to the same applet. This way, the presentation of the text can be adapted to the language of the current session while the visual applet part stays the same.

4 Conclusion

This paper describes some solutions to challenges regarding the representation and usage of MMLOs. The Chinese co-authors provided explanations how to deal with large-scale learning environments, while the German co-authors focused on semantic

representation and usage of such a representation. Further research will investigate how to marry these solutions, i.e., how to use semantic representation to support the needs of many efficiently.

Acknowledgements

Some of the authors were supported by the LeActiveMath project, funded under the 6th Framework Program of the European Community (Contract Nr IST-2003-507826). The authors are solely responsible for its content, it does not represent the opinion of the European Community and the Community is not responsible for any use that might be made of data appearing therein.

References

1. Dublin Core Metadata Initiative Usage Board. DCMI Metadata Terms Specification. DCMI, 2005.
2. IEEE Learning Technology Standards Committee. 1484.12.1-2002 IEEE Standard for Learning Object Metadata. 2002.
3. IMS Global Learning Consortium. IMS Learning Design Specification. 2003.
4. Kohlhase M. OMDOC: Towards an Internet Standard for Mathematical Knowledge. In Proc. of Artificial Intelligence and Symbolic Computation, AISC'2000, Eugenio Roanes Lozano (eds.), Springer Verlag, 2001.
5. Melis E et. al. Knowledge Representation and Management in ActiveMath. Annals of Mathematics and Artificial Intelligence, Special Issue on Management of Mathematical Knowledge. 2003, 41(1-3):47-64.
6. Doan A, Noy N F, Halevy A Y. Introduction to the special issue on semantic integration. *SIGMOD Rec.*,2004, 33(1):11-13.
7. Ullrich C. The learning-resource-type is dead, long live the learning-resource-type!. *Learning Objects and Learning Designs*, 2005, 1(1):7-15.
8. Erol K, Hendler J, Nau D. HTN Planning: Complexity and Expressivity. In Proc. of the Twelfth National Conference on Artificial Intelligence (AAAI-94), 1994.
9. Melis E, Andres E. Global Feedback in ActiveMath. *Journal of Computers in Mathematics and Science Teaching,* 2005, 24:197-220

An Intelligent Platform for Information Retrieval*

Fang Li[1] and Xuanjing Huang[2]

[1] Dept.of Computer Science and Engineering, Shanghai Jiaotong University, Shanghai 200240 China
[2] Dept of Computer Science and Engineering, Intelligent Information Processing Lab, Fudan University, Shanghai 200433 China
fli@sjtu.edu.cn, xjhuang@fudan.edu.cn.

Abstract. Information Retrieval (IR) has played a very important role in our modern life. However, the results of search engines are not satisfactory for human intelligent activities. The platform proposed in this paper tried to solve the problems from three aspects: One is to provide domain specific IR, which consists of task oriented searching and topic related filtering. The second is to provide open-domain, concept-based retrieval to reduce irrelevant pages and overcome ambiguous keywords. The third is to provide the exact answer by question answering. The intelligent platform will facilitate searching ability on the Web. It will be easier for users to locate the interest information by keywords or questions. Researches described in the paper are developed at Shanghai Jiaotong universities and Fudan university. The experiments have showed a promising result from each aspect. The integration of these three aspects is one of the challenges of IR in the near future.

Keywords: Information Retrieval, Web-based service, Web Search, Text Filtering, Question Answering.

1 Introduction

Since the Internet has become a rich information resource, a huge amount of people search the Internet every day for information of interest such as all kinds of news, goods, papers and so on. How to make effective use of the information available on the Internet is the main issue of Web-based information retrieval and information extraction.

Although there are many search engines like Google, Yahoo and even many meta-search engines like Vivisimo, it is sometimes not easy to find the information one is interested in. One reason for this is that many users search the Internet with keywords while the search engines are mostly based on various keyword matching and ranking algorithms. Most keywords are ambiguous in the rich semantic world. Therefore, search engines produce low precision and recall. The other reason is that the user needs to view a lot of relevant documents returned by search engines in order to find one's

* Supported by the Committee of Science and Technology of Shanghai Municipal Government under Grant No.045107035, and the Natural Science Foundation of China under grant No. 60435020.

R. Lu et al. (Eds.): Chinese-German Ws on Cognitive Systems 2005, LNAI 4429, pp. 45 – 57, 2007.
© Springer-Verlag Berlin Heidelberg 2007

specific information. It costs time and labor. For example, someone searches for information about the city *New York* and queries a search engine with the keywords "New York" he or she will get pages belonging to different areas of interest like pages about the city itself, about the newspaper *The New York Post*, about the basketball team *New York Knicks* and so on. In this case, users will have to either look through a lot of irrelevant pages, or refine their queries to better match the topic he is interested in.

Our Intelligent Platform will solve these problems from three aspects. One aspect is to provide domain specific IR, the second is to provide open domain, concept-based IR and the third is to facilitate IR by question answering (QA). In the following, these three aspects will be discussed in details.

2 Domain Specific IR

Domain specific IR means that the task or the topic has already been given. In the following, two examples are presented to describe how the platform supports task oriented IR.

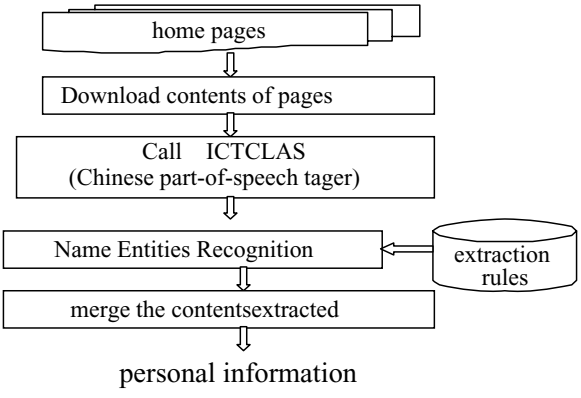


Fig. 1. Process of extraction

2.1 Experts Information Searching

We implemented an experimental system for homepages on the Internet, which combines IR and information extraction (IE). There are three steps to automatically extract personal information on the Internet:

1. Searching homepages: crawl the Internet for any personal homepages using URL features and content features. Many URLs of personal homepages carry the special sign "~"or some special strings like "homepage", "user=", and so on. Content features refer to keywords appearing in homepages like "research interest", "publication" and so on.
2. Identifying experts in a specific domain: filter homepages which are not in the Computational Linguistics domain based on two level keywords. Some top 50 words on the high frequency list have been chosen as the first level keywords. Some terms

in the domain of computational linguistics have become the second level keywords, such as, "information extraction", "question answering" and so on.

3. Extracting personal information: acquire personal information such as name, title, affiliation and so on based on predefined rules. The extraction process is depicted in Fig.1[1].

A prototype has been implemented in JAVA with the average precision of 82% [1] for our virtual information center [2].

2.2 Adaptive Text Filtering

The task of text filtering is to monitor a document stream to find those documents that are of interest to the user. The system is given a brief description of the topic and a few training documents, and then proceeds to look for documents on that topic. As the document stream is processed, the system may be provided with relevance judgments for the retrieved documents. An adaptive system can learn from this feedback, so that it becomes more and more accurate over time.

Our research on filtering focuses on how to create the initial filtering profile and set the initial threshold, and then modify them adaptively. Fig.2 shows the architecture of the training in adaptive filtering. At first, feature vectors are extracted from positive and pseudo-positive document samples. The initial profile is the weighted sum of positive and pseudo-positive feature vectors. Then we compute the similarity between the initial profile and all the training documents to find the optimal initial threshold for every topic.

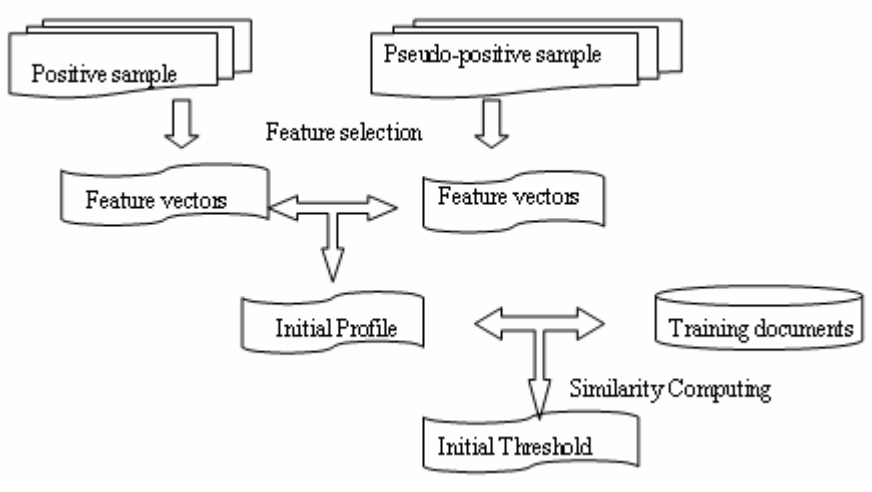


Fig. 2. Architecture of the training in adaptive filtering

Since the total number of all words is very large and it costs too much time in similarity computation, we decided to select some important words from them. First,

[1] Chinese part-of-speech tagger is from the Chinese academy of Science, Beijing, China.

we carry out a morphological analysis and stop word removing. Then we compute the logarithm Mutual Information between remaining words and topics:

$$\log MI(w_i, T_j) = \log\left(P(w_i \mid T_j) \Big/ P(w_i) \right) . \tag{1}$$

Where, w_i is the ith word and T_j is the jth topic. Higher logarithm Mutual Information means w_i and T_j are more relevant. P(w_i) and P($w_i|T_j$) are both estimated by *maximal likelihood method*. For each topic, we select those words with logarithm Mutual Information higher than 3.0 and occurs more than once in the relevant documents. Logarithm Mutual Information is not only used as the selection criterion, but also as the weight of feature words.

The similarity between the profile and training documents is computed by the cosine formula:

$$Sim(d_i, p_j) = Cos\theta = \sum_k d_{ik} * p_{jk} \Big/ \sqrt{(\sum_k d_{ik}^2)(\sum_k p_{jk}^2)} . \tag{2}$$

Where, p_j is the profile of the jth topic and d_i is the vector representation of the ith document. d_{ik}, the weight of the kth word in d_i, is computed as such:

$$d_{ik} = 1 + \log(tf_{ik} * avdl / dl) . \tag{3}$$

where tf_{ik} is the frequency of the kth word in the ith document, *dl* is the average number of different tokens in one document, *avdl* is the average number of tokens in one document.

Each topic profile is represented by a vector which is the weighted sum of the feature vector from positive (relevant) documents and the feature vector from pseudo relevant documents with the ratio of 1: X_0.

To make use of the hierarchy of categories, those documents of the same high-level category are considered as pseudo relevant documents. Since the number of low-level categories is different among different high-level categories, we set different X_0 for different categories. After combining the positive and pseudo-positive feature vectors, we get the initial profile. Once the initial profiles are acquired, the initial thresholds should be set to those values that can lead to better performance, for example, T10F and T10SU [3].

2.2.1 Adapting Threshold and Profiles

For adaptive filtering, we use an adaptation procedure to modify the initial profile and threshold during filtering documents. Fig.3 shows the architecture for the adaptation.

1) Adjustment of threshold

We adjust the threshold once a positive document is retrieved. Let:

- *t*: denote the sequence number of document, since the documents are processed by temporal order, *t* also can be considered as time.
- *n*(*t*): denote the number of documents processed up to *t*.

- $n_R(t)$: denote the relevant documents retrieved up to t.
- $n_N(t)$: denote the irrelevant documents retrieved up to t.
- $T(t)$: denote the threshold at time t.
- $S^-(t_{k+1}, t_k)$: denote the average similarity of the document been rejected in (t_k, t_{k+1}) interval.
- $P(t_{k+1}, t_k)$: denote the precision of system in (t_k, t_{k+1}) interval, here

$$P(t_{k+1}, \ t_k) = n_R(t_{k+1}) - n_R(t_k) \Big/ n(t_{k+1}) - n(t_k) \ . \tag{4}$$

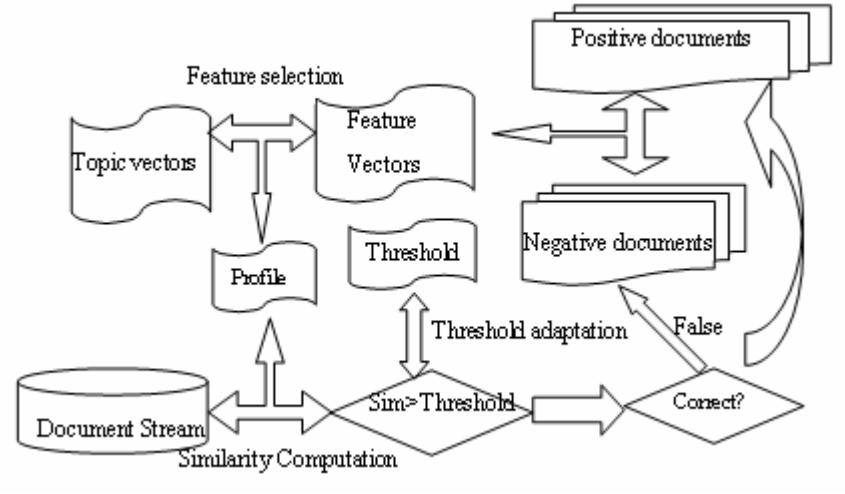


Fig. 3. Architecture for the adaptation

Intuitively, we should increase the threshold if the precision is too low and lower the threshold if too few documents are retrieved. So we can use $S^-(t_{k+1}, t_k)$ and $P(t_{k+1}, t_k)$ to decide whether to increase or decrease the threshold. When the precision is lower than expected, we should increase the threshold. Otherwise, we can decrease the threshold. In particular, when the threshold is too higher than the similarity with the rejected documents, the threshold should be decreased quickly.

The above strategy of threshold adjusting can be written as below:

If $P(t_{k+1} - t_k) \leq EP(t_{k+1})$ then

$$T(t_{k+1}) = T(t_k) + \alpha(t_{k+1}) * (1 - T(t_k))$$

Else If $S^-(t_{k+1} - t_k) < T(t_{k+1}) * D$

$$T(t_{k+1}) = T(t_k) * A + S^-(t_{k+1} - t_k) * (1 - A)$$

Else $\quad T(t_{k+1}) = (1 - \beta(t_{k+1})) * T(t_k)$

Where $\alpha(t_k)$ is the coefficient for increasing the threshold and $\beta(t_k)$ is the coefficient for lowering the threshold, both of them can be considered as the function of $n_R(t)$.

In our experiment, we use the following linear functions shown as:

$$\alpha(t_k) = \begin{cases} \alpha_0 * (\mu - n_R(t_k))/\mu, & n_R(t_k) \le \mu \\ 0, & n_R(t_k) > \mu \end{cases}. \tag{5}$$

$$\beta(t_k) = \begin{cases} \beta_0 * (\mu - n_R(t_k))/\mu, & n_R(t_k) \le \mu \\ 0, & n_R(t_k) > \mu \end{cases}. \tag{6}$$

Where α_0 and β_0 are the initial parameter. The parameter of μ indicates the maximum number of positive documents should be used to adjust the threshold and modify the profile. Here we set $\alpha_0 = 0.02$, $\beta_0 = 0.1$ and $\mu = 300$.

The introduction of parameter D aims at increasing the recall. Since the actual number of relevant documents of every topic cannot be observed, we can only acquire some indirect estimation. We believed when the average similarity between the profile and those rejected documents are too small, the similarity threshold should be decreased in order to enhance the recall. In our experiment, we set $D = 0.1$ and $A = 0.8$.

$EP(t_k)$ means the precision, which we wish the system to reach. At first, we regarded this parameter as constant and tried several different values, but the results are not very satisfactory. Since it is impractical to require the system to reach the desired high precision at the beginning of filtering, we adopt a gradual-ascent function. The function is showed as:

$$EP(t_{k+1}) = \begin{cases} P_0 + (P_{final} - P_0) * n_R(t_{k+1})/\mu, & n_R(t_k) \le \mu \\ 0, & n_R(t_k) > \mu \end{cases}. \tag{7}$$

Where, P_0 and P_{final} are the desired initial and final precision. In our experiment, $P_0 = 0.2$ and $P_{final} = 0.6$.

2) Adaptation of profile

Once a retrieved document has been judged relevant, it is added to the positive document set otherwise it is added to the negative document set. During profile adaptation, feature vectors are extracted from positive documents and negative documents. The new topic profile is the weighted sum of the feature vector from positive documents and negative documents with the ratio of $1:X_1$ (Here $X_1 = -0.25$). For effectiveness and efficiency reason, we adjust the topic profile only after $L(L = 5)$ positive documents have been retrieved.

2.2.2 Evaluation Results

We have participated in the TREC's filtering track for 3 years. Table 1 summarizes our adaptive filtering runs in TREC9. Four evaluation criteria are calculated, including T10SU, and T10F [3]. Underlined value means that the run is optimized for the

Table 1. Adaptive Filtering Results

Run	T10SU	T10F	Comparison with median		
			>	=	<
FDUT10AF1	0.215	0.404	64	5	15
FDUT10AF4	0.213	0.414	71	4	10

corresponding criterion. The last columns give the number of topics in which our runs perform better, equal and worse than median ones according to the criteria for which our runs are optimized.

3 Concept-Based Open-Domain IR

Currently, most concept-based information retrieval systems are using some kind of a conceptual structure like an ontology or thesaurus and try to assign a user's query to a predefined category. In this paper we use a new notion of concepts. It needs neither a conceptual structure nor domain knowledge that requires human supervision. Concepts are defined as sets of words, which appear frequently in the context of a certain semantic meaning. They can be derived from their co-occurrence on results of search engines. The method consists of three steps: frequent word sets mining, concept deriving and concept itemset post-processing.

3.1 Frequent Word Sets Mining

The first step is to mine the word sets that frequently appear together. This task is equivalent to the Data Mining task of searching re-occurring patterns in data collections. Using Apriori Algorithm [4] with keyword vectors derived from the web pages as an input, all frequent word itemsets can be obtained.

3.2 Concept Itemset Deriving and Post-processing

The second step is to build concept itemsets from these frequent word itemsets. The algorithm for generating the concept itemsets works as follows: First we divide the collection of frequent word sets into two separate sets by taken a certain percentage of these itemsets with the highest confidence values and put them into one set (we call these sets the *seeds*), the remaining itemsets will form the second set (the *sprouts*). Our experiments showed the best result when taking the 70% of the frequent itemsets as *seeds*. Every element in the set of *seeds* is possible candidates for a concept itemset. Starting with the *sprouts* of cardinality two, each *seed* will be compared to each *sprout*. If they have a big enough overlap then the part of the *sprout* that does not overlap with the *seed* will be put in a separate set and after a *seed* has been compared to all *sprouts* this set will be added to the *sprout*. For a *sprout* of cardinality $\boldsymbol{k}$, the overlap to a *seed* is considered big enough if exactly $\boldsymbol{k-1}$ items of the *sprout* also appear in the *seed*. After all the *seeds* have been expanded with the *sprouts* of cardinality two, those itemsets that

have a superset among the other itemsets will be erased. Then we repeat with the next higher level of frequent sets. Finally, we erase redundant words within each set and then the sets, which have a superset. The pseudo code for the algorithm for building the concepts from the frequent itemsets is shown in Fig. 4.

Let I be the set of *sprouts* and h be the cardinality of the largest mined frequent itemset. Then the output Ω of the algorithm are the Concept Itemsets.

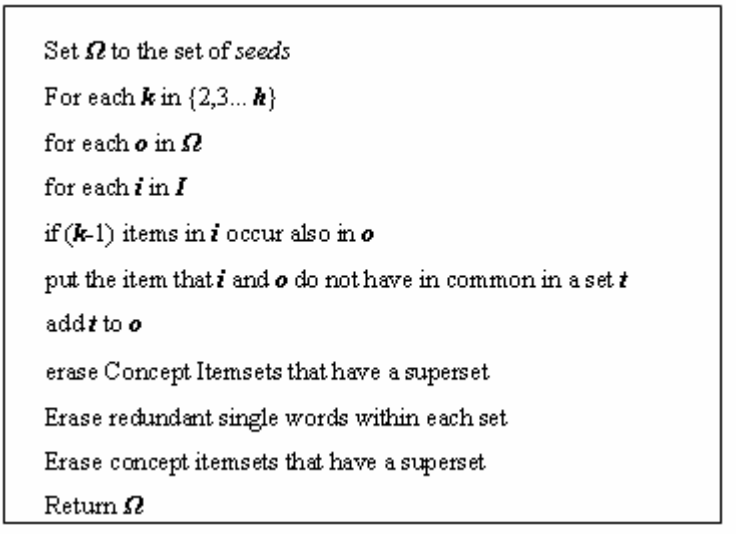


Fig. 4. Algorithm of concept generation

After generation, post-processing is needed. There are two tasks in post-processing: grouping similar sets together and then building sub-concepts by extracting overlap.

Let I1 and I2 be two itemsets with cardinality n1 and n2. Let further be n1≥n2. I1 and I2 are *similar* if:

$$\text{overlap}(I1,I2) \geq d*n2 \tag{8}$$

Where d is a real value between zero and one. In our experiment we set d=0.5.

Taking as an example someone who searches for information by using the query "Shanghai", the created concept itemsets before post-processing might look like this:

Concept1 = [china, city, east, guide, hotel, tour]
Concept2 = [business, china, city, east, economy, trade]
Concept3 = [airport, china, city, east, pudong]
Concept4 = [embassy, us]

After post-processing, the result will be:

Concept1 = [
Sub1 = [china, city, east]
Sub2 = [guide, hotel, tour]
Sub3 = [business, economy, trade]

Sub4 = [airport, pudong]
]
Concept2 = [embassy, us]

3.3 Concept-Based IR

We implemented a prototype system of mining concept itemset for IR. Given a query, this prototype system downloads the first 200 search results returned by the Google search engine and extracts keywords for each page using the publicly available KEA software[2]. Frequent itemsets are mined from these 200 keyword vectors with the Apriori algorithm and finally concept itemsets are generated based on the method we described.

In order to evaluate the precision of concept mining, queries to search engines are divided into three categories: ambiguous terms, named entities and general terms. We chose some samples for each category and evaluate how many useful concept itemsets

Query	Concept Itemsets	Precision @5	Precision @10	Precision @20
Matrix	neo	0.6	0.4	0.2
	film	1	1	1
	world	-	-	-
	dvd, news	0.4	0.4	0.3
	movie, reloaded, smith, wachowski	0.8	0.8	0.85
	linear algebra, matrices	0.6	0.8	0.9
	average	0.68	0.68	0.65
Population	contents alerting, information, journal, number table of contents, volume	1	1	0.8
	end, publication, science, springer	0.6	0.5	0.25
	bereau, census, statistics, u.s., world	-	-	-
	cities, countries, growth, human	0.8	0.8	0.8
	development, family planing, report, reproductive health	0.6	0.6	0.6
	center, demography, research, training	1	0.8	0.75
	united states	0.8	0.7	0.6
	average	0.8	0.73	0.63
New York	nyc, theater, world	0.2	0.1	0.25
	calendar, city, events	0.8	0.7	0.7
	books, travel	-	-	-
	business, health	-	-	-
	campus, faculty, school, students	1	0.9	0.95
	apr, season	-	-	-
	john paul ii, pope john paul	1	1	1
	information, ny, state	0.6	0.6	0.55
	average	0.72	0.66	0.69
total	***average***	0.73	0.69	0.66

Fig. 5. Precision of concept-based information retrieval

[2] www.nzdl.org/Kea

are produced. The average precision of our method – the number of meaningful word set over the total number of created word sets – is 75%. One query term from each category has been chosen for the experiment to evaluate the precision of IR with concept itemsets. For each of these queries, the concept itemsets were mined, then the pages were re-ranked with these itemsets and then we looked through the ranking results. We check how many of the first five, first ten and first twenty pages are related to the concept that is reflected by the concept itemset. The precision for each concept itemset of the three queries is shown in Fig.5. It shows that values deviate only slightly from the average for all three queries. Our method can deal with a wide range of queries and the result can cover a variety of subtopics even if pages returned to a query have a greatly varying vocabulary. The precision values are in average between 60% and 70%. This result is comparable to the method where pages are clustered with salient phrases [5].

4 Intelligent Question Answering on the Web

Typical QA system can be divided into three modules [6]. The first module is question analysis, where questions should be translated into query words, and the type of answer is estimated. The second module is passage retrieval, where some relevant passages are provided for further analysis. The final module is answer extraction, where the exact answers are located from relevant passages.

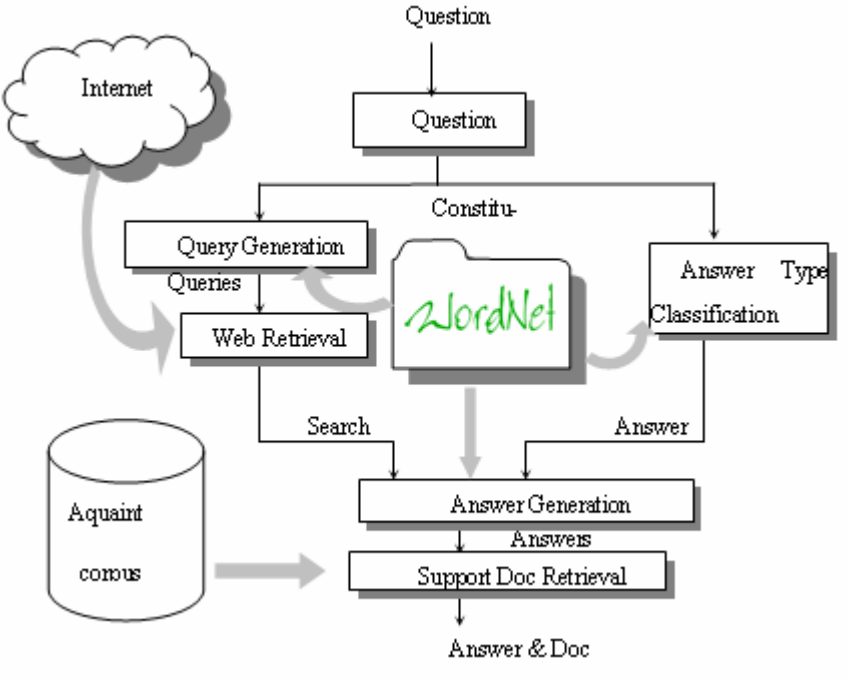


Fig. 6. Flow Chart of FDUQA on Factoid Question

4.1 Algorithm for Factoid QA

Factoid question means those questions, which can be answered succinctly and exactly, for example, *"When was the Hale Bopp comet discovered?"* Fig. 6 describes the process of our factoid question answering, and the details about each module will be introduced in the following.

4.2 Process of the QA

1) Question analysis
During question analysis, constituents are extracted from the question with the help of LinkParser [7], an English parser based on link grammar. For example: *"When was the Hale Bopp comet discovered?"* Its constituents are: *"the Hale Bopp comet"* – subject; *"was discovered"* – predicate. Constituent information, as we will see in the following, is used in answer type classification, query generation, as well as answer generation.
2) Answer type classification
In this step, our system determines the answer type of the input question based on some ordered rules obtained by machine learning. We adopt a thirty-one-class answer type classification system, illustrated in table 2. Different types of the questions are treated differently in the following answer generation module.

The classifier and focus word decision algorithm are both based on Transformation-Based learning (TBL) [8]. TBL is an approach to corpus-based natural language processing. In this approach, the learned linguistic information is represented in a concise and intelligible form. There are two components of a transformation: a rewrite rule and a triggering environment.

Table 2. Answer Type Concepts

NAMEBASIC	PRN	LCN	ORG
PIECEOF-WORK	QUOTATION	POSTADDR	ABBR
TIM	DAT	NUMBASIC	NUMBER
ORDINAL	AGE	MEASURE-BASIC	LENGTH
PCT	MNY	INTEGER	CODEBASIC
URL	TELEPHONE	POSTCODE	EMAILAD-DRESS
BNP	DESP_OF_ABBR	TRANSLATION	MANNER
REASON	CONSEQUENCE	OTHER	

By using TBL, a rule-based question classifier is developed to determine the question focus and answer type. The question focus is a word or a phrase, which indicates what the question is looking for, or what the question is all about. After searching the Wordnet and getting the focus word's sense, all its Hyponyms in WordNet and other information are input into another TBL classifier, then we get the category of a question.
3) Query Generation and Web Retrieval
We find answers from Internet using Google search engine. Query generation is subject to the characteristics of Google Search such as phrase search. To utilize this charac-

teristic, constituents are combined together. For example, *"When was the Hale Bopp comet discovered?"* has constituents: *"the Hale Bopp comet"* – subject, *"was discovered"* – predicate. Queries for this question are: *"the Hale Bopp comet was discovered"* and *"the Hale Bopp comet" "was discovered"*.

Upon the basic queries, some query expansion are also done according to the following rules:

- Synonym expansion – to replace the noun in the query with its synonym, which is found in WordNet.
- Preposition expansion – to add some prepositions to queries when the question asks for location or time.
- Unit expansion – to add some units to queries when the question asks for measure.

4) Answer Generation and Document Retrieval
Answer generation is based on the answer type categorization. The support documents are retrieved in Aquaint Corpus with the answers obtained and the question target as queries. By measuring the distance between the answer and key words, we will get the best answer and its support document.

4.3 Experimental Results

We submitted the results of our QA system to TREC13 and our system ranked the fifth among all, a detailed description can be found in [9]. Therefore, we can draw the conclusion that the algorithm we use to answer factoid questions is quite promising.

5 Conclusion

The paper introduced an intelligent platform for information retrieval[3,4], where the search performance can be enhanced with three different techniques. One is to support task oriented IR using adaptive text filtering. The second is to implement an open domain concept-based IR. Concepts are derived from frequent word sets, which are mined directly on results of search engines. The third is to facilitate users by getting the exact answer using QA. The integration of these three aspects is considerably more challenging in the near future. Some problems still remain:

1. How to represent the information need or the user goal is a key point in information retrieval. Query expansion can not solve all problems now existing in IR. The query, whether a simple bag of conjoined keywords or complex Boolean filter, essentially specifies the indices that should be examined to find matching documents. Keywords-based methods will make retrieval vulnerable to the word mismatch problem if the authors have chosen to lexicalize their ideas in a different way [10].
2. Concept–based IR can solve ambiguous keywords. Relying on a conceptual structure, the precision of concept-based IR will be higher than the method we use, but it may have some constraints when encountered unknown keywords or

[3] Research on 2.1 and 3 are done at Jiaotong University.
[4] Research on 2.2 and 4 are developed at Fudan University.

new words. Concepts derived directly from frequent word sets can be used in any domain and any searching tasks, however some mined concept itemsets are meaningless. How to erase those meaningless words sets and increase the precision of concept building is a big obstacle.

3. Question answering and text filtering can help people to locate their required information on the Internet. How to combine QA with search engines or how to use text filtering in ranking documents is still a big open challenge in IR

Our future research will focus on the following:

1) Implement a user model to describe all kinds of searching tasks and define their information need.
2) Increase the precision of concept-based information retrieval by integrating some existing general ontologies like Wordnet, Hownet and so on.
3) Make TREC experimental systems applicable in real life.
4) Discover all kinds of knowledge for different real applications based on the platform.

References

1. Fang Li, Huanye Sheng.: Personal Information Extraction from their Homepage. In the Proceeding of 8th JSCL 2005, NanJing China.
2. Fang Li, et al.: Virtual Information center for human interaction in the *Human Interaction with Machines* edited by Hommel, G. and Huanye Sheng, Proceedings of the 6th International Workshop held at the Shanghai JiaoTong University, March 15-16, 2005 by Spring Verlag.
3. S. Robertson, D. Hull.: The TREC-9 Filtering Track Final Report In Proceeding of the Ninth Text Retrieval Conference (TREC-9) Feb. (2001)
4. Agrawal R, Imielinski T, Swami A.: Mining association rules between sets of items in large databases. In the Proceeding of the 1993 ACM SIGMOD International Conference on Management of data.
5. Chen Z, He Q, Ma J, Ma W, Zeng H.: Learning to cluster web search results. In the Proceeding of ACM SIGIR (2004)
6. Hirschman.: Natural Language Question Answering, the view from here, Journal of Natural Language Engineering, Special Issue on Question Answering, Fall-Winter (2001)
7. Daniel Sleator and Davy Temperley.: *Parsing English with a Link Grammar.* In the Proceeding of Third International Workshop on Parsing Technologies. (1993)
8. Brill, Eric.: Transformation-Based Error-Driven Learning and Natural Language Processing: A Case Study in Part of Speech Tagging. Computational Linguistics, Dec. (1995)
9. E.M. Voorhees.: Overview of the TREC 2004 Question Answering Track, in the Proceeding of the Thirteenth Text Retrieval Conference (TREC-13) Feb.(2005)
10. Tony Veale.: The challenge of Creative Information Retrieval in A. Gelbukh (Ed.) Computational Linguistics and Intelligent Text Processing Lecture Notes in Computer Science, Vol. 2945. Springer-Verlag, Berlin Heidelberg New York (2004) 457-467

P-Terse: A Peer-to-Peer Based Text Retrieval and Search System
(Extended Abstract)

Weining Qian[1,2], Feibo Chen[1], Bei Du[1], and Aoying Zhou[1,2]

[1] Department of Computer Science and Engineering, Fudan University
[2] Intelligent Information Processing Laboratory, Fudan University
{wnqian, chenfeibo, beidu, ayzhou}@fudan.edu.cn

Abstract. P-Terse, a peer-to-peer (P2P) text retrieval and search prototype system is introduced in this paper. Compared with existing P2P systems, P-Terse has three novel features: 1) The text content of the shared documents is searchable. 2) The system is open for extensions. 3) Our search and query processing techniques are implemented in the system. These techniques are designed for achieving high efficiency and scalability. The presentation of the system includes the design strategies of the system and the technologies that are implemented. We also discuss the on-going research and development work related to P-Terse.

1 Introduction

With the maturity of data management and Internet technology, providing complex data query and retrieval functions in a network environment becomes a hotspot of both industry and research community nowadays. Text search is one of the most popular complex query and retrieval functions. The success of several popular search engines enable users to search billions of documents world-widely. Recently, some efforts have been paid to equip the personal computers with light-weight search engines, which are called *desktop search* tools. The desktop search tools have the advantage that they are highly customizable, and are being daily accessaries to traditional information management tools such as tree-style file or resource managers and email clients. However, user requirements often lie in-between of world-wide general-purpose search engines and desktop search tools.

We present our effort on providing text search and retrieval in a network environment without a centralized server in this paper. We developed a prototype system, called P-Terse (for **P**eer-to-peer **TE**xt **R**etrieval and **SE**arch). Each node participating the system is equipped with a local text retrieval engine. The nodes are organized in a peer-to-peer (P2P) scheme. Thus, the system does not require a centralized file server.

Peer-to-peer computing emerges as a new distributed computing paradigm. It has advantage over previous models. First, it does not require a central server so that single point failure would not happen. Furthermore, nodes are equivalent, which means each node can serve as a server or client, while nodes have

R. Lu et al. (Eds.): Chinese-German Ws on Cognitive Systems 2005, LNAI 4429, pp. 58–69, 2007.
© Springer-Verlag Berlin Heidelberg 2007

autonomy, which means a node can join or leave the system on will. Third, personal computers are majority participants in most P2P systems. The resource of those computers are usually not fully utilized. P2P computing model provides a simple yet efficient way for sharing the edge resources. Last but not the least, since large amount of nodes are allowed to participate the system, P2P systems usually have extraordinary storage or processing capability that is hard to be achieved by centralized systems or traditionally distributed systems.

P2P computing has succeeded in file sharing and instant message applications. Several most popular softwares used over Internet are P2P systems, including file sharing softwares, and instant message systems. Current technologies used in these systems are mainly for searching resource based on their addresses or resource names. Thus, they are not suitable for our text search and retrieval task.

P-Terse is not the first effort to support text retrieval in a peer-to-peer system. Previous work include PSearch, which is mainly designed for providing similarity search for vector space model [1] over a structured P2P network [2], and PlanetP, which tends to provide text search in unstructured P2P systems [3]. However, our work has different objectives. First, we focus on extensibility and open of the system. Since P2P is a fast evolving technology, we hope our system is open for function extensions and implementation extensions for more advanced techniques. Secondly, P-Terse is not a pure research project. Our aim is to develop a *runnable* system that users can share and search the *content* of the documents. Therefore, we try to choose techniques that are effective, simple and compatible to other ones in the development. Last but not the least, we try to make the system a testbed for our researchwork.

The rest part of this paper is organized as follows. The system architecture is introduced in Section 2. In Section 3, the core technologies employed in P-Terse are introduced, including our research result on overlay network, message routing, search and query processing, and distributed management. The related projects are introduced in Section 4. Finally, Section 5 is for discussion on our future work and for concluding remarks.

2 System Architecture

The architecture of each node in P-Terse is shown in Figure 1. There are five main modules:

Overlay communication module. Being the lowest level of our system, it is responsible for handling the communication with other nodes in the system. Two types of communication are supported, i.e. message-based and large object based communication. The communication supported in this level is *identity-based*.

Search module. Similar to overlay communication module, search module is in charge of the communication with other nodes. The difference is that search module is *content-based* or *semantic-based*. To achieve this purpose,

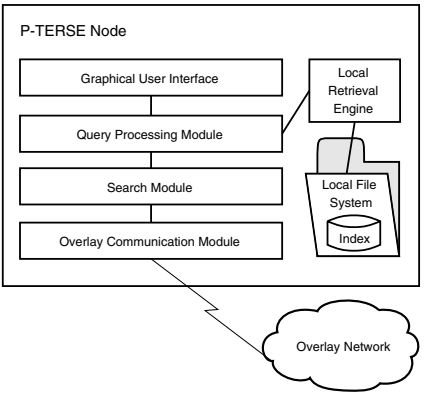


Fig. 1. The architecture of P-TERSE

it maintains summary information about its local corpus and its neighboring nodes. When a search request arrives, it determines whether its local corpus need to be searched and which neighboring nodes should the query be forwarded to.

Query processing module. This is the control center of a P-TERSE node. It receives query from user, preprocesses the query, and forwards it to local retrieval engine and search module. It also merges results sent back by local retrieval engine and other nodes.

Local retrieval engine. This module is responsible to the maintenance of index on local corpus that is stored in the local file system. When a query is received, it searches the index and returns matched results.

Graphical user interface (GUI). GUI accepts configuration commands and queries from user, and shows the configuration setting and query results in a graphical way.

Our main research focus is on the first three modules listed above. Lucene, an open source text retrieval engine is adopted as the local retrieval engine. We extends Lucene with the function of file system monitoring and logging for maintenance purpose. As for the GUI, it is designed as Web based, and developed upon Apache HTTP server.

2.1 Protocols

Overlay Interface. The overlay interface provides a separation of the underlying overlay protocol (address-based) and the search strategy (content-based). Basically, two kinds of application program interfaces (APIs) are supported.

Maintenance APIs. This kind of APIs includes the interfaces for *join*, *leave*, and *re-stablization* of the overlay network.

Data transmission APIs. Currently, two data transmitting APIs are supported, including **sendMsg**(SID *sourceID*, PID *destID*, *string msg*) and **sendFile**(SID *sourceID*, PID *destID*, $File$ *file*). Note that since files are usually of larger sizes than messages, we use a different interface to support file transmission. Meanwhile, a **listen**() interface is provided for any node ready to receive data from other nodes.

Search Interface. Search interface takes four types of parameters. *Keywords* are the list of terms interested by the user. *Search parameter* is the search constraint on the documents to be searched, including document type, document size, and creation and modification date or time etc. *Search scope* is the scope of network to be searched. Currently, two types of scopes are supported, i.e. *neighborhood* and *all*. At last, *quality of service (QoS)* indicates the constraints on the quality of the search. QoS parameters include the response time, number of results to be returned, and the similarity of the result documents to the query. QoS parameters have direct impact on the search strategy in search module and routing configuration in overlay communication module.

Local Indexing and Text Retrieval Engine. The text retrieval engine accepts keyword-based queries, and returns result documents most similar to the query. This kind of interfaces is provided by Lucene directly. Lucene also accepts indexing request indicating the location of documents to be indexed. We extend Lucene's interfaces with maintenance interfaces to satisfy our requirements. The user can set the time interval to re-scan the file system for detecting the creation, deletion or modification of the documents. Furthermore, the access control level of the shared files can be set via the interfaces.

User Interface (non-graphical). The user interface provides a set of APIs for setting system parameters, setting indexing configurations, and posing queries.

3 Core Technologies

3.1 Overlay Network

The overlay network is responsible for the maintenance of the connectivity, locating of peers, routing to the destinations and data transmitting. It is totally identity based. There are several popular overlay network protocols, including Chord [4], CAN [5], Pastry [6], and Tapestry [7].

In P-Terse, an adapted version of C^2 is implemented. C^2 is an overlay network designed to optimize the routing efficiency [8]. The overall C^2 identity space is organized as a multi-dimensional torus, while each node maintains one Chord-alike finger table on each dimension. P-Terse uses a two-dimensional C^2 as the overlay network, which is illustrated in Figure 3.1.

The basic routing strategy of C^2 is to follow the link in current finger table that jumps to the node closest to destination. Given a network with N nodes,

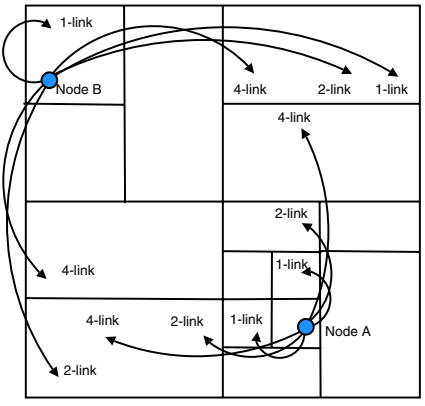


Fig. 2. A C^2 overlay network used in P-Terse, with finger tables of two nodes illustrated

this routing strategy can achieve a routing path with maximally $O(\log N)$ hops for any pair of nodes.

The original C^2 protocol is range-based, which means each node is identified by the range it maintains. However, nodes are identified by their identities in P-Terse. Therefore, a little modification is applied in our implementation. First, the identity of a node, which is a 128-bit bit-string, is divided into two 64-bit bit-strings. Each bit-string is treated as a number, so that one point in the two-dimensional torus is determined. Second, a range is assigned to a node if the node's identity (point) locates in the range. Third, when two nodes' identities locate in the same range, the range is partitioned on the dimension where two nodes' identities (points) have maximum difference. Furthermore, two points should have same distance to the partition line. We believe that the modification would not affect the routing efficiency when the nodes' identities are normally distributed.

In P-Terse, C^2 is used for two purposes. First, C^2 is used to maintain the connectivity of the nodes. When a new node joins the system, or a node finds it cannot connect certain nodes, it invokes the API provided by C^2 to refresh its finger table. So that all nodes are guaranteed to be connected to other nodes. The second task that C^2 takes is keyword indexing. In information retrieval, both keyword weighting and result ranking need information of global statistics of keywords. The keywords are used as keys, while the document that containing it and the frequency of its occurence are treated as values of the keys. These information is published in a traditional DHT scheme over C^2. And they can be queried by any peer. The details of indexing scheme is introduced in [9] and demonstrated in ICDE 2006 [10].

3.2 Search Algorithm

Different from the overlay network, which is identity-based, in the search module, the nodes are located and visited based on their content.

According to the study on real-life P2P networks, the networks tend to have the *small-world phenomena* [11]. The small-world networks are intensively studied in physics community [12,13]. Two major characteristics of small-world networks are 1) each pairs of nodes are connected by short path(s), and 2) the nodes are highly clustered. Furthermore, Barabasi et al. stated that real-life small-world networks are scale free networks, which means the degrees of vertices (nodes) satisfy the power-law distribution [13]. Kleinberg introduces an algorithm to construct small-world networks in which the short path between two nodes can be found [14]. However, this search algorithm is identity-based, which is not feasible in our content-based context.

Two natural search strategies to traverse the network are *depth-first search (DFS)* and *breadth-first search (BFS)*. However, both have obvious shortcomings. The former is infeasible in a large scale network because the trace back process costs too much overhead, and results in too long response time. The later would cause message flooding. Therefore, they neither can be directly applied in our search module.

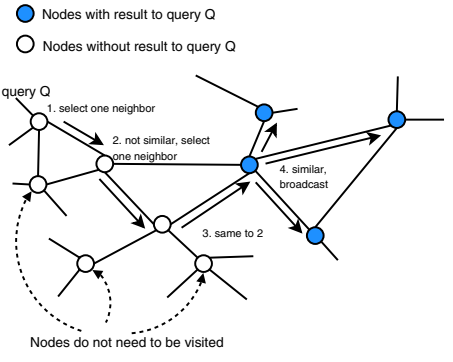


Fig. 3. The basic idea of SHINOV

We propose a heuristic based search strategy, called SHINOV, for **S**imilarity-based **H**eur**I**stic search with **NO**de-to-**V**isit control [15]. Its basic idea is to forward a query to more neighbors when the query arrives at a node *deep in a cluster*. This process is illustrated in Figure 3.2.

The search module maintains its own list of neighboring nodes. The nodes in the list are maintained dynamically. Each entry in the list consists of the node's identifier and its physical address, i.e. IP address in current implementation. Thus, the neighboring nodes can be visited without going through the overlay network. Furthermore, an online neighboring node is guaranteed to be available by the overlay network. When a node containing many high quality result to a query is found, it is maintained as a neighboring node. The number of neighboring nodes is specified by user.

To judge whether a node locates within a cluster, the similarity between the query and the local corpus is evaluated. If the similarity exceeds a predefined threshold, the query is broadcast to all neighboring nodes. If the similarity is too low, the node randomly selects one neighboring node to forward the query. Otherwise, the number of nodes the query to be forwarded to is in proportion to the similarity.

The similarity of a query to the corpus is estimated using the following method. First, the corpus is clustered into several several groups. Each group is represented by a vector which is the mean of all documents falling in that group. When a query arrives, it is first transformed to the vector representation of that node. Then, its similarity to the most similar representative vector is used as the similarity of the query and the corpus. Since the clustering process is conducted offline, and the number of groups is limited[1], this process will not affect the query processing efficiency.

Another novel feature of SHINOV is that the life time of a query in the network is controlled by *NOV* (**NO**de-to-**V**isit) instead of the number of hops called TTL (**T**ime-**T**o-**L**ive) that is frequently used in previous systems. NOV denotes the upper bound for the number of nodes to visit. The intuition is that the query processing cost, as part of the QoS, is composed of two parts, i.e. the response time and the network transmitting consumption. The former can be controled by the query initiator by stop receiving query results, while the later is determined just after the query is sent out by the query initiator. It is difficult to set a reasonable value of TTL, since the second part of the cost is determined not only by TTL but also by topology of the network. Furthermore, slightly difference on TTL can result in huge difference on the cost in real-life P2P networks. Taking Gnutella for example, studies show that using broadcast, seven hops can traverse the majority of millions of nodes in the network. This small-world characteristic limits the number of possible settings of TTL.

In SHINOV, a predefined NOV is assigned to a query when it is initiated. When a node is visited, and it is visited by the same query before, the NOV is sent back to the query initiator, and the search procedure stops. Otherwise, NOV is decreased by 1. For NOV larger than zero, it is uniformly assigned to all nodes the query is forwarded to. When NOV arrives 0, the node identity is sent back to register on the query initiator. As for the query initiator, if both extra NOV and registering node identity are received, the extra NOV is sent to the node with the identity so that the search process can continue from that registering node.

Our SHINOV algorithm has several advantages. First, it is a *state-free* algorithm, which means each node does not need to store information about other nodes. Thus, the cost for maintaining neighboring nodes is low. Second, it is designed for small-world networks. The message transmitting cost is saved, since no messages are wasted in nodes out of clusters. Furthermore, NOV obviously provides a much richer control over the query processing cost than seven-level-control of the regular TTL based methods. Last but not the least, SHINOV

[1] It is set to ten in our system implementation.

fits well in our P-Terse system. C^2 overlay network serves as a indexing mechanism. SHINOV uses C^2 for off-line keyword weighting and clustering. SHINOV's search does not rely on the C^2 for efficiency consideration. However, its search is guaranteed to find the neighboring nodes by C^2.

3.3 Result Ranking

The vector space model [1] is adopted in text retrieval of P-Terse. Each document is represented by a vector, in which each dimension denotes one term. The similarity of two documents are measured as the cosine of the angle between the vectors. The problem of using this scheme in peer-to-peer environments is that the terms are weighted using term-frequency/inverse-document-frequency (TF/IDF) based method, which relies on global information of all peers. For term t in document d, its weight is defined as $w_{t,d} = tf_{t,d} \times idf_d = freq_{t,d} \times \log \frac{N}{N_t}$, where $freq_{t,d}$ denotes the frequency of term t in document d, N denotes the number of documents in the corpus, and N_t is the number of documents containing term t. Here, N and N_t are hard to be obtained in a large scale P2P system.

A distributed hash table (DHT) based method is chosen to index the IDFs of the terms [16]. For each term t, a node construct a triple (t, n_t, N'), in which n_t is the number of local documents containing term t, and N' is the number of local documents. This tuple is sent to node with identity responding to $h(t)$, which is the hash result on the term. The information can be accessed by a similar process.

Though this method cannot guarantee the consistency of all nodes, we argue that the slightly inconsistency on the statistics does not affect the ranking result much.

3.4 Access Control and User Management

Access control is an important problem in file management. However, to the best of our knowledge, it is not mentioned in most research on peer-to-peer file sharing. We designed a naïve approach to provide simple access control over the shared files.

In P-Terse, there are three levels of sharing. They are *shared for all* (A), *shared for neighborhoods* (N) and *for owner only* (O), which are encoded in three bits. A document is identified by its owner identity and its local document identity on that node. The share level of a document is stored in the index. When a search query is received by the search module, the identity of the query initiator is checked. Only the files whose share level matched with the query initiator's belonging are searched by the local retrieval engine.

To manage the users and their basic information, super-peers are used. A super-peer is a node pre-determined. They are responsible for the identity assignment and management. When a node joins the system for the first time, an

identity is assigned, which is the hash result on the super-peer's identity and the new node's basic information. Furthermore, each node can set a password. In future joins to the system, the identity and password should be provided.

The neighboring nodes are managed by users[2]. The list of neighboring nodes are stored on the super-peer. Each time the node re-joins the system, the list is retrieved from the super-node. Then, the nodes in the list are informed for the re-join of this node.

Note that there are several super-peers. These peers are equipped with reliable hardware, and do not leave the system except by accidents. Furthermore, super-peers backup information for each other. Therefore, no single point failure would affect the running of the system. Though super-peers are used in P-Terse's current implementation. They can be replaced by usual peers by using aggressive replication.

4 Related Work

There are several research projects whose technical focuses are related to P-Terse. PlanetP tries to provide text-based search functions over an unstructured peer-to-peer network [3], while PSearch takes another approach for support of search based on vector-space model [2] over CAN [5]. Different to these two prototype systems, P-Terse is designed to provide a series of technologies to enable text search in a real-world applications. Therefore, the issues of user management and access control are considered and implemented.

Resource or collection selection is widely studied in centralized and distributed information retrieval community. Some of the research focus on distributed Web search, such as the distributed PageRank computation [17]. P-Terse has a different application scenario. We believe that in document sharing applications, linkage information cannot be utilized, since PDF or DOC files usually contain few or none hyperlinks.

Compared to other resource or collection selection strategies, such as those presented in MINERVA [18], Pepper [19], and Odissea [20], our SHINOV-based search module has one advantage that no additional information exchange or storage is needed on each peer. We show that the effectiveness of this search scheme is guaranteed by the small-world phenomenon. We are studying more powerful search schemes that can utilize information provided by neighboring peers while being independent to it.

GridVine shares the similar design strategy of P-Terse in that it also seperates the physical routing module with logical search module [21]. However, GridVine is designed for applications in which the semantic information or metadata of each document is available in RDF form. P-Terse has a different motivation. It is designed for general purpose document sharing.

[2] Note that neighboring nodes are managed in the search module, which are different from the nodes pointed by links in the finger table that are managed by the overlay communication module.

5 Discussion and Conclusion

The design and preliminary implementation of P-Terse is introduced in this paper. It is developed aiming at the file content sharing in a large scale network environment. The peer-to-peer based approach is taken to implement the system. The component-based design enables future extensions on the prototype system. Our on-going research and development lies in the following aspects.

- We are studying on technologies to provide more complex query and retrieval functions. Top-k query processing, for example, is an important function that are provided by most search engines. However, as result ranking, it also needs global information. This is the main challenge of this part of research.
- Currently, P-Terse supports *pull-based* search, which means query should be issued each time some documents need to be retrieved. We plan to support *push-based* search in our system. It means user can *subscribe* some content. After a subscription query is issued, each time a new document containing the required content appears, the document is sent to the node that needs it. The maintenance of the subscriptions and the document routing strategies are the main research topics of this issue.
- We are working on the new technologies that can improve the efficiency, fault-tolerance, and scalability of the system. This part of research includes the studying on general technologies of peer-to-peer computing systems, such as caching and replication. It also includes the effort to adapt the existing methods to fit in our P-Terse framework.
- The last but not the least issue is to provide more useful management functions. For example, current access control in P-Terse is still weak in terms of file management. We plan to extend it to enable versioning and trust management.

Acknowledgement

The authors would like to thank Dr. Beng Chin Ooi and Dr. Kian-Lee Tan for valuable suggestion on the development of P-Terse, Dr. Bo Ling and Zhiguo Lu for their involvement in the development of P-Terse, and anonymous reviewers for valuable comments on related work.

This work is partially supported by the National Natural Science Foundation of China under Grant No.60496325, Grant No.60496327, and Grant No.60503034, and the Shanghai Rising-Star Program under Grant No. 04QMX1404.

References

1. Ricardo Baeza-Yates and Berthier Ribeiro-Neto. *Modern Information Retrieval*. Addison Wesley Longman Publishing Co. Inc., 1999.
2. Chunqiang Tang, Zhichen Xu, and Mallik Mahalingam. Psearch: Information retrieval in structured overlays. In *Proceedings of the 1st ACM Workshop on Hot Topics in Networks (HotNets-I)*, 2002.

3. Francisco Matias Cuenca-Acuna, Christopher Peery, Richard P. Martin, and Thu D. Nguyen. Planetp: Using gossiping to build content addressable peer-to-peer information sharing communities. In *Proceedings of the 12th International Symposium on High Performance Distributed Computing (HPDC'2003)*, 2003.
4. Ion Stoica, Robert Morris, David Karger, M. Frans Kaashoek, and Hari Balakrishnan. Chord: a scalable peer-to-peer lookup service for internet applications. In *Proceedings of the ACM SIGCOMM 2001 Conference on Applications, Technologies, Architectures, and Protocols for Computer Communication (SIGCOMM'2001)*, pages 149–160. ACM Press, 2001.
5. Sylvia Ratnasamy, Paul Francis, Kark Handley, Richard Karp, and Scott Shenker. A scalable content-addressable network. In *Proceedings of the ACM SIGCOMM 2002 Conference on Applications, Technologies, Architectures, and Protocols for Computer Communication (SIGCOMM'2001)*, 2001.
6. A. Rowstron and P. Druschel. Pastry: Scalable, distributed object location and routing for large-scale peer-to-peer systems. In *Proceedings of the IFIP/ACM International Conference on Distributed Systems Platforms (Middleware'2001)*, pages 329–350, 2001.
7. Ben Y. Zhao, John Kubiatowicz, and Anthony D. Joseph. Tapestry: a fault-tolerant wide-area application infrastructure. *ACM SIGCOMM Computer Communication Review*, 32(1):81, January 2002.
8. Wenyuan Cai, Shuigeng Zhou, Linhao Xu, Weining Qian, and Aoying Zhou. C^2: A new overlay network based on can and chord. In *Proceedings of the Second Grid and Coperative Computing Conference (GCC'2003)*, pages 42–50, 2004.
9. Zheng Zhang, Shuigeng Zhou, Weining Qian, and Aoying Zhou. Keynote: Keyword search using nodes selection for text retrieval on dht-based p2p networks. In *In Proceedings of the 11th International Conference on Database Systems for Advanced Applications (DASFAA'2006)*, 2006.
10. Shuigeng Zhou, Zheng Zhang, Aoying Zhou, , and Weining Qian. Sipper: Selecting informative peers in structured p2p environment for content-based retrieval. In *In Proceedings of the 22nd International Conference on Data Engineering (ICDE'2006)*, 2006.
11. Adriana Iamnitchi, Matei Ripeanu, and Ian T. Foster. Locating data in (small-world?) peer-to-peer scientific collaborations. In *Proceedings of the first International Workshop on Peer-to-Peer Systems (IPTPS'2002)*, pages 232–241, 2002.
12. Ducan J. Watts. *Small Worlds: The Dynamics of Networks between Order and Randomness*. Princeton University Press, 1999.
13. A. Barabasi and R. Albert. Emergence of scaling in random networks. *Science*, 286:509–512, 1999.
14. Jon Kleinberg. The small-world phenomenon: An algorithmic perspective. In *Proceedings of the 32nd Annual ACM Symposium on Theory of Computing (STOC'2000)*, pages 163–170, 2000.
15. Yi Ren, Chaofeng Sha, Weining Qian, Aoying Zhou, Beng Chin Ooi, and Kian-Lee Tan. Explore the "small world phenomena" in a p2p information sharing system. In *Proceedings of the third IEEE/ACM International Symposium on Cluster Computing and the Grid (CCGrid'2003)*, 2003.
16. Zhiguo Lu, Bo Ling, Weining Qian, Wee Siong Ng, and Aoying Zhou. A distributed ranking strategy in peer-to-peer based information retrieval systems. In *Proceedings of the 6th Asia-Pacific Web Conference (APWeb'2004)*, 2004.
17. Yuan Wang and David J. DeWitt. Computing pagerank in a distributed internet search engine system. In *Proceedings of the 30th International Conference on Very Large Databases (VLDB'2004)*, pages 420–431.

18. M. Bender, S. Michel, P. Triantafillou, and G. Weikum a nd C. Zimmer. Minerva: Collaborative p2p search. In *Proceedings of the 31th International Conference on Very Large Databases (VLDB'2005)*, pages 1263–1266.
19. Henrik Nottelmann and Norbert Fuhr. Combining cori and the decision-theoretic approach for advanced resource selection. In *Proceedings of the 26th European Conference on Information Retrieval (ECIR'2004)*, pages 138–153, 2004.
20. T. Suel and J. Zhang. Efficient query evaluation on large textual collections in a peer-to-peer environment. In *Proceedings of the 5th IEEE International Conference on Peer-to-Peer Computing*, 2005.
21. K. Aberer, P. Cudré-Mauroux, M. Hauswirth, and Tim Van Pelt. Gridvine: Building internet-scale semantic overlay networks. In *Proceedings of the International Semantic Web Conference 2004*, pages 107–121.

Identifying Semantic Relations Between Named Entities from Chinese Texts

Tianfang Yao[1] and Hans Uszkoreit[2]

[1] Department of Computer Science and Engineering, Shanghai Jiao Tong University,
800 Dong Chuan Road,
200240 Shanghai, China
yao-tf@cs.sjtu.edu.cn
[2] Department of Computational Linguistics and Phonetics, Saarland University,
Postfach 15 11 50,
66041 Saarbrücken, Germany
uszkoreit@coli.uni-sb.de

Abstract. COLLATE is a project dedicated to building up a German authority center for language technology in Saarbrücken. Under this project, a computational model with three-stage pipeline architecture for Chinese information extraction has been proposed. In this paper, we concentrate on the presentation for the third stage, viz., the identification of named entity relations (NERs). A learning and identification approach for NERs called positive and negative case-based learning and identification is described in detail. It pursues the improvement of the identification performance for NERs through simultaneously learning two opposite cases, automatically selecting effective multi-level linguistic features for each NER and non-NER, and optimally achieving an identification tradeoff etc. The experimental results have shown that the overall average recall, precision, and F-measure for 14 NERs are 78.50%, 63.92% and 70.46% respectively. In addition, the above F-measure has been enhanced from 63.61% to 70.46% due to adoption of both positive and negative cases.

Keywords: Chinese information extraction, machine learning, named entity relation identification.

1 Introduction

1.1 COLLATE Project

COLLATE (Computational Linguistics and Language Technology for Real World Applications) [1] is a project dedicated to establishing a German authority center for language technology in Saarbrücken, Germany. The motivations of this project are to convert the research into a strong impact and speed up the ways of the research results of the laboratory into commercial products considerably. The effort extends from the largest information and intelligence service of the subject on the Internet over the interactive demonstration of a broad spectrum of applications for language technology up to the intensive consultation over technologies and systems. Besides in the

R. Lu et al. (Eds.): Chinese-German Ws on Cognitive Systems 2005, LNAI 4429, pp. 70 – 83, 2007.
© Springer-Verlag Berlin Heidelberg 2007

authority center innovative research results in principal areas of the language technology were acquired, improved processing for the purposeful extraction of information from large text quantities helps company and organization users to get the daily flood of information into the grasp.

The international adviser and the consultants certified the project its success. Currently, the center is world-wide admits and serves in Germany as the first approach place for scientists, users, decision makers, journalists and the interested public.

1.2 A Computational Model for Chinese Information Extraction

The investigation for Chinese information extraction (IE) is one of the topics in the project COLLATE. During intensively studying Chinese IE, primarily word processing, named entity (NE) recognition, and named entity relation (NER) identification, we have proposed a Chinese IE computational model. Our motivations concerning the establishment of the above model are:

- To combine the different effective techniques in IE model, such as knowledge-base, statistical, machine learning techniques etc;
- To design a novel IE computational model that can be suitable to different Chinese IE tasks, such as NE and NER identification; and
- To make IE systems more efficient and effective, especially in reliability, portability, and performance.

This model has a pipeline architecture with three stages shown in Fig. 1. In the first stage, word processing includes word segmentation and part-of-speech tagging. In general, its processing quality has considerable influence on the performance of the consequent two stages (It has been proved by our experiments [2].). In order to reduce unfavorable influence, we utilize a trainable approach to automatically generate effective rules, by which the word processing component can repair different errors caused by word segmentation and part-of-speech tagging. At the second stage, there are two kinds of NE constructions to be processed. One is the NEs which involve trigger words; the other those without trigger words. For the former NEs, a knowledge engineering technique, i.e. finite-state cascades (FSC) as a shallow parsing mechanism, is adopted for reliably identifying different categories of NEs. For the latter NEs, however, some special strategies, such as the valence constraints of domain verbs, the constituent analysis of NE candidates, the global context clues and the analysis for preposition objects etc., are designed for identifying them. After the recognition for NEs, NER identification is performed in the last stage. Because of the diversity and complexity of NERs, at the same time, considering portability requirement in the identification, we suggest a novel supervised machine learning approach used in this stage. In the model, in addition to the above techniques, we have adopted an annotation technique for machine learning in the last stage and developed a lexical ontology for chunking NE, categorizing NEs without trigger

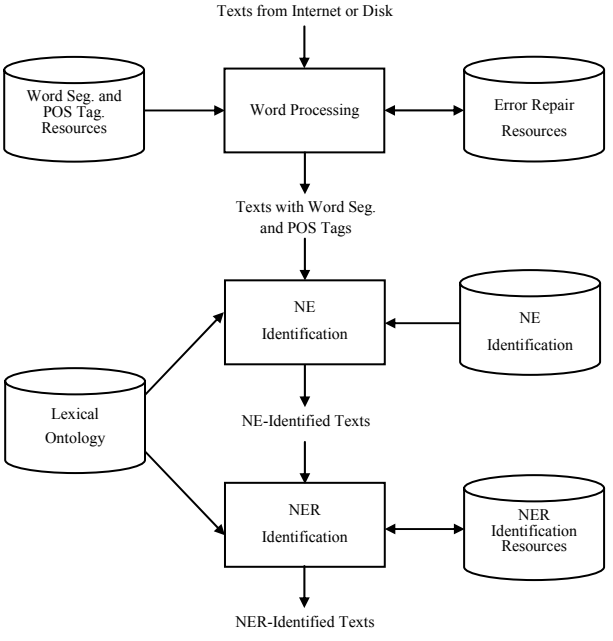


Fig. 1. A three-stage Chinese information extraction computational model

words and computing NER semantic similarity in the later two stages. Moreover, the integration of the existing system, e.g., word segmentation and part-of-speech tagging system, is also a strategy employed in our model.

In the rest of this paper, we will focus on the elaboration of the last stage in the model.

2 Definition of Relations

An named entity relation (NER) represents a binary semantic relationship between NEs, which exists within a sentence or even across sentence. The task with regard to NE recognition has been accomplished, the new task - the investigation of identification issues for NERs confronted us. Based on Chinese NE recognition system [3], we define 14 different NERs related to six identified NEs in a sports domain, which are shown in the Table 1. Before learning NERs, we annotate the output texts from this system with XML. Meanwhile, the annotated NERs are also added in the above XML texts. Because sometimes the type of two NEs in a NER is the same, e.g., HT_VT, this procedure is executed by interactive mode. Thus, users can assign a subcategory tag for TN (Team Name), which is either host team or visiting team. To further indicate the positions of NEs in an NER, we define a general frame for NERs and give the following example using this description.

Table 1. Named entity relation category

NER category	Explanation
CP_DA	The staged date for a sports competition
CP_LOC	It gives the location where a sports match is held
CP_TI	The staged time for a sports competition
DT_DT	The names of two teams which draw a match
HT_VT	The home and visiting teams in a sports competition
ID_TM	The position of a person employed by a sports team
LOC_ CPC	The location ownership (LOC belongs to CPC)
PS_CP	A person takes part in a sports competition
PS_CPC	The origin location of a person
PS_ID	A person and her/his position in a sports team or other occasions
PS_TM	The membership of a person in a sports team
TM_CP	A team participates in a sports competition
TM_CPC	It indicates where a sports team comes from
WT_LT	The winning and losing team name in a sports match

Definition 1 (Named Entity Relation General Frame). NamedEntityRelation (NamedEntity1, ParagraphSentenceNamedEntityNo1; NamedEntity2, ParagraphSentenceNamedEntityNo2).

Example 1. 广东宏远队[1]客场以 3 比 0 击败广州太阳神队。The Guangdong Hongyuan Team defeated the Guangzhou Taiyangshen Team by 3: 0 in the guest field.

There exist two NERs, i.e., HT_VT and WT_LT, in the above sentence. According to the general frame, the first NER is defined as HT_VT (广州太阳神队(Guangzhou Taiyangshen Team), 1-1-2; 广东宏远队(Guangdong Hongyuan Team), 1-1-1) and the other is defined as WT_LT (广东宏远队(Guangdong Hongyuan Team), 1-1-1; 广州太阳神队(Guangzhou Taiyangshen Team), 1-1-2).

3 Positive and Negative Case-Based Learning

The learning approach we propose is a supervised statistical learning [4]. Actually, it is a variant of memory-based learning [5, 6, 7]. Unlike memory-based learning, positive and negative case-based learning (PNCBL) does not simply store cases in memory but transforms case form into NER and non-NER patterns. Additionally, it stores not only positive but also negative cases. Here, it should be clarified that the negative case we mean is a case in which two or more than two NEs have no relationships, that is, they have non-relationships. Such relationships are also investigated objects we are interested.

During the learning, depending on the average similarity of features and the self-similarity of NERs (also non-NERs), it automatically selects general or individual-character features (GCFs or ICFs) to construct an effective feature set for different NERs or non-NERs, and then based on these feature sets it can decide different

[1] The underlining of Chinese words means that an NE consists of these words.

feature weights as well as NER and non-NER identification thresholds. Thus, the learning results provide identification references for the forthcoming NER identification.

3.1 Relation Features

Relation features, by which we can effectively identify different NERs or non-NERs, are defined for capturing critical information of the Chinese language. With the features, we can easily define NER or non-NER patterns. The following essential factors motivate our definition for relation features:

- The relation features should be selected from multiple linguistic levels, i.e., morphology, grammar and semantics [8];
- They can help us to identify NERs using PNCBL approach, e.g., their information deals with not only NERs but also non-NERs;
- They should embody the crucial information of Chinese language processing [9], such as word order, the context of words, and particles etc.

There are a total of 13 relation features shown in Table 2, which are empirically determined in terms of the above motivations. It should be explained that to distinguish feature names from element names of the NER and non-NER patterns (See Definition 2 and 3), we add a capital letter "F" in the ending of feature names. In

Table 2. Feature category

Feature category	Explanation
NECF	The context of named entities. The context only embodies a word or a character preceding or following the current named entity.
NECPF	The parts-of-speech of the context for the named entities associated with a relation.
NECTF	The concepts of the named entities of a relevant relation from HowNet[10].
NEOF	The order of the named entities of a relevant relation.
NEPF	The parts-of-speech of the named entities of a relevant relation.
NEPPOF	The relative order between parts-of-speech of particles and named entities. The particles occur within the sentences where the relation is embedded.
NESPF	The named entities of a relevant relation are located in the same sentence or different sentences.
NEVPF	The relative position between the verbs and the named entities of a relevant relation. The verbs of a relevant relation mean that they occur in a sentence where the relation is embedded.
SGTF	The type of a sentence group in which there exists a relation.
SPF	The sequence of parts-of-speech for all sentence constituents within a relation range.
VCTF	The concepts of the verbs of a relevant relation from HowNet.
VSPF	The verbs are located in the same sentence or different sentences in which there is a relevant relation.
VVF	The valence expression of verbs in the sentence(s) where there is a relation embedded.

addition, a sentence group in the following definitions can contain a sentence or multiple sentences. In other words, a sentence group must end with a stop, semicolon, colon, exclamation mark, or question mark.

In 13 features, three features (NECF, NECPF and NEPF) belong to morphological features, three features (NEOF, SGTF and SPF) are grammatical features, four features (NEPPOF, NESPF, NEVPF and VSPF) are associated with not only morphology but also grammar, and three features (NECTF, VCTF and VVF) are semantic features.

Every feature describes one or more properties for a relation. For instance, NECF can capture the noun 客场 (the guest field) and also determine that the closest NE by this noun is 广东宏远队 (the Guangdong Hongyuan Team), which is a visiting team. On the other hand, NEOF can fix the sequence of two relation-related NEs., namely, another NE 广州太阳神队 (the Guangzhou Taiyangshen Team) can be determined as a host team. Consequently, these two features reflect the properties of the relation HT_VT.

3.2 Relation and Non-relation Patterns

A relation pattern describes the relationships between an NER and its features. In other words, it depicts the linguistic environment in which NERs exist.

Definition 2 (Relation Pattern). A relation pattern (RP) is defined as a 14-tuple: RP = (no, RE, SC, sgt, NE, NEC, VERB, PAR, NEP, NECP, SP, VV, NECT, VCT) where *no* represents the number of a RP; *RE* is a finite set of relation expressions; *SC* is a finite set for the words in the sentence group except for the words related to named entities; *sgt* is a sentence group type; *NE* is a finite set for named entities in the sentence group; *NEC* is a finite set that embodies the context of named entities; *VERB* is a finite set that includes the sequence numbers of verbs and corresponding verbs; *PAR* is a finite set of particles; *NEP* is a finite set of named entities and their POS tags; *NECP* is a finite set which contains the POS tags of the context for named entities; *SP* is a finite set in which there are the sequence numbers as well as corresponding POS tags and named entity numbers in a sentence group; *VV* is a finite set comprehending sno of verbs and its valence constraints from Lexical Sports Ontology which is developed by us; *NECT* is a finite set that has the concepts of named entities in a sentence group; and *VCT* is a finite set which gives the concepts of verbs in a sentence group.

Example 2. 据新华社北京３月２６日电全国足球甲Ｂ联赛今天进行了第二轮赛事的５场比赛，广东宏远队客场以３比０击败广州太阳神队，成为唯一一支两战全胜的队伍，暂居积分榜榜首。 According to the news from Xinhua News Agency Beijing on March 26th: National Football Tournament (the First B League) today held five competitions of the second round, The Guangdong Hongyuan Team defeats the Guangzhou Taiyangshen Team by 3: 0 in the guest field, becoming the only team to win both matches, and temporarily occupying the first place of the entire competition.

Relation Pattern. *no* = 34; *RE* = {(CP_DA, NE1-3, NE1-2), …}; *SC* = {(1, 据, according_to, Empty, AccordingTo), …}; *sgt* = multi-sentences; *NE* = {(NE1-1, 3,

LN, {(1, 北京)}), ...}; *NEC* = {(NE1-1, 新华社, 3), ...}; *VERB* = {(8, 进行), ...}; *PAR* = {(1, 据), ...}; *NEP* = {(NE1-1, {(1, N5)}), ...}; *NECP* = {(NE1-1, N, M), …}; *SP* = {(1, P), ...}; *VV* = {(V_8, {Agent|fact/compete|CT, -Time|time|DT}), ...}; *NECT* = {(NE1-1, place/ capital/ProperName/China), …}; *VCT* = {(V_8, GoForward/GoOn/Vgoingon), …}.

Analogous to the definition of the relation pattern, a non-relation pattern is defined as follows:

Definition 3 (Non-relation Pattern). A non-relation pattern (NRP) is also defined as a 14-tuple: NRP = (no, NRE, SC, sgt, NE, NEC, VERB, PAR, NEP, NECP, SP, VV, NECT, VCT), where *NRE* is a finite set of non-relation expressions which specify the nonexistent relations in a sentence group. Excepting that, the definitions of other elements are the same as the ones of the relation pattern.

For example, if we want to build an NRP for the above sentence group in Example 2, *NRE* = {(CP_LOC, NE1-3, NE1-1), ...}. In this sentence group, the named entity (CT) 全国足球甲B联赛 (National Football Tournament (the First B League)) doesn't have the relation CP_LOC to the named entity (LN) 北京 (Beijing). This LN only indicates the release location of the news from Xinhua News Agency.

3.3 Similarity Calculation

In the learning, the similarity calculation is a kernel measure for the selection of effective features. First of all, let us look at the definition of self-similarity and how to calculate it for the same kind of NERs.

Definition 4 (Self-similarity). The self-similarity of a kind of NERs or non-NERs in the corresponding library can be used to measure the concentrative degree of this kind of relations or non-relations. The value of the self-similarity is between 0 and 1. If the self-similarity value of a kind of relation or non-relation is close to 1, we can say that the concentrative degree of this kind of relation or non-relation is very "tight". Conversely, the concentrative degree of that is very "loose".

The calculation of the self-similarity for the same kind of NERs is equal to the calculation for the average similarity of the corresponding relation features. Suppose *R(i)* is an NER in the NER set ($1 \le i \le 14$). The average similarity for this kind of NERs is defined as follows:

$$\mathrm{Sim}_{\mathrm{average}}(\mathrm{R(i)}) = \frac{\sum_{1 \le j,\, k \le m;\, j \ne k} \mathrm{Sim}\,(\mathrm{R(i)_j}, \mathrm{R(i)_k})}{\mathrm{Sum}_{\mathrm{relation\_pair}}(\mathrm{R(i)_j}, \mathrm{R(i)_k})} \tag{1}$$

where $Sim\ (R(i)_j,\ R(i)_k)$ denotes the relation similarity between the same kind of relations, $R(i)_j$ and $R(i)_k$. $1 \le j, k \le m$, $j \ne k$; m is the total number of the relation $R(i)$ in the NER pattern library. $Sim(R(i)_j,\ R(i)_k)$ is calculated in terms of different features (See below). $Sum_{relation\_pair}(R(i)_j,\ R(i)_k)$ is the sum of calculated relation pair number. They can be computed using the following formulas:

$$\mathrm{Sim}(R(i)_j, R(i)_k) = \frac{\sum_{t=1}^{Sumf} \mathrm{Sim}(R(i)_j, R(i)_k)(f_t)}{Sum_f} \tag{2}$$

$$Sum_{relation\_pair}(R(i)_j, R(i)_k) = \begin{cases} 1 & m = 2 \\ \dfrac{m!}{(m-2)! * 2!} & m > 2 \end{cases} \tag{3}$$

In the formula (2), f_t is a feature in the feature set ($1 \leq t \leq 13$). Sum_f is the total number of features. The calculation of $Sim\ (R(i)_j, R(i)_k)\ (f_t)$ is executed depending on different features. For example, if f_t is equal to NECF, $Sim\ (R(i)_j, R(i)_k)\ (f_t)$ is shown as follows:

$$\mathrm{Sim}(X(i)_j, X(i)_k)\ (\mathrm{NECF}) = \begin{cases} 1 & \text{if all contexts of named entities for two relations are the same} \\ 0.75 & \text{if only a preceding or following context is not the same} \\ & \text{if two preceding and/or following contexts are not the same} \\ 0.25 & \text{if three preceding and/or following contexts are not the same} \\ 0 & \text{if all contexts of named entities for two relations are not the same} \end{cases} \tag{4}$$

Notice that the similarity calculation for non-NERs is the same as the above calculations.

Before giving the learning algorithm, we predefine some fundamental conceptions related to the algorithm as follows:

Definition 5 (General-Character Feature). If the average similarity value of a feature in a relation is greater than or equal to the self-similarity of this relation, it is called a General-Character Feature (GCF). This feature reflects a common characteristic of this kind of relation.

Definition 6 (Individual-Character Feature). An Individual-Character Feature (ICF) means its average similarity value in a relation is less than or equal to the self-similarity of this relation. This feature depicts an individual property of this kind of relation.

Definition 7 (Feature Weight). The weight of a selected feature (GCF or ICF) denotes the important degree of the feature in GCF or ICF set. It is used for the similarity calculation of relations or non-relations during relation identification.

$$f(s)_w(R(i)) = \frac{Sim_{average}f(s)(R(i))}{\sum_{t=1}^{n} Sim_{average}f(t)(R(i))} \tag{5}$$

where *R(i)* is a defined relation in the NER set ($1 \leq i \leq 14$); n is the size of selected features, $1 \leq s, t \leq n$; and

$$Sim_{average}f(s)(R(i)) = \frac{\sum_{1 \leq j, k \leq m; j \neq k} Sim\,(R(i)_j, R(i)_k)\,(f(s))}{Sum_{relation\_pair}(R(i)_j, R(i)_k)} \tag{6}$$

Sim $(R(i)_j, R(i)_k)$ $(f(s))$ computes the feature similarity of the feature *f(s)* between same kinds of relations, $R(i)_j$ and $R(i)_k$. $1 \leq j, k \leq m, j \neq k$; m is the total number of the relation *R(i)* in the NER pattern library. $Sum_{relation\_pair}(R(i)_j, R(i)_k)$ is the sum of calculated relation pair numbers, which can be calculated by the formula (3).

Definition 8 (Identification Threshold). If a candidate relation is regarded as a relation in the relation pattern library, the identification threshold of this relation indicates the minimal similarity value between them. It is calculated by the average of the sum of average similarity values for selected features:

$$IdenThrh(R(i)) = \frac{\sum_{t=1}^{n} Sim_{average}f(t)(R(i))}{n} \tag{7}$$

where n is the size of selected features, $1 \leq t \leq n$.

Finally, the PNCBL algorithm is described as follows:

1) Input annotated texts;
2) Transform XML format of texts into internal data format;
3) Build NER and non-NER patterns;
4) Store both types of patterns in hash tables and construct indexes for them;
5) Compute the average similarity for features and self-similarity for NERs and non-NERs;
6) Select GCFs and ICFs for NERs and non-NERs respectively;
7) Calculate feature weights for selected features;
8) Decide identification thresholds for every NER and non-NER;
9) Store the above learning results.

4 Relation Identification

The approach of the NER identification is based on PNCBL, it can utilize the outcome of the learning for recognizing NERs and removing non-NERs. However, it is also confronted with some problems, for example, relation identification tradeoff, relation conflicts, and relation omissions. Hence, the procedure of the NER identification offers a solusion for the above problems.

4.1 Achieving an Optimal Identification Tradeoff

During the NER identification, the GCFs of NER candidates are matched with those of all of the same kind of NERs in the NER pattern library. Likewise, the ICFs of NER candidates are compared to those of non-NERs in the non-NER pattern library. The computing formulas for both cases are listed as follows:

$$\mathrm{Sim}\,(\mathrm{R(i)_{can}}, \mathrm{R(i)_{j1}}) = \sum_{k1=1}^{\mathrm{Sum(GCF)_i}} \{\, w_i\,(\mathrm{GCF_{k1}}) * \mathrm{Sim}(\mathrm{R(i)_{can}}, \mathrm{R(i)_{j1}})\,(\mathrm{GCF_{k1}}) \,\} \tag{8}$$

and

$$\mathrm{Sim}\,(\mathrm{R(i)_{can}}, \mathrm{NR(i)_{j2}}) = \sum_{k2=1}^{\mathrm{Sum(ICF)_i}} \{\, w_i\,(\mathrm{ICF_{k2}}) * \mathrm{Sim}(\mathrm{R(i)_{can}}, \mathrm{NR(i)_{j2}})\,(\mathrm{ICF_{k2}}) \,\} \tag{9}$$

where *R(i)* represents the $\mathrm{NER_i}$, and *NR(i)* expresses the non-$\mathrm{NER_i}$, $1 \le i \le 14$. $R(i)_{can}$ is defined as a $\mathrm{NER_i}$ candidate. $R(i)_{j1}$ and $NR(i)_{j2}$ are the j1-th $\mathrm{NER_i}$ in the NER pattern library and the j2-th non-$\mathrm{NER_i}$ in the non-NER pattern library. $1 \le j1 \le Sum\ (R(i))$ and $1 \le j2 \le Sum\ (NR(i))$. *Sum (R(i))* and *Sum (NR(i))* are the total number of *R(i)* in the NER pattern library and that of *NR(i)* in non-NER pattern library respectively. $w_i\ (GCF_{k1})$ and $w_i\ (ICF_{k2})$ mean the weight of the k1-th GCF for the $\mathrm{NER_i}$ and that of the k2-th ICF for the non-$\mathrm{NER_i}$. $Sum\ (GCF)_i$ and $Sum\ (ICF)_i$ are the total number of GCF for $\mathrm{NER_i}$ and that of ICF for non-$\mathrm{NER_i}$ separately.

In matching results, sometimes the similarity values of a number of NERs or non-NERs matched with NER candidates are all more than the identification threshold. Accordingly, we have to use a voting mode to achieve an identification tradeoff. For an optimal tradeoff, the final identification performance embodies two aspects: i.e., recall and precision. To enhance recall, correct NERs should be captured as many as possible; on the other hand, to increase precision, misidentified non-NERs should be removed as accurately as possible.

The voting refers to the similarity calculation results between an NER candidate and NER / non-NER patterns. It pays special attention to circumstances in which both results are very close. If it happens, multiple calculation results are used to measure and arrive at a final decision. Additionally, notice that the impact of non-NER patterns is to restrict possible misidentified non-NERs. For instance, if an NER candidate's similarity result matched with a non-NER pattern is equal to or more than

0.75. Obviously, this NER candidate is very similar to the non-NER. In this situation, the NER candidate is forbidden to pass by means of one of the examination conditions. The thresholds in the conditions refer to [11, 12] and our experiments.

On the other hand, the voting assigns different thresholds to different NER candidates (HT_VT, WT_LT, and DT_DT or other NERs). Because the former three NERs have the same kind of NEs, i.e., they all have two TNs, the identification for these NERs is more difficult than for that of other NERs. Consequently, when voting, the corresponding threshold should be set more strictly.

4.2 Resolving NER Conflicts

In fact, although the voting is able to use similarity comparison to achieve an optimal tradeoff, there still remain some problems to be resolved. The relation conflict is one of the problems, which means that contradictory NERs occur in identification results. For example:

1) The same kind of relations with different argument positions: e.g., the relations HT_VT:
   HT_VT(ne1, no1; ne2, no2) and HT_VT(ne2, no2; ne1, no1) occur in an identification result at the same time.
2) The different kinds of relations with same or different argument positions: e.g., the relations WT_LT and DT_DT:
   WT_LT(ne1, no1; ne2, no2) and DT_DT(ne1, no1; ne2, no2) appear simultaneously in an identification result.

The reason for relation conflict lies in the simultaneous and successful matching of a pair of NER candidates, whose NEs are the same kind. They do not compare and distinguish themselves further. Considering the impact of NER and non-NER patterns, we enumerate examination conditions to remove one of relations, which has lower average similarity value with NER patterns or higher average similarity value with non-NER patterns.

4.3 Inferring Missing NERs

Due to a variety of reasons, some relations that should appear in an identification result may be missing. Fortunately, we can utilize some of identified NERs to infer them. Of course, the prerequisite of the inference is that we suppose identified NERs are correct and non-contradictory. For all identified NERs, first, we should examine whether there are missing NERs within them. After determining the type of missing NERs, we may infer them - containing the relation name and its arguments. For instance, in an identification result, two NERs are:

PS_ID (ne1, no1; ne2, no2) and PS_TM (ne1, no1; ne3, no3)

In the above NER expressions, ne1 is a personal name, ne2 is a personal identity, and ne3 is a team name, because if a person occupies a position, i.e., he / she has a corresponding identity in a sports team, that means the position or identity belongs to this sports team. Accordingly, we can infer the following NER:

ID_TM (ne2, no2; ne3, no3)

5 Experimental Results and Evaluation

The main resources, which are used for the learning and identification, are NER and non-NER patterns. Before the learning, we have completed the annotation for the more than 50 texts from the Jie Fang Daily[2] in 2001 based on the NE identification. During the learning, both pattern libraries are established in terms of the annotated texts and Lexical Sports Ontology. They have 142 (534 NERs) and 98 (572 non-NERs) sentence groups respectively.

To test the performance of this approach, we randomly choose 32 sentence groups from the Jie Fang Daily in 2002 (these sentence groups are out of either NER or non-NER pattern library), which embody 117 different NER candidates.

For evaluating the effects of negative cases, we made two experiments to compare its impact on the above two sides. Table 3 shows the average and total average recall, precision, and F-measure for 14 different NERs only by positive case-based learning and identification respectively. Table 4 demonstrates those by PNCBL&I separately. Comparing these two experimental results, among 14 NERs, the F-measure values of the seven NERs (PS_ID, ID_TM, CP_TI, WT_LT, PS_CP, CP_DA, and DT_DT) in Table 4 are higher than those of corresponding NERs in Table 3; the F-measure values of three NERs (LOC_CPC, TM_CP, and PS_CP) have no variation; but the F-measure values of other four NERs (PS_TM, CP_LOC, TM_CPC, and HT_VT) in Table 4 are lower than those of corresponding NERs in Table 3. This shows the performances for half of NERs are improved due to the adoption of both positive and negative cases. Moreover, the total average F-measure is enhanced from 63.61% to 70.46% as a whole.

Table 3. Performance for 14 NERs (only by positive case-based learning and identification)

Relation Type	Average Recall	Average Precision	Average F-measure
LOC_CPC	100	91.67	95.65
TM_CP	100	87.50	93.33
PS_ID	100	84.62	91.67
PS_TM	100	72.73	84.21
CP_LOC	88.89	69.70	78.13
ID_TM	90.91	66.67	76.93
CP_TI	83.33	71.43	76.92
PS_CP	60	75	66.67
TM_CPC	100	42.50	59.65
HT_VT	71.43	38.46	50
WT_LT	80	30.77	44.45
PS_CPC	33.33	66.67	44.44
CP_DA	0	0	0
DT_DT	0	0	0
Total Ave.	**71.99**	**56.98**	**63.61**

[2] This is a local newspaper in Shanghai, China.

Table 4. Performance for 14 NERs (by PNCBL&I)

Relation Type	Average Recall	Average Precision	Average F-measure
LOC_CPC	100	91.67	95.65
TM_CP	100	87.50	93.33
CP_TI	100	75	85.71
PS_CPC	100	68.75	81.48
ID_TM	90.91	68.19	77.93
PS_ID	72.22	81.67	76.65
CP_LOC	88.89	66.67	76.19
PS_TM	80	65	71.72
CP_DA	100	50	66.67
DT_DT	66.67	66.67	66.67
PS_CP	60	75	66.67
WT_LT	60	37.50	46.15
HT_VT	42.86	30	35.30
TM_CPC	37.50	31.25	34.09
Total Ave.	**78.50**	**63.92**	**70.46**

6 Conclusion

In this paper, we introduce the project COLLATE and a computational model for Chinese information extraction under this project. Especially, we focus on the elaboration of the third stage for this model. At this stage, we propose a novel machine learning and identification approach, positive and negative case-based learning and identification. This approach includes the following advantages: (i) The defined negative cases are used to improve the NER identification performance as compared to only using positive cases; (ii) All of the tasks, building of NER and non-NER patterns, feature selection, feature weighting and identification threshold determination, are automatically completed. It is able to adapt the variation of NER and non-NER pattern library; (iii) The information provided by the relation features deals with multiple linguistic levels, depicts both NER and non-NER patterns, as well as satisfies the requirement of Chinese language processing; (iv) Self-similarity is a reasonable measure for the concentrative degree of the same kind of NERs or non-NERs, which can be used to select general-character and individual-character features for NERs and non-NERs respectively; (v) The strategies used for achieving an optimal NER identification tradeoff, resolving NER conflicts, and inferring missing NERs can further improve the performance for NER identification; (vi) It can be applied to sentence groups which can contain multiple sentences. Thus, identified NERs are allowed to cross sentences. The experimental results have shown that it is appropriate and effective for improving the identification performance of Chinese NERs.

Acknowledgments. This work is a part of the COLLATE project, which is supported under contract no. 01INA01B by the German Ministry for Education and Research.

References

1. COLLATE: Computational Linguistics and Language Technology for Real Life Applications. DFKI, Saarbrücken, Germany. http://collate.dfki. de/ (2002)
2. Yao T., Ding W., Erbach G.: Correcting Word Segmentation and Part-of-Speech Tagging Errors for Chinese Named Entity Recognition. In: Hommel G., Sheng H. (eds.): The Internet Challenge: Technology and Applications. Kluwer Academic Publishers, Dordrecht, The Netherlands (2002) 29-36
3. Yao T., Ding W., Erbach G.: CHINERS: A Chinese Named Entity Recognition System for the Sports Domain. In: Proc. of the Second SIGHAN Workshop on Chinese Language Processing (ACL 2003 Workshop), Sapporo, Japan (2003) 55-62
4. Nilsson N.: Introduction to Machine Learning: An Early Draft of a Proposed Textbook. http://robotics.stanford.edu/people/nilsson/mlbook.html (1996) 175-188
5. Stanfill C., Waltz D.: Toward memory-based reasoning. Communications of the ACM, Vol. 29, No. 12. (1986) 1213-1228
6. Daelemans W.: Memory-based lexical acquisition and processing. In: Steffens P. (ed.): Machine Translation and the Lexicon. Lecture Notes in Artificial Intelligence Vol. 898. Springer-Verlag, Berlin Heidelberg New York (1995) 85-98
7. Daelemans W., Bosch A., Zavrel J., Van der Sloot K., Vanden Bosch A.: TiMBL: Tilburg Memory Based Learner, Version 3.0, Reference Guide. Technical Report ILK-00-01, ILK, Tilburg University. Tilburg, The Netherlands. http://ilk.kub.nl/~ilk/papers/ ilk0001.ps.gz (2000)
8. Cardie C.: Automating Feature Set Selection for Case-Based Learning of Linguistic Knowledge. In: Proc. of the Conference on Empirical Methods in Natural Language Processing. University of Pennsylvania, Philadelphia, USA (1996) 113-126
9. Dang H., Chia C., Palmer M., Chiou F.: Simple Features for Chinese Word Sence Disambiguation. In: Proc. of the 19th International Conference on Computational Linguistics (COLING 2002). Taipei, Taiwan (2002) 204-210
10. Dong Z., Dong Q.: HowNet. http://www.keenage.com/zhiwang/e_zhiwang.html (2000)
11. Cover T., Hart P.: Nearest neighbor pattern classification. Transactions on Information Theory Vol 13. Institute of Electrical and Enginees (1967) 21-27
12. Duda R., Hart P.: Pattern Classification and Scene Analysis. Wiley & Sons, New York (1973)

Research on English-Chinese Bi-directional Cross-Language Information Retrieval

Yuejie Zhang[1] and Tao Zhang[2]

[1] Department of Computer Science & Engineering,
Shanghai Key Laboratory of Intelligent Information Processing,
Fudan University, Shanghai 200433, P.R. China
[2] School of Information Management & Engineering,
Shanghai University of Finance & Economics, Shanghai 200433, P.R. China
yjzhang@fudan.edu.cn, taozhang@mail.shufe.edu.cn

Abstract. With the rapid growing amount of information available to us, the situations that a user needs to use a retrieval system to perform querying a multilingual document collection are becoming increasingly emerging and common. Thus an important problem is formed, to match the user queries specified in one language against documents in another different language, i.e. Cross-Language Information Retrieval (CLIR). Based on the work in CLIR evaluation task in the 9th Text Retrieval Conference (TREC-9), we have constructed an English-Chinese bi-directional CLIR system. In this system, we adopt English-Chinese bi-directional query translation as the dominant strategy, use English and Chinese queries as translation objects, and utilize English and Chinese machine readable dictionaries as the important knowledge source to acquire correct translations. By combining English and Chinese monolingual IR systems constructed by us, the complete English-Chinese bi-directional CLIR process can be implemented successfully.

Keywords: Information retrieval, cross-language information retrieval, machine translation, machine readable dictionary, Chinese segmentation.

1 Introduction

Information Retrieval (IR) mainly refers to a process that users can find their required information or knowledge from corpus including different kinds of information. Traditional IR system is monolingual system. However, with the rapid growing amount of information available to us, the situations that a user needs to use a retrieval system to perform querying a multilingual document collection are becoming increasingly emerging and common. This tendency causes the difficulty of information acquisition. Meanwhile, language barriers become a serious problem. For many users using this kind of multilingual data resource, they must have some knowledge about foreign language [1]. If the users do not have a good command of some foreign languages, users cannot narrate their query explicitly in order to express their information requirement [2]. Cross-Language Information Retrieval (CLIR) provides a convenient way that can solve the problem of crossing the language boundary, and users can submit queries which are written in their familiar language

R. Lu et al. (Eds.): Chinese-German Ws on Cognitive Systems 2005, LNAI 4429, pp. 84 – 95, 2007.
© Springer-Verlag Berlin Heidelberg 2007

and retrieve documents in another language [3]. In present information society, CLIR is a key problem which is in need of solution in the scope of the world [4]. English and Chinese are two languages used extensively in the world, so English-Chinese bi-directional CLIR has attracted many interests in the research field of IR.

Text Retrieval Conference (TREC) is an international conference which is organized by National Institute of Standard Technology (NIST) of USA. Its goal is to facilitate the research in the field of large-scale retrieval, accelerate the conversion from research results to commercial products, and promote the exchange and cooperation among academic research institutes, commercial organizations and governmental departments. CLIR was a new task proposed in the 6th Text Retrieval Conference (TREC-6). And in CLIR evaluation task of the 9th Text Retrieval Conference (TREC-9), Chinese was taken as the text description language for the first time. So based on the participation in CLIR evaluation task of TREC-9, our Natural Language Processing Research Group develops the research on English-Chinese bi-directional CLIR. We adopt English-Chinese bi-directional query translation as the dominant strategy, in which queries in source language are taken as the translation objects, and English and Chinese electronic dictionaries are taken as the important knowledge source for the acquisition of translation knowledge. Meanwhile, combining with the constructed English and Chinese Monolingual IR systems, the whole English-Chinese bi-directional CLIR process can be implemented successfully. The basic framework of the whole system is shown in Fig.1.

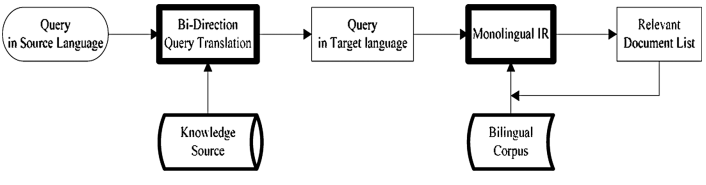


Fig. 1. Basic framework of English-Chinese bi-directional CLIR system

Our research concentrates on three aspects including Chinese word segmentation and indexing, query translation and monolingual retrieval. Through the experimental test, it can be seen that using a more complete electronic dictionary and some basic linguistic analysis tools in CLIR is a practical manner. However, there is a remarkable gap between monolingual IR performance and CLIR performance. Through the study of CLIR methods, we find that if a CLIR system with perfect function is built under ideal status, then the cost of time and resource will be considerable. Thus our research goal is limited to focus on English-Chinese bi-directional CLIR. We try to understand the basic requirement of the efficient CLIR and the resulting problems in the process of system construction.

2 Chinese Word Segmentation and Indexing

No matter for monolingual IR or CLIR, and no matter for Chinese or English, the adopted techniques and methods are mainly based on index wordlist, i.e., using keyword

(or index word) search and word frequency comparison. The content of each document in corpus is represented and indexed by combining words in index wordlist. And a user's query requirement is expressed in a certain kind of form composed of words in index wordlist. Therefore, word segmentation is the most important basic processing procedure and it affects the performance of the whole system. English belongs to Indian-Europe phylum, in which there is a kind of natural separation between two words. Thus, it's very easy to solve the problem of word segmentation in English [5]. However, different with English, there is no obvious border between two words in Chinese. So Chinese word segmentation becomes a much difficult problem [6].

In order to complete the indexing process for query and document in Chinese successfully, Chinese word segmentation must be performed first. That is, the initial Chinese character sequence is segmented into individual words or ngrams. Therefore documents in corpus are processed by using the combination of two different indexing manners.

(1) In word-based indexing manner

Given a document, it is first divided into a sequence of sentences by punctuation such as full stop, comma, etc. Each sentence then passes through the word segmentor and is segmented into a sequence of words. Then a text-layer post-processor will act on the word sequence and generate the final segmentation result. This document post-processor based on the cache is necessary to detect missed or mistakenly segmented words before.

(2) In ngram-based indexing manner

Ngram-based approach takes ngram as an index unit and does not require any linguistic knowledge. It segments texts into strings containing one (unigram) or two (bigram), or more characters. Bigram and trigram are often used in IR. Bigrams take all the two adjacent characters in document as an index term. For example, given a character string $C_1C_2C_3C_4$, then the index terms generated by it are C_1C_2, C_2C_3 and C_3C_4. Since 75% of all available and commonly used Chinese words are made up of two characters, bigram-based approach is an effective approach. The most obvious advantage is its simplicity and ease of application. On the other hand, it can skip the unknown word problem. For example, for proper noun that is not in the dictionary, such as "大亚湾" (a place in southern China), word segmentor will segment it into three characters, i.e. "**大**', "亚', and "**湾**'. When using overlapping bigrams, it will be segmented into two bigrams, i.e. "大亚" and "亚湾". If both bigrams occur in the same document, there is a higher probability that the document concerns "大亚湾" than the documents where the three single characters occur.

Every document in corpus is cut into no more than 64k fragment to make indexing procedure more robust and normalize the document length. After being segmented, information about text id, term frequency, document frequency and term position is stored for the task. No stopword is removed from the inverted file, since the corpus is rather small. In order to optimize the disk space storage and I/O operation in retrieval time, we have also implemented inverted file compression. The file was then decreased to about one half of its original size.

3 English-Chinese Bi-directional Query Translation

Before term matching and relevant document ranking, query or document translation process must be applied first [7]. So CLIR is more complex than monolingual IR. At present, query translation has become the most popular technique for CLIR [8]. Through query translation, CLIR task can be converted into monolingual IR task. A kind of dictionary-based query translation strategy is adopted, assisted with some basic linguistic processing tools. Each query term in source language is replaced with one or several translations in target language extracted from the constructed English and Chinese electronic dictionaries. And the correct translation knowledge in target language corresponding with query in source language can be acquired. Hence the final query form in target language is generated.

3.1 Knowledge Source Construction

The knowledge source used in English-Chinese bi-directional CLIR system mainly includes bilingual dictionary and synonym dictionary. In addition, stopword list and word morphological resumption list are also utilized in our system. In fact, dictionary is a carrier of knowledge expression and storage, which involves almost all information about vocabulary, namely static information.

(1) Bilingual Dictionary

This dictionary is mainly used in translation processing procedure. And it consists of three kinds of dictionary component as follows:

a. Basic Dictionary -- It is a basic knowledge source independent of specific domain, and records basic linguistic vocabulary;

b. Technical Terminology Dictionary -- It records terminology knowledge in a specific technical field and is incorporated into the basic dictionary, which mainly refers to Hong Kong commercial terminology knowledge;

c. Idiom Dictionary -- It records familiar fixed matching phenomena, such as idiom and phrase.

The whole bilingual dictionary involves almost 360,000 lexical entries. And each entry is constructed as the following data structure:

Lexical Information in Source Language	Part-of-Speech Information	Subcategory Information	Concept Number	Matching Information	Semantic Class Code	Translation in Target Language

Two simple examples of particular entry representation form in the bilingual dictionary are listed as the following:

*happiness||n||ng||0||M:种/个;[U];||bbaaa||幸福||||

*实践||v||vv3||2||vv30||2322131||carry out||||

(2) English and Chinese Synonym Dictionary

Actually, this dictionary is an English and Chinese thesaurus, which involves nearly 180,000 entries. All entries are arranged according to specified semantic relations in English and Chinese. It is mainly used in expanding translation that has passed through translation processing, namely query expansion.

While the stopword list is used in tagging the stopwords in queries and documents in English and Chinese, and the morphological resumption list which describes all irregular varieties about vocabulary in English is used in morphological resumption of words with irregular variety forms in English.

3.2 Translation Processing

The English-Chinese bi-directional translation process is mainly composed of the following three parts:

(1) Preprocessing -- including segmentation in sentence level and word level, punctuation tagging, and capital-to-lower letter conversion (for query in English).

(2) Pre-analysis -- including stopwords tagging, word morphological resumption (for query in English) and part-of-speech tagging.

Considering that translation processing is related with some stopwords, the stopwords must be tagged by stopword list. Because there are some words with variety forms in query in English, translation knowledge cannot be induced correctly. So by using bilingual dictionary, morphological resumption lists for irregular variety and heuristics for regular variety, we get English words' original form from the process called "morphological resumption". In order to analyze word part-of-speech in English and Chinese, we have developed a HMM-based (Hidden Markov Mode) Part-of-Speech Tagger [9].

(3) Translation processing -- including translation processes in word level and phrase level.

Word Level Translation. By using the basic vocabulary part of bilingual dictionary, this process mostly implements translation word by word. For word disambiguation, a word may correspond with several kinds of different sense. Word sense is related with particular word, and cannot be given without particular linguistics context. The condition of linguistics context may be syntactic and semantic parameters. This difference mark represents a certain syntactic and semantic feature, and identifies the sense of word uniquely, namely Concept Code. When selecting a particular word, the difference mark of this word should be chosen. The concept code together with the lexical entry can decide a certain word sense to accomplish word sense disambiguation. For machine translation, word disambiguation should be an important problem. But in our CLIR system, in some degree, word disambiguation has not taken some obvious effect to retrieval efficiency. At the same time, in order to provide more query information to retrieval system, by using "English and Chinese Synonym Dictionary", expansion operation is performed for translation knowledge through translation processing. According to various synonymous relations described in the dictionary above, all synonyms corresponding with translation knowledge are listed, namely query expansion process. Thus, more affluent query information can be provided to retrieval system. So the retrieval efficiency is increased greatly, and the retrieval performance is improved.

Phrase Level Translation. This process is implemented based on the idiom dictionary part of bilingual dictionary. The recognition of near distance phrase and far distance phrase in English and various phrase identification in Chinese are the important problems. By adopting Greedy Algorithm, the recognition and

translation processing of near distance phrase in English and noun phrase in Chinese is mainly completed. The phrase identification algorithm is shown simply in the following:

a. Acquiring phrase set taken current query word as head word from bilingual dictionary;
b. Establishing some phrases which take current word as head word and involve the same word number as the member in phrase set;
c. Comparing each one of the established phrases and every member in the corresponding phrase set and obtaining the matched phrase with the maximum length.

4 English and Chinese Monolingual IR

After English-Chinese bi-directional query translation, monolingual IR techniques for English and Chinese are utilized to get relevant document list. Here, a variant of MIT's method and probabilistic method is adopted as the monolingual IR algorithm. And both weight values in title and description fields of each query are available in processing.

4.1 MIT's Search Engine

This is our first statistic-based search engine, tuned by corpus from the 5th Text Retrieval Conference (TREC-5). It scores document by maximum likelihood ratio, put forward by Spoken Language Systems Group in MIT [10].

They propose to use the relative change in the document likelihood, expressed as the likelihood ratio of the conditional and the prior probabilities, as the metric for scoring and ranking the documents in response to query Q, shown as follows:

$$S(D_i, Q) = \frac{p(D_i \mid Q)}{p(D_i)} \ . \tag{1}$$

The term's weight thus can be defined as follows:

$$S_l(D_i, Q) = \sum_{t \in Q} q(t) \log\left(\frac{\alpha * p_{ml}(t \mid D_i) + (1-\alpha) * p_{gt}(t)}{p_{gt}(t)}\right) \cdot \tag{2}$$

where $q(t)$ is the weight of term t in the query, typically its frequency in the query.

$$S_l(D_i, Q) = \sum_{t \in Q} q(t) \log\left(\frac{\alpha * p_{ml}(t \mid D_i) + (1-\alpha) * p_{gt}(t)}{p_{gt}(t)}\right) \cdot \tag{3}$$

where $d_i(t)$ is the occurrence time of term t in document D_i, and k is the number of distinct terms in the document collection.

The Turing-Good estimate for $p(t)$ is given by the following formula:

$$p_{gt}(t) = p_r(t) = r^*/N \,. \tag{4}$$

where $r^* = (r+1)\frac{N_{r+1}}{N_r}$, r is the occurrence time of term t in the document collection, N_r is the number of terms that occur exactly r times in the document collection, and N is the total number of terms observed in the document collection.

Therefore, we could rank every document related with the original query by the formulae above.

4.2 Automatic Relevance Feedback

Automatic feedback is a proven method for improving information retrieval performance. We use a variant of MIT's approach to select terms from the pilot search.

First, if $\frac{S(D_i, Q)}{\max_{D_i} S(D_i, Q)} \le \gamma$, we select D_i and then merge those documents to create a joined document D', if terms in D' satisfy the inequality below:

$$\frac{p(Q' \mid D)}{p(Q')} \ge \frac{p(Q \mid D')}{p(Q)} \,. \tag{5}$$

That is, select terms if $\frac{p_{ml}(t \mid D')}{p_{gt}(t)} \ge 1$, then add them to original query with the weight $-\log(\frac{p(t \mid D')}{p(t)})$ and perform search again.

However, we find that many resulting documents are not related to the query. We further investigate automatic feedback process, and find that the weights of newly added terms should not be equal to that of the original query words, so we set the weight of original query words to the maximum weight of the feedback words and thus optimize the precision and recall of retrieval process.

5 Experimental Evaluation

Based on our work in CLIR evaluation task of TREC-9, English-Chinese bi-directional system has been established. Its whole performance is evaluated on the

basis of English-Chinese bilingual query set, English-Chinese bilingual corpus and a series of evaluation standard provided by TREC.

5.1 Construction of Bilingual Query Set

As the processing object, English-Chinese bilingual query set includes 25 topics totally. And each original query topic in source language includes not only title field, description field and narrative field, but also the corresponding query translation in target language. According to length and content, query is divided into title query (including only words in title field), middle query (including words in description field) and long query (including words in both title and description fields). Details about the expression form of original query are shown in Fig. 2.

<title> World Trade Organization membership
<title-chn> 世界貿易組織(WTO)成員國
<desc> Description:
What speculations on the effects of the entry of China or Taiwan into the World Trade Organization (WTO) are being reported in the Asian press?
<desc-chn> 描述:
亞洲國家新聞對中國或台灣加入世界貿易組織(WTO)的影響持什麼看法?
<narr> Narrative:
Documents reporting support by other nations for China's or Taiwan's entry into the World Trade Organization (WTO) are not relevant.
<narr-chn> 附註:
非亞洲國家支持中國或台灣加入世界貿易組織(WTO)的文章是與此不相關的.

Fig. 2. Expression form of original query

5.2 Construction of Bilingual Corpus

As the basis of Chinese monolingual IR and English-Chinese CLIR, Chinese corpus comes from three news sets including Hongkong Commercial Daily (HKCD), Hongkong Daily (HKD) and TakungPao (TKP). The whole corpus is composed of 127,938 documents. Because news in Hongkong is generally encoded in Big-5 character set, but our Chinese processing tools are mainly applied for code based on GB character set. So for the initial document set, it is necessary to convert Big-5 code to GB code. The Chinese corpus is summarized in Table 1.

Table 1. Date and Size of the Chinese corpus

Source	*Date*	*Size*
HKCD	8/98--7/99	~100MB
HKD	2/99--7/99	~80MB
TKP	9/98--9/99	~80MB

While as the basis of English monolingual IR and Chinese-English CLIR, English corpus comes from Associated Press Newswire 88-90. The whole corpus is composed of 242,918 documents. The English corpus is summarized in Table 2.

Table 2. Date, Size and Document Number of the English corpus

Source	*Date*	*Size*	*Document Number*
AP Newswire 1988	2/88--12/88	~237MB	79,919
AP Newswire 1989	1/89--12/89	~254MB	84,678
AP Newswire 1990	1/90--12/90	~237MB	78,321

5.3 Performance Analysis

In our retrieval experiment, we use corpus provided by TREC-5 as the "training" data set. And eight runs are completed for testing. The details about the eight runs are shown below.

(1) *E-CCLIR1* -- using English long query and without pseudo relevance feedback;
(2) *E-CCLIR2* -- using English long query and pseudo relevance feedback;
(3) *E-CCLIR3* -- using English middle query and pseudo relevance feedback;
(4) *C-CIR4* -- using Chinese long query and pseudo relevance feedback, i.e. Chinese monolingual IR run;
(5) *C-ECLIR1* -- using Chinese long query and without pseudo relevance feedback;
(6) *C-ECLIR2* -- using Chinese long query and pseudo relevance feedback;
(7) *C-ECLIR3* -- using Chinese middle query and pseudo relevance feedback;
(8) *E-ECIR4* -- using English long query and pseudo relevance feedback, i.e. English monolingual IR run.

where (1)-(3) and (5)-(7) runs aim at English-Chinese CLIR and Chinese-English CLIR performance tests respectively, and the other two runs are monolingual IR performance tests used as a baseline of CLIR performance evaluation.

Our best run has achieved the MAP (Mean Average Precision) of 0.3869, which is acquired by using corpus of TREC-5. CLIR runs are all automatic query translation run, using the combination of word-based segmentation approach and ngram-based approach for indexing, while the monolingual IR runs use ngram-based segmentation. Although the results are not as good as that of training results, the run of "E-CCLIR2" still can achieve the MAP of near 0.30. Fig.3 and Fig.4 give the performance description of eight runs respectively.

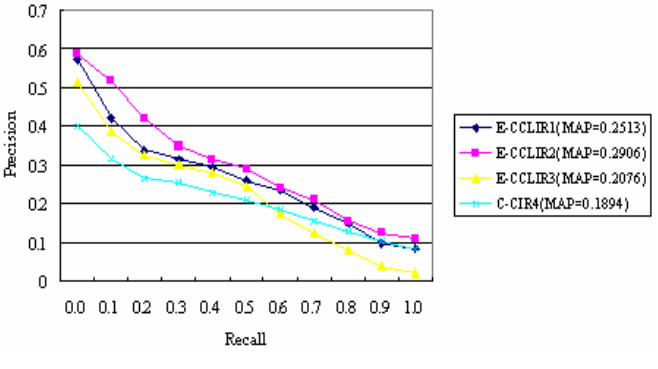


Fig. 3. Precision-Recall curves and MAP values for English-Chinese CLIR performance test

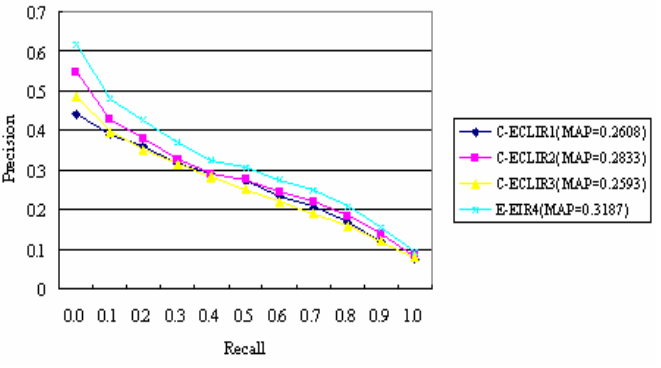


Fig. 4. Precision-Recall curves and MAP values for Chinese-English CLIR performance test

It can be seen from Fig.3 that the result of Chinese monolingual IR run does not run better. We speculate that it may due to our sophisticated segmentation method, which could correctly segment the names of people, place and organization and so on. In other word, indexer based on the combination of two indexing manners performs much better than based on ngram alone. It may also due to Chinese queries used in monolingual IR run. These queries use words and phrases in the style of Mainland China and does not match those from Hong Kong corpus. While for English-Chinese CLIR, using our bilingual knowledge mainly derived from news report of Hong Kong and Taiwan, we are able to make sound translation for the query than that made by human.

It can be seen from Fig.4 that English monolingual IR run is superior to three Chinese-English CLIR runs, and better than Chinese monolingual IR run as shown in Fig.3. Because by comparing Chinese monolingual IR and English-Chinese CLIR, it can be found that indexing manner is no longer an important factor which affects monolingual IR performance. In addition, English-Chinese CLIR runs perform better than Chinese-English CLIR runs. The main reason mainly attributes to that standard query set and corpus provided by TREC are used in English-Chinese CLIR runs, but in Chinese-English runs, the relevant degree between query set and corpus is not better enough.

On the other hand, comparisons between every run and Median are shown in Table 3 and Table 4 respectively.

Table 3. Comparisons between every run and Median in English-Chinese CLIR

	> Median	*= Median*	*< Median*
E-CCLIR1	14	0	11
E-CCLIR2	17	4	4
E-CCLIR3	14	3	8
C-CIR4	15	2	8

It can be observed from Table 3 that "E-CCLIR2" run performs best in English-Chinese CLIR. Specifically, for 25 English query topics, results of 17 testing topics exceed Median.

Table 4. Comparisons between every run and Median in Chinese-English CLIR

	> Median	*= Median*	*< Median*
C-ECLIR1	13	1	11
C-ECLIR2	16	0	9
C-ECLIR3	13	2	10
E-EIR4	18	2	5

It can be observed from Table 4 that "C-ECLIR2" run gets the best performance in Chinese-English CLIR. Specifically, for 25 Chinese query topics, results of 16 testing topics exceed Median.

6 Conclusion

This paper presents some research work on English-Chinese bi-directional CLIR. In the established system, we focus on finding the efficient strategy to implement Chinese word segmentation and indexing, query translation and monolingual IR.

In English-Chinese bi-directional query translation, knowledge source is mainly composed of English and Chinese electronic dictionaries. But except the problem of dictionary integrality, we have to solve the problem of selecting the best translation of a word from dictionary. Therefore, the translation processing patterns in word level and phrase level are established. Combining kinds of information including part-of-speech tag, concept code and so on, the final correct translation knowledge can be acquired. Assisted by query expansion operation, affluent query information can be provided for the subsequent monolingual IR process. The monolingual search engines are accomplished based on MIT's method and probabilistic method. And at the same time, an automatic relevance feedback technique is used. Each of the techniques above can improve the retrieval performance in some degree.

Experimental tests based on the established system give promising results. At present, our research on CLIR mainly focuses on English and Chinese. In the future, we will do some extension of the system, in order to implement multilingual CLIR and cross-language WEB IR processing.

Acknowledgments. This paper is supported by National Natural Science Foundation of China (No. 60533100, No. 70501018).

References

1. Christian Fluhr.: Multilingual Information Retrieval. In: Ronald A. Cole, Joseph Mariani, Hans Uszkoreit, Annie Zaenen, and Victor Joe Zue (eds.): Survey of the State of the Art in Human Language Technology. (1995) 291-305

2. Carol Peters.: Cross-Language Information Retrieval and Evaluation. In: Lecture Notes in Computer Science 2069. Springer-Verlag, Germany (2001)
3. V. A. Pigur.: Multilanguage Information-Retrieval Systems: Integration Levels and Language Support. In: Automatic Documentation and Mathematical Linguistics, Vol.13(1). (1979) 36-46
4. J. Xu, R. Weischedel, and C. Nguyen.: Evaluating a Probabilistic Model for Cross-Lingual Information Retrieval. In: Proc. of 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. New Orleans: ACM Press, (2001) 105-110
5. Chung hsin Lin, and Hsinchun Chen.: An Automatic Indexing and Neural Network Approach to Concept Retrieval and Classification of Multilingual (Chinese-English) Documents. In: IEEE Transaction on Systems, Man and Cybernetics, Vol.26(1). (1996) 75-88
6. W. Oard, and F. Ertunc.: Translation-Based Indexing for Cross-Language Retrieval. In: Proc. of ECIR 2002. (2002) 324-333
7. Ellen M. Voorhees.: Variations in Relevance Judgments and the Measurement of Retrieval Effectiveness. In: Proc. of 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, ACM Press. (1998) 315–323
8. V. Lavrenko, M. Choquette, and W. B. Croft.: Cross-Lingual Relevance Models. In: Proc. of 25th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, ACM. (2002) pp.175-182
9. Wu Li-de, et al.: Large Scale Chinese Text Processing. Fudan University Press. (1997)
10. Kenney Ng.: A Maximum Likelihood Ratio Information Retrieval Model. In: Proc. of 8th Text Retrieval Conference (TREC-8). (1999)

Analyzing Image Texture from Blobs Perspective

Yi Yi Huang, Cun Lu Xu, and Yan Qiu Chen*

Shanghai Key Laboratory of Intelligent Information Processing
School of Information Science and Engineering
Fudan University, Shanghai 200433, China
chenyq@fudan.edu.cn

Abstract. We introduce in this article a blobs perspective for understanding image texture, and the subsequent motivation to characterize texture through analyzing the blobs in the textured image. Three texture description schemes arising from this motivation are discussed and their performance is experimentally evaluated. The experiment results show that a 94.9% correct classification rate on the entire Brodatz set of 112 different types of texture is achieved, which is the highest classification performance to date among the published methods according to the literature survey carried out by the authors of this article.

Keywords: Image texture analysis, Blobs perspective, Statistical Geometrical Features, Statistical Landscape Features.

1 Introduction

Texture is one of the most important issues in computational vision. Decades of research on texture analysis has created hundreds of methods. Further research effort, however, is still needed to discover better approaches to meet the stringent requirements of real-world applications.

A textured monochrome image can be perceived as being made up of constituent bright and dark blobs. The visual appearance of the texture is jointly determined by the shape, gray-level values, and spatial distribution of the blobs. This blobs perspective to view a textured image motivates the belief that texture analysis can be achieved by analyzing the blobs and their spatial configuration.

The structural approach[5] is likely the earliest published method relating to the blobs perspective. It first attempts to find the recurring primitives (subimages) and then describe them and their spatial configuration using a formal language. Such an approach works well on images that show clear-cut primitives. Its performance quickly deteriorates when the primitives are not well defined, which is often the case for natural images.

Statistical Geometrical Features (SGF)[4] adopts a different strategy. Instead of trying to identify the primitives and their configuration, it uses a variable threshold to transform a grey-level textured image into a binary image stack, and whereby automatically reveal the blobs (Figure 1). This process has avoided the difficulties encountered by structural approaches when dealing with natural

* Corresponding author.

R. Lu et al. (Eds.): Chinese-German Ws on Cognitive Systems 2005, LNAI 4429, pp. 96–107, 2007.
© Springer-Verlag Berlin Heidelberg 2007

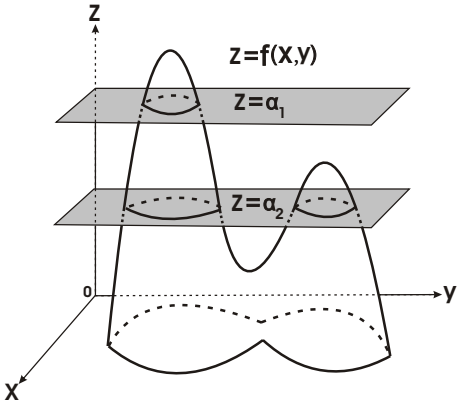


Fig. 1. Binary Image Stack

textures not exhibiting clear-cut blobs. SGF has been reported to have achieved good performance for texture classification[1,2].

Our recent research has discovered that the classification performance of SGF can be improved by directly using the four feature signatures (functions of the threshold value) instead of the 16-dimensional feature vector derived from the signatures. Performance improvement has also been achieved by using a new measure of blob irregularity- Normalized Rotational Inertia.

Another deficiency of Statistical Geometrical Features is its isolated utilization of the topological and geometrical attributes of the connected regions in a binary image, that is, it does not take into account the relationship among the regions in adjacent binary images. This inadequacy has been overcome by Statistical Landscape Features (SLF)[3] which extracts information from the graph (three-dimensional surface) of a textured image (two-dimensional function) by using a horizontal plane at a variable height to intersect the surface to cut out the hills above the planes and the inverted hills below the plane. These hills correlate the corresponding connected regions in SGF as shown in Figure 1, and measuring the hills thus provide more information than measuring the connected regions.

The remaining part of this article is organized into five sections. The next Section 2 introduces Statistical Geometrical Features. The extensions to SGF are discussed in the subsequent Section 3. Section 4 introduces Statistical Landscape Features. Section 5 presents the results of classification experiments using the entire Brodatz texture benchmark. This article is ended with Section 6 giving concluding remarks.

2 Statistical Geometrical Features

The key idea of Statistical Geometrical Features (SGF) is to transform a given gray-scale image into a stack of binary images and then measure the black and white blobs in each binary image.

An $n_x \times n_y$ digital image with n_l grey levels can be represented by a function $f(x, y)$, where $(x, y) \in \{0, 1, ..., n_x - 1\} \times \{0, 1, ..., n_y - 1\}$, and $f(x, y) \in \{0, 1, ..., n_l - 1\}$.

Thresholding an image $f(x, y)$ with a threshold value $\alpha \in \{1, ..., n_l - 1\}$ creates a binary image

$$f_b(x, y; \alpha) = \begin{cases} 1 & \text{if } f(x, y) \geq \alpha \\ 0 & \text{otherwise} \end{cases}. \tag{1}$$

For a given gray-scale image, there are $n_l - 1$ potentially different binary images, i.e., $f_b(x, y; 1)$, $f_b(x, y; 2)$,..., $f_b(x, y; n_l - 1)$. This set of binary images can be imagined as a pile of binary images vertically ordered according to the threshold value, and is therefore termed in this paper as a binary image stack.

For each binary image $f_b(x, y; \alpha)$, we group 1-valued pixels into a number of connected regions. The number of connected regions $NOC_1(\alpha)$ is an important topological attribute indicating the granularity of the binary image. Also informative is the geometrical shape of the regions. An irregularity measure is used to extract shape information for each region:

$$Irregularity = \frac{1 + \sqrt{\pi} \cdot \max\limits_{i \in I} \sqrt{(x_i - \overline{x})^2 + (y_i - \overline{y})^2}}{\sqrt{|I|}} - 1 \tag{2}$$

where $\overline{x} = \frac{\sum_{i \in I} x_i}{|I|}$, $\overline{y} = \frac{\sum_{i \in I} y_i}{|I|}$, I is the set of indices to all pixels in the connected region concerned, $|I|$ denotes the number of elements in I. $(\overline{x}, \overline{y})$ is the center of mass of the region assuming that all the pixels are of identical weight.

Alternatively, the compactness measure can be used:

$$compactness = \frac{4\sqrt{|I|}}{perimeter}, \tag{3}$$

where

$$\begin{aligned} perimeter = \sum_{i \in I} & (f_b(x_i, y_i) \oplus f_b(x_i - 1, y_i) \\ & + f_b(x_i, y_i) \oplus f_b(x_i + 1, y_i) \\ & + f_b(x_i, y_i) \oplus f_b(x_i, y_i - 1) \\ & + f_b(x_i, y_i) \oplus f_b(x_i, y_i + 1)), \end{aligned} \tag{4}$$

where $\oplus$ denotes the logic XOR operator.

Let the irregularity of the i-th connected region of 1-valued pixels of the binary image $f_b(x, y; \alpha)$ be denoted by $IRGL_1(i, \alpha)$. The average (weighted by size) of irregularity of the regions of 1-valued pixels in the binary image $f_b(x, y; \alpha)$ is defined to be

$$\overline{IRGL_1}(\alpha) = \frac{\sum_{i \in I} (NOP_1(i, \alpha) \cdot IRGL_1(i, \alpha))}{\sum_{i \in I} NOP_1(i, \alpha)}, \tag{5}$$

where $NOP_1(i, \alpha)$ is the number of pixels in the i-th connected region of 1-valued pixels of the binary image $f_b(x, y; \alpha)$.

We have by now obtained two functions of the threshold value α. $NOC_1(\alpha)$ describes how the number of connected regions varies with α. $\overline{IRGL_1}(\alpha)$ measures how the average irregularity of the connected regions changes with the threshold value. Following the same procedures, we can obtain $NOC_1(\alpha)$ and $\overline{IRGL_0}(\alpha)$ to describe the connected regions of 0-valued pixels.

To obtain a concise feature vector, each of the four functions is further statistically characterized. By regarding each of the four functions $g(\alpha)$ as a collection of samples $g(1), g(2), ..., g(n_l - 1)$, two statistics, the maximum and the average values, can be obtained:

$$max(g) = \max_{1 \le \alpha \le n_l - 1} g(\alpha), \tag{6}$$

$$E(g) = \frac{1}{n_l - 1} \sum_{\alpha=1}^{n_l - 1} g(\alpha), \tag{7}$$

The function $g(\alpha)$ may also be considered as a probability density function after normalization of its integral to 1. The mean and standard deviation of α are the other two statistics used to characterize the function:

$$E(\alpha) = \frac{1}{\sum_{\alpha=1}^{n_l - 1} g(\alpha)} \sum_{\alpha=1}^{n_l - 1} \alpha \cdot g(\alpha) \tag{8}$$

$$SD(\alpha) = \sqrt{\frac{1}{\sum_{\alpha=1}^{n_l - 1} g(\alpha)} \sum_{\alpha=1}^{n_l - 1} (\alpha - E(\alpha))^2 \cdot g(\alpha)} \tag{9}$$

where $g(\alpha)$ is one of the four functions $NOC_1(\alpha)$, $NOC_0(\alpha)$, $\overline{IRGL_1}(\alpha)$, and $\overline{IRGL_0}(\alpha)$.

By now, we have obtained sixteen feature measures for a textured image, four statistics obtained from $NOC_1(\alpha)$, four from $NOC_0(\alpha)$, four from $\overline{IRGL_1}(\alpha)$, and another four from $\overline{IRGL_0}(\alpha)$.

3 Extended Statistical Geometrical Features

Our recent study has discovered that the direct use of the four functions $NOC_1(\alpha)$, $NOC_0(\alpha)$, $\overline{IRGL_1}(\alpha)$, $\overline{IRGL_0}(\alpha)$ with $L1$ distance metric

$$\rho(f, g) = \int |f(x) - g(x)| \, dx \tag{10}$$

gives better classification results than the further-extracted 16 statistics, indicating that the extraction of the 16 statistics from the four functions has led to inadvertent loss of discrimination power.

In reminiscence, the original motivation was to employ the second stage feature extraction to reduce the dimensionality of the texture descriptor. This

reduction of dimensionality, intriguingly, does not seem to have a positive effect on classification performance as long as the k-Nearest Neighbor Classifier is concerned.

Another aspect for improvement is the irregularity measure. Experimental investigation using the entire Brodatz texture set shows that the texture signatures $NOC_1(\alpha)$ and $NOC_0(\alpha)$ alone can achieve a correct classification rate of 86.3% while $\overline{IRGL_1}(\alpha)$ with $\overline{IRGL_0}(\alpha)$ only achieves 70.70%. We have recently developed a new irregularity measure, Normalized Rotational Inertia, that gives better performance.

Being a physical concept, rotational inertia measures the resistance of an object to change of rotation speed. For an object consisting of discrete mass points m_i at distances r_i from the rotation axis, the rotational inertia is

$$J = \sum_i r_i^2 m_i \tag{11}$$

To apply rotational inertia to images, we assume each pixel possesses one unit of mass. For a connected region R in an image, the rotational inertia with the center of mass being the axis is:

$$J = \sum_{i \in R} \left((x_i - \overline{x})^2 + (y_i - \overline{y})^2 \right) \tag{12}$$

where $\overline{x} = \frac{\sum_{i \in R} x_i}{|R|}$ and $\overline{y} = \frac{\sum_{i \in R} y_i}{|R|}$

For a region with a given number of pixels, rotational inertia reflects its irregularity, that is, the more irregular the region is, the higher rotational inertia it will possess. A solid disk has the least rotational inertia among the regions of the same size. For regions of the same shape, rotational inertia is proportional to the size of the region. To eliminate the size dependency, we normalize it against the rotational inertia of the solid disk of the same size:

$$\tilde{J} = \frac{J(R)}{J\left(D(R)\right)} \tag{13}$$

where $D(R)$ is the solid disk of the same size as region R.

4 Statistical Landscape Features

To overcome the deficiency of isolated utilization of the topological and geometrical attributes of the connected regions in binary images by Statistical Geometrical Features, we develop Statistical Landscape Features to extract information from the graph (a three-dimensional surface) of a textured image (two-dimensional function).

An $n_x \times n_y$ digital image with n_l gray levels can be represented by a function

$$z = f(x, y), \tag{14}$$

where $(x, y) \in \{0, 1, \cdots, n_x - 1\} \times \{0, 1, \cdots, n_y - 1\}$ and $z \in \{0, 1, \cdots, n_l - 1\}$. The graph of this function is bounded in the box $\Omega = \{(x, y, z) \in \mathbb{Z}^3 : 0 \leq x \leq n_x - 1, 0 \leq y \leq n_y - 1, 0 \leq z \leq n_l - 1\}$ and divides the box Ω into a lower part

$$\begin{aligned} \Omega_L^f = &\{(x, y, z) \in \mathbb{Z}^3 : 0 \leq x \leq n_x - 1, \\ &0 \leq y \leq n_y - 1, 0 \leq z \leq f(x, y)\} \end{aligned} \tag{15}$$

and an upper part

$$\begin{aligned} \Omega_U^f = &\{(x, y, z) \in \mathbb{Z}^3 : 0 \leq x \leq n_x - 1, \\ &0 \leq y \leq n_y - 1, f(x, y) \leq z \leq n_l - 1\} \end{aligned} \tag{16}$$

as illustrated in Fig. 2(a). The part of the graph within the box Ω is denoted by $\Omega_{z=f(x,y)}$. This box Ω can also be cut by the plane $z = \alpha$, $\alpha \in \{0, 1, \cdots, n_l - 1\}$ into a lower part

$$\begin{aligned} \Omega_L^\alpha = &\{(x, y, z) \in \mathbb{Z}^3 : 0 \leq x \leq n_x - 1, \\ &0 \leq y \leq n_y - 1, 0 \leq z \leq \alpha\} \end{aligned} \tag{17}$$

and an upper part

$$\begin{aligned} \Omega_U^\alpha = &\{(x, y, z) \in \mathbb{Z}^3 : 0 \leq x \leq n_x - 1, \\ &0 \leq y \leq n_y - 1, \alpha \leq z \leq n_l - 1\} \end{aligned} \tag{18}$$

as shown in Fig. 2(b). The part of the plane $z = \alpha$ within the box Ω is denoted by $\Omega_{z=\alpha}$.

The proposed features are extracted from the two intersections $A^\alpha = \Omega_L^f \cap \Omega_U^\alpha$ and $B^\alpha = \Omega_U^f \cap \Omega_L^\alpha$ as illustrated in Fig. 2(c). As shown in the figure, the first intersection is composed of hills above the plane $z = \alpha$ while the second of inverted hills below the plane. To be precise, the set A^α consists of a number n_A^α of solids (the hills) A_i^α, $i = 1, 2, \cdots, n_A^\alpha$, that is, A^α is partitioned into a number n_A^α of subsets, in such a way that

$$A^\alpha = \bigcup_{i=1}^{n_A^\alpha} A_i^\alpha, \tag{19}$$

where for any A_i^α and A_j^α, $A_i^\alpha \neq \Phi$, $A_i^\alpha \cap A_j^\alpha = \Phi$ for $i \neq j$, Φ denotes the empty set. The above mentioned procedure also applies to B^α.

The numbers of solids A_i^α and B_i^α reflect the granularity of a landscape. The height of a solid also carries shape information. We will discuss the height descriptor in detail in the following paragraphs.

For a given coordinate triple (x, y, z), its 6-neighborhood is defined to be the set

$$\begin{aligned} N_6(x, y, z) = &\{(x + 1, y, z), (x - 1, y, z), \\ &(x, y + 1, z), (x, y - 1, z), \\ &(x, y, z + 1), (x, y, z - 1)\}. \end{aligned} \tag{20}$$

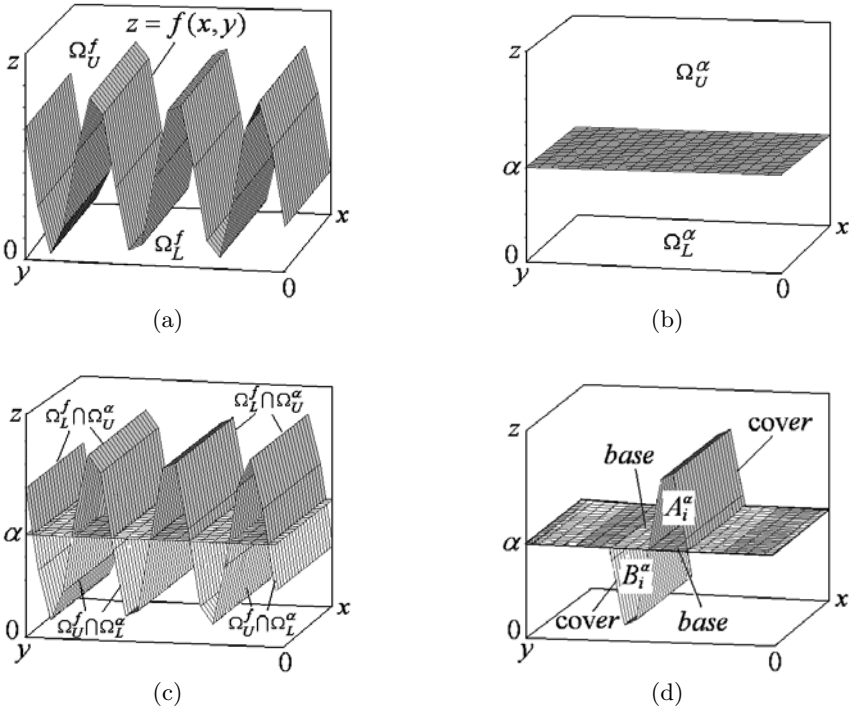


Fig. 2. Partition of bounding box. (a) $z = f(x, y)$ divides Ω into Ω_L^f and Ω_U^f. (b)The plane $z = \alpha$ cuts Ω into Ω_L^α and Ω_U^α. (c) The proposed features are extracted from $A^\alpha = \Omega_L^f \cap \Omega_U^\alpha$ and $B^\alpha = \Omega_U^f \cap \Omega_L^\alpha$. (d) Two components of the surface of a solid: a base and a cover.

A point p at (x, y, z) is a 6-neighbor of p' at (x', y', z') if and only if $(x, y, z) \in N_6(x', y', z')$. Two points p and p' are 6-neighbors if and only if p is a 6-neighbor of p'. A 6-connected path from p_i to p_j is a sequence of points $\{p_{l_k}\}_{k=1}^m$ such that p_{l_k} and $p_{l_{k+1}}$, $k = 1, \cdots, m-1$, are 6-neighbors and $p_i = p_{l_1}$ and $p_j = p_{l_m}$. The length of this path is $m - 1$. Apparently, for any pair of points in the same A_i^α or B_i^α, there is at least one 6-connected path between them in A_i^α or B_i^α. A set is said to be a 6-connected set if and only if there is at least one 6-connected path for each pair of points in the set. So both A_i^α and B_i^α are 6-connected sets.

A point in a 6-connected set is an inner point if and only if all of its 6-neighbors belong to this set; otherwise it is a boundary point. The boundary points of a solid A_i^α is composed of a base, denoted by $Base(A_i^\alpha)$, and a covering surface, denoted by $Cover(A_i^\alpha)$, as shown in Fig. 2(d); that is,

$$\{(x, y, z) \in A_i^\alpha : N_6(x, y, z) \not\subseteq A_i^\alpha\} = Base(A_i^\alpha) \cup Cover(A_i^\alpha). \tag{21}$$

Those boundary points of a solid located at the plane $z = \alpha$ comprise its base, while the remaining points form its cover. $Cover(B_i^\alpha)$ is similarly defined.

We now consider the height of a solid. For a given plane $z = \alpha$ and a point $p(x, y, z) \in A_i^\alpha$, the perpendicular distance from p to the plane $z = \alpha$ is the length of the vertical 6-connected path

$$\{(x, y, z), (x, y, z-1), \cdots, (x, y, \alpha)\}. \tag{22}$$

If we regard the plane $z = \alpha$ as the horizon, the altitude of a point p is the perpendicular distance from p to the horizon. So the altitude of point $p(x, y, z)$, based on the horizon or the plane $z = \alpha$, can be computed by the formula:

$$l(x, y, z; \alpha) = |z - \alpha|. \tag{23}$$

Those points in $Cover(A_i^\alpha) \cap \Omega_{z=f(x,y)}$ are of concern. If $p(x, y, z) \in Cover(A_i^\alpha) \cap \Omega_{z=f(x,y)}$, then

$$l(x, y, z; \alpha) = card\{\{(x, y, z') \in A_i^\alpha\}\} - 1. \tag{24}$$

So for all points in $Cover(A_i^\alpha) \cap \Omega_{z=f(x,y)}$, the sum of their altitudes is

$$\sum_{j=1}^{card\{Base(A_i^\alpha)\}} l(x_j, y_j, z_j; \alpha) = card\{A_i^\alpha \setminus Base(A_i^\alpha)\}, \tag{25}$$

where $p_j(x_j, y_j, z_j) \in Cover(A_i^\alpha) \cap \Omega_{z=f(x,y)}$ and the symbol $\setminus$ denotes the set difference operation. The average height $h_{A_i^\alpha}^\alpha$ of A_i^α can therefore be computed by

$$h_{A_i^\alpha}^\alpha = \frac{card\{A_i^\alpha \setminus Base(A_i^\alpha)\}}{card\{Base(A_i^\alpha)\}}. \tag{26}$$

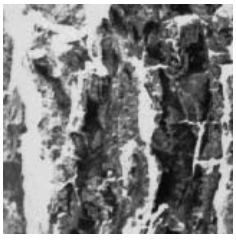
Then the average height of A^α is

$$h_A^\alpha = \frac{1}{n_A^\alpha} \sum_{i=1}^{n_A^\alpha} h_{A_i^\alpha}^\alpha. \tag{27}$$

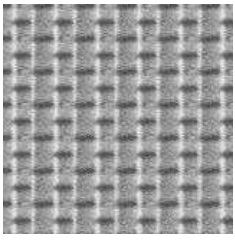
The above definition also applies to the weighted average height h_B^α.

The average height of a 6-connected solid A_i^α or B_i^α is different from the height as defined in solid geometry. The conventional height of a pyramid in geometry is the perpendicular distance from the vertex to the base plane, which is independent of its volume and the area of its base. The definition adopted here is related to its volume and the area of its base, therefore, it is more informative for describing the shape of a 6-connected solid. A solid that appears a right prism and is associated with certain texture, differs from the one represented by a solid like a right pyramid. When the shape of this solid is a right prism, its average height is equal to the height in solid geometry; when the shape of this solid is a right pyramid, its average height is less than the height in solid geometry.

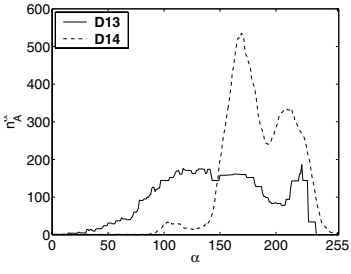
By now, for a given textured image, four functions of α, n_A^α, n_B^α, h_A^α and h_B^α, have been obtained, with each being one feature curve. We use these four feature curves to characterize the texture. Fig. 3 shows two sample textured images and their feature curves. The original images, given in Fig. 3(a) and (b), are both of size 160×160 with 256 gray levels and are sub-images of D13 and D14 taken from the Brodatz texture set [6] respectively. Their feature curves are shown in Fig. 3(c)-(f).



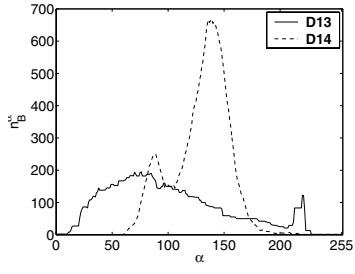
(a) D13



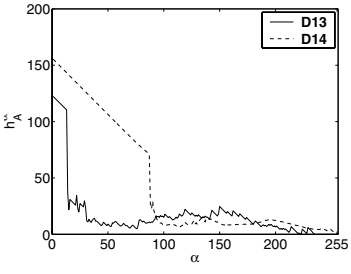
(b) D14



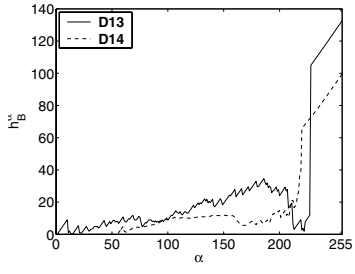
(c) Solid count n_A^α



(d) Solid count n_B^α



(e) Average height h_A^α



(f) Average height h_B^α

Fig. 3. Sample sub-images of Brodatz texture D13 and D14 and their feature curves

The above mentioned process has produced a set of six feature curves for a given textured image as a texture description. These feature curves can be directly used for texture classification and segmentation. There are many choices to measure the distance between two curves. We select the L_1 distance (or city

block distance) based on its good experimental results. Let f_1, f_2, f_3, f_4 respectively denote the feature curves n_A^α, n_B^α, h_A^α and h_B^α of a textured image I, and let f_1', f_2', f_3', f_4' respectively denote the six feature curves of another textured image I'. Then the difference between two textured images I and I' can be calculated by the following formula:

$$d(I, I') = \sum_{i=1}^{4} \sum_{\alpha=0}^{n_l - 1} |f_i(\alpha) - f_i'(\alpha)|. \tag{28}$$

5 Experimental Comparison and Discussions

We now compare the three texture description schemes, Statistical Geometrical Features, Extended Statistical Geometrical Features, and Statistical Landscape Features in classification performance using Brodatz textures[6]– a de-facto standard benchmark. The entire set of 112 pictures of texture was used to make the comparison comprehensive and fair. Each 640×640 image is segmented into sixteen non-overlapping 160×160 sub-images.

The k-Nearest Neighbor Classifier and the "leave one out" evaluation procedure are employed to compare the performance of the three schemes. Each component of the feature signature is normalized against its mean over all the training samples to equalize the contribution:

$$\tilde{x} = \frac{|S|}{\sum_{i \in S} x_i} x \tag{29}$$

where S is the sample set, $|S|$ denotes the number of elements in S.

For each texture description scheme, 16 rounds of classifications are conducted. In each round, one sub-image was taken from the sixteen sub-images of each Brodatz image as test sample, and the other fifteen sub-images as training samples. There are 112 test sub-images and 112×15 training sub-images for each classification round. The correct classification rate averaged over the 16 rounds is used as the performance index. The results as shown in Table 1 indicate:

1. The direct use of feature functions and the new irregularity measure (Normalized Rotational Inertia) has increased the correct classification rate to 91.9% from the original 85.6% for Statistical Geometrical Features.
2. Statistical Landscape Features has achieved a correct classification rate of 94.9% which is higher than the existing published methods according to the literature survey done by the authors.

Since the three texture descriptors discussed in this article all use a sequence of thresholds to extract features from textured images, their computational complexity is proportional to the number of thresholds (slices) used. We now investigate the correlation between the performance and the number of thresholds to find out the optimum threshold number. Figure 5 shows the correlation. It is observed that the performance initially rises as the number of thresholds increase

Table 1. Correct Classification Rates (CCR) on the entire Brodatz set of textured images

	CCR(%)
NOC_1 and NOC_0 signatures (SGF)	86.3
$\overline{IRGL_1}$ and $\overline{IRGL_0}$ signatures (SGF)	70.7
SGF four signatures	87.5
$\overline{IRGL_1}$ and $\overline{IRGL_0}$ signatures (NRI)	76.7
Extended SGF	91.9
h_A^α and h_B^α signatures (SLF)	90.8
SLF	94.9

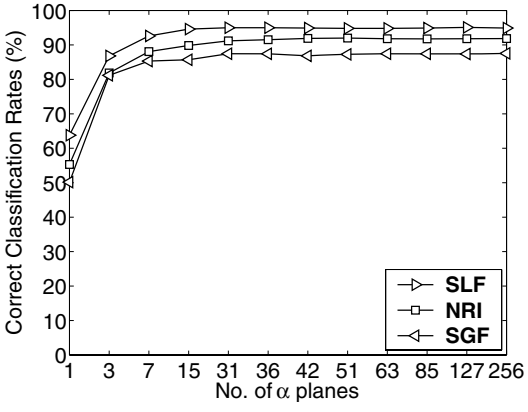


Fig. 4. Correct classification rates versus number of thresholds

from 1 through 32 and stabilizes afterwards. This suggests that 32 thresholds is a good choice.

6 Concluding Remarks

We have discussed in this article the blobs perspective to view textured images, and three texture description schemes, Statistical Geometrical Features, Extended Statistical Geometrical Features, and Statistical Landscape Features motivated from this idea. Their performance is experimentally evaluated. It is noted that Statistical Landscape Features achieves a 94.9% correct classification rate on the set of all 112 Brodatz textures. This is the highest classification performance to date among the published methods according to the literature survey carried out by the authors.

Acknowledgments. This research work is supported by National Natural Science Foundation of China, Grant No. 60575022; Specialized Research Fund for

the Doctoral Program of Higher Education, Grant No. 20050246063; and Science and Technology Commission of Shanghai Municipality, Grant No. 04JC14014

References

1. Singh M, Singh S, Spatial Texture Analysis: A Comparative Study. Proc. 15th International Conference on Pattern Recognition (ICPR'02) 1, pp. 676 - 679, 2002
2. Dai XY, Maeda J, Unsupervised Segmentation of Natural Images, Optical Review, Vol. 9, No. 5, pp 197-201, 2002
3. Xu CL, and Chen YQ, Statistical Landscape Features for Texture Classification, Proc. 17th International Conference on Pattern Recognition, Oxford, UK, Vol. 1, pp 676-679, 2004
4. Chen YQ, Nixon MS, Thomas DW, Statistical geometrical features for texture classification. Pattern Recognition 28(4), pp. 537-552, 1995
5. Haralick RM. Statistical and structural approaches to texture. Proc. IEEE 67(5), pp. 786-804, 1979
6. Brodatz P, Textures: A Photographic Album for Artists and Designers, Dover, Paris, 1966

Author Biographies

Yi Yi Huang is a final year student at Department of Computer Science and Engineering of School of Information Science and Engineering, Fudan University, Shanghai, China

Cun Lu Xu is a postgraduate student pursing the PhD degree at the Department of Computer Science and Engineering of School of Information Science and Engineering, Fudan University, Shanghai, China

Yan Qiu Chen received his PhD from Southampton University, UK in 1995; and his BEng and MEng from Tongji University, Shanghai, China in 1985 and 1988 respectively. Dr Chen joined School of Information Science and Engineering of Fudan University, Shanghai, China in 2001 where he is currently a full professor. He was an assistant professor with School of Electrical and Electronic Engineering of Nanyang Technological University, Singapore from 1996 through 2001; and was a postdoctoral research fellow with Glamorgan University, UK in 1995. Dr Chen's current research interest includes optical imaging, 2-D and 3-D image analysis and understanding. He has authored over 70 research papers, of which more than 30 are published in international research journals.

Access to Content

Dietrich Klakow

Saarland University, 66041 Saarbrücken, Germany
Dietrich.Klakow@LSV.Uni-Saarland.De

Abstract. Enabling everybody to access content in a very simple and user friendly way becomes an increasingly important topic as content availability still increases exponentially. In this paper we will first address the problem of content annotation. The core part of the paper is a description of the LISy system which analyses TV content and allows the user to access it using a natural language dialogue system. Finally we will give a brief overview of the SmartWeb system which extends the capabilities of LISy to the access to the semantic web.

Keywords: Content Analysis, Multimedia Information Systems, User Interfaces, Intelligent Web Services and Semantic Web.

1 Introduction and Motivation

Computers provide us with more information than we can reasonably consume. This phenomenon of information overflow is well known for a couple of years. However, it now also enters our daily lives.

For example in the living room we are going to face this problem. On the one hand side, hard disk recorders recently appeared that can hold about 100 hours of video nowadays. As Moore's law continues to be valid we will have devices that can hold about 1000 hours of video in about five years from now. Moreover these devices become networked and users can also access the content stored on their own PCs or information from the internet.

Also mobile phones can now provide access to the internet. The new generation of smart phones and PDAs has powerful displays and hence can easily display web pages. However, they lack the key board that is necessary to search for information. Methods for robust speech input would make using the internet on a smart phone much easier.

The topic of accessing content has different aspects to it. First of all the content has to be analyzed and structured [1][2]. This can be done either manually or automatically. In particular given the exponential growth of content, manual annotation is not really an option.

The second aspect is the indexing and retrieval. Google is sufficient for normal computers. On devices with smaller displays or no keyboards at all more specific mechanisms that do not require inspecting several web pages but provide the desired information directly in condensed form. Here question answering comes into play [3][4].

R. Lu et al. (Eds.): Chinese-German Ws on Cognitive Systems 2005, LNAI 4429, pp. 108 – 117, 2007.
© Springer-Verlag Berlin Heidelberg 2007

Finally in absence of keyboards suitable mechanisms for users to simply ask for information have to be provided. Speech is very expressive and flexible enough to provide such a user interface. Presently however speech recognition lacks the necessary robustness [5][6].

The paper is structured as follows: we will first present some algorithmic work in TRECVID concept annotation based purely on speech recognition output. The main part of the paper is devoted to a description of the LISy project. Finally we briefly describe the Smart Web project.

2 Annotating Multimedia Content

One of the first steps in accessing content is very often its annotation. The annual TRECVID evaluation organized by NIST tries to address this issue. The TRECVID task consists of different subtasks. In this chapter we will describe one specific activity in that area: we will focus on the TRECVID concept annotation task.

2.1 The Video TREC Concept Annotation Task

The task is to assign multiple labels to each video shot. These labels are usually motivated from the images and examples are “car”, “outdoors” or “horse”. An example of a key frame ant is annotation can be seen in figure 1. Overall there are 75 different concept labels. On average each shot has 3.5 annotations.

For each shot there is a key-frame and the transcript from a speech recognizer (ASR) available for extracting features. All participants so far focused on using visual features to annotate these concepts.

The ASR output has not been used so far. When watching and listening to typical shots the idea of using the speech part seems strange. For example the recognizer

Fig. 1. Example key frame from a shot from TREVID annotated with the labels *car, transportation, vehicle, outdoors, non-studio setting, snow*

output "FROM A. B. C. THIS IS WORLD NEWS TONIGHT WITH PETER JENNINGS REPORTING TONIGHT FROM" and the corresponding key-frame (not shown here on purpose to illustrate the challenge of the task) should be annotated with `text_overlay, non-studio_setting, outdoors, man_made_scene, building`. In this section we will show that a reasonable performance can be achieved by only considering the ASR output and that it is not falling far behind image based approaches.

2.2 Feature Extraction from Speech Recognition Output

The first step in any classification task is the feature extraction. For this we apply standard language processing tools. However, they are not directly applicable to the ASR output. We first have to restore proper capitalization and also to insert sentence boundaries. Now we run a named entity tagger to extract the names of people, locations and organizations. As this provides a very restricted feature set we also run a part-of-speech tagger on the data and extract all the nouns. To generalize the nouns we also look up hypernyms in WordNet [7]. Overall, we used about 39000 features, 22000 of them are the words from the ASR transcript.

2.3 Classification Schemes and Results

These features are used in different classifiers all being variants of the Naïve Bayes classifier.

Naïve Bayes: This is the standard version and we used the Weka toolkit for it [8].

Maximum Entropy: Here we used a maximum entropy toolkit to estimate the probabilities used for classification. The features are used as constraints. A Gaussian prior is used for smoothing. For this purpose the Stanford toolkit was used.

Language Model (LM): This version uses smoothing techniques from statistical language modeling to estimate the probabilities. The probability is estimated using absolute discounting

$$P(x_i \mid \omega_k) = \begin{cases} \dfrac{N_{\omega_k}(x_i) - d}{N_{\omega_k}} + \alpha \dfrac{1}{V} & \text{if } N_{\omega_k}(x_i) > 0 \\ \alpha \dfrac{1}{V} & \text{else} \end{cases}$$

Where x_i is the feature, ω_k is the class label, d is the discounting parameter, d the backing off-weight and V the number of features. Note that the backing-off weight can be determined from the normalization of the probabilities. The discounting parameter can be estimated using leaving-one-out.

Table 1. Comparison of the average precision for a selection of concepts

Concept \ Classifier	Max. Ent.	Naïve B.	LM
Weather	0.34	0.51	0.44
Basketball	0.34	0.27	0.39
Face	0.53	0.57	0.58
Sky	0.11	0.11	0.12
Indoors	0.41	0.44	0.47
Beach	0.02	0.01	0.01
Vehicle	0.11	0.10	0.11
Car	0.09	0.07	0.09

Table 1 gives a comparison of the three classifiers for a couple of selected concepts. We observe that there is always a concept where one of the classifiers has an advantage over the other two.

Table 2 provides an overview of the mean average precision over all 75 concepts. The maximum entropy approach and the Weka implementation of Naïve Bayes have a comparable performance. The language model approach outperforms the other two. The only difference is the use of absolute discounting as the smoothing method. However, in speech recognition absolute discounting is also the best performing smoothing technique. Hence this result is reasonable. Just as a reference point we also used the SVM implementation of Weka. It is better than maximum entropy and naïve bayes but not as good as the language model.

Table 2. Comparison of the introduced classifiers and a SVM as provided by Weka

Method	**Mean Average Precision**
Maximum Entropy	0.100
Naïve Bayes	0.102
SVM	0.116
Language Model	0.125

2.4 Future Work

As one of the next steps we want to combine this ASR based approach of concept annotation with the image based approach. Purely based on images a mean average precision of 0.16 can be achieved. It is unclear how complementary the information from the speech and the images are. Naively we would expect so but only experiments will show whether this is really true.

3 Access to Multimedia Content

As outlined in the introduction, there are various systems that analyze content and make it accessible to the user like e.g. the Informedia project at CMU. In this paper we describe the LISy project. LISy stands for Living room Information System. The project was done by a group eight researchers at Philips research from 2001 until 2003 and coordinated by the author.

3.1 The LISy-Project

The goal of LISy was to provide an intelligent user interface to access automatically annotated multimedia content. In this sense it goes well beyond other projects: because of the large interaction bandwidth of speech it uses a full fledge dialog system for interaction with the user. It can provide answers to questions like “What happened in New York today?”, “Give me recent information about China.” or “Give me today’s baseball results.”.

The hardware platform is a hard disk recorder. To simplify system development and integration the Siemens-Fujitsu Multitainer has been selected because it can be run under the Linux operating system and allows for simple integration of other PC hardware. For example at the time of the project, the Multitainer was available with a 40 GByte hard disk. For the purpose of the project we integrated a 120 GByte hard

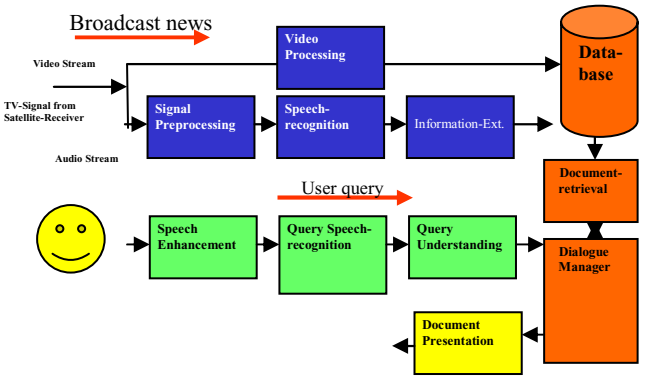


Fig. 2. Architecture of the living room information system LISy

disk. The processor is a 1.3 GHz Celeron. Connected to the system are a remote control, a touchpad and a plasma screen. Providing robust real time performance was essential in the project. In the next section the various software components of the system are described. You can find an overview in figure 2.

3.2 Processing the TV-Broadcast

The processing of the broadcast news is focusing on the speech part. Except for a very simple key frame extraction, no video processing is done. The reason for this is that during the course of the project we discovered that the essential information for later retrieval is contained in the speech part. Very often the images just illustrated the story without providing any additional explicit information.

3.2.1 Segmentation

TV broadcasts are a continuous stream. The very first task is to segment it. For this purpose we used and algorithm based on the Bayesian information criterion [9]. This is a standard procedure in speech recognition. It hypothesizes a shot boundary based on a significant change in the audio signal. This hypothesis is then tested by training Gaussian models for the complete segment as well as for the two possible subsegments. Depending on the likelihood ratio the hypothesized boundary is accepted or rejected.

Experiments with the running system have shown that this segmentation is sufficient to provide the user with the stories he or she wants even so it is not always perfect. In the long run integration of video information may be interesting.

3.2.2 Broadcast News Speech Recognition

The speech recognizer uses a core vocabulary of 64000 words from the broadcast news corpus as provided by LDC [10]. This corpus is from the mid nineties. Hence there is a small mismatch. To reduce the rate of out-of-vocabulary words even further, we augmented the vocabulary with frequent words from the web pages of American TV stations.

The acoustic model is trained on the 130 hours of official broadcast news training data. It is a triphone model with no speaker clustering or adaptation. The search of the system was tuned such as to provide about real time recognition of the broadcast [11].

3.2.3 Information Extraction

In the course of the project a maximum entropy based named-entity tagger was developed jointly with Aachen University.

3.3 Understanding the User

Parallel to the processing of the TV broadcast is the processing of the user utterance. It is also divided into three sub-tasks which go from signal processing to extracting meaning. However, the individual modules differ significantly because of very different constraints.

3.3.1 Improving Signal Quality

Living room acoustics is particularly challenging. If the user is not using a close talking microphone the user's speech signal is disturbed by all kinds of noises like the sound of the TV set or other people speaking in the room. Even reverberation - usually hardly noticed by humans - causes sever degradation of the speech recognizer's performance.

To address this problem we have done initial experiments using a microphone array made out of four simple and cheap microphones. For a large vocabulary task we observed an improvement of about 50% relative in word error rate for a distance of two meters from the array.

3.3.2 User Speech Recognition

The speech recognizer for the user utterances is based on the Philips dictation speech recognizer because in case the recognition of user utterances it is important to deliver results immediately after the end of the utterance. Also, here the user is known and hence speaker adaptation can be used. We used a combination of MLLR and MAP.

For language model training a corpus of typical user utterances was collected. As this corpus could not cover all possible queries being asked to the system we used a rule based scheme to generate automatically new questions based on templates derived from the corpus and a collection of topics about which questions could be asked.

This extended corpus was then used to train a 1k vocabulary and the language model, a trigram using marginal backing-off.

3.3.3 Speech Understanding

The speech understanding is done by island parsing. This has the advantage that the module is robust against spontaneous speech as well as recognition errors. The rules of the grammar where written such that they cover the collected data as good as possible. However, all items in the corpus that can be asked for are replaced by general classes (e.g. PERSON_NAME, LOCATION). These rules are later extended automatically to cover names of other people not present in the training data as well.

Certain rules are written independent of the corpus. For time expressions, we used the grammar that has been created for the Philips train timetable information system.

Independent of the understanding of the question is the grammar to process clarification and selection questions. This grammar was written completely independent of any data collection only anticipating what a user may want to say. Practical experiments with the system have shown that this approach provides sufficient coverage because this part of the application is well-structured and possible user utterances can be enumerated in principle.

3.4 Database and Retrieval

This is the back-end of the system. It stores the data in a suitable format for multimedia documents, provides an adjusted retrieval module as well as a dialogue manager that handles the interaction with the user.

3.4.1 Data Base

We use an SQL-database. For each segment of the video data we store in the data base a lattice from ASR as well as the best hypothesis. The lattice is annotated with confidence values which can be used later in retrieval. Also, the key-frames and the results of the information extraction are stored here. Finally a time stamp for the beginning of the segment and a link to the video stream are stored for later replay.

3.4.2 Fuzzy Retrieval

Traditional IR assumes that the document and the query are textual input. In spoken document retrieval the documents are the result of a speech recognizer but the query is still written. The situation here is that both the document and the query are spoken. In particular the uncertainties induced by the spoken queries require special treatment.

To this end, the standard tf-idf approach is extended. We calculate the expectation value of tf-idf with respect to the confidence values provided by the speech recognizer. It can be shown that this results in a separable expression. Hence, retrieval can be performed as efficiently as with tf-idf. We get a consistent improvement for different settings of the speech recognizer.

3.4.3 Dialogue Management

The dialogue manager's task is to handle the interaction with the user. It is written in HDDL developed at Philips for the train time table information system [11].

3.5 Updating Language Resources

The TV-broadcast domain does not really exist when considering life data. The domain is constantly shifting and evolving. New people or organizations appear on the scene. Topics become fashionable and disappear again. Hence, the system has to be constantly changing without manual interference.

Several modules in the system need a continuous update of the vocabulary: the document speech recognizer, the query speech recognizer as well the query speech understanding.

We have developed an additional module - not shown in the overview of the system architecture - that harvests the specific parts of internet each night for new items. It turned out to be most useful to crawl web pages of broadcasting and news agencies. Those pages are then analyzed and new words and entities are extracted. Based on a simple salience criterion, we pick additional terms. It turns out that five to ten new terms per day are sufficient.

For the update of the speech recognizer we use an automatic grapheme-to-phoneme conversion tool. The language model is updated using the context provided by the internet pages providing the new term. For the speech understanding we identify automatically a suitable terminal of the grammar. This automatic update of the grammar often results in rules that contradict a normal linguistic grammar. In case of doubt, we add the new term which results in a grammar that covers all user utterances but is over generating at the same time. Because we use a separate language model for speech recognition this over generation does not affect system performance.

3.6 Implementation

The system is implemented using a multi-blackboard architecture. To this end, we extended the SmartKom architecture [12]. Changes where done in particular to increase the efficiency of the architecture.

Two specific blackboards are of interest. One is the interface between the understanding module and the dialogue manager. In figure 2 only the speech understanding module is shown. However, the selection from a list of alternatives can also be done by using the remote control or a touch pad. Both modalities have theirs own interpretation modules writing to the same blackboard as the speech understanding.

Similarly on output, different devices read the blackboard on which the dialogue manager puts its responses to the user's utterance. Each output device will generate its own specific representation from this data: the plasma screen will only show the five best results from a query whereas the touchpad can also display intermediate results (e.g. the best hypothesis from the speech recognizers) as well as buttons for help or reset. Retrieved video segments will only be displayed on the plasma screen.

4 Access to the Semantic Web

In the LISy project, all the data was annotated automatically and no special structure of the content was provided. The SmartWeb project [13] goes beyond this. It not only allows accessing unstructured data but also semi-structured and structured content, which often is referred to as knowledge of which the semantic web is a specific case.

The core idea of the SmartWeb project is to enable the access to the unstructured content as well as to the semantic web through a natural language dialogue system. Four different streams provided potential answers to the user's questions. A handcrafted ontology will be able to provide high quality answers to a very specific and small set of questions. There will be a mechanism to convert a few thousand web pages automatically into a semantic web structure. A question answering component will allow to answer questions of restricted complexity but in any domain. Finally interfaces to existing structured information services (e.g. weather forecast) are established.

The system runs several dialogue engines which can be located either on a server or on a mobile device like a car or a PDA. In particular robustness and flexibility are a key concern in this part of the project.

5 Conclusion

This paper gives examples of the different aspects of accessing content. In the area of concept annotation of video we have shown that the ASR output can indeed be used for concept annotation and the performance is comparable with image based methods. We expect that image based method have a large potential for future improvement. The LISy system and SmartWeb are two examples of complete systems that allow the access to content using a natural language dialogue system.

Building complete systems turns out to be very valuable to discover the weak points of a system.

Acknowledgements

This paper is an overview of various projects and activities. Special thanks go to the author's previous research team at Philips Research, the SmartWeb team as well as Matt Krause and Giri Iyengar.

References

1. Wactlar, H. D., Kanade, T., Smith, M. A., and Stevens, S. M. 1996. Intelligent Access to Digital Video: Informedia Project. Computer 29, 5 (May. 1996), 46-52.
2. Smeaton, A. F., Over, P., and Kraaij, W. 2004. TRECVID: evaluating the effectiveness of information retrieval tasks on digital video. In Proceedings of the 12th Annual ACM international Conference on Multimedia (New York, NY, USA, October 10 - 16, 2004). MULTIMEDIA '04. ACM Press, New York, NY, 652-655.
3. Garofolo, J., Auzanne, C., Voorhees, E. and Fisher D 1999. TREC Spoken Document Retrieval Track: A Success Story. TREC Proceedings
4. Voorhees, E. 2004. Overview of the TREC 2003 Question Answering Track. TREC Proceedings
5. Aust, H., Oerder, M., Seide, F., and Steinbiss, V. 1995. The Philips automatic train timetable information system. Speech Commun. 17, 3-4 (Nov. 1995), 249-262.
6. Zue, V.W.; Glass, J.R.. 2000. Conversational interfaces: advances and challenges, Proceedings of the IEEE Volume 88, Issue 8, 166 – 1180.
7. Miller, G. A. 1995. WordNet: a lexical database for English. Commun. ACM 38, 11 (Nov. 1995), 39-41.
8. Witten, I. and Frank, E. 2005. Data Mining: Practical machine learning tools and techniques, 2nd Edition, Morgan Kaufmann, San Francisco.
9. Wegmann, S., Zhan, P., Gillick, L. 1999 Progress in broadcast news transcription at Dragon Systems. ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing.
10. Fiscus, J., Garofolo, J., Przybocki, M., Fisher, W. and Pallett D. 1997. English Broadcast News Speech (Hub-4), Linguistic Data Consortium.
11. Aubert, X. L. 2000. A brief overview of decoding techniques for large vocabulary continuous speech recognition, ASR-2000, 91-97.
12. Wahlster, W.,Reithinger, N. and Blocher, A. 2001. SmartKom: Multimodal Communication with a Life-Like Character, Eurospeech.
13. See http://smartweb.dfki.de

Content-Based Image and Video Indexing and Retrieval

Hong Lu[1], Xiangyang Xue[1], and Yap-Peng Tan[2]

[1] Shanghai Key Laboratory of Intelligent Information Processing, Department of Computer Science & Engineering, Fudan University, Shanghai 200433, China
[2] School of Electrical & Electronic Engineering, Nanyang Technological University, Singapore 639798
{honglu,xyxue}@fudan.edu.cn, eytan@ntu.edu.sg

Abstract. This paper surveys some of the existing techniques and systems for content-based indexing and retrieval of two main types of multimedia data – images and videos. Furthermore, some of our recent work are reported. Specifically, in content-based image retrieval, we have proposed multi-scale color histograms by incorporating color and spatial information. In content-based video retrieval, we have proposed a new level of program between story/event and video sequence levels. Video summarization results are given based on the scene clustering results.

Keywords: Content-based indexing and retrieval, image, video.

1 Introduction

The advances in computing technologies, broadband communication networks, mass storage devices, and consumer electronics have resulted in large amounts of multimedia data being generated and made accessible in digital form throughout the world. However, finding the desired multimedia information from a large number of distributed databases remains a very difficult and time-consuming task. This is largely due to the fact, that a common standard for indexing and cataloging multimedia data according to their content is not yet available. As a result, the main challenge for getting the desired multimedia data is not due to the availability of the data itself, but our ability in locating them within a reasonable amount of processing time. To remedy this situation, substantial research efforts have focused on developing effective and efficient techniques to manipulate, analyze, index, represent, browse, and retrieve multimedia data over the past several years.

There is a wide variety of multimedia data available today, including text, speech, audio, images, graphics, animations, videos, and their various combinations. In order to search and retrieve these data in an automated fashion, video content needs to be annotated and indexed in formats usable by a search engine. Today, most of these data do not carry much descriptive information with regard to their semantic content. In particular, annotations for audio-visual data are mainly restricted to textual descriptions, such as title (e.g., "Gone with the Wind"), genre (e.g., drama, civil war), summary or abstract (e.g., movie abstract or review), and name of the subject (e.g., Clark Gable, Vivien Leigh), etc. Conceivably, these textual descriptions can hardly be

R. Lu et al. (Eds.): Chinese-German Ws on Cognitive Systems 2005, LNAI 4429, pp. 118 – 129, 2007.
© Springer-Verlag Berlin Heidelberg 2007

exhaustive or comprehensive in depicting the audio-visual data. Moreover, as generally agreed-upon vocabularies and syntax are not yet in place for the description of video data, these textual descriptions could be rather subjective and may not generally agree with each other. More objective descriptions and representations of video data, especially those that can be automatically and unambiguously extracted from audio-visual content, are therefore desirable and can complement the capability of textual descriptions.

In general, there are three main approaches to content-based multimedia description or indexing: text-based approach, feature-based approach, and semantic-based approach. **In a text-based approach**, keywords or phrases are used to describe the media content in various possible ways. However, these textual descriptions are incapable of detailing the rich information embedded in audio-visual data. **In a feature-based approach**, low-level audiovisual features such as color, texture, shape, motion, and audio are extracted from the data and used as indexing keys. The underlying rationale for this approach is that video data that are similar in terms of their audiovisual features are likely to be similar in content.

In a semantic-based approach, multimedia data are annotated with their high-level semantic meanings. Substantial recent research efforts have focused on extracting these high-level semantics with automated or computer-assisted methods. Given the vast varieties of video contents and the difficulty of the problem itself, this approach is currently the most challenging and often requires some extent of human intervention. However, by restricting the analysis to a specific class of video, useful high-level semantics can be derived, based on some primitive audiovisual features and domain-specific knowledge. Examples are news items in news programs, play events of sports games and rating of motion pictures.

2 Content-Based Image Indexing and Retrieval

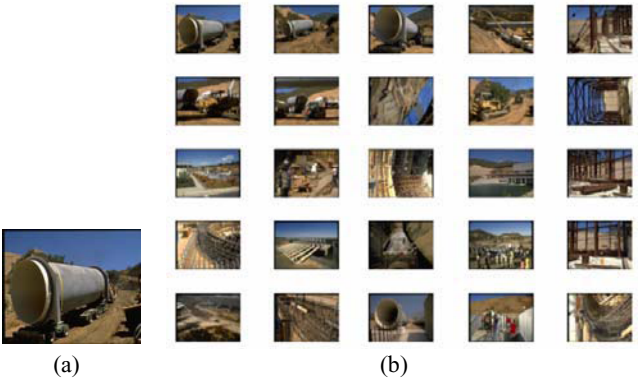
Images record scenes and/or objects that are of interest to us or useful to some automated systems. Although people can easily comprehend the high-level semantic meanings of images (e.g., humor of comics), it is very difficult to use a computing machine to automatically extract or understand these high-level semantic information from the images. Hence, most of the current automated image analysis techniques are restricted to the study of low-level image features, such as color, texture, shape, and spatial layout. Among these features, color is the most prominent visual attribute for image content and has been commonly used in many Content-Based Image Retrieval (CBIR) systems, such as QBIC [1], Virage [2], Visualseek [3], etc., while the color histogram proposed in [4] is the most popular color descriptor. We call it as conventional color histogram (CCH) method.

CCH represents an image based on its color distribution. To compute an image's CCH, the number of pixels within each color bin is counted. It is easy to compute but lacks image spatial information. Many other approaches have been proposed, including the generalized color histogram (GCH) [5], to overcome this drawback. The GCH is an extension of CCH and it is computed by using a window sliding over the given image. It addresses not only the color distribution in the whole image but also

the color composition of the area covered by each color within a window and records the number of windows containing pixels with several fixed percent of area for each color. GCH has been shown capable of capturing the aggregation information of colors in an image [5].

However, both CCH and GCH have some limitation. They only evaluate the pixel counts within each color bin of the images to be compared and do not take into account the color similarity (or dissimilarity) among different color bins. Moreover, since many bins in CCH and GCH are normally empty in practice, they may not efficiently represent the images. On the other hand, the number of bins representing each individual image can not be too small either, so that enough color information can be captured to attain good image discrimination. Consequently, CCH and GCH need a relatively large amount of storage space and are difficult to be indexed efficiently. Although GCH can capture some spatial information of an image, this capability is restricted by the size of its sliding window. Its discrimination power degrades when the window size is too small or too large when compared to the sizes of the objects or scenes in an image. Other approaches such as color coherence vector [6] and spatial color histograms [7] have also been proposed to exploit the spatial information of images in various ways based on CCH. However, the above limitations still remain.

In [8], we propose a new spatial-color descriptor referred to as the multi-scale spatial color histogram (MSCH) to overcome some of the limitations mentioned above. First, the dominant colors of an image are obtained by a fuzzy c-means clustering method. Second, an image can be partitioned into a sequence of non-overlapping blocks of various sizes. A color histogram can be constructed from each image partition. A set of color histograms computed from all different partitions of interest constitutes the multi-scale color histograms (MSCH) of the image. Experimental results show that the proposed method can achieve better performance than that of the CCH and GCH when they use the similar number of bins. Fig. 1 shows the retrieval results of one query image “big pipe” by using our proposed method.



(a) (b)

Fig. 1. Retrieval results of one query image “big pipe”. (a) the query image and (b) the retrieved images.

3 Content-Based Video Indexing and Retrieval

In content-based video analysis, video can be analyzed in six structured levels as shown in Fig. 2. For the nature of the analysis involved in these six structured levels and towards the higher structured levels, the domain dependency tends to increase and the computability/analysis accuracy tends to decrease. Specifically, in frame level analysis, low-level features such as color, texture, shape, motion, and audio are generally used and the analysis requires no or minimum domain knowledge. In this level, many shot boundary detection (SBD) methods have been proposed to segment video into shots, each of which can then be represented by one or a few key frames from the shot.

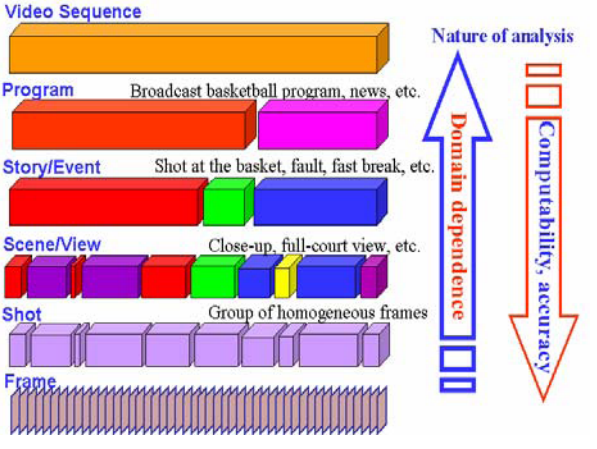


Fig. 2. Six structured video levels for content-based video analysis

As human users are more interested in the semantic levels concerning the underlying scenes or views, stories or events, or plots of a video program, higher level of analysis is generally required to analyze video shots for more compact or meaningful representation. The analysis includes, for example, scene clustering (clustering visually similar shots into scenes) and scene segmentation (grouping related shots into scenes each featuring a dramatic event). Based on the detected shots and clustered or segmented scenes, one or more key frames can be extracted. Afterwards, image features such as color, shape, and texture are used to index these key frames. In addition, high-level representations such as regions, objects, and motion can also be used to infer high-level semantic events and help summarize the content of the whole video sequence. Finally, videos can be browsed and retrieved based on the similarity of the features of the query video sequence and the video sequences in the database.

3.1 Shot Boundary Detection

Shot boundary detection (SBD) is commonly the first step in the process of indexing, characterizing, and retrieving of video. As shown in Fig. 3, a shot is a temporal

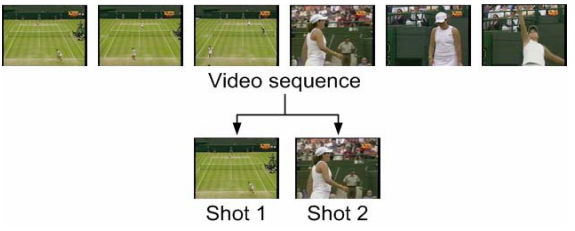


Fig. 3. Shot boundary detection and shot representation

sequence of frames generated and recorded continuously by a single camera act, which usually depicts a continuous action without having major content changes.

To form a video sequence, shots are joined together during video sorting and post editing with either abrupt cuts or gradual visual effects according to the nature of the scene changes or story sequences. As shown in Fig. 4, there are two types of shot transitions: abrupt shot boundary (ASB), also known as shot cut or hard cut, where the change of video content occurs over a single frame, and gradual shot boundary (GSB), such as fade in, fade out, dissolve, and wipe, where the change takes place gradually over a short sequence of frames.

Fig. 4. Different types of shot boundaries: (a) a shot cut, (b) a fade out followed by a fade in, (c) a dissolve, and (d) a wipe

The main technique in SBD is to detect the change between each consecutive frame pair. If the change exceeds a pre-determined threshold, a shot boundary is assumed. In the literature, the change is usually measured by using such low-level video features as pixel intensities, color histograms, edges, and compressed video attributes [9-12,16]. For example, in [16], Zhang et al. propose a pair-wise comparison between each pixel in one frame with the corresponding pixel in the next frame. This method has been known to be sensitive to camera and object motion, and may cause excessive false detections. Color histogram is another type of feature used

in SBD which is rather insensitive to small changes over consecutive frames [10,11,16]. Edge differences have also been used to counter problems caused by fades, dissolves, etc. [11]. In the compressed domain, much work has been done by using Discrete Cosine Transform coefficients, motion vectors, and macroblock types of MPEG video for SBD [9-11].

3.2 Scene Clustering and Segmentation

As one hour video can easily contain hundreds of shots, it is rather inefficient to represent or retrieve a video in the shot level. To reduce the number of shots of a video to a manageable size, a common approach is to group the shots into different scene clusters [17-19], each consisting of shots with similar contents. One or a few shots can then be selected from each scene cluster, for example, for compact representation and indexing of video content [20], or for automatic identification of video events of interest [20-24].

In video scene clustering, conventional k-means clustering and hierarchical clustering methods have been exploited in the literature [17-19,25,26]. Zhong et al. [17,25] use a fuzzy hierarchical clustering approach to group video shots into classes of similar visual content. Here the term fuzzy indicates that a feature can be assigned to several clusters according to different degrees of belonging, and not to only one cluster like that in hard clustering. First, each video shot is denoted by visual features such as the color histogram of its representative frame and the temporal variance of frame color histograms within the video shot, as well as motion features such as the directions of motion and the speed at each motion direction around a representative frame. Based on these low level features, a fuzzy k-means clustering method and a conventional clustering method such as k-means clustering are used in turn to cluster the video shots into classes in different hierarchical levels. Each class of video shots is represented by the representative frame which is nearest, in terms of the low level features used, to the centroid of the class. A top-down hierarchical representation of the video can then be constructed to permit non-linear browsing and retrieval of video segments.

In [19], Yeung et al. propose to group video shots into shot clusters based on the visual similarity (e.g., color histogram intersection) of shot representative frames as well as the local time-constraints of neighboring video shots. Each shot cluster is represented by a node in a directed graph. The temporal relations among different shot clusters are indicated by directed edges connecting the nodes in the graph. For instance, a directed edge from shot cluster A to shot cluster B indicates that some video shot in shot cluster A is followed by some video shot in shot cluster B. The video graph can provide a better view of the overall video structure depicting the visual similarity and temporal locality of shots. It is assumed that high-level video semantics, such as scenes and stories, can be readily identified from the video graph with proper human intervention.

In [18], k-means clustering is applied to all the frames in the video sequence to find the cluster centers. The obtained cluster centers are then used for selecting key frames and concatenating representative shots, to which the key frames belong, to form the preview of the video sequence.

In the work of scene clustering, graph-theoretic method has also been proposed [19,27-30]. Specifically, in [27], the graph-theoretical clustering algorithm of minimum-spanning tree is performed on key frames from each video shot to identify the anchorperson shot. The anchorperson shots are further distinguished from other news video shots. In [31], four measures of minimum cut, average cut, normalized cuts, and a variation of normalized cuts are investigated for image segmentation. Experimental results show that the minimization of the average-cut and the normalized-cuts measures, using recursive bi-partitioning, can result in better segmentation accuracy on average. More recently, a normalized-cuts based graph partitioning method is widely used in image segmentation [32] and video scene clustering [28-30].

In content-based video analysis, another type of scene segmentation is grouping of related shots by incorporating some temporal information or correlation.

In [19], time constrained clustering of video shots is performed, i.e. two shots are considered similar if their content similarity is high enough and the difference of their occurrences in time is small. After clustering, shots belonging to the same cluster can be represented by the same label, such as cluster label A, B, or C, etc. To detect scene boundaries, a Scene Transition Graph based on the scene clustering results is proposed in [19], where the scene segmentation problem is considered as segmentation of cluster label pattern. For example, if the cluster labels of a sequence of shots are ABABCDCD, then the corresponding graph for the temporal transitions is:

$$A - B \Rightarrow C - D$$

However, this method has the limitations that it does not consider the shot length and depends upon the scene clustering threshold parameters which need to be manually tuned. In [26], a more general time-adaptive grouping approach based on the shots' visual similarity and temporal locality is proposed to cluster shots into groups. Semantically related groups are then merged into scenes. The method relies on predefined thresholds on group clustering and scene merging. To overcome these limitations, Kender and Yeo propose a continuous video coherence measure that gauges the extent of the current shot reminds the viewer of a previous shot if the two shots are similar [13]. In [14,15], Sundaram and Chang propose a first-in-first-out (FIFO) memory model to detect the audio and video scenes separately.

3.3 Structured Video Analysis

As stated, in the six structured levels of video analysis, the domain dependency tends to increase and the computability and analysis accuracy tend to decrease, respectively, towards the higher levels. Specifically, many existing SBD methods, which examine the change between each consecutive frame pair, can be used on different types of videos. However, for the analysis in scene and story levels, one research direction is to restrict to specific types of video so that suitable domain knowledge can be exploited.

There are many kinds of videos such as movie, news, cartoon, sports, commercial, etc. Among them, structured video is the type of video for which we have some prior knowledge of either the frame structure (what and where objects are likely to appear) or the video content (when and what events are likely to occur), or both. The representative types of structured video are sports video and news video. For example, as shown in Figure 5 (a), in sports video such as tennis, the camera mainly keeps track of the ball and players. After focusing on a serving ball event of one player with a close-up view, the camera switches to the full court view to show the ball movement. As shown in Figure 5 (b), in news video, after the reporter tells a news item, the video content switches to the story of the news and then back to the reporter again. Furthermore, in the frames of reporter telling the story, the location of the reporter is usually fixed. Such similar scenes repeat regularly in sports and news videos, and present as well defined video content structure. However, in other types of videos such as movie and commercials, viewers usually cannot predict what will happen and appear next in the video as the contents are generally less structured.

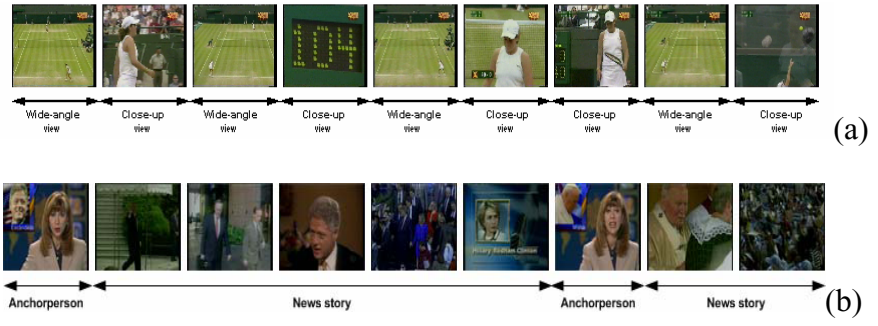


Fig. 5. (a) A typical tennis clip with interleaving wide angle and close-up view shots. (b) A typical news clip in which anchorperson shots introduce a news story and news story shots present the details of the news.

3.4 Program Segmentation

Since TV videos of specific channels normally contain a large number of scenes, which makes the task of browsing and retrieval video content of interest time-consuming. Further, most generally viewers watch or specify TV videos by programs and not by scenes or shots. Thus, we propose to construct a new video structure level, referred to as program, between the video and scene levels to enhance the hierarchical structure for analyzing or representing the content of TV videos [33]. There are a number of possible definitions for the term program in the encyclopedia, and we cite below the one that closely matches what we are dealing with.

A program is the content of a television broadcast and may be a one-off broadcast or, more usually, part of a periodically returning television series.

Our proposed program segmentation technique is based on the intrinsic characteristics of TV videos.

- For a TV channel, programs appear and end at a relatively fixed time every day; slight variations do exist sometimes.
- For programs of the same type, they have stable or similar starting and ending clips even when they appear in different days.

Take the news of CCTV-1 (*China Central Television 1*) for example. It always begins around 7:00pm with a fixed prologue (i.e. the start video clip), and ends around 7:30pm with another fixed video clip, every day.

As such, our proposed approach consists of two steps: model construction and program segmentation. First, we construct the program boundary models for the selected TV channel by detecting the repeat shots in different days. Then, based on the obtained models, videos recorded from the same TV channel can be segmented into programs. Fig. 6 illustrates the detected program represented by two key frames of the beginning and ending shot in this program.

Fig. 6. Detected programs represented by two key frames of the beginning and ending shot in this program

3.5 Video Summarization

Since the final objective of video analysis is to construct the summarization of video content, Fig. 7 shows the summarization result by using the video scene clustering method for the Tennis1 test video, which has 17982 frames and 137 shots. Fig. 7 (a) and (b) show the result before and after the summarization. In Figure 8 (a), the first frame of each shot is used to represent the shot, and the frame number of each shot is also given. Fig. 7 (b) gives the summarization result of the two dominant scenes and one miscellaneous scene obtained by our proposed unsupervised clustering method. The summarization results can be shown by one shot for each scene which has the highest similarity to other shots in the scene. It can be seen from Fig. 7 that the summarization result can capture the main content of the sports video, i.e., the wide angle scene, the close-up scene, and the miscellaneous scene. The original video has 17982 frames and after the summarization, the video can be represented by 3 shots from 3 scenes with a total of 448 frames, i.e., the temporal length is reduced from 11.98 minutes to 0.30 minutes. The summarization ratio (SR) between the lengths of the original video sequence and the summarization result is 39.93.

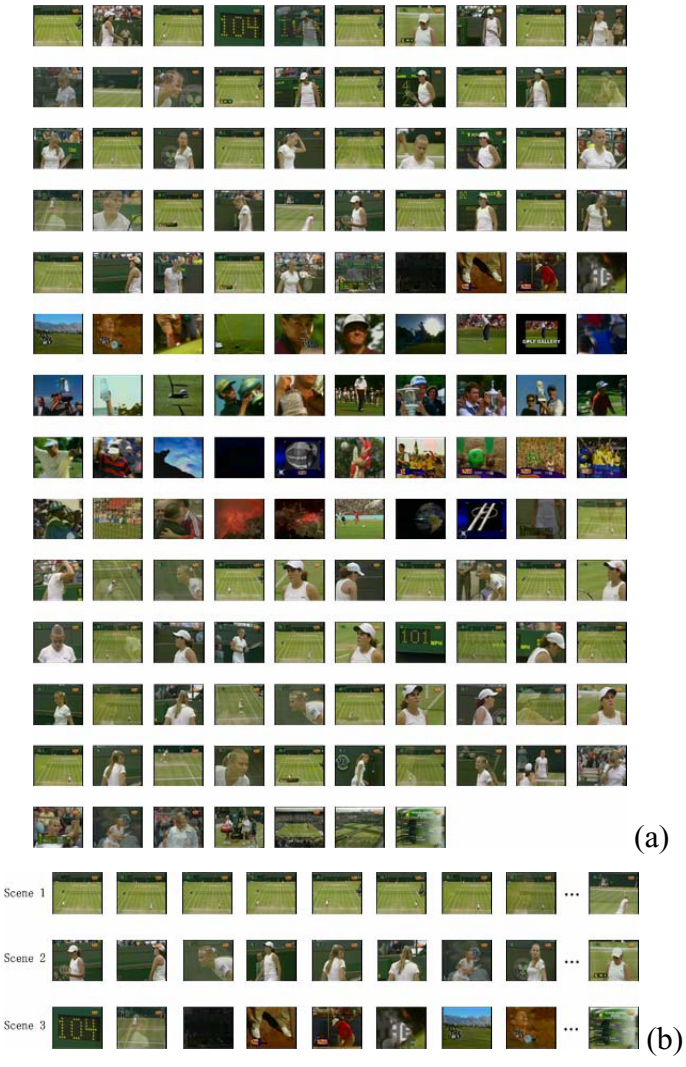


Fig. 7. (a) The original video shots and (b) the summarization result by using our proposed unsupervised clustering method for Tennis1 test video

4 Conclusion

This paper surveys some of the existing techniques and systems for content-based image and video indexing and retrieval. Furthermore, some of our recent work is reported. Specifically, multi-scale color histograms are proposed for content-based image retrieval. Also, the methods on six levels of structured video analysis are reviewed and we have proposed a new level of program between story/event and

video sequence levels. By using scene clustering, more compact and dominant information are kept for the video clip under analysis.

Acknowledgments. This work was supported in part by Natural Science Foundation of China under contracts 60533100, 60373020, and 60402007, Shanghai Municipal R&D Foundation under contract 05QMH1403, and Open Research Foundation of Shanghai Key Lab of Intelligent Information Processing.

References

1. M. Flickner, H. Flickner, W. Niblack, J. Ashley, Q. Huang, B. Dorn, M. Gorkani, J. Hafner, D. Lee, D. Steel and D. Yanker, "Query by image and video content: the QBIC system," IEEE Computer, Sept. 1995, vol. 28 (9), pp. 23-32.
2. J. Bach, C. Fuller, A. Gupta, A. Hampapar, B. Harawitz, R. Humpherey, F. Jain, and C. Shu, "The Virage image search engine: An open frame work for image management," in *Proc. SPIE: Storage and Retrieval for Image Video Databases*. Feb. 1996, Vol. 2670, pp. 76-87.
3. John R. Smith and Shih-Fu Chang, "VisualSeek: a fully automated content-based image query system," *ACM Multimedia*, 1996, pp. 87-95.
4. M. J. Swain and D. H. Ballard, "Color Indexing," *International Journal of Computer Vision*, Jan. 1991, vol. 7(1), pp. 11-32.
5. "Core Experiment CT3 on Spatial Information Embedding Color Descriptors," Technical Report ISO/IEC JTC1/SC29/WG11/M5223, Oct.1999.
6. G. Pass, R. Zabih, "Histogram refinement for content-based image retrieval," *IEEE Workshop on Application of Computer Vision*, 1996, pp. 96-102.
7. Aibibg Rao, Rohinik Srihari, Zhongfei Zhang, "Spatial color histograms for content-based image retrieval," in *Proc. 11th IEEE International Conf. on Tools with Artificial Intelligence*, 1999, pp. 183-186.
8. Anning Ouyang and Yap-Peng Tan, "A novel multi-scale spatial-color descriptor for content-based image retrieval," *Seventh International Conference on Control, Automation, Robotics and Vision*, 2002, pp. 1204-1209.
9. I. Koprinska and S. Carrato, "Temporal video segmentation: a survey," *Signal Processing: Image Communication*, 2001, vol. 16, pp. 477-500.
10. U. Gargi, R. Kasturi, and S. H. Strayer, "Performance characterization of video-shot-change detection methods," *IEEE Transactions on Circuits and Systems for Video Technology*, 2000, 10(1), pp. 1-13.
11. P. Browne, A. F. Smeaton, N. Murphy, N. O'Connor, S. Marlow, and C. Berrut, "Evaluating and combining digital video shot boundary detection algorithms," *Irish Machine Vision and Image Processing Conference*, 2000.
12. D. Zhong, "Segmentation, Index and Summarization of Digital Video Content," *Doctoral Dissertation*, Graduate School of Arts and Sciences, Columbia University, 2001.
13. J. R. Kender and B.-L. Yeo, "Video scene segmentation via continuous video coherence," IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 1998, pp. 367-373.
14. H. Sundaram and S.-F. Chang, "Computable scenes and structures in films," IEEE Transactions on Multimedia, 2002, 4(4), pp. 482-491.
15. H. Sundaram, Segmentation, Structure Detection and Summarization of Multimedia Sequences. Ph.D. thesis, 2002, Columbia University.

16. B. Furht, S. W. Smoliar, and H. J. Zhang, Video and image processing in multimedia systems, *Kluwer Academic Publisher*, Boston.
17. Di Zhong, Hong J. Zhang, and Shih-Fu Chang, "Clustering Methods for Video Browsing and Annotation," *SPIE Proceeding Volume 2670, Storage and Retrieval for Still Image and Video Databases IV*, 1996, pp. 239-246.
18. A. Hanjalic and H. J. Zhang, "An integrated scheme for automated video abstraction based on unsupervised cluster-validity analysis," *IEEE Transactions on Circuits and Systems for Video Technology*, 1999, vol. 9, no. 8, pp. 1280-1289.
19. M. Yeung, and B.-L. Yeo and Bede Liu, "Extracting story units from long programs for video browsing and navigation," *IEEE International Conference on Multimedia Computing and Systems*, 1996, pp. 296-305.
20. P. Aigrain, H. J. Zhang, and D. Petkovic, "Content-based representation and retrieval of visual media: a state-of-the-art review," *Multimedia Tools and Applications*, Kluwer Academic, 1996, vol. 3, pp. 179-202.
21. H. J. Zhang, Y. Gong, S. W. Smoliar, and S. Y. Tan, "Automatic parsing of news video," *IEEE International Conference on Multimedia Computing and Systems*, 1994, pp. 45-54.
22. Y.-P. Tan, D. D. Saur, S. R. Kulkarni, and P. J. Ramadge, "Rapid estimation of camera motion from MPEG video with application to video annotation," *IEEE Transactions on Circuits and Systems for Video Technology*, 2000, vol. 10, no. 1, pp. 133-146.
23. P. Xu, L. Xie, S.-F. Chang, A. Divakaran, A. Vetro, and H. Sun, "Algorithms and System for Segmentation and Structure Analysis in Soccer Video," *IEEE International Conference on Multimedia and Expo*, 2001.
24. Y. Gong, L. T. Sin, C. H. Chuan, H. J. Zhang, and Masao Sakauchi, "Automatic parsing of TV soccer programs," *IEEE International Conference on Multimedia Computing and Systems*, 1995, pp. 167-174.
25. Di Zhong, "Segmentation, Index and Summarization of Digital Video Content," *Ph.D. thesis*, Columbia University, 2001.
26. Y. Rui and T. S. Huang and S. Mehrotra, "Constructing Table-of-Content for Videos," *ACM Journal of Multimedia Systems*, 1999, vol. 7, no. 5, pp. 359-368.
27. X. B. Gao, J. Li, and B. Yang, "A graph-theoretical clustering based anchorperson shot detection for news video indexing," *Fifth International Conference on Computational Intelligence and Multimedia Applications*, 2003, pp. 108-113.
28. C. W. Ngo, Y. F. Ma, and H. J. Zhang, "Automatic video summarization by graph modeling," *Ninth IEEE International Conference on Computer Vision*, 2003, pp. 104-109.
29. Zeeshan Rasheed and Mubarak Shah, "A graph theoretic approach for scene detection in produced videos," Multimedia Information Retrieval Workshop 2003 in conjunction with the 26th annual ACM SIGIR Conference on Information Retrieval, 2003.
30. Zeeshan Rasheed, "Video Categorization using Semantics and Semiotics," *Ph.D. thesis*, University of Central Florida, 2003.
31. P. Soundararajan and S. Sarkar, "Investigation of measures for grouping by graph partitioning," *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2001, vol. 1, pp. I-239-I-246.
32. J. Shi and J. Malik, "Normalized cuts and image segmentation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2000, vol. 22, no. 8, pp. 888-905.
33. L. Liang, H. Lu, X. Xue, and Y.-P. Tan, "Program segmentation for videos from TV channels" *IEEE International Symposium on Circuits and Systems*, 2005, pp. 1549-1552.

Shape Recognition with Coarse-to-Fine Point Correspondence Under Image Deformations

Huixuan Tang and Hui Wei

Department of Computer Science, Lab of Algorithm for Cognitive Model, Intelligent Information Processing Laboratory, Fudan University, Shanghai 200433, P.R. China
{0124143,weihui}@fudan.edu.cn

Abstract. Matching techniques are part-and-parcel of shape recognition. A coarse-to-fine method is presented which finds point correspondence between open or closed curves and is invariant to various image deformations, including affine transformation, perspective distortion, non-rigid motion and so forth. The method is inspired by the idea to use point correspondences established at one level to generate a priori information, which is either topological or geometric, to match features at finer levels. This has all been achieved through an analysis of the curve topology and a synthesis of the B-spline interpolation techniques. This is in contrast to existing multi-scale methods for curve matching that use pure feature correlation or 3D structure recovery at a fixed scale. The presented method proves to be robust and accurate and can serve as a powerful aid to measure similarity of shape, as demonstrated in various experiments on real images.

1 Introduction

In the computer vision literature, curve matching techniques finds applications in a wide range of areas, especially in the shape recognition tasks. However there exists no general solution for this problem due to external noise, change of viewpoint, and occlusion.

Among the most popular methods that have been proposed so far are proximity matching and syntactic matching. The basic idea behind proximity matching [2, 25, 26] is to minimize the distance between the matched curves. A prerequisite of this type of methods is the design of invariants against noise and deformation, such as geometric invariants [20], corners [19] and curvature [15]. However, they are only invariant under RST (Rotation, Scaling and Translation) transformations, and are sensitive to noise. Syntactic matching [8, 11, 24] interprets a curve as a symbolic description and utilizes various criteria, such as editing distance, to match curves. It is more robust to noise, but does not provide as dense correspondences as proximity matching. Besides, this style of methods takes only topological constrains into consideration.

Many authors have stressed that correspondence should be established at multiple levels with matches between physically salient features [6, 13, 14]. However, most existing multi-scale techniques [1, 7, 22], whether syntactic or proximity matching,

R. Lu et al. (Eds.): Chinese-German Ws on Cognitive Systems 2005, LNAI 4429, pp. 130 – 144, 2007.
© Springer-Verlag Berlin Heidelberg 2007

consider the matching processes at different scales separately. We argue instead that these processes are otherwise interrelated. Since features extracted at different levels express the same curve, they share the same topological and geometric property. We have designed a coarse-to-fine framework to realize this idea by exploiting results at a coarser level as a guidance to finer matches.

The objective of this paper is to obtain multi-scale correspondence between curves from intensity images for object recognition. It is in contrast to existing multi-scale methods for curve matching since it computes estimations of correspondences at a coarser scale and then generates a priori information for finer matching processes. Moreover, it gives rise to a natural way to fuse feature correlation and motion estimation into an integrated framework by synthesizing topological analysis and B-spline interpolation techniques, thus resolving the inherent difficulty of 3D structure recovery.

When matching features, the computed correspondences generate a sketch of the curves and can be used to describe the shape context of a given point. Shape context can be topological or geometric. The topological shape context, composed of the order-preserving constraint and the sided-ness constraint, are used to prune the feature matching process, hence reducing not only the computational costs but mismatches as well.

The geometric shape context is in the form of a hybrid similarity measure combining feature correlation and motion estimation. Existing techniques for estimating camera motion are mostly derived from projective geometry [4, 10, 17], which models the camera motion with the fundamental matrix. However, 3D structure recovery has always become a bottleneck in such methods. Moreover, such methods assume consistent motion of all points in the scene, and therefore do not satisfy our purpose of handling nonrigidity. Because of the above reasons, we exploit a method derived from B-spline interpolation. B-spline functions are suitable for representing curves because they are independent of the affiliated coordinate system, and are stable under affine transformations. However, they are rarely used in recognition problems because the representation is not unique [3]. Nevertheless, by fixing the control points with the correspondence estimation the B-spline representations can be fixed to the same anchor point and therefore become comparable. Grounded on this comparison a combined similarity measure fusing feature correlation and motion estimation is derived.

This paper is organized as follows. Section 2 gives an overview of the method. The topological constraints and the similarity measure is discussed respectively in Section 3 and Section 4. In Section 5 the algorithm is extended to open curve matching and measuring similarity of curves. We will also demonstrate that robust estimation of correspondence can be achieved against noise, rigid and nonrigid transformations using tests with real images. Section 6 concludes the paper.

2 Overview of the Method

In this section we first formulate the problem and then give an overview of the coarse-to-fine framework.

2.1 Problem Formulation and Notations

The goal of the curve matching algorithm is, given two intensity images I and $\hat{I}$ which have undergone an arbitrary transformation, to align a template curve C in I and a target $\hat{C}$ curve in $\hat{I}$.

Denote the parametric representation of the template curve $C(u) = [x(u), y(u)]^T \ (0 \le u \le L)$ and the target curve $\hat{C}(v) = [\hat{x}(v), \hat{y}(v)]^T \ (0 \le v \le \hat{L})$, and their curvatures as $k(u)$ and $\hat{k}(v)$ respectively. The presented method operates on a coarse-to-fine pyramid and seeks for corresponding features, parametrically represented as $M_i \subset \tilde{U} \times \tilde{V}$ at different scales w, where $\tilde{U} = \{\tilde{u}_j\}$ and $\tilde{V} = \{\tilde{v}_j\} \ (1 \le j \le \tilde{n})$ are matched features in the last iteration. For convenience, two associative symbols are defined as $U = \{u_j\}$ and $V = \{v_j\} \ (1 \le j \le n)$ to denote the parametric form of all matched features in the current estimation M . Without loss of generality, it is assumed that $\tilde{U}$ and U are arranged in an ascendant order.

Table 1. Detecting salient features in the curve $C(u)$ at a scale w

1. Compute first and second derivative along the curve.
2. Smooth the curve's orientation with a Guassian kernel of size w . $$g(u) = \frac{1}{\sqrt{2\pi}\,w} e^{-u^2/2w^2}$$
3. Compute the curvature of each point of $C(u)$ with the following equation: $$k(u) = \frac{\dot{x}(u)\ddot{y}(u) - \ddot{x}(u)\dot{y}(u)}{\left\lvert \dot{x}(u)^2 + \dot{y}(u)^2 \right\rvert^{\frac{3}{2}}}$$
4. Return locally maximums in curvature as salient features in C .

2.2 The Coarse-to-Fine Framework

We start by casting the matching problem as the derivation of an approximate representation of the target curve $\hat{C}$ with salient features in both C and $\hat{C}$. The algorithm starts by briefly sketching the target curve with a few correspondences computed from feature correlation and gradually refines the sketch by progressively detecting and matching more features at finer levels. Each iteration is composed of four stages: feature detection, curve sketching, feature matching, and correspondence refinement.

The first issue to address is the introduction of a multi-scale feature detector. Following the smooth-derivation algorithm introduced in [1], features are extracted as local maximums in curvature that is computed from the first and second derivations at each pixel. They can be easily derived during the edge tracking process which sequentially extracts pixels on the edges. To obtain a multi-scale representation, the curvature discontinuities are convolved by different size Gaussian kernels. The main steps of this process are summarized in Table 1.

During the second stage, a shape context is established from the current estimation and used to generate a brief sketch of the target curve. The shape context can be topological or geometric, and is used either as matching constraints or as matching heuristics in the feature matching stage. Note that the representation of the shape context is local, therefore the new method is able to handle very complex deformations such as non-rigid motions.

Subsequent to that, candidate matches are firstly filtered by the topological matching constraints, and then get scored by a hybrid similarity measure combining feature correlation and the geometric shape context. Pairs of features with greatest similarity are put into correspondences. In latter sections, we formally address the generation of the shape context and describe the feature matching stage in details.

During the last stage, matches computed in the feature matching stage are validated at first, then exploited to refine the correspondence estimations. Let define three kinds of relationships between pairs of correspondence estimations.

Let $m_1 = [C(u_1), \hat{C}(v_1)]$ and $m_2 = [C(u_2), \hat{C}(v_2)]$ be a pair of correspondence in two sets of estimations M^1 and M^2 respectively, then their relationship is defined as:

- Related: m_1 and m_2 are related if $|C(u_1) - C(u_2)| \le \varepsilon_w$ or $|\hat{C}(v_1) - \hat{C}(v_2)| \le \varepsilon_w$. Without loss of generality we assume that $|C(u_1) - C(u_2)| \le \varepsilon_w$ so that $C(u_1)$ and $C(u_2)$ roughly correspond to the same point in the scene within the tolerance of acceptable alignment error.
- Compatible: Given that m_1 and m_2 are related, they are compatible if $|C(u_1) - C(u_2)| \le \varepsilon_w$ and $|\hat{C}(v_1) - \hat{C}(v_2)| \le \varepsilon_w$. Following our assumption above, $\hat{C}(v_1)$ and $\hat{C}(v_2)$ are roughly the same and their difference can be ignored.
- Contradictory: m_1 and m_2 are contradictory if m_1 and m_2 are related but incompatible. The difference of two estimations is beyond the tolerance of acceptable error, therefore there is at least one mismatch in the two correspondences.

Table 2. Summary of the Coarse-to-Fine Framework

1. Start the matching process at $w = w_{\max}$, set $i = 1$ and $M_0 = \phi$.
2. Derive a map M^1 from $\hat{C}$ to C and another one M^2 from C to $\hat{C}$: (a) Extract salient features along the curve at the current scale w. (b) Filter candidate correspondence pairs with the topological constraints derived from M_{i-1}. (c) Score all candidate matches with the similarity measurement derived from M_{i-1} and put those with greatest similarity into correspondences.
3. Apply reciprocal constraint to M^1 and M^2 so as to achieve a reliable correspondence M.
4. Compare M and M_{i-1} to refine the current estimation M_i.
5. Set $w = w - 2$, $i = i + 1$ and goto step 2 if $w \geq w_{\min}$

Here ε_w is a small constant that represents acceptable alignment error at a scale. In experiments we have set the value of ε_w 1/2 of the current scale w.

Now we use the above definitions to explain the match elimination stage, during which we adopt a two pass approach called reciprocal constraint pruning that is somewhat similar to the IRCP method as described by [18]. In the first pass, points in C are matched with points in $\hat{C}$, and in turn points in $\hat{C}$ are also matched with points in C. As the matching process is not symmetric, the initial matchings work out different results and can be used to cull points in both curves that have no corresponding point in the opposing image. The second pass refines this correspondence by retaining the pairs that do not have contradictory counterparts.

Note that, for purpose of providing a safeguard against mismatches, the reciprocal constraint applies strict restrictions to the matching process and risks discarding a lot of obviously good matches. [1] mentioned that the adopted detector does not guarantee stable feature detection at all scales. Actually, a feature detected at a coarser level typically splits into several features at a finer level, yet in case that they are mismatched or eliminated by the reciprocal constraint, these minor features sometimes do not get matched in the iteration at a finer level. Therefore we try to retain useful correspondences in M_i by comparing M and M_{i-1} carefully.(See Table 2). The assumption here is, since correspondences at a finer level is based on more priories, they are more reliable than estimations at a coarser level. Following this assumption,

only those estimations that do not have related correspondence in M are propagated to the next iteration.The framework of the proposed algorithm is summarized in Table 2.

3 Topological Constraints

In order to limit the computational cost in feature matching, candidate matches are pruned with two kinds of topological constraints. For simplicity we assume that both C and $\hat{C}$ are closed curves extracted in the same direction. In Section 5, the method will be extended to open curve matching .

3.1 The Order-Preserving Constraint

To ensure the consistency of the target and template curve, the correspondence relationship should be order-preserving. We have assumed that the parameters in U are consecutively distributed in the template curve. According to the order-preserving constraint, we also require that all parameters in V are consecutively distributed in $\hat{C}$.

$$\exists p : 0 \le v_{p+1} < v_{p+2} < \ldots < v_n < v_1 < \ldots < v_p \le \hat{L}$$

Observe that U and V divide C and $\hat{C}$ into n parametric intervals, i.e., $[u_1, u_2]$, $[u_2, u_3]$, ..., $[u_{n-1}, u_n]$ and $[u_n, L] \cup [0, u_n]$ for the template curve, and that, $[v_{p+1}, v_{p+2}]$, $[v_{p+2}, v_{p+3}]$, ..., $[v_{n-1}, v_n]$, $[v_n, v_1]$, $[v_1, v_2]$, ..., $[v_{p-1}, v_p]$ and $[v_p, \hat{L}] \cup [0, v_{p+1}]$ for the target curve, respectively, the order-preserving constraint is

$$v \in \begin{cases} [v_j, v_{j-1}] & if\ u_j \le u < u_{j+1} \\ [v_p, v_1] & if\ 0 \le u < u_1\ or\ u_n \le u \le L \\ [v_p, \hat{L}] \cup [0, v_{p+1}] & if\ u_p \le u < u_{p+1} \end{cases}$$

where v is the corresponding parameter of the given parameter u , $1 \le j < n$, $j \ne p$.

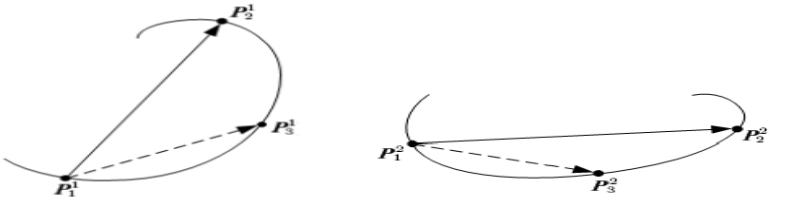


Fig. 1. P_3 is always on the same side of the line $\overrightarrow{p_1 p_2}$

Fig. 2. The sidedness constraint does not hold globally

The major contribution of the order constraint is decreasing the computational cost by greatly reducing the range of searching. Assume that all features are uniformly distributed in the parametric space and get matched during each iteration of the feature matching, only $O(N \log N)$ feature correlations are computed during the feature matching stage, where N is the amount of matched features. This can be derived directly following the method introduced in [12]. In comparison, the computational cost goes up to as expensive as $O(N^2)$ if we use a brute-force method that computes feature correlations pointwise.

3.2 The Sidedness Constraint

Consider a triple (p_1, p_2, p_3) of points in the scene, and let p_j^k be the corresponding point of p_j in the kth image. As illustrated in Fig.1, the sidedness constraint put up by[5] states that p_3 is always on the same side of the directed Line $\overrightarrow{p_1 p_2}$ in both images. This topological property can be easily computed from scalar triple products of vectors and is stable unless the camera translates perpendicularly across the 3D plane containing these points, which only happens to a minority of triples[5].

The sidedness definition given in [5] is derived from 3D coordinates yet we try to avoid using 3D coordinates here In the left image, p_3 is to the left of $\overrightarrow{p_1 p_2}$, while it lies on the righthand side in the right image.

so as to avoid reconstructing the 3D structure. Therefore the sidedness property is redefined as

$$Side2D(p_1^k, p_2^k, p_3^k) = \mathrm{sgn}((\overrightarrow{p_1 p_2} \times \overrightarrow{p_1 p_3}) \cdot \vec{n})$$

where p_j^k are 2D vectors and $\vec{n}$ is the unit normal of the image plane.

[5] also invented a global method to eliminate mismatches subsequent to the matching process. In our approach, however, a local constraint is applied ahead of the time-consuming process of computing feature correlations. The main reason of this modification is that we want to handle non-rigidity, yet in this case the sidedness constraint does not hold globally. Fig.2 shows a counterexample where p_3 moves from one side of $\overrightarrow{p_1 p_2}$ to the other because of a non-rigid deformation. Therefore the sidedness constraint is only valid in local cases, in other words, for all candidate matches (u,v), let

$$l = \sup_j \{\tilde{u}_j < u, \tilde{u}_j \in \tilde{u}\} \text{ and } r = (l+1)\%\tilde{n}$$

The following equation states the sidedness constraint

$$Side2D(C(\tilde{u}_l), C(\tilde{u}_r), C(u)) = Side2D(\hat{C}(\tilde{v}_l), \hat{C}(\tilde{v}_r), \hat{C}(v))$$

where n is the amount of obtained correspondences.

In experiments, we found that the sidedness constraint helpfully discards spurious matches and therefore improves the accuracy of the established correspondences between curves more efficiently.

4 Similarity Measurement

As explained above, both feature correlation and shape context are taken into account to design the synthesized similarity measurement. Feature correlation is used as a fundamental measure to depict the neighborhood of a point of interest, and can be derived from intensity, texture, and optical flow, etc. For brevity, normalized intensity cross-correlation is used here

$$Corr\ (u,v) = \frac{\sum_{i,j} (I_{ij} - EI_{ij})(\hat{I}_{ij} - E\hat{I}_{ij})}{\sqrt{\sum_{i,j} (I_{ij} - EI_{ij})^2} \sqrt{\sum_{i,j} (\hat{I}_{ij} - E\hat{I}_{ij})^2}}$$

where I_{ij} and $\hat{I}_{ij}$ are respectively the points in the neighborhood of $C(u)$ and $\hat{C}(v)$.

The shape context, though suffering from alignment error and mismatches, provides strong implications about the motion between I and $\hat{I}$. A method based on B-spline interpolation has been developed to represent this context by deriving a curve sketch from correspondences .

In short, B-spline functions are piecewise polynomial functions of finite supports. Nonrational B-spline functions of order k are recursively generated by the Cox-deBoor formulas[21]

$$N_{j,1}(u) = \begin{cases} 1, & if\ u_j \le u < u_{j+1} \\ 0, & else \end{cases}$$

And

$$N_{j,k}(u) = \frac{(u-u_j)N_{j,k-1}(u)}{u_{j+k-1}-u_j} + \frac{(u_{j+k}-u)N_{j+1,k-1}(u)}{u_{j+k}-u_{j+1}}.$$

where $u_1, u_2, ...u_{n+k}$ are consecutive parameters and k is the rank of the spline functions.

A 2D Non-Uniform Rational B-spline(NURBS) curve is the projection of a nonrational B-spline representation in 3D

$$C(u) = \frac{\sum_{j=0}^{n} W_j C(u_j) N_{j,k}(u)}{\sum_{j=0}^{n} W_j N_{j,k}(u)} = \sum_{j=0}^{n} C(u_j) R_{j,k}(u)$$

where $C(u_j)$ is the jth control point on the curve and W_j are weights that measures the importance of $C(u_j)$. In experiments we assign W_j =1 for all control points $C(u_j)$. NURBS representations of rank three are good enough to approximate complex curves and therefore we use cubic NURBS to represent curves in our method.

An important property of the B-spline representation is the property of affine invariance

$$A[C(u)] = A[\sum_{j=0}^{n} C(u_j) R_{j,k}(u)] = \sum_{j=0}^{n} A[C(u_j)] R_{j,k}(u)$$

where A is some kind of affine transformation, i.e., rotation, scaling, translation or shear. Observe that $A[C(u_j) = \hat{C}(v_j)]$, we have

$$A[C(u\ )] = \sum_{j=0}^{n} A[C(u_j)] R_{j,k}(u) = \sum_{j=0}^{n} \hat{C}(v_j) R_{j,k}(u).$$

This makes a lot of sense: First, the curve sketch is derived from 2D points, hence it is invariant to affine transformations but avoids reconstructing the 3D scene. Second, to compute a point in the target curve, the number of required parameters is reduced to four pairs of correspondences. Comparatively, eight pairs of 3D correspondences are required in the popular fundamental matrix estimation methods. Third, affine invariance can model local motions in the neighborhood of a point. Consequently, by aggressively increasing the density of control points, NURBS can be a powerful tool for modelling very complex motions, especially non-rigid motions.

In the coarsest scale, points are matched according to feature correlations. With the scale decreasing, more correspondences are obtained and a more complex measurement combining feature correlation and shape context is used in scoring candidate correspondences

$$Similarity(u,v) = Corr(u,v) \exp[-\frac{\left\| \hat{C}(v) - A[C(u)] \right\|}{\alpha}]$$

In experiments, we found that α = 80 is an optimal choice.

The feature matching stage is summarized in Table 3. In experiments, we have shown that the use of a combined measure is more effective than using feature correlation alone.

Table 3. Summary of the Feature Matching Stage

1. Arranged all points $C(u)$ extracted from the template curve C in a descending order in curvature.
2. For each point $C(u)$ in the template curve C (a). If there are more than four pairs of correspondences, find the correspondence v of u $v = \arg\max Similarity(u, v')$ (b) If there are less than four pairs of correspondences, find the correspondence v of u $v = \arg\max Corr(u, v')$
3. Add (u, v) to the current estimation M

5 Experiments and Applications in Shape Recognition

In this section, we present some experimental evaluation of the new curve matching technique. There are three aspects to this study, each highlighting one facet of the presented method. The first part of this study validates the ability of the new method to handle rigid and non-rigid deformations. Secondly, we provide some indication on how the new method outperforms its alternatives using less constraints. Finally we extend the method to open curve matching and test its capability of comparing shapes.

To test the robust performance under rigid motions, we commence with matching closed curves using images taken from the COIL-20 database[16]. The example depicts a piglet saving box corresponding to different camera viewing directions. There is a obvious difference between the two silhouettes due to feature degradation and occlusion. Fig.4 shows the achieved correspondence connected with dashed lines. These results show that the new method returns considerable correspondences. We repeated this set of experiments using a set of human face images to demonstrate the ability of the method to handle non-rigid motion. The images used here are shown in the bottom row of Fig.3, which depicts different expression of a Chinese girl. The first example is furnished by matching features of her left eye. The result is shown in Fig.5.

We have introduced various constraints in the preceding sections. To indicate how the new method benefit from them, Fig.6 shows the proportion of computed correspondence when each of them is not used. The experiment is conducted on the piglet saving

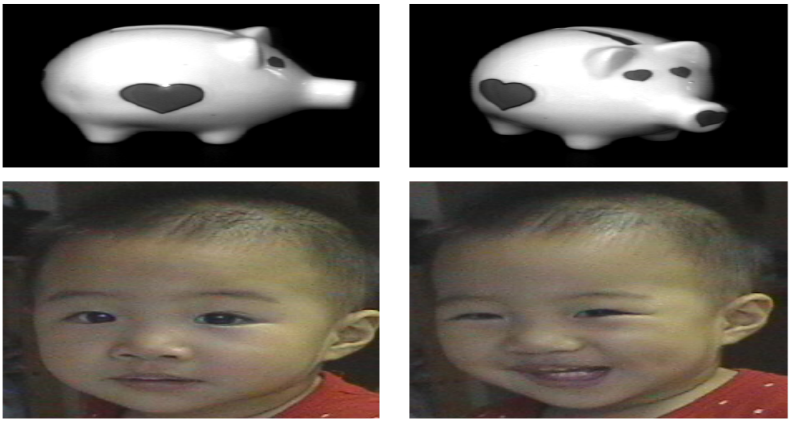


Fig. 3. Two pairs of images used in the experiments Top row: piglet saving box from different views. Bottom row: faces with different expressions.

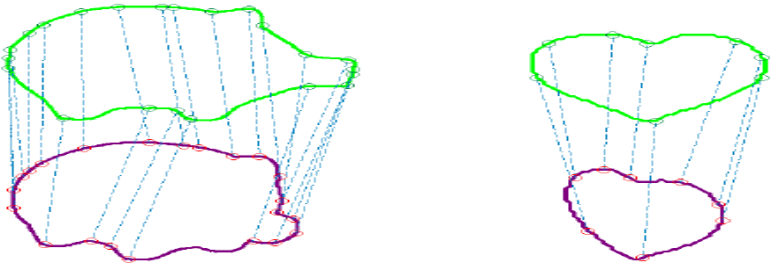


Fig. 4. Matching result of the piglet saving box. Left column: silhouette. Right column: heart shape.

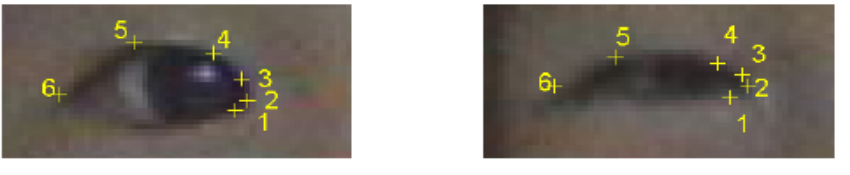


Fig. 5. Matching result of the girl's left eye. The matching result is 100% right though some non-rigid motion has occurred between two images.

box images. Due to space limitations, we do not present the results in details here, but it is worth pointing out that the new method always returns more accurate results though the amount of correspondence sometimes varies little.

In the final experiment, we have matched features in the left ear of the Chinese girl. The aim here is to validate the capacity of the new method to match open curves and achieve curve correspondence. The concept of open curve matching is two-fold. We

first explain the case of occlusion, i.e., a part of the contour is not visible. By connecting both ends of a curve, the problem can be settled in the same way as introduced in the preceding sections. In the second case, both curves are visible. Consequently, we add both ends of the curve as feature points for three times to extend the new method to match open curves. The correspondence results are shown in Fig.7, and the amount of correspondence is shown in Table 4. At a time, the match with the highest score for both curves is put into correspondence and removed from all other considerations. The result is compared with manual checks and finally we found that the algorithm achieves 100 percent correctness in curve correspondence and 94 percent correctness (32 out of 34) in point correspondence by matching the most similar curves at a time. More specifically, the 2 errors are very slight alignment errors. Note that the correspondence of curve 20-22 and 33-34 rank second in the similarity matrix. This results mainly from uniform intensity distribution of the skin and the relatively straightness of the curves.

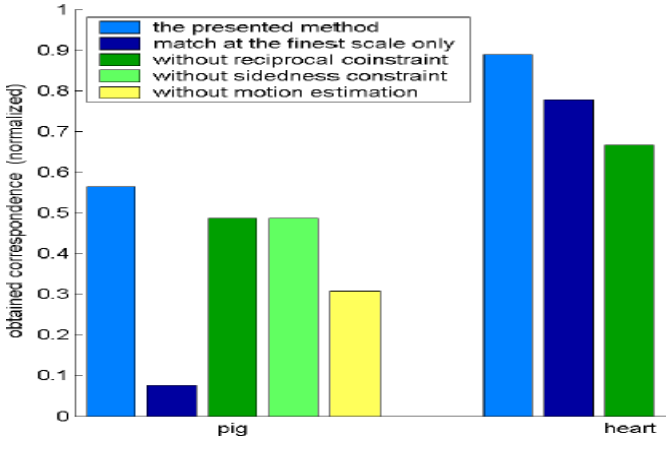


Fig. 6. Comparative results of using different subsets of constraints. Our method outperforms all alternatives using less constraints.

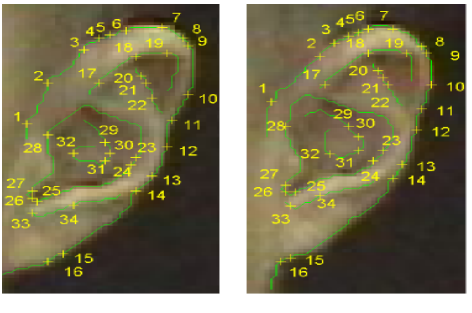


Fig. 7. Point correspondence between the right ears that have undergone perspective deformation. The correctness of curve correspondence is 100% and that of point correspondence is 94%.

Table 4. Similarity measurement of curves extracted from Fig.7. The italic fonts marks the item referring to the matches obtained by our algorithm. Each row represents a curve in the left image and each column represents a curve in the right image.

	1-16	17-19	20-22	23-28	29-32	33-34
1-16	*47*	1	0	3	5	2
17-19	7	*16*	2	10	6	8
20-22	0	2	*6*	8	4	3
23-28	3	7	6	*36*	9	1
29-32	2	6	6	4	*9*	7
33-34	7	0	2	4	4	*4*

Interestingly, note that similar parts of the curves tend to achieve more correspondences. Therefore, unlike most similarity metrics of curves, our method provides evidence for the similarity score by telling similarity and dissimilarity of the curves. Consequently, the new method gives rise to a similarity metric with good comprehensibility.

6 Conclusions and Future Works

A coarse-to-fine approach for matching curves is presented. It is inspired by the idea of using the correspondence estimation to generate the topological and geometric priors at a finer level. This has been achieved through the analysis of the curve topology and the synthesis of the B-spline interpolation techniques. This is in contrast to existing multi-scale methods for curve matching that use pure feature correlation or 3D structure recovery at a fixed scale. The contribution of this work are summarized by:

- provided a natural way to fuse image intensity distribution and shape context into an integrate framework
- resolved the inherent difficulty associated with 3D structure recovery using the B-spline interpolation techniques
- defined a multi-scale metric of similarity between two curves with good comprehensibility.

We would specially point out that the proposed method is more than simply a novel method of contour matching, but an attempt to integrate knowledge into image analysis. There has been a wide acknowledgement that the usage of knowledge is the core of human and machine vision [23], yet related works on this topic is very limited. The presented method implies how to acquire knowledge automatically from intensity images and integrate them into an image analyzing system.

The method presented here can be extended in several directions. Interesting future work could redesign the interpolation algorithm to estimate the correspondences between surfaces. Following established work on triangular B-spline representations [9], this can be done in a straightforward and elegant way by interpreting the problem as matching Delaunay triangulations. Besides, the localness of the metrics and

representations used here indicates the potential of this method to solve the problem of occlusion. Therefore extending our method to match occluded curves will be investigated in the future as well.

Acknowledgment

This work was supported by the NSFC under grant No. 60303007 and China Basic Research Project (973) under contract No. 2001CB309401.

References

1. H. Asada and M. Brady, "The curvature primal sketch," IEEE Trans. on Pattern Analysis and Machine Intelligence, Vol.8, No.1, pp. 2-14, 1986.
2. N. Ayache, O. D. Faugeras, "HYPER: A New Approach for the Recognition and Positioning of two-dimensional objects", IEEE Trans. on Pattern Analysis and Machine Intelligence, Vol.8, No.1, pp. 44-54, 1986.
3. F. S. Cohen, Z. Huang and Z. Yang, "Invariant Matching and Identification of Curves Using B-Spline Curve Representation," IEEE Trans. on Image Processing, Vol. 4, No. 1, pp.110, 1995.
4. O. Faugeras, "What Can Be Seen in Three Dimensions with an Uncalibrated Stereo Rig", European Conference on Computer Vision, pp. 563-578, 1992.
5. V. Ferrari, T. Tuytelaars and L. V. Gool, "Wide-baseline Multiple-view Correspondence", IEEE Computer Society Conference on Computer Vision and Pattern Recognition, Vol.1, pp.718-725, 2003
6. D. A. Forsyth and J. Ponce, Computer Vision: A Modern Approach, Prientice Hall, 2002.
7. Y. Gdalyahu and D. Weinshall, "Flexible Syntactic Matching of Curves and Its Application to Automatic Hierarchical Classification of Sihouettes", IEEE Trans. on Pattern Analysis and Machine Intelligence, Vol. 21, No. 12, pp. 1312-1328, 1999.
8. P. G. Gottschalk, J. L. Turney, and T. N. Mudge, "Efficient Recognition of Partially Visible Objects Using a Logarithmic Complexity Matching Technique," The International Journal of Robotics Research, Vol.8, No.2, pp. 110-131, 1989.
9. G. Greiner and H. -P. Seidel, "Modeling with Triangular B-Splines", *IEEE Graphics and Applications*, Vol.14, No.2, pp.56-60, 1994.
10. R. I. Hartley, R. Gupta, and T. Chang. "Stereo from Uncalibrated Cameras", Conference on Computer Vision and Pattern Recognition, pp. 761-764, 1992.
11. M. H. Han and D.Jang, "The Use of Maxium Curvature Points for the Recognition of Partially Occluded Objects," Pattern Recognition, Vol.23, No.1 ,pp. 21-33, 1990.
12. D. E. Knuth, The Art of Computer Programming, Addison-Wesley ,1973.
13. D. Marr, Vision, Freeman, 1982.
14. D. Marr and T. Poggio, "A Computational Theory of Human Stereo Vision", Proceedings of the Royal Society of London, B-204, pp.301-328, 1979.
15. F. Mokhtarian and M.Bober, Curvature Scale Space Representation: Theory, Applications and MPEG-7 Standardization, Kluwer Academic Publishers, 2003.
16. H. Murase and S. Nayar, "Visual Learning and Recognition of 3-D Objects from Oppearance", International Journal on Computer Vision, Vol. 14, No. 1, pp. 5-24, 1995.

17. C. Orrite S. Blecua, and J. E. Herrero, “Shape Matching of Partially Occluded Curves Invariant under Projective Transformation”, Computer Vision and Image Understanding, Vol.93, pp. 34-64, 2004.
18. T. Pajdla and L. Van Gool, “Matching of 3-D curves using semi-differential invariants”, International Conference on Computer Vision, pp.390-395, 1995.
19. A. Rattarangsi and R. T. Chin, “Scale-based Detection of Corners of Planar Curves”, IEEE Trans. on Pattern Analysis and Machine Intelligence, Vol. 14, No. 4, pp. 430-449, 1992.
20. E. Rivlin and I. Weiss, “Local Invariants for Recognition”, IEEE Trans. on Pattern Analysis and Machine Intelligence, Vol. 17, No. 3, pp. 226-238, 1995.
21. D. F. Rogers and J. A. Adams, Mathematical Elements for Computer Graphics, 2nd ed. New York: McGraw-Hill, 1990.
22. D. Shen, W. Wong and H. H. S. Ip, “Affine-invariant Image Retrieval by Correspondence Matching of Shape,” Image and Vision Computing, Vol. 17, pp. 489-499, 1999.
23. M. Sonka, V. Hlavac and R. Boyle, Image Processing, Analysis and Machine Vision, Chapman & Hall, 1993.
24. W. H. Tsai and S. S. Yu, “Attributed String Matching with Merging for Shape Recognition,” IEEE Trans. on Pattern Analysis and Machine Intelligence, Vol. 7, No. 4, pp. 453462, 1985.
25. Z. Xue, D. Shen, E. K. Teoh, “An Efficient Fuzzy Algorithm for Aligning Shapes under Affine Transformations,” Pattern Recognition, Vol. 34, pp. 1171-1180, 2001.
26. A. Zisserman, D. A. Forsyth, J. L. Mundy, C. A. Rothwell, “Recognizing General Curved Objects Efficiently”, in: Geometric Invariance in Computer Vision, MIT Press, Cambridge, MA, pp. 228-251, 1992.

Towards Efficient Ranked Query Processing in Peer-to-Peer Networks*

Keping Zhao[1], Shuigeng Zhou[1,2], and Aoying Zhou[1,2]

[1] Department of Computer Science and Engineering
[2] Shanghai Key Lab of Intelligent Information Processing
Fudan University, Shanghai 200433, China
{kpzhao, sgzhou, ayzhou}@fudan.edu.cn

Abstract. P2P computing is gaining more and more attention from both academia and industrial communities for its potential to reconstruct current distributed applications on the Internet. However, the basic DHT-based P2P systems support only *exact-match* queries. Ranked queries produce results that are ordered by certain computed scores, which have become widely used in many applications relying on relational databases, where users do not expect exact answers to their queries, but instead a ranked set of the objects that best match their preferences. By combing P2P computing and ranked query processing, this paper addresses the problem of providing ranked queries support in Peer-to-Peer (P2P) networks, and introduces efficient algorithms to solve this problem. Considering that the existing algorithms for ranked queries consume an excessive amount of bandwidth when they are applied directly into the scenario of P2P networks, we propose two new algorithms: *PSel* for ranked selection queries and *PJoin* for ranked join queries. *PSel* and *PJoin* reduce bandwidth cost by pruning irrelevant tuples before query processing. Performance of the proposed algorithms are validated by extensive experiments.

1 Introduction

Recently, Peer-to-Peer (P2P) computing [17] is gaining more and more attention from both academia and industrial communities for its potential to reconstruct current distributed applications on the Internet. In P2P systems, the peers use a dedicated naming space and act as both service providers and service consumers, while keeping themselves fully autonomous. Since most peers in P2P networks are equal in capability and responsibility, traffic is often evenly distributed and congestion can be reduced to some degree. Most P2P networks are also designed

* This work was supported by the National Natural Science Foundation of China (NSFC) under grant numbers 60373019, 60573183, and 90612007, and the Shuguang Scholar Program of Shanghai Education Development Foundation. †Shuigeng Zhou is the correspondence author. ‡Part of this paper has been first appeared in DEXA'05 Workshop – PDMST'05.

R. Lu et al. (Eds.): Chinese-German Ws on Cognitive Systems 2005, LNAI 4429, pp. 145–160, 2007.
© Springer-Verlag Berlin Heidelberg 2007

under the assumption of frequent node failures or off-line, thus such networks are always able to maintain relatively stable performance.

Recent P2P systems employ *distributed hash table* (DHT) [23,20,21,26] to overcome performance challenges (e.g. high routing costs) confronted by previous P2P systems like Gnutella [1]. DHT is a distributed and decentralized mechanism for associating hash values (keys) with objects. Each participant in the DHT stores a fraction of the contents of a hash table. There is a basic interface provided by DHT mechanisms, namely *lookup*(*key*), which returns the IP address of the node currently responsible for the key. DHT provides theoretical bounds on both the numbers of routing hops and maintenance messages needed to manage peer join and departure in the networks. The fundamental limitation of such systems is that they support only *exact-match* queries. However, what the users need is much more than searching objects using their keywords. Recently, the database community has commenced research work on providing complex query facilities for the DHT-based P2P systems. For example, in the paper [10] titled "What can Peer-to-peer do for database, and vice versa?" Gribble et al. outlined some potential challenges to be confronted when P2P systems are exploited for database query processing. In [13], the authors introduced a database-style query engine PIER, which aims to provide database query processing facilities in large scale by using DHT, while leaving data to remain in its natural style (typically a file system).

In this paper, we study the problem of *ranked queries* (or *top-k queries*) in the scenario of P2P networks. In multimedia database systems, ranked queries are widely used to access the data (such as images and videos), whose attributes are inherently fuzzy. Ranked queries can return results that have the highest scores computed by user-defined rank functions. Ranked query model is also prevalent over plain relational data for certain applications where users do not expect exact answers to their queries, but instead a rank set of objects. Thanks to the ability of providing results in information retrieval style, ranked queries have become a vital need for many applications relying on relational databases. The database community has done a lot of work in supporting ranked queries for traditional relational databases in recent years [6,8,25,15,7,12,5], and this paper advances the work to address the problem of supporting ranked queries in P2P networks.

1.1 Ranked Queries in P2P Networks: An Example

A proper example to demonstrate the applications of ranked query in P2P networks is legal music discovery and trading. Music fans may have bought many music tracks and have collected information of a large number of albums and artists of different genres. In order to trade the tracks with others and to find interesting music, the fans can publish and share the information in the P2P networks. The information can be organized in relational data fashion, however the music need not be loaded into the P2P database system. For example, we can organize the data into 3 tables: Track{title, album, length, style}, Album{title, artist, released_date, number_of_tracks} and Artist{name, age}. Suppose that

a user is looking for some country music by old artists, and he prefers long length. He can pose a query to the network using the following SQL statement:

```
SELECT tr.title, al.artist
FROM Track tr, Album al, Artist ar
WHERE tr.style = "country"
        AND tr.album = al.title
        AND al.artist = ar.name
ORDER BY (0.5*tr.length + 0.5*ar.age)
STOP AFTER 10
```

The SQL statement above produces the top-10 results ordered by the score computed by the function in the ORDER BY clause, which expresses the preference of the user for long-playing music and old artists.

How to evaluate efficiently such a query aforementioned in P2P environment where data (relational tables distributed over the peers) is the goal of this paper.

1.2 The Challenges

Although quite a lot work on ranked query processing has been done in traditional database area, the existing methods can not be applied to P2P context straightforward. The different application scenario arouses some new challenges.

In traditional databases area, most of the algorithms for answering ranked queries assume that tuples are ordered on some rank attributes. With such an assumption, the algorithms use the monotone property of rank functions to prune irrelevant tuples to reduce I/O costs. This assumption is reasonable in centralized databases. Even when the assumption is not satisfied, some indices can be used or the table can be sorted before processing. However, in the P2P scenario, because all tuples of a table are stored on different peers in the network, we can not assume the tuples are sorted on some rank attributes. Nor can we sort the tuples before processing, for it is very expensive if not impossible.

Due to the large scale of P2P networks and the wide distribution of data, reducing network traffic is one of the major concerns of P2P applications. This requires querying processing algorithms to consume bandwidth resources as little as possible.

1.3 Our Contributions

Contributions of this paper are as follows:

- We propose two efficient algorithms *PSel* and *PJoin* to answer ranked selection queries and join queries in P2P networks respectively. These two algorithms distinguish themselves from previous work in many aspects. For example, they make no assumption on availability of ordered access to data. Both of the algorithms can save considerable amount of bandwidth cost compared with the straightforward methods.

– Extensive experiments are carried out to evaluate the performance of the algorithms proposed and the impact of different settings.

1.4 Organization of the Paper

Section 2 surveys the related work. Section 3 presents the problem definitions. Section 4 and Section 5 introduce the *PSel* algorithm for ranked selection queries and the *PJoin* algorithm for ranked join queries respectively. Section 6 gives extensive experimental results of synthetic datasets to validate the proposed algorithms. Finally, Section 7 concludes the paper and highlights future work.

2 Related Work

Recent work on P2P computing can be found in a survey [17]. In [20], [23], [21] and [26], the authors present the popular DHT based overlay networks (Chord, CAN, Pastry and Tapestry respectively). There is an emerging set of work which aims to provide complex query facilities and IR-style functionalities in P2P networks. For example, in [22] Schmidt and et al. proposes a framework to support rang query, which maps the high-dimensional data space to Chord namespace using space filling curve. In [24] the authors introduce the *pSearch* information retrieval system based on distributed document indices over the P2P network based on document semantics.

The early paper [10] outlines some potential challenges to be confronted when P2P systems are exploited for database query processing. [13] is closely related to our work, which presents PIER system, a database-style query engine based on DHT overlay networks. The architecture of PIER system consists of three layers, namely the application layer, the PIER query processor and the underlying DHT. The PIER query processor is the core part of PIER, which supports the simultaneous execution of multiple operators. PIER's aim is to provide database query processing facilities in large scale by using DHT, while leaving data to remain in its natural habits (typically a file system). The architecture of the system we considered is similar to that of PIER. However, *our focus is to develop efficient algorithms to answer specifically ranked queries in DHT-based P2P networks*, which has not yet been properly addressed in PIER.

In the traditional database area, there is a variety of work on ranked queries evaluation. Fagin et al. [9], Guntzer et al. [11] and Nepal et al. [19] independently propose the first set of efficient algorithms to answer top-k selection queries. A database consisting of m attributes is considered as m sorted lists. The TA algorithm [9] assumes the availability of random access to object scores in any list besides the sorted access to each list, while the NRA algorithm [9] makes no random accesses. In [6], the authors introduce the *MPro* algorithm to evaluate ranked top-k queries with expensive predicates, which minimizes expensive probes by determining whether a probe is really required. Bruno et al. [7] propose algorithms to evaluate top-k selection queries over web-accessible autonomous databases. They assume that one of the web sources supports ordered access and multiple random-access sources are available.

In [18], the authors introduce the J^* algorithm for incremental join of ranked inputs. By mapping the ranked join problem to a search problem in the Cartesian space of the ranked inputs, J^* uses the A^* class of search algorithm to produce ranked results. Ilyas *et al.* [15,14] propose another two algorithms to answer ranked join queries. *NRA-RJ* [15] requires *key* as the join attributes, thus it supports only the join queries based on key-equality condition. HRNJ [14] is an efficient ranked join query operator based on hash ripple join algorithm, which assumes the inputs ordered individually on the score attributes.

There is also a set of work [4,3,16] which proposes solutions for ranked queries in distributed scenarios. However, the solutions are designed for the case where a table is vertically partitioned onto different servers, that is each server stores a column of the table. In this paper we consider the scenario where a table is horizontally partitioned, that is each server store a subset of the tuples. In [2], the authors present an algorithm for ranked selection queries, which also assume that the table is horizontally partitioned. However the algorithm is restricted to a specific overlay network. Our methods proposed herein make no assumption on the underline overlay network, and moreover address ranked join queries.

3 Problem Statement

In the ranked queries system we consider, the peers share their data in the form of database tuples and relations, where a tuple is stored in the DHT namespace of its corresponding table. We assume that a global schema is known to each peer in the network. The assumption is reasonable in many applications where, for example, data shared are generated by popular software. Though the design of P2P schema and data integration is still an open problem, we agree with the argument in [13] that massively distributed database research can and should proceed without waiting for breakthroughs on the semantic front. We use bandwidth cost as the main performance metric. We argue that bandwidth cost dominates other costs (e.g. cost of computation) in a P2P systems, which is consistent with other published work on P2P systems. In implementation, we adopt Chord [23] as the DHT scheme. However, the algorithms proposed herein are independent of the Chord protocol.

Without loss of generality, we assume the domain of the attributes is in $[0, 1]$. Peers are allowed to pose SQL-like ranked queries to the system. The query statement takes the following form, as proposed in [5]:

SELECT **select-list**
FROM **from-list**
WHERE **qualification**
ORDER BY **rank-function**
STOP AFTER **k**

Although the query can be written by following the new SQL99 standard, we adopt the above concise form for simplicity of discussion. The SELECT, FROM

and WHERE clauses are similar to those of the standard SQL query. The ORDER BY defines the *rank function*, which specifies the scores of result tuples. A rank function $\mathcal{F}(x_1, ..., x_n)$ may take either attributes of the tables in the from-list or complex predicates as parameters. We consider the case of attributes, and the solutions derived can be extended easily to the case of predicates. As the previous work on ranked queries, this paper assumes that the rank function is the *monotone*.

Definition 1. (**Monotone Function**) *A function* $\mathcal{F} : [0, 1]^n \rightarrow R$ *is* monotone *if the following holds: if* $x_i \leq y_i$ *holds for each* i*, then* $\mathcal{F}(x_1, ..., x_n) \leq \mathcal{F}(y_1, ..., y_n)$.

The last clause, STOP AFTER, sets the number of ordered results that are expected to return by the value of **k**.

In the following discussion, we distinguish the *selection* and *join* predicates just as in relational queries, depending on whether the **from-list** consists of one or more tables. For the ranked join queries, we focus on *equi-join* of two relations. We are now ready to formally define the problems we address in this paper.

Definition 2. (**Ranked Selection Query**) *Given a table* R *stored in the namespace of* N_R*, a monotone rank function* $\mathcal{F}(r_1, ..., r_n)$*, where* r_i *(i ranges from 1 to n) is corresponding to an attribute of* R*, and* k *(*$k > 0$*),* ranked selection query *returns the top* k *tuples of* R*, which are ordered by the scores computed by the rank function* $\mathcal{F}$*.*

Definition 3. (**Ranked Join Query**) *Given two relations* R *and* S*, which are stored in the DHT namespace of* N_R *and* N_S *respectively, the join attributes* $R.j$ *and* $S.j$*, the monotone rank function* $\mathcal{F}(r_1, r_2, ..., r_u, s_1, s_2, ..., s_v)$*, where* r_i *(*s_i*) (i ranges from 1 to n) is an attribute of* R *(*S*), and* k *(*$k > 0$*),* ranked join query *returns the top* k *tuples of* $R \bowtie_{R.j=S.j} S$*, which are ordered by the scores computed by the function* $\mathcal{F}$*.*

In the next sections, we present new algorithms to answer ranked queries in P2P networks with small bandwidth cost. To save the space, we introduce the algorithms in natural language. Pseudocodes, proofs and analysis of the algorithms can be found in [27].

4 *PSel*–A P2P Ranked Selection Algorithm

A ranked selection query returns the top-k tuples from a single table R. A straightforward scheme for answering this kind of queries is as follows: first, each node in namespace N_R evaluates the local *top-k* tuples, and sends the k tuples to the query node; then the query node merges all the tuples received to compute the final results. However, there are two major drawbacks with this naive scheme:

– Only a small part of the large number of tuples sent to the query nodes is really relevant to the k final results.

– When the scale of the network is large, the flooding of the massive messages to a single query node leads to a hot spot in the network.

In fact, for the top k results of the query we are seeking, they are stored over *at most* k different peers. If we can limit the query processing to k peers which may have the final results, a large amount of bandwidth cost will be saved. Based on this observation, we propose the *PSel* algorithm, which is as follows.

1. Each node belonging to the namespace N_R computes the tuple of top score, which is the ceiling score value of the R tuples stored on the node. This job can be done by scanning the whole local R table if the scale of the local table is small, or by making use of the algorithm of TA [9] or its variants when the scale is large. The ceiling value along with the *id* of the corresponding node are then sent to the query node, which computes the top k ceiling values and recognize the k corresponding node *id*s.
2. The k nodes with the top k ceiling score values are notified to calculate the local top k tuples and send their results to the query node, where the received results are merged and the final top k tuples are retrieved.

The *PSel* algorithm consists of 2 steps. The first step finds the k possible peers which may store the final results, and the second step evaluates the final results by applying the straightforward scheme to the k peers. In the first step, only the ceiling score values (rather than the tuples themselves) and the *id*s of the node are sent to the query node, which greatly reduces the size of the messages. By restricting the query processing only to k peers, the PSel algorithm uses only about k/n bandwidth cost of the straightforward scheme.

5 *PJoin*–A P2P Ranked Join Algorithm

Because the ranked join query involves tow relations, and the tuples of the relations are scattered over the overlay network, evaluation of ranked join queries is a nontrivial job. A straightforward strategy may generates the complete join results of $R \bowtie S$, afterwards calculates the top-k join results. However, only a small fraction of the complete join results is relevant to the final top-k results. The naive strategy suffers from the expensive cost of computation on the whole join results.

In this section, we present a novel strategy *PJoin* for answering ranked join queries in P2P networks. *PJoin* aims to evaluate the top-k tuples with small bandwidth cost, which is achieved by pruning irrelevant tuples of local nodes before they are sent to be probed for join matches. *PJoin* is based on the symmetric hash join algorithm in parallel databases. It's straightforward to revise the algorithm to adopt the symmetric semi-join strategy to save more bandwidth cost. In *PJoin* the DHT facility serves as both the overlay network and the hash tables for storing tuples. A temporary namespace N_J is used to provide the hash-table facilities. Each node maintains two hash tables for R and S respectively. To probe the join matches, each *relevant* tuple from R or S is

hashed into the namespace N_J on the hash join attribute. And the node responsible for the tuples in N_J performs local probing just as the traditional hash join algorithm does, that is to insert the tuple into its corresponding hash table and probe the other table for join matches. However, different from traditional hash join algorithms, on each node only the matches with the top-k local rank scores are kept and updated when a new match is generated.

The $PJoin$ algorithm is as follows:

1. This step calculates a lower bound on the smallest score of the top-k results by sampling method. First, the query node randomly selects p nodes from N_R and N_S. These selected nodes then hash the tuples in local R table or S table into N_J for join matching. Then the query node calculates the top-k matches from N_J by applying $PSel$ algorithm. Let Γ denote the smallest score of the top-k matches. And we use Γ as the lower bound on the smallest score of the top-k results.
2. The query node sends the original query along with Γ to the remaining nodes in the namespaces of R and S. For a node in the namespace of R, it scans the local table of R, and calculates the ceiling value of each tuple's score by setting all s_is of $\mathcal{F}(r_1, r_2, ..., r_u, s_1, s_2, ..., s_v)$ to 1. The tuples with the ceiling values less than Γ are surely irrelevant to the top-k results. So only the tuples whose ceiling value are greater than Γ are hashed to the namespace N_J for join matching. With the same strategy, the peers in the namespace of S also discard the tuples whose ceiling value are smaller than Γ, and only the remaining tuples are hashed to N_J for join matching. At last, the query node computes the final top-k join results from N_J by making use of $PSel$.

In $PJoin$, the first step calculates a lower bound of the top-k join, then in the second step the remaining nodes use it to prune irrelevant tuples for match probing. Notice that the tuples hashed to N_J in the first step are still remained in hash tables for probing during the second step. The experimental evaluations show that the algorithm can prune a big fraction of tuples in practice.

6 Experimental Evaluation

To evaluate the efficiency of the ranked query algorithms proposed herein and the impact of different settings on performance of the algorithms, we carried out a variety of experimental evaluations by simulation. First we evaluate the $PSel$ ranked selection algorithm, and compare it with the naive strategy. Then performance of the $PJoin$ is studied as different parameters vary, including the number of selected nodes, the number of attributes in rank function and etc.

For the ranked queries over traditional databases, the performance of an algorithm is measured by the time needed to retrieve the results, and the number of pages accessed during query processing. However in the scenario of P2P networks, the performance is greatly impacted by amount of data transferred in networks. With limited bandwidth, data transferred in turn determined the

time needed to return results. Thus, in the evaluations, we take bandwidth cost as the main performance metric. With the DHT-based P2P network considered in this paper, the bandwidth cost can be evaluated by the product of exchanged data's size and the number of routing hops in the network.

We first present the simulation and experimental setup, then the experimental results.

6.1 Simulation and Experimental Setup

We implement the algorithms in C++, and run the experiments on a Linux 2.4 workstation with four 2.8G Intel Xeon processors and 2.5G RAM.

The implementation makes use of Chord as the DHT facility, and the number of nodes in the overlay network varies from 1000 to 10,000. In order to test the performance of the algorithms over different datasets, we employ 3 kinds of synthetic datasets.

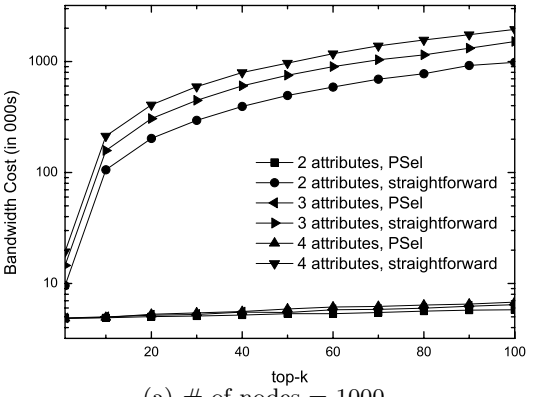
1. Uniform Dataset The values of each attribute of the uniform dataset uniformly distributed in the range of [0,1].
2. Gaussian Dataset The distribution of the value of each attribute follows the Gaussian distribution with mean value 0.5 and standard deviation 0.16.
3. Zipf Dataset The value of each attribute of the Zipf dataset is generated by the Zipf distribution, which is defined by $P(i) = 1/i^{\theta} H_{\theta}(V)$, where $P(i)$ is the occurrence probability of the ith most frequent object, V is the size of the object set, $H_{\theta}(V)$ is harmonic number of order θ of V, and θ is a parameter controlling the skew of the distribution. We set θ to 1.5 in our experiments.

6.2 Results of the *PSel* Algorithm

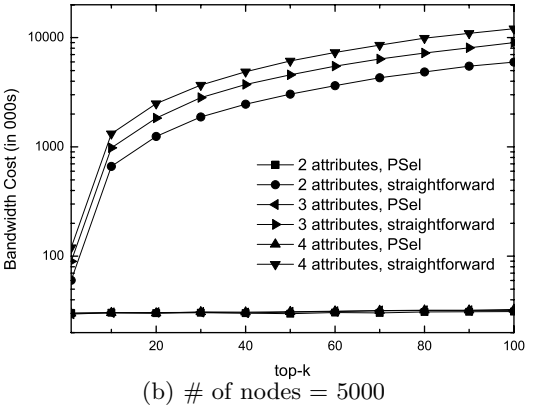
Our first experiment compares the performance of $PSel$ with that of the straightforward scheme. Fig. 1 shows the bandwidth cost of the two different schemes. We uses a uniform dataset of 500,000 tuples, whose's attributes number varies from 2 to 4. Each result in the figures is the averaged bandwidth cost over 100 queries. Because the results of the other two datasets show similar trend, we omit them for brevity. Fig. 1(a) presents the results of a P2P network with 1000 nodes. Notice that the vertical axis is of logarithmic scale. When k is small, the difference of the two schemes' costs is not very large. However, as k increases, the difference becomes much larger, which conforms to our analysis. The results of Fig. 1(b) and Fig. 1(c) show similar changing trend. The costs of both $PSel$ and the straightforward scheme increase along with the network scale. However the cost of the straightforward scheme increases much faster than that of $PSel$.

6.3 Results of the *PJoin* Algorithm

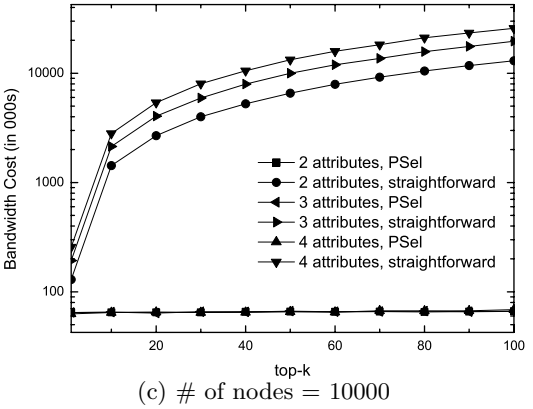
We test the effectiveness of the PJoin ranked join query algorithm in this section. PJoin aims to prune irrelevant tuples before they are sent to other nodes for join



(a) # of nodes = 1000



(b) # of nodes = 5000



(c) # of nodes = 10000

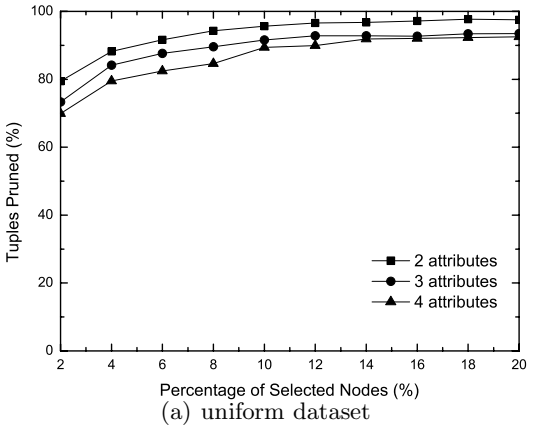
Fig. 1. Performance comparison between PSel and the straightforward scheme

matching. By pruning the irrelevant tuples, the algorithm saves both bandwidth cost and computation cost. Therefore, we use the percentage of tuples pruned before join matching to measure saved bandwidth cost.

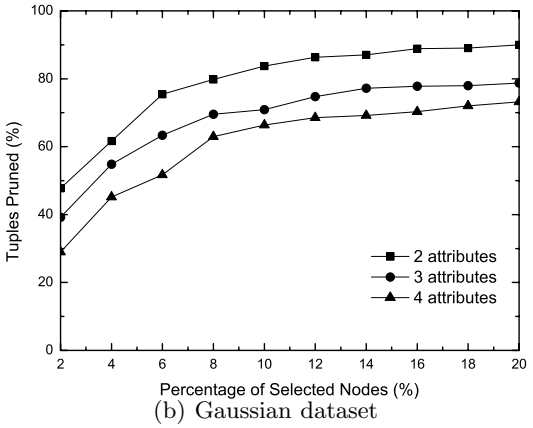
The Impact of the number of selected nodes on performance. Fig. 2 shows the performance changing of *PJoin* when the number of selected nodes is adjusted. Experiments are conducted over the 3 synthetic datasets, each of which consists of two join tables with 1,000,000 tuples. And the total join results also consist of 1,000,000 matches. We vary the percentage of selected nodes in the network of 10,000 peers from 1% to 20%, and generate 100 random queries for every setting under which the averaged performance is tested, that is the percentage of tuples pruned in the whole P2P network before join matching. Each figure consists of 3 curves, each of which corresponds to the results of queries with 2-attribute, 3-attribute or 4-attribute rank functions respectively. Fig. 2(a) and Fig. 2(c) present the results over the uniform dataset and the Zipf dataset respectively. The performance changing patterns over these two datasets are nearly similar. As the number of selected nodes increases, the percentage of pruned tuples grows. It grows a little fast when the sampling percentage is below 8%; the growing speed slows down after the sampling percentage surpasses 8%. When sampling 5% to 8% percent nodes, over 80% percent of the tuples are pruned. The performance is better for the cases of small rank attributes. In *PJoin* we compute the upper bound of the possible matches' rank score of a tuple by setting the value of remaining unknown attributes to 1. When the number of rank attributes in the rank function gets larger, the estimated bound turns looser. Considering that a tuple is pruned only when its estimated upper bound is less than the lower bound of the top-k results computed in the first step, so a loose bound can make an irrelevant tuple remain for join matching.

The results over Gaussian dataset shown in Fig. 2(b) is roughly similar to that over the other two datasets. However, the overall performance over the other two datasets is better. As we mentioned, the value of each attribute of the tables in Gaussian dataset follows a Gaussian distribution whose's mean value is equal to the center of the attribute's domain. Thus in the space formed by the rank attributes, the distribution of the total join results also follows an approximate Gaussian distribution whose's mean value is the *center* of that space. This leads to negative impact on performance during both the sampling step and the pruning step. In the sampling step, the results computed are more likely to lie at the center of the space, so the lower bound is not as tight as that of the other two datasets. And in the pruning step, most tuples lies near the center of the space, so the estimated upper bounds of the possible join matches' rank are also loose, especially when the number of attributes in rank function is large. Therefore, less tuples can be pruned before join matching.

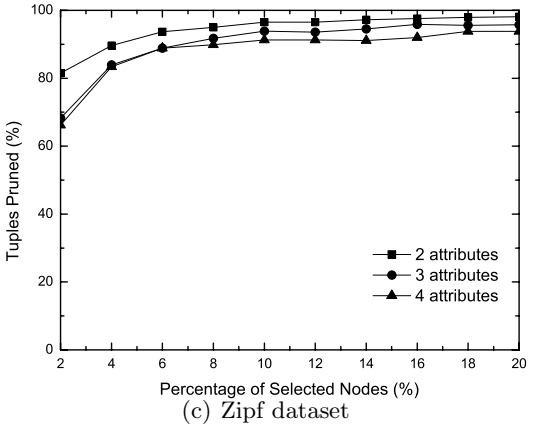
The impact of the number of total join results on Performance. Fig. 3 presents the performance changing of *PJoin* as the number of total join results varies. The experimental setting is similar to that of the last experiment, except that we fix the sampling percentage to 8%. Experiments are conducted over the datasets that generate different number of join results, which varies from 200,000



(a) uniform dataset

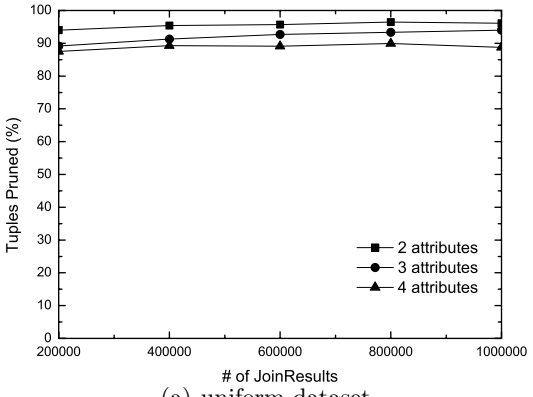


(b) Gaussian dataset

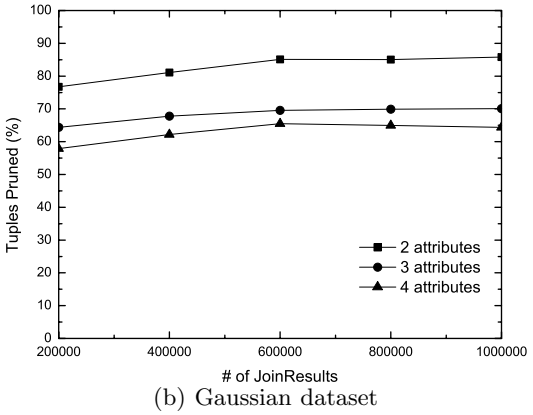


(c) Zipf dataset

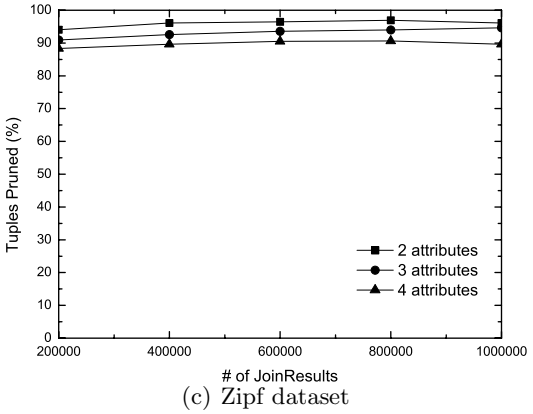
Fig. 2. *PJoin* performance *vs.* the number of selected nodes



(a) uniform dataset



(b) Gaussian dataset



(c) Zipf dataset

Fig. 3. *PJoin* performance *vs.* the number of join results

to 1,000,000. As the results show, the performance turns better slightly when the number of join results increases. However, the changing is very slow.

The impact of k value on Performance. Fig. 4 presents the performance changing of $PJoin$ as the value of k is adjusted over three datasets. The experiments runs over a P2P network of 10,000 nodes, two tables each with 1000000 tuples generate 500000 join results, and 8% nodes are selected. We increment k from 1 to 100 by the step of 10. The results show as k increases, the performance decreases slightly, and when k is large enough, its impact on performance becomes indistinguishable. We can see that the performance over Gaussian dataset decreases faster than those over the other two datasets. The underlying reason is the same as the number of selected nodes impacts performance over Gaussian dataset.

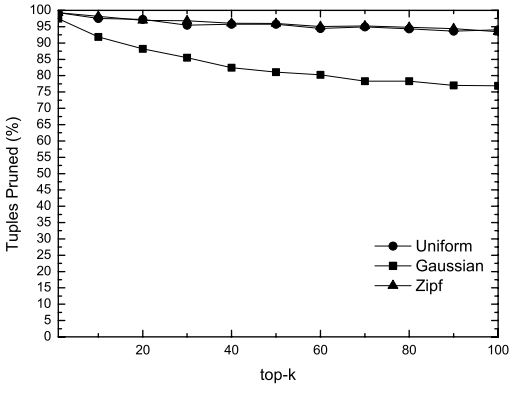


Fig. 4. The impact of k on performance of $PJoin$

7 Conclusion

We have presented new algorithms for efficiently supporting ranked queries in P2P networks. The objective of these algorithms proposed herein is to save the bandwidth when processing ranked queries. We first proposed the $PSel$ algorithm to answer ranked selection queries, which performs several orders of magnitude better than the straightforward method in our experiments. Then we proposed the $PJoin$ algorithm to answer ranked join queries. $PJoin$ selects the nodes in the P2P network to get a lower bound of the top-k results, and use this bound to prune the irrelevant tuples before join matching, thus saves bandwidth cost. More than 80 percent of tuples can be pruned in our experiments over uniform datasets when 6 percent of nodes are selected.

Our future work is mainly in two directions. On one hand, the algorithm for ranked join queries in this paper is based on the symmetric hash join algorithm.

It should be valuable to exploit the Bloom join algorithms for answering ranked join queries. On the other hand, we are planning to use cached query results to answer new user queries.

References

1. Gnutella Home Page. *http://www.gnutella.com/*.
2. W.-T. Balke, W. Nejdl, W. Siberski, and U. Thaden. Progressive distributed peer-to-peer top-k retrieval in peer-to-peer networks. In *Proceedings of ICDE'05*, pages 174–185, 2005.
3. N. Bruno, L. Gravano, and A. Marian. Evaluating top-k queries over web-accessible databases. In *Proceedings of ICDE'02*, pages 369–380, 2002.
4. P. Cao and Z. Wang. Efficient top-k query calculation in distributed networks. In *Proceedings of PODC'04*, pages 206–215, 2004.
5. M. J. Carey and D. Kossmann. On saying "enough already!" in sql. In *Proceedings of SIGMOD'97*, pages 219–230, 1997.
6. K. C. Chang and S. Hwang. Minimal probing: Supporting expensive predicates for top-k queries. In *Proceedings of SIGMOD'02*, pages 346–357, 2002.
7. S. Chaudhuri and L. Gravano. Evaluating top-k selection queries. In *Proceedings of VLDB'99*, pages 397–410, 1999.
8. R. Fagin. Combining fuzzy information from multiple systems (extended abstract). In *Proceedings of PODS'96*, pages 216–226, 1996.
9. R. Fagin, A. Lote, and M. Naor. Optimal aggregation algorithms for middleware. In *Proceedings of PODS'01 (Award talk)*, 2001.
10. S. Gribble, A. Haleby, Z. Ives, M. Rodrig, and D. Suciu. What can peer-to-peer do for database, and vice versa? In *Proceedings of WebDB'01*, 2001.
11. U. Guntzer, W.-T. Balke, and W. Kiebling. Optimizing multi-feature queries for image databases. In *Proceedings of VLDB'00*, pages 419–428, 2000.
12. V. Hristidis, N. Koudas, and Y. Papakonstantinou. Prefer: a system for the efficient execution of multi-parametric ranked queries. In *Proceedings of SIGMOD'01*, pages 259–270, 2001.
13. R. Huebsch, J. Hellerstein, N. Lanham, B. Loo, and S. Shenker. Querying the internet with PIER. In *Proceedings of VLDB'03*, pages 321–332, 2003.
14. I. Ilyas, W. Aref, and A. Elmagarmid. Joining ranked inputs in practice. In *Proceedings of VLDB'02*, pages 950–961, 2002.
15. I. Ilyas, W. Aref, and A. Elmagarmid. Supporting top-k join queries in relational databases. In *Proceedings of VLDB'03*, pages 754–765, 2003.
16. S. Michel, P. Triantafillou, and G. Weikum. Klee: A framework for distributed top-k query algorithms. In *Proceedings of VLDB'05*, pages 637–648, 2005.
17. D. S. Milojicic, V. Kalogeraki, R. Lukose, K. Nagaraja, J. Prunyne, B. Richard, S. Rollins, and Z. Xu. Peer-to-peer computing. Technical Report HPL-2002-57, HP Lab, 2002.
18. A. Natsev, Y.-C. Chang, J. R. Smith, C.-S. Li, and J. S. Vitter. Supporting incremental join queries on ranked inputs. In *Proceedings of VLDB'01*, pages 281–290, 2001.
19. S. Nepal and M. V. Ramakrishna. Query processing issues in image(multimedia) databases. In *Proceedings of ICDE'99*, pages 22–29, 1999.
20. S. Ratnasamy, P. Francis, M. Handley, R. Karp, and S. Schenker. A scalable content-addressable network. In *Proceedings of SIGCOMM'01*, 2001.

21. A. Rowstron and P. Druschel. Pastry: Scalable, decentralized object location, and routing for large-scale peer-to-peer systems. In *Proceedings of Middleware'01*, pages 329–350, 2001.
22. C. Schmidt and M. Parashar. Enabling flexible queries with guarantees in p2p systems. *Internet Computing*, 8:19–26, 2004.
23. I. Stoica, R. Morris, D. Karger, M. F. Kaashoek, and Hari Balakrishnan. Chord: A scalable peer-to-peer lookup service for internet applications. In *Proceedings of SIGCOMM'01*, 2001.
24. C. Tang, Z. Xu, and M. Mahalingam. Psearch: Information retrieval in structured overlays. In *Proceedings of HotNets-I*, 2002.
25. P. Tsaparas, T. Palpanas, Y. Kotidis, N. Koudas, and D. Srivastava. Ranked join indices. In *Proceedings of ICDE'03*, pages 277–290, 2003.
26. B. Y. Zhao, J. D. Kubiatowicz, and A. D. Joseph. Tapestry: An infrastructure for fault-tolerant wide-area location and routing. Technical report, UC Berkeley, 2001.
27. K. Zhao. Supporting ranked queries in peer-to-peer networks. *Master Degree Thesis of Fudan University*, 2005.

Author Index