Part VII

Workshop on Active Mining

Award-Winning Papers (Overview)

Katsumi Nitta

Tokyo Institute of Technology
nitta@dis.titech.ac.jp

On behalf of the program committee (PC) of JSAI 2003, I would like to thank all the chair persons, discussants, and attentive audience who contributed to select these awarded papers.

JSAI 2003 was held from 25 to 27 of June 2003, where 255 technical papers were presented in 42 sessions.

During JSAI 2003, at least two PC members participated in each session, and after the conference, the PC members and the session chair person nominated candidate papers for the award from each session by submitting reports which described the reasons of the nomination. As a result of the first stage, 28 papers were nominated. Then, as the second stage, PC members exchanged their opinions about evaluation of each paper by email, and finally they selected 5 papers for the award by voting. Among them, one author is declined of translating his paper in English, four papers are included in this special edition.

Followings are their overviews.

1. Bus Information System based on User Models and Dynamic Generation of VoiceXML Script by Shinichi Ueno, Fumihito Adachi, Kazunori Komatani, Tatsuya Kawahara, and Hiroshi G. Okuno:

This paper presents flexible dialogue management for various user utterance. Real dialogue data are collected by Kyoto city bus information system, and user models are construected based on the analysis. There are three categories of user models such as skill level, knowledge level of the task domain and the degree of hastiness to generate cooperative response to the user. These user models are constructed by decision tree learning technique. Features specific to spoken dialogue systems as well as semantic attributes are used in the classification. Experimental evaluation shows thet the cooperative responses adaptive to individual users serve as good guidance for novice users without increasing the dialogue durtion for the skilled users.

2. Efficient ALgorithms for Finding Frequent Substructures from Semi-structured Data Streams by Tatsuya Asai, Kenji Abe, Shinji Kawagoe, Hiroshi Arimura, and Setsuo Arikawa:

This paper presents efficient online data mining algorithms from streams of semi-structured data. Recently, new types of application programs on the network have been developed. These programs treat semi-structured data such as XML data. The semi-structured data are modeled as labeled ordered trees. The online algorithm, Stream T, receives fragments of an unseen possibly infinite semi-structured data in the document order through a data stream, and outputs the current set of frequent patterns on request at any time. These algorithms

A. Sakurai et al. (Eds.): JSAI 2003/2004, LNAI 3609, pp. 3–4, 2007.

are implemented in different online mining models and candidate management strategies.

3. Analysis of Heptitis Dataset by Decision Tree based on Graph-Based Induction by Warodom Geamsakul, Takashi Matsuda, Tetsuya Yoshida, Kouzou Ohara, Hiroshi Motoda, Takashi Washio and Katsuhiko Takabayashi:

This paper presents a technique of constructing data classifiers (decision trees) from graph-structured data inwhich attributes and values are not explicitly expressed. Though a decision tree is an effective means of data classification, it is not constructed for the data which is not explicitly expressed with attribute-value pairs. The proposed method, Decision Tree Graph-Based Induction (DT-GBI),generates attribute-value pairs from the graph-structured data by extracting typical patterns and then applies them to construct decision trees. DTGBI is applied to the classification task of hepatitis data, and shows the error rate is satisfactory.

4. Robotic Communication Terminals as a ubiquitous system for improving human mobility by making environment virtually barrier-free by Ikuko Eguchi Yairi, Kentaro Kawamura and Seiji IGI:

This paper presents an overview of Robotic Communication Terminals (RCT) project for disabled, elderly, or ill individuals. RCT consists of three types of terminals; Environment-embedded terminals, user-carried mobile terminals, and user carrying mobile terminals. These terminals communicate each other to connect real world, computer network and users, and support the three elementary behaviours of recognizing things, controlling motor functions, and accessing information, and assist the users to move around in the city.

Analysis of Hepatitis Dataset by Decision Tree Based on Graph-Based Induction

Warodom Geamsakul[1,⋆], Takashi Matsuda[1], Tetsuya Yoshida[1], Kouzou Ohara[1], Hiroshi Motoda[1], Takashi Washio[1], Hideto Yokoi[2], and Katsuhiko Takabayashi[2]

[1] Institute of Scientific and Industrial Research, Osaka University, Japan
{warodom,matsuda,yoshida,ohara,motoda,washio}@ar.sanken.osaka-u.ac.jp
[2] Division for Medical Informatics, Chiba University Hospital, Japan
yokoi@telemed.ho.chiba-u.ac.jp,takaba@ho.chiba-u.ac.jp

Abstract. A machine learning technique called Graph-Based Induction (GBI) efficiently extracts typical patterns from graph-structured data by stepwise pair expansion (pairwise chunking). It is very efficient because of its greedy search. We have expanded GBI to construct a decision tree that can handle graph-structured data. DT-GBI constructs a decision tree while simultaneously constructing attributes for classification using GBI. In DT-GBI attributes, namely substructures useful for classification task, are constructed by GBI on the fly during the tree construction. We applied both GBI and DT-GBI to classification tasks of a real world hepatitis data. Three classification problems were solved in five experiments. In the first 4 experiments, DT-GBI was applied to build decision trees to classify 1) cirrhosis and non-cirrhosis (Experiments 1 and 2), 2) type C and type B (Experiment 3), and 3) positive and negative responses of interferon therapy (Experiment 4). As the patterns extracted in these experiments are thought discriminative, in the last experiment (Experiment 5) GBI was applied to extract descriptive patterns for interferon therapy. The preliminary results of experiments, both constructed decision trees and their predictive accuracies as well as extracted patterns, are reported in this paper. Some of the patterns match domain experts' experience and the overall results are encouraging.

Keywords: Data mining, graph-structured data, Decision Tree Graph-Based Induction, hepatitis dataset analysis.

1 Introduction

Over the last few years there has been a number of research work on data mining in seeking for better performance. Better performance includes mining from structured data, which is a new challenge. Since structure is represented by proper relations and a graph can easily represent relations, knowledge discovery

⋆ Motoda lab., I.S.I.R., Osaka University 8-1 Mihogaoka, Ibaraki, Osaka 567-0047, Japan.

A. Sakurai et al. (Eds.): JSAI 2003/2004, LNAI 3609, pp. 5–28, 2007.

from graph-structured data poses a general problem for mining from structured data. Some examples amenable to graph mining are finding typical web browsing patterns, identifying typical substructures of chemical compounds, finding typical subsequences of DNA and discovering diagnostic rules from patient history records.

Graph-Based Induction (GBI) [13,3] is a technique which was devised for the purpose of discovering typical patterns in a general graph data by recursively chunking two adjoining nodes. It can handle a graph data having loops (including self-loops) with colored/uncolored nodes and links. There can be more than one link between any two nodes. GBI is very efficient because of its greedy search. GBI does not lose any information of graph structure after chunking, and it can use various evaluation functions in so far as they are based on frequency. It is not, however, suitable for graph-structured data where many nodes share the same label because of its greedy recursive chunking without backtracking, but it is still effective in extracting patterns from such graph-structured data where each node has a distinct label (*e.g.,* World Wide Web browsing data) or where some typical structures exist even if some nodes share the same labels (*e.g.,* chemical structure data containing benzene rings etc).

On the other hand, decision tree construction method [6,7] is a widely used technique for data classification and prediction. One of its advantages is that rules, which are easy to understand, can be induced. Nevertheless, as is the case in the majority of data mining methods, to construct a decision tree it is usually required that data is represented by attribute-value pairs beforehand. Since it is not trivial to define proper attributes for graph-structured data beforehand, it is difficult to construct a classifier represented by decision tree for the classification of graph-structured data.

We have proposed a method called Decision Tree Graph-Based Induction (DT-GBI), which constructs a classifier (decision tree) for graph-structured data while simultaneously constructing attributes themselves for classification using GBI [10,11]. A pair extracted by GBI, which consists of nodes and the links between them, is treated as an attribute and the existence/non-existence of the pair in a graph is treated as its value for the graph. Thus, attributes (pairs) that divide data effectively are extracted by GBI while a decision tree is being constructed. To classify graph-structured data after the construction of a decision tree, attributes are produced from data before the classification.

We applied both GBI and DT-GBI for the hepatitis data which was provided by Chiba University Hospital. Viral hepatitis is a very critical illness. If it is left without undergoing a suitable medical treatment, a patient may suffer from cirrhosis and fatal liver cancer. The progress speed of condition is slow and subjective symptoms are not noticed easily, hence, in many cases, it has already become very severe when they are noticed. Although periodical inspection and proper treatment are important in order to prevent this situation, there are problems of expensive cost and physical burden on a patient. Although there is an alternative much cheaper method of inspection (blood test), the amount of data becomes enormous since the progress speed of condition is slow. In our analysis temporal records in the provided dataset was converted into graph structured data with respect to time correlation so that both intra-correlation of individual inspection and inter-correlation among inspections are represented in graph-structured data. Three classification problems were solved in five experiments. Target classes are the stages of fibrosis in the first problem (Experiments 1 and 2), the types of hepatitis (B and C) in the second one (Experiment 3) and responses of interferon therapy in the third one (Experiment 4 and 5). In the first 4 experiments, DT-GBI was applied to build decision trees. Patterns extracted in these experiments are thought discriminative. In order to extract descriptive patterns, GBI was applied to interferon therapy problem in the last experiment (Experiment 5). The results of experiments are reported and discussed in this paper.

There are some other analyses already conducted and reported on this dataset. [12] analyzed the data by constructing decision trees from time-series data without discretizing numeric values. [2] proposed a method of temporal abstraction to handle time-series data, converted time phenomena to symbols and used a standard classifier. [9] used multi-scale matching to compare time-series data and clustered them using rough set theory. [5] also clustered the time-series data of a certain time interval into several categories and used a standard classifier. These analyses examine the temporal correlation of each inspection separately and do not explicitly consider the relations among inspections. Thus, these approaches do not correspond to the structured data analysis.

Section 2 briefly describes the framework of GBI. Section 3 explains our method called DT-GBI for constructing a classifier with GBI for graph-structured data and illustrates how decision tree is constructed with a simple example. Section 4 describes the details of how we applied GBI and DT-GBI for the hepatitis dataset and reports the results of experiments, both constructed decision trees and their predictive accuracies as well as extracted patterns. Section 5 concludes the paper with summary of the results and the planned future work.

2 Graph-Based Induction (GBI)

2.1 Principle of GBI

GBI employs the idea of extracting typical patterns by stepwise pair expansion as shown in Figure 1. In the original GBI an assumption is made that typical patterns represent some concepts/substructure and "typicality" is characterized by the pattern's frequency or the value of some evaluation function of its frequency. We can use statistical indexes as an evaluation function, such as frequency itself, Information Gain [6], Gain Ratio [7] and Gini Index [1], all of which are based on frequency. In Figure 1 the shaded pattern consisting of nodes 1, 2, and 3 is thought typical because it occurs three times in the graph. GBI first finds the 1→3 pairs based on its frequency, chunks them into a new node 10, then in the next iteration finds the 2→10 pairs, chunks them into a new node 11. The resulting node represents the shaded pattern.

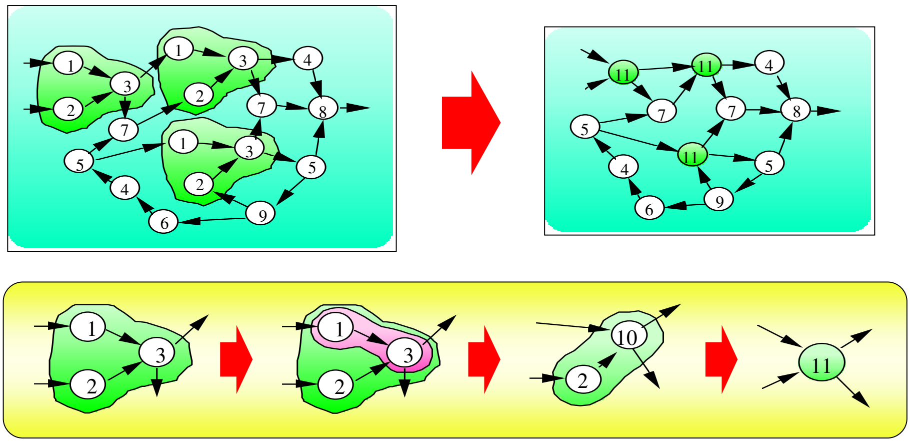


Fig. 1. The basic idea of the GBI method

It is possible to extract typical patterns of various sizes by repeating the above three steps. Note that the search is greedy. No backtracking is made. This means that in enumerating pairs no pattern which has been chunked into one node is restored to the original pattern. Because of this, all the "typical patterns" that exist in the input graph are not necessarily extracted. The problem of extracting all the isomorphic subgraphs is known to be NP-complete. Thus, GBI aims at extracting only meaningful typical patterns of a certain size. Its objective is not finding all the typical patterns nor finding all the frequent patterns.

As described earlier, GBI can use any criterion that is based on the frequency of paired nodes. However, for finding a pattern that is of interest any of its subpatterns must be of interest because of the nature of repeated chunking. In Figure 1 the pattern 1→3 must be typical for the pattern 2→10 to be typical. Said differently, unless pattern 1→3 is chunked, there is no way of finding the pattern 2→10. Frequency measure satisfies this monotonicity. However, if the criterion chosen does not satisfy this monotonicity, repeated chunking may not

GBI(G)
  Enumerate all the pairs P_{all} in G
  Select a subset P of pairs from P_{all} (all the pairs in G)
    based on typicality criterion
  Select a pair from P_{all} based on chunking criterion
  Chunk the selected pair into one node c
  G_c := contracted graph of G
  while termination condition not reached
    $P := P \cup \text{GBI}(G_c)$
  return P

Fig. 2. Algorithm of GBI

find good patterns even though the best pair based on the criterion is selected at each iteration. To resolve this issue GBI was improved to use two criteria, one for frequency measure for chunking and the other for finding discriminative patterns after chunking. The latter criterion does not necessarily hold monotonicity property. Any function that is discriminative can be used, such as Information Gain [6], Gain Ratio [7] and Gini Index [1], and some others.

The improved stepwise pair expansion algorithm is summarized in Figure 2. It repeats the following four steps until chunking threshold is reached (normally minimum support value is used as the stopping criterion).

Step 1. Extract all the pairs consisting of connected two nodes in the graph.

Step 2a. Select all the typical pairs based on the typicality criterion from among the pairs extracted in Step 1, rank them according to the criterion and register them as typical patterns. If either or both nodes of the selected pairs have already been rewritten (chunked), they are restored to the original patterns before registration.

Step 2b. Select the most frequent pair from among the pairs extracted in Step 1 and register it as the pattern to chunk. If either or both nodes of the selected pair have already been rewritten (chunked), they are restored to the original patterns before registration. Stop when there is no more pattern to chunk.

Step 3. Replace the selected pair in Step 2b with one node and assign a new label to it. Rewrite the graph by replacing all the occurrence of the selected pair with a node with the newly assigned label. Go back to Step 1.

The output of the improved GBI is a set of ranked typical patterns extracted at Step 2a. These patterns are typical in the sense that they are more discriminative than non-selected patterns in terms of the criterion used.

2.2 Beam-Wise Graph-Based Induction (B-GBI)

Since the search in GBI is greedy and no backtracking is made, which patterns extracted by GBI depends on which pair is selected for chunking. There can be many patterns which are not extracted by GBI. A beam search is incorporated to GBI, still, within the framework of greedy search [4] in order to relax this problem, increase the search space, and extract more discriminative patterns while still keeping the computational complexity within a tolerant level. A certain fixed numbers of pairs ranked from the top are selected to be chunked individually in parallel. To prevent each branch growing exponentially, the total numbers of pairs to chunk (the beam width) is fixed at every time of chunking. Thus, at any

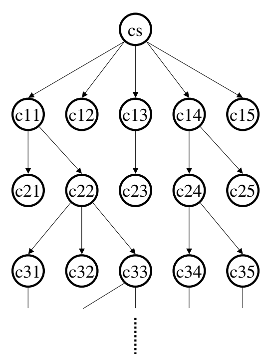


Fig. 3. Beam search in B-GBI (beam width = 5)

iteration step, there is always a fixed number of chunking that is performed in parallel.

Figure 3 shows how search is conducted in B-GBI when the beam width is set to five. First, five frequent pairs are selected from the graphs at the starting state in search (cs in Figure 3). Graphs in cs are then copied into the five states (c11 ∼ c15), and each of five pairs is chunked in the copied graphs at the respective state. At the second cycle in search, pairs in graphs are enumerated in each state and five frequent pairs are selected from all the sates. In this example, two pairs are selected from c11, one pair from c13, and two pairs from c14. At the third cycle in search, graphs in c11 are copied into c21 and c22, graphs in c13 are copied into c23, and graphs in c14 are copied into c24 and c25. As in the second cycle, the selected pairs are chunked in the copied graphs. The states without the selected pairs (in this example c12 and c15) are discarded.

3 Decision Tree Graph-Based Induction (DT-GBI)

3.1 Decision Tree for Graph-Structured Data

Since the representation of decision tree is easy to understand, it is often used as the representation of classifier for data which are expressed as attribute-value pairs. On the other hand, graph-structured data are usually expressed as nodes and links, and there is no obvious components which corresponds to attributes and their values. Thus, it is difficult to construct a decision tree for graph-structured data in a straight forward manner.

We formulate the construction of decision tree for graph-structured data by defining attributes and attribute-values as follows:

– attribute: a pattern/subgraph in graph-structured data
– value for an attribute: existence/non-existence of the pattern in a graph

Since the value for an attribute is yes (the classifying pattern exists) and no (the classifying pattern does not exist), the constructed decision tree is represented as a binary tree. Data (graphs) are divided into two groups, namely, the one with the pattern and the other without the pattern. The above process is summarized in Figure 4. One remaining question is how to determine classifying patterns which are used as attributes for graph-structured data. Our approach is described in the next subsection.

3.2 Feature Construction by GBI

In our Decision Tree Graph-Based Induction (DT-GBI) method, typical patterns are extracted using GBI[1] and used as attributes for classifying graph-structured data. When constructing a decision tree, all the pairs in data are enumerated and one pair is selected. The data (graphs) are divided into two groups, namely, the one with the pair and the other without the pair. The selected pair is then chunked

[1] We use GBI to mean B-GBI.

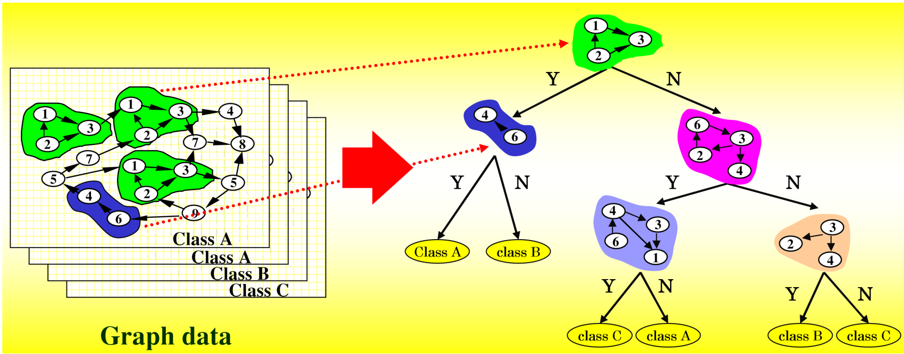


Fig. 4. Decision tree for classifying graph-structured data

DT-GBI(D)
  Create a node DT for D
  if termination condition reached
    return DT
  else
    P := GBI(D) (with the number of chunking specified)
    Select a pair p from P
    Divide D into D_y (with p) and D_n (without p)
    Chunk the pair p into one node c
    D_{yc} := contracted data of D_y
    for D_i := D_{yc}, D_n
      DT_i := DT-GBI(D_i)
      Augment DT by attaching DT_i as its child along yes(no) branch
  return DT

Fig. 5. Algorithm of DT-GBI

in the former graphs and these graphs are rewritten by replacing all the occurrence of the selected pair with a new node. This process is recursively applied at each node of a decision tree and a decision tree is constructed while attributes (pairs) for classification task is created on the fly. The algorithm of DT-GBI is summarized in Figure 5. As shown in Figure 5, the number of chunking applied at one node is specified as a parameter. For instance, when it is set to 10, chunking is applied for 10 times to construct a set of pairs P, which consists of all the pairs by applying chunking $1 \sim 10$ times. Note that the pair p selected from P is not necessarily constructed by applying chunking at depth 10.

DT-GBI has the characteristic of constructing the attributes (pairs) for classification task on-line while constructing a decision tree. Each time when an attribute (pair) is selected to divide the data, the pair is chunked into a larger node in size. Thus, although initial pairs consist of two nodes and the link between them, attributes useful for classification task are gradually grown up into larger pair (subgraphs) by applying chunking recursively. In this sense the proposed

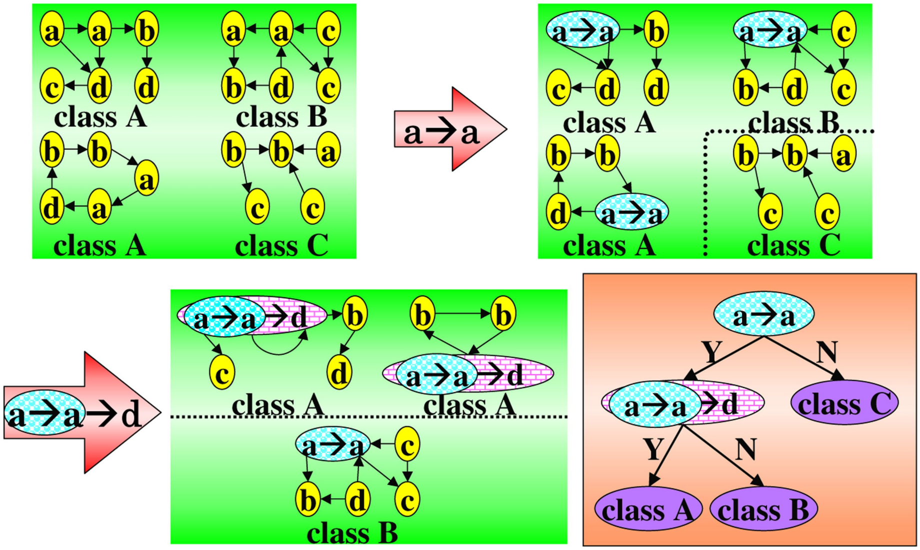


Fig. 6. Example of decision tree construction by DT-GBI

DT-GBI method can be conceived as a method for feature construction, since features, namely attributes (pairs) useful for classification task, are constructed during the application of DT-GBI.

Note that the criterion for chunking and the criterion for selecting a classifying pair can be different. In the following experiments, frequency is used as the evaluation function for chunking, and information gain is used as the evaluation function for selecting a classifying pair.

3.3 Working Example of DT-GBI

Suppose DT-GBI receives a set of 4 graphs in the upper left-hand side of Figure 6. The number of chunking applied at each node is set to 1 to simplify the working of DT-GBI in the following example. By enumerating all the pairs in these graphs, 13 kinds of pair are extracted from the data. These pairs are: a→a, a→b, a→c, a→d, b→a, b→b, b→c, b→d, c→b, c→c, d→a, d→b, d→c. The existence/non-existence of the pairs in each graph is converted into the ordinal table representation of attribute-value pairs, as shown in Figure 7. For instance, for the pair a→a, graph 1, graph2 and graph 3 have the pair but graph 4 does not have. This is shown in the first column in Figure 7.

Next, the pair with the highest evaluation for classification (e.g., information gain) is selected and used to divide the data into two groups at the root node. In this example, a pair “a→a” is selected. The selected pair is then chunked in graph 1, graph 2 and graph 3 and these graphs are rewritten. On the other hand, graph 4 is left as it is.

Graph	a→a	a→b	a→c	a→d	b→a	b→b	b→c	b→d	c→b	c→c	d→a	d→b	d→c
1 (class A)	1	1	0	1	0	0	0	1	0	0	0	0	1
2 (class B)	1	1	1	0	0	0	0	0	0	1	1	1	0
3 (class A)	1	0	0	1	1	1	0	0	0	0	0	1	0
4 (class C)	0	1	0	0	0	1	1	0	1	0	0	0	0

Fig. 7. Attribute-value pairs at the first step

Graph	a→a→b	a→a→c	a→a→d	b→a→a	b→a	···	d→c
1 (class A)	1	0	1	0	0	···	1
2 (class B)	1	1	0	0	0	···	0
3 (class A)	0	0	1	1	1	···	0

Fig. 8. Attribute-value pairs at the second step

The above process is applied recursively at each node to grow up the decision tree while constructing the attributes (pairs) useful for classification task at the same time. Pairs in graph 1, graph 2 and graph 3 are enumerated and the attribute-value tables is constructed as shown in Figure 8. After selecting the pair "(a→a)→d", the graphs are separated into two partitions, ech of which contains graphs of of a single class. The constructed decision tree is shown in the lower right-hand side of Figure 6.

3.4 Pruning Decision Tree

Recursive partitioning of data until each subset in the partition contains data of a single class often results in overfitting to the training data and thus degrades the predictive accuracy of decision trees. To avoid overfitting, in our previous approach [10] a very naive *prepruning* method was used by setting the termination condition in DT-GBI in Figure 5 to whether the number of graphs in D is equal to or less than 10. On the other hand, a more sophisticated *postpruning* method, is used in C4.5 [7] (which is called "pessimistic pruning") by growing an overfitted tree first and then pruning it to improve predictive accuracy based on the confidence interval for binomial distribution. To improve predictive accuracy, pessimistic pruning in C4.5 is incorporated into the DT-GBI by adding a step for postpruning in Figure 5.

4 Analysis of Hepatitis Data

The hepatitis data set provided by Chiba University Hospital contains long time-series data (from 1982 to 2001) on laboratory examinations of 771 patients of hepatitis B and C. The data can be broadly split into two categories. The first

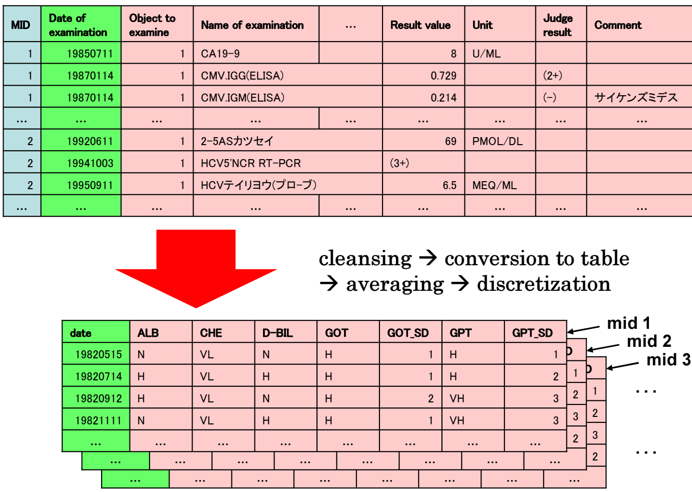


Fig. 9. Example of graph conversion in phase 1-2

category includes administrative information such as patient's information (age and date of birth), pathological classification of the disease, date of biopsy, and result. The second category includes temporal record of blood test and urinalysis. It contains the results of 983 types of both in- and out-hospital examinations.

To apply DT-GBI, we use the same two criteria for selecting pairs. One is frequency for selecting pairs to chunk, and the other is information gain [6] for finding discriminative patterns after chunking.

4.1 Data Preprocessing

In phase 1, a new reduced data set is generated because the data of visit is not synchronized across different patients and the progress of hepatitis is considered slow. The data set provided is cleansed [2], and the numeric attributes are averaged over two-month interval and for some of them, standard deviations are calculated over six month interval and added as new attributes. Numerical average is taken for numeric attributes and maximum frequent value is used for nominal attributes over the interval. Further, numerical values are discretized when the normal ranges are given. In case there are no data in the interval, these are treated as missing values and no attempt is made to estimate these values. At the end of this phase, reduced data is divided into several files so that each file contains the data of each patient.

In phase 2, data in the range from 500 days before to 500 days after the first biopsy of each patient were converted into a graph. Here, the date of first biopsy and the result, to be treated as class [3], of each patient are searched from the

[2] Letters and symbols such as H, L, +, or - are deleted from numeric attributes.
[3] Activity, progress of fibrosis, hepatitis type, etc. can be taken as class.

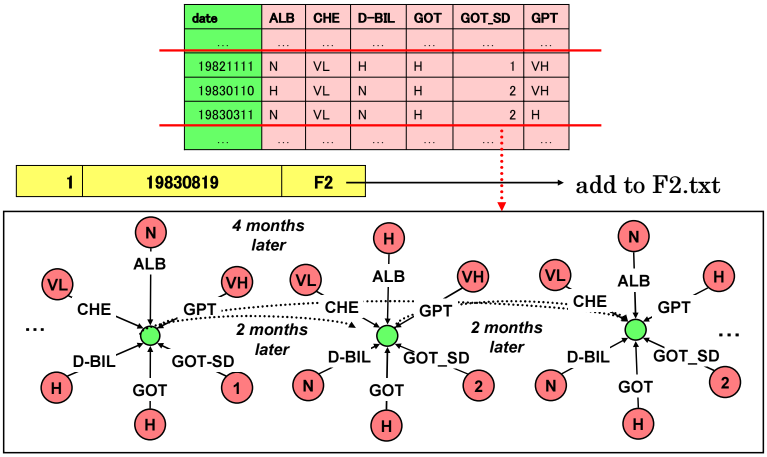


Fig. 10. An example of graph conversion in phase 3

biopsy data file. In case that the result of the second biopsy or after differs from the result of the first one, the result from the first biopsy is defined as the class of that patient for the entire 1,000-day time-series. Figure 9 illustrates these two phases.

In the last phase of data preparation, one patient record is mapped into one directed graph as illustrated in Figure 10. Assumption is made that there is no direct correlation between two sets of pathological tests that are more than a predefined interval (here, two years) apart. Hence, time correlation is considered only within this interval. Figure 10 shows an example of conversion of data to graph. In this figure, a star-shaped subgraph represents values of a set of pathological examination in the two-month interval. The center node of the subgraph is a hypothetical node for the two-month interval. An edge pointing to a hypothetical node represents an examination. The node connected to the edge represents the value (processed result) of the examination. And the edge linking two hypothetical nodes represents time difference.

4.2 Classifying Patients with Fibrosis Stages

Fibrosis stages are categorized into five stages: F0 (normal), F1, F2, F3, and F4 (severe). We constructed decision trees which distinguish the patients at F4 stage from the patients at the other stages. In the following two experiments, we used 32 attributes. These attributes are: ALB, CHE, D-BIL, GOT, GOT_SD, GPT, GPT_SD, HBC-AB, HBE-AB, HBE-AG, HBS-AB, HBS-AG, HCT, HCV-AB, HCV-RNA, HGB, I-BIL, ICG-15, MCH, MCHC, MCV, PLT, PT, RBC, T-BIL, T-CHO, TP, TTT, TTT_SD, WBC, ZTT, and ZTT_SD. Table 1 shows the size of graphs after the data conversion.

As shown in Table 1, the number of instances (graphs) in cirrhosis (F4) class is 43 while the number of instances (graphs) in non-cirrhosis ({F0+F1+F2+F3})

Table 1. Size of graphs (classified by fibrosis stage)

Stage	F0	F1	F2	F3	F4	Total
No. of graphs	4	125	53	37	43	262
Avg. No. of node	303	304	308	293	300	303
Max. No. of node	349	441	420	414	429	441
Min. No. of node	254	152	184	182	162	152

Table 2. Average error rates (%) in exp. 1 and 2

	Experiment 1		Experiment 2	
run	N_r=20	N_e=20	N_r=20	N_e=20
1	14.81	11.11	27.78	25.00
2	13.89	11.11	26.85	25.93
3	15.74	12.03	25.00	19.44
4	16.67	15.74	27.78	26.68
5	16.67	12.96	25.00	22.22
6	15.74	14.81	23.15	21.30
7	12.96	9.26	29.63	25.93
8	17.59	15.74	25.93	22.22
9	12.96	11.11	27.78	21.30
10	12.96	11.1	27.78	25.00
average	**15.00**	**12.50**	**26.67**	**23.52**
SD	**1.65**	**2.12**	**1.80**	**2.39**

class is 219. Imbalance in the number of instances may cause a biased decision tree. In order to relax this problem, we limited the number of instances to the 2:3 (cirrhosis:non-cirrhosis) ratio which is the same as in [8]. Thus, we used all instances from F4 stage for cirrhosis class and select 65 instances from the other stages for non-cirrhosis class, 108 instances in all. How we selected these 108 instances is describe later.

A decision tree was constructed in either of the following two ways: 1) apply chunking N_r=20 times at the root node and only once at the other nodes of a decision tree, 2) apply chunking N_e=20 times at every node of a decision tree. Note that n_r and n_e are defined along the depth in Figure 3. Thus, there is more chunking taking place during the search when the beam width is larger. The pair (subgraph) that is selected for each node of the decision tree is the one which maximizes the information gain among all the pairs that are enumerated. Decision tree pruning is conducted by postpruning: conduct pessimistic pruning by setting the confidence level to 25%.

Experiment 1: F4 Stage Vs {F0+F1} Stages. All four instances in F0 and 61 instances in F1 stage for non-cirrhosis class were used in this experiment. We performed 10 runs of 10 fold cross-validation. To determine an optimal beam width, we first conducted 9-fold cross-validation for one randomly chosen 9 folds

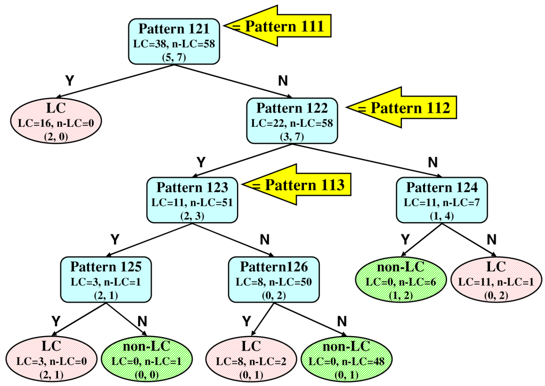


Fig. 11. One of trees from the best run in exp.1 (N_e=20)

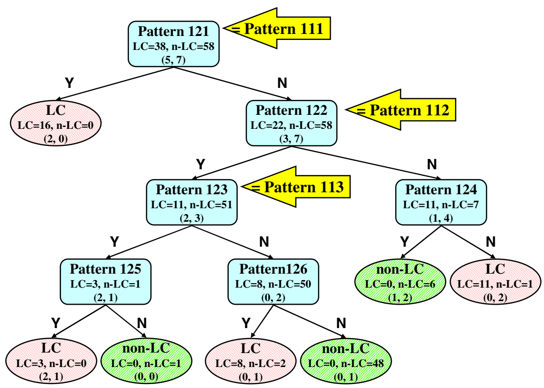


Fig. 12. One of trees from the worst run in exp.1 (N_e=20)

data (90% of all data) of the first run of 10-fold cross-validation varying the beam width from 1 to 15. Consequently, since we observed that for both methods (N_r=20 and N_e=20) the narrowest beam width that brings to the lowest error rate was 15 where the error rate was levelled off, we set the beam width to 15 for this experiment. In the following experiments 2, 3, and 4, we determined the beam width in the same manner.

The overall result is summarized in the left half of Table 2, in which SD means the standard deviation of the average error rates of the 10 runs of 10-fold cross validation. The average error rate was 15.00% for 1) (N_r=20) and 12.50% for 2) (N_e=20). Figure 11 and Figure 12 show one of the decision trees each from the run with the lowest error rate (run 7) and from the run with the highest error rate (run 8) respectively. Comparing the both decision trees, there are three pairs of identical patterns appeared at the upper level of each tree.

Experiment 2: F4 Stage vs {F3+F2} Stages. In this experiment, we used all instances in F3 and 28 instances in F2 stage for non-cirrhosis class. As in

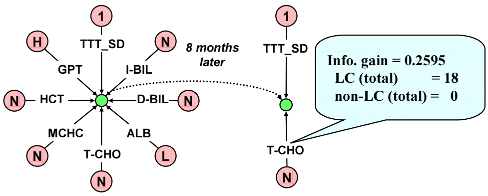


Fig. 13. Pattern 111 = Pattern 121 (if exist then LC)

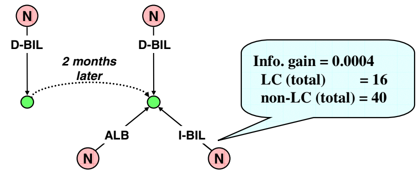


Fig. 14. Pattern 112 = Pattern 122

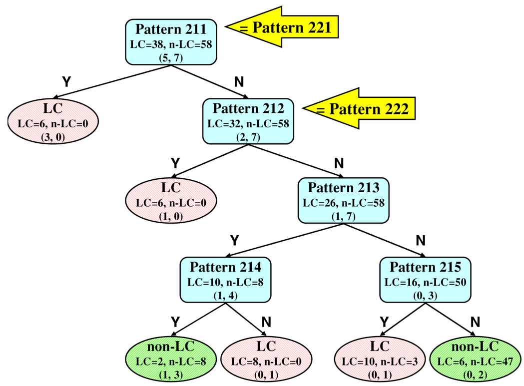


Fig. 15. One of trees from the best run in exp.2 (N_e=20)

experiment 1, we performed 10 runs of 10 fold cross-validation. The beam width was set to 14. The overall result is summarized in the right half of Table 2. The average error rate was 26.67% for 1) (N_r=20) and 23.52% for 2) (N_e=20). Figure 15 and Figure 16 show examples of decision trees each from the run with the lowest error rate (run 3) and the run with the highest error rate (run 4) respectively. Comparing the both decision trees, there are two pairs of identical patterns appeared at the upper level of each tree.

Discussion. The average prediction error rate in the first experiment is better than that in the second experiment, as the difference in characteristics between data in F4 stage and data in {F0+F1} stages is intuitively larger than that between data in F4 stage and data in {F3+F2}. The averaged error rate of 12.50% in experiment 1 is fairly comparable to that of 11.8% obtained by the decision tree reported in [8].

Patterns shown in Figure 13, 14, 17, and 18 are sufficiently discriminative since all of them are used at the nodes in the upper level of all decision trees. The certainty of these patterns is ensured as, for almost all patients, they appear after the biopsy.

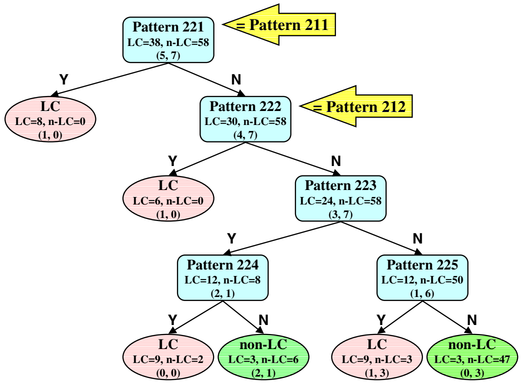


Fig. 16. One of trees from the worst run in exp.2 (N_e=20)

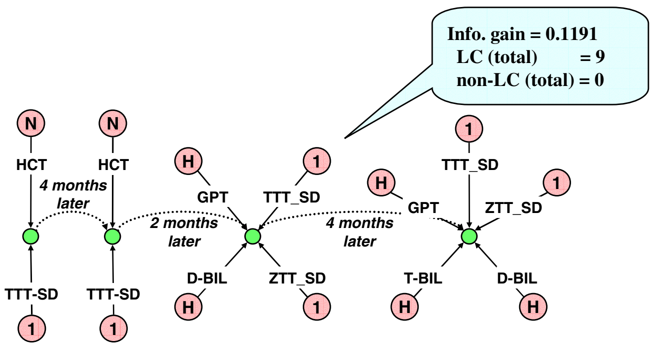


Fig. 17. Pattern 211 = Pattern 221 (if exist then LC)

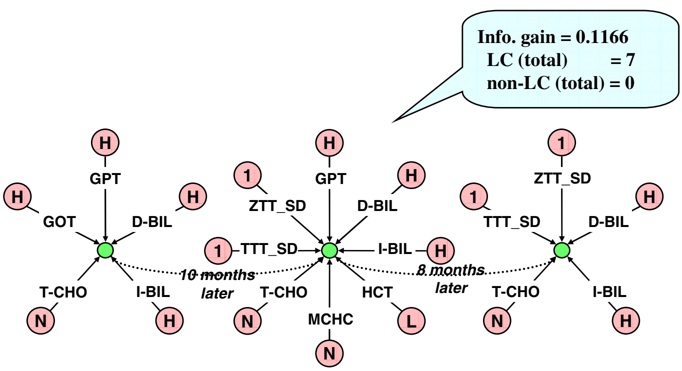


Fig. 18. Pattern 212 = Pattern 222

date	ALB	D-BIL	GPT	HCT	I-BIL	MCHC	T-CHO	TTT_SD	…
19930517	L	N	H	N	N	N	N	1	…
19930716	L	L	H	N	N	N	N		…
19930914	L	L	H	N	N	N	N	1	…
19931113	L	N	H	N	N	N	N		…
19940112	L	L	H	N	N	N	N	1	…
19940313	L	N	N	N	N	N	N	1	…
19940512	L	N	H	N	N	N	N	1	…
19940711	L	N	H	N	N	N	N	1	…
19940909	L	L	H	N	N	N	N	1	…
19941108	L	N	N	N	N	N	N	1	…
19950107	L	N	N	L	N	N	N	1	…
19950308	L	N	N	N	N	N	N	1	…
19950507	L	N	H	N	N	N	N	1	…
19950706	L	N	N	L	N	N	N	1	…
19950904	L	L	N	L	N	L	N	1	…
19951103	L	L	N	N	N	N	N	1	…

Fig. 19. Data of No.203 patient

These patterns may appear only once or several times in one patient. Figure 19 shows the data of a patient for whom pattern 111 exists. As we did no attempt to estimate missing values, the pattern was not counted even if the value of only one attribute is missing. At data in the Figure 19, pattern 111 would have been counted four if the value of TTT_SD in the fourth line had been "1" instead of missing.

4.3 Experiment 3: Classifying Patients with Types (B or C)

There are two types of hepatitis recorded in the data set; B and C. We constructed decision trees which distinguish between patients of type B and type C. The attributes of antigen and antibody (HBC-AB, HBE-AB, HBE-AG, HBS-AB, HBS-AG, HCV-AB, HCV-RNA) were not included as they obviously indicate the type of hepatitis. Table 3 shows the size of graphs after the data conversion. To keep the number of instances at 2:3 ratio, we used all of 77 instances in type B as "Type B" class and 116 instances in type C as "Type C" class. Hence, there are 193 instances in all. In this experiment, the beam width was set to 5.

The overall result is summarized in Table 4. The average error rate was 23.21% for 1) (N_r=20) and 20.31% for 2) (N_e=20). Figure 20 and Figure 21 show a sample of decision trees from the run with the lowest error rate (run 1) and the run with the highest error rate (run 6) respectively. Comparing the both decision trees, two patterns (shown in Figure 22 and 23) were identical and used at the upper level nodes. There patterns also appeared at almost all the decision trees and thus are considered to be sufficiently discriminative.

4.4 Experiment 4: Classifying Interferon Therapy by DT-GBI

An interferon is a medicine to get rid of the hepatitis virus and it is said that the smaller the amount of virus is, the more effective interferon therapy is. Unfortu-

Table 3. Size of graphs (classified by type)

Stage	Type B	Type C	Total
No. of graphs	77	185	262
Avg. No. of node	238	286	272
Max. No. of node	375	377	377
Min. No. of node	150	167	150

Table 4. Average error rates (%) in exp. 3

run	Experiment 3	
	N_r=20	N_e=20
1	21.76	18.65
2	21.24	19.69
3	21.24	19.17
4	23.32	20.73
5	25.39	22.80
6	25.39	23.32
7	22.28	18.65
8	24.87	19.17
9	22.80	19.69
10	23.83	21.24
average	**23.21**	**20.31**
SD	**1.53**	**1.57**

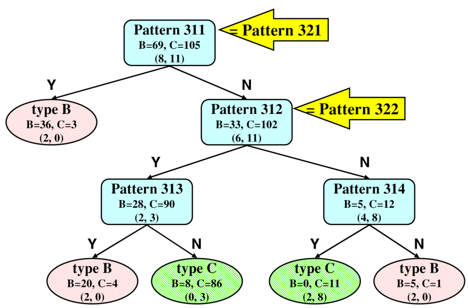


Fig. 20. One of trees from the best run in exp.3 (N_e=20)

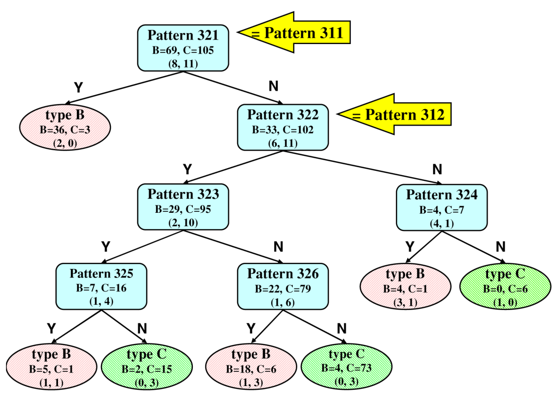


Fig. 21. One of trees from the worst run in exp.3 (N_e=20)

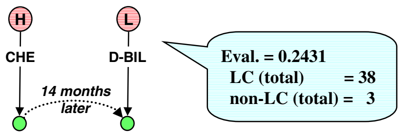


Fig. 22. Pattern 311 = Pattern 321 (if exist then type B)

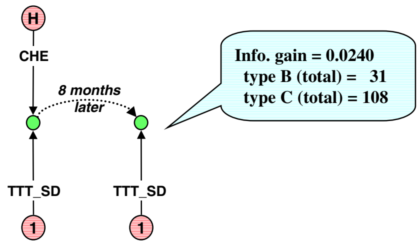


Fig. 23. Pattern 312 = Pattern 322

nately, the dataset provided by Chiba University Hospital does not contain the examination record for the amount of virus since it is expensive. However, it is believed that experts (medical doctors) decide when to administer an interferon by estimating the amount of virus from the results of other pathological examinations. Response to interferon therapy was judged by a medical doctor for each patient, which was used as the class label for interferon therapy. The class labels specified by the doctor for interferon therapy are summarized in Table 5. Note that the following experiments (Experiment 4) were conducted for the patients with label R (38 patients) and N (56 patients). Medical records for other patients were not used.

To analyze the effectiveness of interferon therapy, we hypothesized that the amount of virus in a patient was almost stable for a certain duration just before the interferon injection in the dataset. Data in the range of 90 days to 1 day before the administration of interferon were extracted for each patient and average was taken for two-week interval. Furthermore, we hypothesized that each pathological condition in the extracted data could directly affect the pathological condition just before the administration. To represent this dependency, each subgraph was directly linked to the last subgraph in each patient. An example of converted graph-structured data is shown in Figure 24.

As in subsection 4.2 and subsection 4.3, feature selection was conducted to reduce the number of attributes. Since the objective of this analysis is to predict the effectiveness of interferon therapy without referring to the amount of virus,

Table 5. Class label for interferon therapy

label	
R	virus disappeared (Response)
N	virus existed (Non-response)
?	no clue for virus activity
R?	R (not fully confirmed)
N?	N (not fully confirmed)
??	missing

Table 6. Size of graphs (interferon therapy)

effectiveness of interferon therapy	R	N	Total
No. of graphs	38	56	94
Avg. No. of nodes	77	74	75
Max. No. of nodes	123	121	123
Min. No. of nodes	41	33	33

Table 7. Average error rate (%) (interferon therapy)

run	n_e=20
1	18.75
2	23.96
3	20.83
4	20.83
5	21.88
6	22.92
7	26.04
8	23.96
9	23.96
10	22.92
average	**22.60**
SD	**1.90**

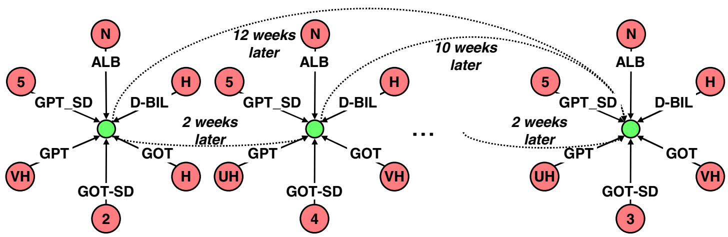


Fig. 24. An example of graph-structured data for the analysis of interferon therapy

the attributes of antigen and antibody (HBC-AB, HBE-AB, HBE-AG, HBS-AB, HBS-AG, HCV-AB, HCV-RNA) were not included. Thus, as in subsection 4.3 we used the following 25 attributes: ALB, CHE, D-BIL, GOT, GOT_SD, GPT, GPT_SD, HCT, HGB, I-BIL, ICG-15, MCH, MCHC, MCV, PLT, PT, RBC, T-BIL, T-CHO, TP, TTT, TTT_SD, WBC, ZTT, and ZTT_SD. Table 6 shows the size of graphs after the data conversion. The beam width was set to 3 in experiment 4.

The results are summarized in Table 7 and the overall average error rate was 22.60%. Two examples of decision trees selected out of 100 constructed are shown in Figure 25 and Figure 26. Figure 25 is selected from the decision trees with the best prediction accuracy and Figure 26 is selected from the ones with time-correlated patterns, prediction accuracy being about the average of 100. Patterns at the upper nodes in these trees are shown in Figures 27, 28, 29, 30. Although the structure of decision tree in Figure 25 is simple, its prediction accuracy was actually good (error rate=10%). Furthermore, since the pattern shown in Figure 27 was used at the root node of many decision trees, it is considered sufficiently discriminative for classifying patients for whom interferon therapy was effective (with class label R).

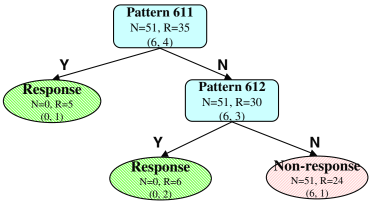


Fig. 25. One of trees from the best run in exp.4

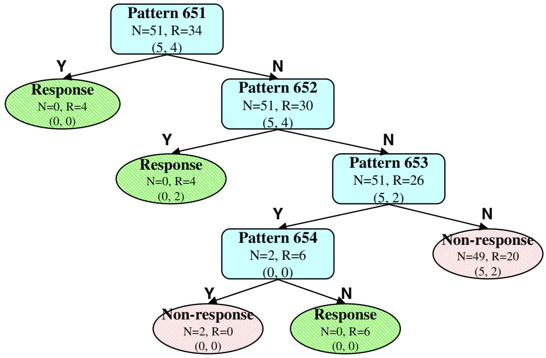


Fig. 26. One of trees with time-correlated patterns in exp. 4

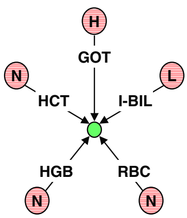


Fig. 27. Pattern 611 (if exist, then R)

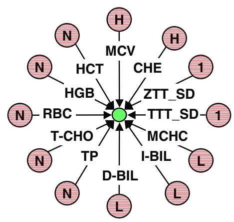


Fig. 28. Pattern 612 (if exist, then R)

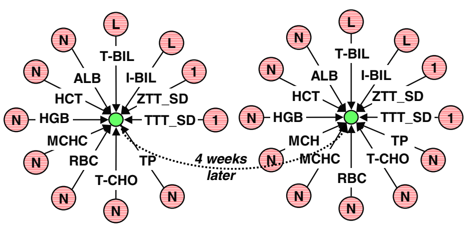


Fig. 29. Pattern 651 (if exist, then R)

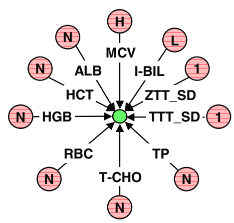


Fig. 30. Pattern 652 (if exist, then R)

Unfortunately, only several patterns contain time interval edges such as the one shown in Figure 29 and it was unable to investigate how the change or stability of blood test will affect the effectiveness of interferon therapy.

4.5 Experiment 5: Analysis of Interferon Therapy by B-GBI

We applied B-GBI to extract descriptive patterns from the patients analyzed in subsection 4.4. B-GBI was terminated when the support of all the extracted

Table 8. Extracted patterns

	94 Graphs in Table 6				41 Graphs with patterns in Figures 31,32,33			
	only R	only N	common	Total	only R	only N	common	Total
No. of patterns	2604	3468	5467	11539	1392	4587	3678	9657
Max. No. of nodes per pattern	35	38	38	38	39	42	39	42
Ave. No. of nodes per pattern	15.4	15.0	10.9	13.1	15.9	16.3	10.8	14.1

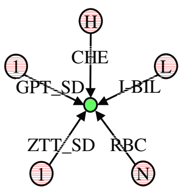


Fig. 31. Pattern a1

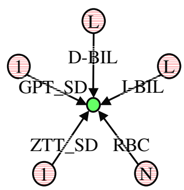


Fig. 32. Pattern a2

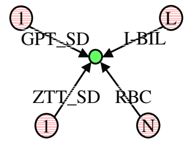


Fig. 33. Pattern a3

patterns [4] became less than 0.1. The beam width was set to 3 as in subsection 4.4. The extracted patterns were divided into 3 groups: 1) patterns included only in the patients with class label R, 2) those with class label N, and 3) those with both label R N (these groups are called only R, only N, common, respectively). The number and size of the extracted patterns from the graphs in Table 6 are summarized in the left-hand side of Table 8.

The decision tree constructed by DT-GBI is rather unbalanced as shown in Figure 26. This is because the patterns with large discriminative power (information gain) have a relatively small support. Small support means that these patterns are specific to some data and does not have sufficient generalization capability.

We, therefore, searched for the patterns by B-GBI in terms of not only the discriminative power but also the support. Patterns which were included both in the data with R and N were sorted in descending order of information gain to reflect their discriminative power. Extracted patterns with the largest information gain are shown in Figure 31, Figure 32 and Figure 33. These patterns are included in 10 patients with label R and 31 patients with label N.

The patients having these three patterns were further analyzed by B-GBI and the extracted patterns were also divided into 3 groups as before. The number and size of the extracted patterns from these graphs are summarized in the right-hand side of Table 8. Examples of extracted patterns with a large information gain are shown in Figure 34 and Figure 35. (actually the pattern in Figure 34

[4] The support of a pattern is defined as the number of graphs with the pattern divided by the total number of graphs.

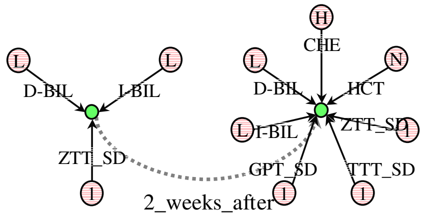


Fig. 34. Pattern r1 (in 4 patients out of 10 with R)

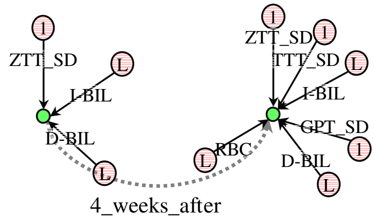


Fig. 35. Pattern n9 (in 8 patients out of 31 with N)

had the largest information gain). Although it was difficult to extract patterns with time interval edges by DT-GBI, these patterns contain a time interval edge and still have sufficient discriminative power in the filtered data by the patterns in Figure 31,32 and 33.

Extracted patterns are still hard to interprete as sufficiently descriptive. One encouraging observation is that the value of HGB might be some clue, because the results show that HBG is N (normal) in all the patterns with class label R but it is L (low) in patterns with class label N. Thus, investigating the effect of HGB is a future direction for the analysis of interferon therapy by B-GBI.

5 Conclusion

This paper explained a method called DT-GBI, which constructs a classifier (decision tree) for graph-structured data and reported the preliminary results of analyzing the hepatitis data set from Chiba University Hospital by using B-GBI and DT-GBI. DT-GBI constructs attributes, namely substructures useful for classification task, by applying repeated chunking in B-GBI on the fly while constructing a decision tree.

Three classification problems were solved in five experiments. Target classes are the stages of fibrosis in the first problem (Experiments 1 and 2), the types of hepatitis (B and C) in the second one (Experiment 3) and responses of interferon therapy in the third one (Experiment 4 and 5). In the first 4 experiments, DT-GBI was applied to build decision trees. Constructed decision trees were simple and the patterns extracted in these experiments are indeed discriminative. In the last experiment (Experiment 5) B-GBI was applied to interferon therapy problem in order to extract descriptive patterns.

Both B-GBI and DT-GBI were able to extract patterns that reflect both intra-correlation of individual inspection and inter-correlation among inspections at different time points. All of these patterns are not necessarily interpretable by domain experts, but some of them match their experience. The obtained prediction error rate results are thought satisfactory considering the various noises that are embedded in the data. We thus believe that B-GBI and DT-GBI are useful tools for practicing evidence-based medicine.

Immediate future work includes to incorporate more sophisticated method for determining the number of cycles to call B-GBI at each node to improve prediction accuracy. Utilizing the rate of change of information gain by successive chunking is a possible way to automatically determine the number. Effectiveness of B-GBI and DT-GBI against the hepatitis data set with another way of preparing data should be examined, *e.g.,* estimating missing values, randomly selecting instances from non-cirrhosis class both for training and testing, etc. The validity of extracted patterns is now being carefully evaluated and discussed by the domain experts (medical doctors).

Acknowledgment

This work was partially supported by the grant-in-aid for scientific research on priority area "Active Mining" (No. 13131101, No. 13131206) funded by the Japanese Ministry of Education, Culture, Sport, Science and Technology.

References

1. L. Breiman, J. H. Friedman, R. A. Olshen, and C. J. Stone. *Classification and Regression Trees*. Wadsworth & Brooks/Cole Advanced Books & Software, 1984.
2. T. B. Ho, T. D. Nguyen, S. Kawasaki, S.Q. Le, D. D. Nguyen, H. Yokoi, and K. Takabayashi. Mining hepatitis data with temporal abstraction. In *Proc. of the 9th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 369–377, August 2003.
3. T. Matsuda, T. Horiuchi, H. Motoda, and T. Washio. Extension of graph-based induction for general graph structured data. In *Knowledge Discovery and Data Mining: Current Issues and New Applications (Springer Verlag LNAI 1805)*, pages 420–431, 2000.
4. T. Matsuda, T. Yoshida, H. Motoda, and T. Washio. Knowledge discovery from structured data by beam-wise graph-based induction. In *Proc. of the 7th Pacific Rim International Confernce on Artificial Intelligence (Springer Verlag LNAI2417)*, pages 255–264, 2002.
5. M. Ohsaki, Y. Sato, S. Kitaguchi, and T. Yamaguchi. A rule discovery support system. In *Project "Realization of Active Mining in the Era of Information Flood" Report*, pages 147–152, March 2003.
6. J. R. Quinlan. Induction of decision trees. *Machine Learning*, 1:81–106, 1986.
7. J. R. Quinlan. *C4.5:Programs For Machine Learning*. Morgan Kaufmann Publishers, 1993.
8. E. Suzuki, T. Watanabe, Y. Yamada, F. Takeuchi, Y. Choki, K. Nakamoto, S. Inatani, N. Yamaguchi, M. Nagahama, H. Yokoi, and K. Takabayashi. Toward spiral exception discovery. In *Project "Realization of Active Mining in the Era of Information Flood" Report*, pages 153–160, March 2003.
9. S. Tsumoto, K. Takabayashi, M. Nagira, and S. Hirano. Trend-evaluation multiscale analysis of the hepatitis dataset. In *Project "Realization of Active Mining in the Era of Information Flood" Report*, pages 191–197, March 2003.

10. G. Warodom, T. Matsuda, T. Yoshida, H. Motoda, and T. Washio. Classifier construction by graph-based induction for graph-structured data. In *Proc. of the 7th Pacific-Asia Conference on Knowledge Discovery and Data Mining (Springer Verlag LNAI2637)*, pages 52–62, 2003.
11. G. Warodom, T. Matsuda, T. Yoshida, H. Motoda, and T. Washio. Performance evaluation of decision tree graph-based induction. In *Proc. of the 6th Pacific-Asia Conference on Discovery Science (Springer Verlag LNAI2843)*, pages 128–140, 2003.
12. Y. Yamada, E. Suzuki, H. Yokoi, and K. Takabayashi. Decision-tree induction from time-series data based on a standard-example split test. In *Proc. of the 12th International Conference on Machine Learning*, pages 840–847, August 2003.
13. K. Yoshida and H. Motoda. Clip : Concept learning from inference pattern. *Journal of Artificial Intelligence*, 75(1):63–92, 1995.

Efficient Algorithms for Finding Frequent Substructures from Semi-structured Data Streams

Tatsuya Asai*, Kenji Abe**, Shinji Kawasoe***, Hiroki Arimura†, and Setsuo Arikawa

Kyushu University
6–10–1 Hakozaki Higashi-ku, Fukuoka 812–8581, Japan
{t-asai,k-abe,s-kawa,arim,arikawa}@i.kyushu-u.ac.jp

Abstract. In this paper, we study an online data mining problem from streams of semi-structured data such as XML data. Modeling semi-structured data and patterns as labeled ordered trees, we present an online algorithm StreamT that receives fragments of an unseen possibly infinite semi-structured data in the document order through a data stream, and can return the current set of frequent patterns immediately on request at any time. We give modifications of the algorithm to other online mining models. Furthermore we implement our algorithms in different online models and candidate management strategies, then show empirical analyses to evaluate the algorithms.

1 Introduction

Recently, a new class of data-intensive applications such as network monitoring, web site management, sensor networks, and e-commerce emerged with the rapid growth of network and web technologies. In these applications, the data are modeled not as static collections but as transient *data streams*, where the data source is an unbounded stream of individual data items, e.g., transaction records or web page visits, which may arrive continuously in rapid, time-varying way [19].

Particularly in data communication through internet, it is becoming popular to use *semi-structured data*-based communication technologies [2], e.g., SOAP [20], to send heterogeneous and ill-structured data through networks. Since traditional database technologies are not directly applicable to such data streams, it is important to study efficient information extraction methods for semi-structured data streams.

In this paper, we model such semi-structured data streams by sequences of *labeled ordered trees*, and study the frequent pattern discovery problem in online setting. We model a semi-structured data stream as an infinite sequence of the

* Presently, He is working at Fujitsu Laboratories Ltd.
** Presently, He is working at Sharp Corp.
*** Presently, He is working at NTT Comware Corp.
† Presently, He is working at Hokkaido University.

A. Sakurai et al. (Eds.): JSAI 2003/2004, LNAI 3609, pp. 29–45, 2007.

nodes generated by the depth-first scanning of a possibly infinite data tree. An online algorithm has to continuously work on the data stream, and has to quickly answer queries on request based on the portion of the data received so far. This formulation captures typical situations for web applications reading a sequence of XML tags or SAX events element by element from a data stream. Since this is a finest-grained online model, the results of this paper can be easily generalized to coarser-grained models where, e.g., XML documents are processed page by page.

We present an online algorithm StreamT for discovering labeled ordered trees with frequency at least a given minimum threshold from unbounded data streams [5]. A difficulty lies in that we have to continuously work with unbounded data streams using only bounded resources. A key idea is a technique of sweeping a branch, called the *sweep branch*, over the whole virtual data tree to find all embeddings of candidate patterns intersecting it. As another idea, we adopt a candidate management policy similar to Hidber [11] for online association mining to limit the number of candidate patterns as small as possible. We discuss several strategies in managing candidate patterns.

We also use the enumeration technique for labeled ordered trees that we recently proposed in [4], a generalization of a technique by Bayardo [7]. Combining these ideas, our algorithm StreamT works efficiently in both time and space complexities in online manner. Furthermore, we extend our algorithm to the *sliding window model* and the *forgetting model* of online data stream mining, where the effect of a past data item decaysin its age. Finally, we run experiments on real XML data in different online strategies and candidate management strategies, to evaluate the efficiency of our algorithms and to compare the performance of the algorithms in several strategies.

The rest of this paper is organized as follows. In Section 2, we give basic notions and definitions. In Section 3, we present our online algorithm StreamT. In Section 4, we discuss management of candidate patterns. In Section 5, we modify this algorithm in the sliding window model and the forgetting model. In Section 6, we report experimental results, and in Section 7, we conclude.

1.1 Related Works

Emerging technologies of semi-structured data have attracted wide attention of networks, e-commerce, information retrieval and databases [2,20]. In contrast, there have not been many studies on *semi-structured data mining* [1,4,6,9,15,16,21,23]. There are a body of researches on online data processing and mining [10,14,19]. Most related work is Hidber [11], who proposed a model of continuous pattern discovery from unbounded data stream, and presented adaptive online algorithm for mining association rules. Parthasarathy *et al.* [18] and Mannila *et al.* [14] studied mining of sequential patterns and episode patterns. Yamanishi *et al.* [22] presented an efficient online-outlier detection system *SmartSifter* with a forgetting mechanism.

Zaki [23] and Asai *et al.* [4] independently developed efficient pattern search techniques, called *rightmost expansion*, for semi-structured data, which is a generalization of the set-enumeration tree technique [7]. Although our algorithm partly uses this technique, its design principle is different from previous semi-structured data mining algorithms [4,6,9,15,16,21,23].

2 Preliminaries

2.1 Model of Semi-structured Data

Semi-structured data are heterogeneous collections of weakly structured data [2], which are typically encoded in a markup language such as XML [20]. We model semi-structured data and patterns [2] by labeled ordered trees. For the basic terminologies on trees, we refer to, e.g. [3].

Labeled Ordered Trees. We define labeled ordered trees according to [4,12]. Let $\mathcal{L} = \{\ell, \ell_0, \ell_1, \ldots\}$ be a fixed alphabet of *labels*. Then, a *labeled ordered tree on* $\mathcal{L}$ (*tree*, for short) is a labeled, rooted, connected directed acyclic graph $T = (V, E, B, L, v_0)$ with the following properties [3]. Each node $v \in V$ of T is labeled by an element $L(v)$ of $\mathcal{L}$, and all node but the root v_0 have the unique parent by the child relation $E \subseteq V^2$. For each node v, its children are ordered from left to right by an indirect sibling relation $B \subseteq V^2$ [3]. Note that the term *ordered* means the order *not* on labels but on children.

The size of a tree T is the number of its nodes $|T| = |V|$. Throughout this paper, we assume that a tree of size k has the node set $V = \{1, \ldots, k\}$ and the nodes are ordered in the pre-order by the depth-first search order on T. We refer to the node i as the *i-th node* of T. This assumption is crucial in our discussion. By this assumption, the root and the rightmost leaf of T, denoted by $root(T)$ and $rml(T)$, are always 1 and k, respectively. For a tree T of size k, the *rightmost branch* of T, denoted by $RMB(T)$, is the path from the root 1 to the rightmost leaf k of T.

We denote by $\mathcal{T}$ the class of all labeled ordered trees on $\mathcal{L}$. We also refer to V, E, B and L as V_T, E_T, B_T and L_T, respectively, if it is clear from context.

Matching and Occurrences. Next, we define the notion of *matching* between two labeled ordered trees T and D. A *pattern tree* T *matches* a *data tree* D if T can be embedded in D with preserving the labels, the (direct) child relation, the (indirect) sibling relation by a non-collapsing mapping, that is, there exists some function $\varphi : V_T \to V_D$ that satisfies the following (i)–(iv) for any $v, v_1, v_2 \in V_T$:

(i) φ is one-to-one.
(ii) $L_T(v) = L_D(\varphi(v))$.
(iii) $(v_1, v_2) \in E_T$ iff $(\varphi(v_1), \varphi(v_2)) \in E_D$.
(iv) $(v_1, v_2) \in B_T$ iff $(\varphi(v_1), \varphi(v_2)) \in B_D$.

Then, we say that φ is a *matching function of* T *to* D, or T *occurs in* D. We assume the *empty tree* $\perp$ such that $|\perp| = 0$ and $\perp$ matches to any tree at any node. An *embedding* of T in D w.r.t. φ is the image $\varphi(T) \subseteq V_D$ of T into D, whose induced subgraph is isomorphic to T. We define the *root occurrence* and the *rightmost leaf occurrence* of T in D w.r.t. φ by the nodes $\varphi(1)$ and $\varphi(k)$ of

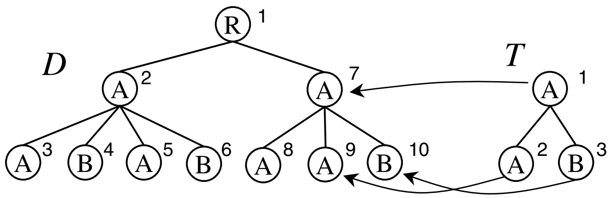


Fig. 1. A data tree D and a pattern tree T on the set $\mathcal{L} = \{A, B\}$ of labels

D to which the root and the rightmost leaf of T map, respectively. If φ is not irrelevant then we simply omit φ.

For example, Fig. 1 shows examples of labeled ordered trees D and T on $\mathcal{L} = \{A, B\}$. We see that the pattern tree T matches the data tree D, where the matching is designated with a set of arrows from T to D. The root occurrences of T in D are 2 and 7, while the rightmost occurrences are $4, 6$, and 10.

Semi-structured Data Streams. Let D be a labeled ordered tree, called a *data tree* with finite depth and possibly infinite width. Given a collection of trees as a data source, we always treat them as a single tree by combining trees with appending the imaginary common root. Recall that the nodes of D are numbered in the preorder traversal of D.

We introduce a convenient sequential representation of labeled ordered trees. The *depth* of node v of tree T is the number of edges on the path from the root to v. The *depth-label representation* of a node v of D is the pair $(d, \ell) \in \mathbf{N} \times \mathcal{L}$ of the depth d and the label ℓ of v. Then, a data tree D is encoded as the sequence $\pi = ((d_1, \ell_1), (d_2, \ell_2), \ldots)$ of depth-label pairs corresponding to the nodes on the pre-order traversal of T. This depth-label representation π also linearly related to the *open-close parentheses* representation as in XML [20].

Conversely, we can uniquely decode a depth-label representation π into a labeled ordered tree as follows.

Definition 1. **([4,17,23])** *Let S be a tree of size $k \geq 1$. Then, a* rightmost expansion *of S is any tree T of size $k+1$ obtained from S by (i) attaching a new node w with a label in $\mathcal{L}$ as a child of a parent node p on the rightmost branch of S so that (ii) w is the rightmost sibling of p. Then, we say that T is a* successor *of S, or S is a* predecessor *of T. If the depth and the label of w is $d \in \mathbf{N}$ and $\ell \in \mathcal{L}$, resp., then T is called the (d, ℓ)*-expansion *of S. The $(0, \ell)$-expansion of $\perp$ is the single node tree with label ℓ.*

Thus, the empty sequence ε transforms to the empty tree $\perp$, and if the sequence π transforms to a tree S, then the sequence $\pi \cdot (d, \ell)$ to the (d, ℓ)-expansion of S. The notion of depth-label representation is motivated by the tree expansion technique [4,17,23], and plays an important role in the following discussion.

For example, in the previous example of Fig. 1, the data tree D transforms to the depth-label representation $\pi = (0, R), (1, A), (2, A), (2, B), (2, A), (2, B),$ $(1, A)$, $(2, A), (2, A), (2, B)$, and *vice versa*.

We model a semi-structured data stream as an infinite sequence of the nodes generated by the depth-first scanning of a possibly infinite data tree as follows. For a set A, we denote by A^∞ the sets of all infinite sequences on A.A *semi-structured data stream* for D is an infinite sequence $\mathcal{S} = (\boldsymbol{v}_1, \boldsymbol{v}_2, \ldots, \boldsymbol{v}_i, \ldots) \in (\mathbf{N} \times \mathcal{L})^\infty$, where for every $i \geq 1$, the i-th element $\boldsymbol{v}_i = (d_i, \ell_i)$ is the depth-label representation of the i-th node $v_i = i$ of D. Then, $\boldsymbol{v}_i$ is called the i-th *node* of $\mathcal{S}$ and i is called the *time stamp*. The *i-th left-half tree*, denoted by D_i, is the labeled ordered tree that is the induced subgraph of D consisting of the first i nodes $(v_1, v_2, \ldots, v_i)$ of the traversal.

Online Data Mining Problems. Now, our online data mining problem is stated as follows. The definition of the frequency of a pattern T at time i will be specified later.

Definition 2.(Online Frequent Pattern Discovery from Semi-structured Data Streams) *Let $0 \leq \sigma \leq 1$ be a nonnegative number called the* minimum support. *In our online mining protocol, for stages $i = 1, 2, \ldots$, an* online mining algorithm *$\mathcal{A}$ iterates the following process: $\mathcal{A}$ receives the i-th node $\boldsymbol{v}_i$ from the stream $\mathcal{S}$, updates its internal state based on the first i nodes $\boldsymbol{v}_1, \ldots, \boldsymbol{v}_i$ received so far, and then on request by a user $\mathcal{A}$ reports a set $\mathcal{F}_i$ of frequent patterns that appears in D_i with frequency no less than σ.*

The goal of an online algorithm is to continuously work on unbounded stream for arbitrary long time with bounded resources, and to quickly answer user's queries at any time.

We define the models of the frequency of patterns as follows. Let $i \geq 1$ be any time. For every $1 \leq j \leq i$, we define the indicator function $hit_j^{(i)}(T) = 1$ if the pattern T has a root occurrence at the node v_j in D_i. Otherwise, we define $hit_j^{(i)}(T) = 0$. For a technical reason, we require not only $\varphi(1)$ but also the whole $\varphi(T)$ to be contained in D_i.

Definition 3. *Let $\mathcal{S}$ be a given semi-structured data stream and $T \in \mathcal{T}$ be any pattern. Below, $count_i(T)$ and $freq_i(T)$ denote the* count *and the* frequency *of T at time i, resp.*

- **Basic Online Model (BO).** *In this model motivated by Hidber [11], we count the number of distinct root occurrences of T in D_i. The* frequency *of T at time i is:*

$$freq_i(T) \;=\; \tfrac{1}{i} count_i(T) \;=\; \tfrac{1}{i} \textstyle\sum_{j=1}^{i} \, hit_j^{(i)}(T) \qquad (1)$$

- **Sliding Window Model (SW).** *Let $w \geq 1$. In this model motivated by Mannila* et al. *[14], we count the number of the distinct root occurrences in the i-th* sliding window *$W_i = (\boldsymbol{v}_{i-w+1}, \ldots, \boldsymbol{v}_i)$ of most recent w nodes of $\mathcal{S}$. If we define $\delta_k^w = 1$ if $0 \leq k < w$ and $\delta_k^w = 0$ otherwise, then:*

$$freq_{w,i}^{sw}(T) = \tfrac{1}{w} count_{w,i}^{\rm sw}(T) = \tfrac{1}{w} \textstyle\sum_{j=1}^{i} \, \delta_{i-j}^{w} \, hit_j^{(i)}(T). \qquad (2)$$

Algorithm StreamT

Input: A set $\mathcal{L}$ of labels, a data stream $(\boldsymbol{v}_1, \boldsymbol{v}_2, \ldots, \boldsymbol{v}_i, \ldots)$ of a data tree D, and a frequency threshold $0 \leq \sigma \leq 1$.

Output: A sequence $(\mathcal{F}_1, \mathcal{F}_2, \ldots, \mathcal{F}_i, \ldots)$ of sets of frequent patterns, where $\mathcal{F}_i$ is the set of frequent patterns for every i.

Variables: The candidate pool $\mathcal{C} \subseteq \mathcal{T}$, and the bucket stack $B = (B[0], \ldots, B[Top])$.

Method:

1. $\mathcal{C}$:= the class of all single node patterns;
   $B := \emptyset$ and $Top = -1$; $i := 1$;
2. While there is the next node $\boldsymbol{v}_i = (d, \ell)$, do the followings:
   (a) Update the bucket stack $B[0] \cdots B[d-1]$:
   $(B, EXP) :=$ UpdateB$(B, \mathcal{C}, (d, \ell), i)$;
   (b) Update the candidate pool $\mathcal{C}$ and the bucket $B[d]$:
   $(B, \mathcal{C}) :=$ UpdateC$(EXP, B, \mathcal{C}, (d, \ell), i)$;
   (c) Output the set $\mathcal{F}_i = \{\, T \in \mathcal{C} \mid freq_i(T) \geq \sigma \,\}$ of frequent patterns; $i = i + 1$;

Fig. 2. An online mining algorithm for semi-structured data stream

– **Forgetting Model (FG).** *In the forgetting model, e.g., [22], the contribution of the past event decays exponentially fast. For positive number* $0 < \gamma < 1$ *called a* forgetting factor*, the* frequency *of* T *is defined by:*

$$freq^{fg}_{\gamma,i}(T) = \tfrac{1}{Z_i} \sum_{j=1}^{i} \gamma^{i-j}\, hit^{(i)}_j(T). \quad (3)$$

Although we used a simplified normalization factor $Z_i = i$ *instead of a more precise one* $Z_i = \sum_{j=1}^{i} \gamma^{i-j}$*, most of the discussion in the later sections also holds.*

A difference of above models is the speed of decay. Since the effect of a past event decays exponentially faster in FG than in BO, the former is more trend conscious than the latter. Also, FG is easier to implement than SW.

3 Online Mining Algorithms

In this section, we present an online algorithm StreamT for solving the online frequent pattern discovery problem from semi-structured data stream.

3.1 Overview of the Algorithm

In Fig. 2, we show our algorithm StreamT in the online model. Let $\mathcal{S} = (\boldsymbol{v}_1, \boldsymbol{v}_2, \ldots, \boldsymbol{v}_i, \ldots) \in (\mathbf{N} \times \mathcal{L})^\infty$ be a possibly infinite data stream for a data tree D. Through the stages $i = 1, 2, \ldots$, StreamT receives the i-th node $\boldsymbol{v}_i = (d_i, \ell_i)$ from $\mathcal{S}$, updates a pool $\mathcal{C} \subseteq \mathcal{T}$ of candidate patterns and the internal state, and on request reports a set of frequent labeled ordered trees $\mathcal{F}_i \subseteq \mathcal{T}$ with frequency no less than a given threshold $0 \leq \sigma \leq 1$.

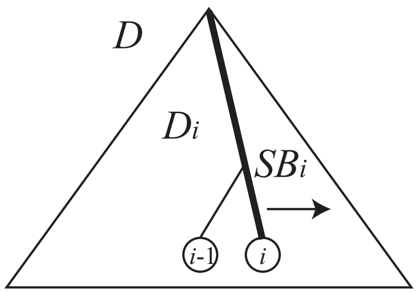


Fig. 3. The i-th left-half tree D_i and i-th sweep branch SB_i for the data tree D

To continuously compute the set of frequent patterns on an unbounded stream, the algorithm uses a technique, similar to *plane sweeping* in computational geometry [8], to find all root occurrences of candidate patterns in D. A basic idea of our algorithm is explained as follows. To detect all embeddings of a set of patterns in D, we sweep a path from the root to the currently scanned node v_i, called the *sweep branch*, rightwards over the data tree D by increasing the stage $i = 1, 2, \ldots$. While we sweep the plane, we keep track of all embedding of patterns that intersect the current sweep branch.

The algorithm incrementally maintains the following data structures during the computation.

– A set $\mathcal{C} \subseteq \mathcal{T}$ of patterns, called the *candidate pool*.
– A stack $B = (B[0], B[1], \ldots, B[Top])$ of buckets, called the *sweep branch stack* (*SB-stack*, for short).

For each candidate $T \in \mathcal{C}$, the following features are associated: A counter $count(T)$ of the root occurrences of T in D_i. A vector $Rto_T = (Rto_T[0], Rto_T[1], \ldots)$ of the latest root occurrences $Rto_T[d] = \rho$ of T with depth d.

3.2 Incremental Pattern Discovery Using Tree Sweeping

To keep track of all embeddings of candidate patterns, we do not need the whole information on them. Instead, we record the information on the intersections of these embedding and the current sweep branch at every stage.

Let $i \geq 1$ be any stage. In what follows, we denote by v_i, D_i and SB_i the present node, the left-half tree and the sweep branch at stage i. In other words, SB_i is the rightmost branch of D_i.

For pattern T, let $\varphi(T)$ be an embedding of T with some matching $\varphi : V_T \to V_D$ of T to D_i. Since an embedding of a tree is also an ordered tree in D_i, we can define the rightmost branch, denoted by $RMB(\varphi(T)) \subseteq V_D$, of $\varphi(T)$ in D_i. During the scan of D, the sweep branch $SB_i \subseteq V_D$ may have nonempty intersection $SB_i \cap RMB(\varphi(T))$ with $RMB(\varphi(T))$.

Lemma 1. *For any embedding $\varphi(T)$ of a pattern T and the sweep branch SB_i, the intersection $SB_i \cap RMB(\varphi(T))$ is a consecutive path in D.*

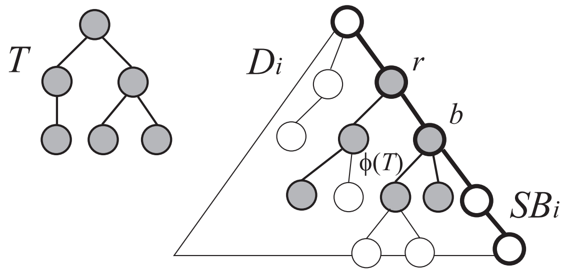


Fig. 4. The root and the bottom occurrences r and b of pattern T on D_i w.r.t the sweep branch SB_i with matching ϕ

From the above lemma, we define the *root* and the *bottom occurrences* of T w.r.t. φ to be the highest and the lowest nodes in the intersection $SB_i \cap RMB(\varphi(T))$ (Fig. 4). We can easily see that if the intersection $SB_i \cap RMB(\varphi(T))$ is contained in SB_i then the corresponding bottom occurrence becomes the rightmost occurrence of T w.r.t. φ. The next lemma gives an incremental characterization of the rightmost occurrences, which enables us to detect all rightmost occurrences of candidate patterns by maintaining all bottom occurrences of their predecessors on the sweep branch using the SB-stack.

Lemma 2. *Let $T \in \mathcal{T}$ be any pattern of size $k > 1$. At every time $i \geq 1$, T has a rightmost occurrence at the current node v_i in D_i iff there exists some pattern S of size $(k-1)$ that has a bottom occurrence at the parent of v_i in D_i and such that T is the (d, ℓ)-expansion of S, where d is the depth of the rightmost leaf k of T from its root and $\ell = L(v_i)$ is the label of v_i. This is also true even if $k = 1$.*

To implement this idea, we use the sweep branch stack $B = (B[0], B[1], \ldots, B[Top])$ to record the intersections of embeddings of patterns with the current sweep branch SB_i. $Top \geq 0$ is the length of SB_i. Each bucket $B[b]$ $(0 \leq b \leq Top)$ contains a set of triples of the form $\tau = (T, r, b)$ such that pattern T has the root and the bottom occurrences of the depths r and b, respectively, on SB_i. For each bucket $B[d]$, the time stamp $B[d].time \in \mathbf{N}$ of the last time is associated with the bucket.

For any stage $i \geq 1$, the SB-stack $B = (B[0], B[1], \ldots, B[Top])$ is said to be *up-to-date w.r.t.* the present sweep branch SB_i if Top is the length of SB_i, and for every $0 \leq b \leq Top$, the bucket $B[b]$ contains all triples $(T, r, b) \in \mathcal{T} \times \mathbf{N} \times \mathbf{N}$ for some $T \in \mathcal{C}$ and $r \in \mathbf{N}$ such that the pattern T appears in D_i and has the root and the bottom occurrences on SB_i of the depths r and b, respectively (Fig. 4). Then, we also say that each bucket $B[b]$ is up-to-date if no confusion arises. Note that the up-to-date stack is unique at time i. Now, we give a characterization of the contents of the current SB-stack B_i inductively.

Fig. 5 illustrates how to update the sweep branch stack B_{i-1}[5]. Suppose that we receive the i-th node (d, ℓ) from a data stream. Then, the triples in UNCHANGE buckets, i.e., $B[0] \cup \cdots \cup B[d-1]$, stay unchanged. The buckets in

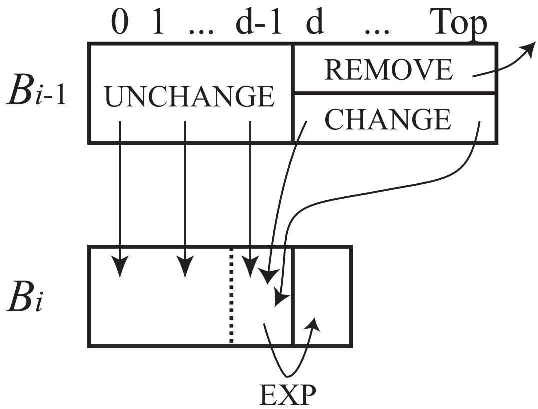


Fig. 5. SB-stacks from time $i-1$ to i

$B[d] \cup \cdots \cup B[Top]$ are partitioned into REMOVE and CHANGE. The triples in REMOVE buckets are discarded, and triples in CHANGE buckets move to the bucket $B[d-1]$. For all triples in $B[d-1]$, we apply the rightmost expansion and then insert obtained expansions into EXP.

For every time invoked in the while loop in StreamT of Fig. 2 at time $i \geq 1$ with the current node $\boldsymbol{v}_i = (d, l)$, the algorithm UpdateB$(B, \mathcal{C}, (d, \ell), i)$ returns the followings [5]: (i) The sequence $B[0], \ldots, B[d-1]$ of buckets that are up-to-date at time i up to depth $d-1$. (ii) The set EXP of all triples corresponding to the bottom occurrences on SB_i whose depth is d and predecessors belong to $\mathcal{C}$. For further details, see our previous paper [5].

4 Candidate Management

The algorithm StreamT stores candidate patterns in a candidate pool $\mathcal{C} \subseteq \mathcal{T}$. Fig. 6 shows an algorithm CandMan for managing $\mathcal{C}$ by updating the frequency count of each patterns. A root occurrence has monotonicity that if pattern T is a rightmost most expansion of pattern S then the root count of S is greater than or equal to the root count of T. Based on this observation, the algorithm CandMan uses a candidate management policy similar to Hidber [11], which is summarized as follows.

- **Initialize.** We insert all single node patterns into $\mathcal{C}$. This is done at Step 1 of the algorithm StreamT.
- **Increment.** We increment $count(T)$ for all pattern trees $T \in \mathcal{C}$ that has the rightmost occurrence at the current node v_i, i.e., $count(T) = count(T) + 1$.
- **Insert.** We insert a pattern $T \notin \mathcal{C}$ into $\mathcal{C}$ if T is detected at the current time i and the predecessor S of T is frequent. In Subsection 4.1, we describe the insert phase in detail.

Algorithm CandMan($CUR, \mathcal{C}, i$)
Input: A set CUR of patterns detected at time i, a candidate pool $\mathcal{C}$,
and the current time i;
Output: The updated candidate pool $\mathcal{C}$;
Method:
1. /* *Increment candidates* */
   For each pattern $T \in CUR \cap \mathcal{C}$, do:
   – If $T \in \mathcal{C}$ then $count(T) := count(T) + 1$;
2. /* *Insert candidates* */
   For each pattern $T \in CUR \backslash \mathcal{C}$, do:
   – If the predecessor of T is frequent at time i and $i > Lk^2$ holds, then $count(T) := init(T)$ and $\mathcal{C} := \mathcal{C} \cup \{T\}$;
3. /* *Delete candidates* */
   For each pattern $T \in \mathcal{C}$, do:
   – If the predecessor of T is infrequent at time i, then $\mathcal{C} = \mathcal{C} \backslash \{T\}$;
4. Return $\mathcal{C}$;

Fig. 6. An algorithm for updating the candidate pool

– **Delete.** We delete a pattern T from $\mathcal{C}$ when its unique predecessor S becomes infrequent. To be consistent to the initialization and the insertion policy, we do not delete any single nodes.

As summary, our algorithm StreamT tries to maintain the *negative border* [18], the set of all patterns that are infrequent but whose predecessors are frequent. The subprocedure UpdateC of StreamT in Fig. 2, which updates a candidate pool $\mathcal{C}$ and the d-th bucket $B[d]$, are constructed similarly to CandMan. For details of UpdateC, please consult [5].

4.1 Insert Candidates

As described above, we insert a pattern $T \notin \mathcal{C}$ into $\mathcal{C}$ if T is detected at the current time i and the predecessor S of T is frequent. Unfortunately, it is difficult to identify the accurate number $count_i(T)$ of the occurrences of T at time i. Thus we define the *initial count* of T, denoted by $init(T)$, as a positive integer $1 \leq init(T) \leq count(S)$ used as a initial value of the counter $count(T)$ of T, where S is the predecessor of T and $count(S)$ is the current value of the counter of S. Since $count_i(T) \leq count_i(S)$ holds for any pattern T and its predecessor S, it is reasonable to assume $1 \leq init(T) \leq count(S)$.

There are several varieties of candidate management strategies. In this paper, we focus on the following two strategies, namely the *lazy* strategy and the *eager* strategy, which respectively give $init(T) = 1$ and $init(T) = count(S)$ to a new inserted pattern T as its initial count.

Now we give characterizations of the lazy and the eager strategies. Let $(\mathcal{A}_1, \ldots, \mathcal{A}_i, \ldots) \subseteq \mathcal{T}^*$ and $(\mathcal{F}_1, \ldots, \mathcal{F}_i, \ldots) \subseteq \mathcal{T}^*$ be sequences of the sets of patterns, where $\mathcal{A}_i$ and $\mathcal{F}_i$ are the answer set that the algorithm StreamT computes and the set of frequent patterns at stage i, respectively. The following theorem indicates the soundness of the algorithm in the lazy strategy.

Theorem 1. *If the algorithm* StreamT *adopts the lazy strategy in candidate management, then* $\mathcal{A}_i \subseteq \mathcal{F}_i$ *holds for any stage* $i \geq 1$.

Corollary 1. *Omitting the delete phase in candidate management from* StreamT *with the lazy strategy,* $\mathcal{A}_i \subseteq \mathcal{F}_i$ *also holds for any* $i \geq 1$.

On the other hand, the next theorem says that the algorithm in the eager strategy becomes complete, if omitting the delete phase in candidate management.

Theorem 2. *If the algorithm* StreamT *adopts the eager strategy and omits the delete phase in candidate management, then* $\mathcal{F}_i \subseteq \mathcal{A}_i$ *holds for any* $i \geq 1$.

Finally, we introduce a method for delaying insertion of patterns into the candidate pool used in TDAG [13]. In this method, we do not insert a pattern T into $\mathcal{C}$ if $i \leq Lk^2$ holds, where L be a given integer and k is the size of T. By using this, we can avoid explosion of candidate patterns in the beginning of the computation.

5 Modified Algorithms for Other Online Models

The algorithm StreamT in Fig. 2 is an online algorithm for the online frequent pattern discovery problem in the *basic online model* of Definition 3. Now we present modification of StreamT to the *sliding window model* and the *forgetting model* also introduced in Definition 3.

5.1 Sliding Window Model

It is easy to modify StreamT into a mining algorithm in the sliding window model. In this model, we replace the frequency $freq_i(T) = count_i(T)/i$ with $freq^{\mathrm{sw}}_{i,w}(T) = count_i(T)/w$, where i is the current time and w is the given window size. Therefore, the modified algorithm for the sliding window model computes the frequent patterns in the *i-th window* $W_i = (\boldsymbol{v}_{i-w+1}, \ldots, \boldsymbol{v}_i)$ *of width* w for any stage i. To remove an old root occurrence out of the window, we store the list $RTO(T)$ of the root occurrences in the window W_i for all trees T. Maintenance of the root occurrence lists $RTO(T)$ is done in the algorithm CandMan in Fig. 6 as follows: (i) remove the node from the occurrence lists that is out of the window W_i (ii) add the root occurrence ρ of T to $RTO(T)$ for all $T \in \mathcal{C}$ if $\rho \neq rto(T)$, (iii) create the occurrence list $RTO(T)$ consisting of the only element ρ for any new candidate T, where ρ is the root occurrence of T.

5.2 Forgetting Model

Recall that in the forgetting model, the contribution of the past event decays exponentially fast. For a forgetting factor $0 < \gamma < 1$, the frequency of T at time

---

$\mathsf{IncFreq}(T, i)$
- If $hit_i(T) = 1$ then $fr(T) := \frac{lt(T)}{i-1}\gamma^{i-lt(T)} fr(T) + \frac{1}{i}hit_i(T)$ and $lt(T) := i$;
- If $T \notin \mathcal{C}$ then $ft(T) := i$;

$\mathsf{GetFreq}(T, i)$
- If $hit_i(T) = 1$ then $\mathsf{IncFreq}(T, i)$ and return $fr(T)$;
- Otherwise, return $\frac{lt(T)}{i-1}\gamma^{i-lt(T)} fr(T)$;

Fig. 7. Updating and Computing the Frequency

i is given by Eq. 3 in Section 2. At first look, implementation of the forgetting model is as easy as the online model above because they only differ in the definition of the frequency. In the forgetting model, however, the frequency at time i depends on all of the weights γ^{i-j} of past events changing as $i \geq 1$ goes larger. Thus, it seems that an algorithm have to maintain all of these weights at every time. Fortunately, this is not true.

We abbreviate the frequency $freq^{\mathrm{fg}}_{\gamma,i}(T)$ and the event $hit^{(i)}_j(T)$, respectively, to fr_i and hit_j. Below, we give an incremental method to compute the frequency. Let $lt(i) = \max\{\, j \leq i \,|\, hit_j = 1 \,\}$ be the last time stamp at which T appeared. Then, we have the following lemma.

Lemma 3. *For every $T \in \mathcal{T}$ and every $i \geq 1$, we have the recurrence*

$$\begin{aligned} fr_0 &= 0, \\ fr_i &= \frac{lt(i)}{i-1}\gamma^{i-lt(i)} fr_{lt(i)} + \frac{1}{i}\, hit_i \quad (i > 0) \end{aligned} \tag{4}$$

Proof. We first derive the recurrence for the consecutive steps. Then, derive the recurrence of the lemma by expanding fr_i using $lt(i)$. Since $hit_u = 0$ for any u with $lt(i) < u < i$, the claim immediately follows.

Now, we present a modified version of $\mathsf{StreamT}$ in the forgetting model. We modify the algorithms $\mathsf{StreamT}$ as follows. $\mathsf{StreamT}$ maintains the three parameters $fr(T), ft(T), lt(T)$, the frequency, the first and the last time stamps of the updates for T. $\mathsf{StreamT}$ uses $\mathsf{IncFreq}$ to update these parameters whenever a root occurrence of pattern T is detected and uses $\mathsf{GetFreq}$ to compute the frequency of T whenever it receives a request.We can see $\mathsf{IncFreq}(T, i)$ and $\mathsf{GetFreq}(T, i)$ can be executed in constant time when invoked by these procedures.

6 Experimental Results

In this section, we present experimental results on real-life datasets to evaluate the performance and the robustness of our algorithms on a range of datasets and parameters.

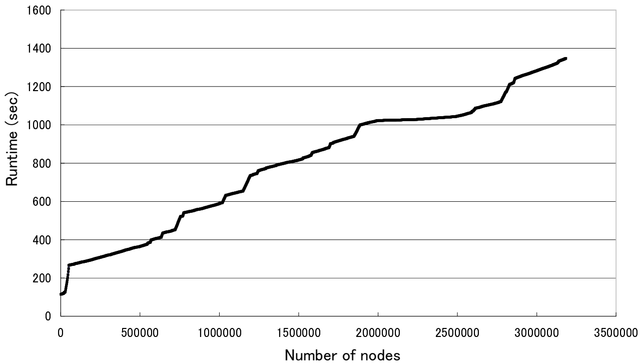


Fig. 8. The online scale-up experiment

6.1 Data

For the experiments, we prepared two datasets called *dblp* and *web-soap*. The dataset *dblp* is an XML document[1] that is available in the online bibliographic archive DBLP. This dataset consists of 130MB and its data tree had 3,185,138 nodes with 22 unique HTML tags.

The dataset *web-soap* is a medium-sized XML document obtained by concatenating two kinds of documents, weblog.xml and soap.xml in the order of weblog, soap, and weblog, where weblog.xml and soap.xml are XML documents of 9,164 (nodes) with 14 (tags) and 4,502 (nodes) with 11 (tags), respectively, consisting of collections of weblogs and soap messages[2], encoded in XML. Total size of weblog-soap dataset is 23,830 (nodes).

6.2 Scalability

In the first experiment, we studied the scalability of StreamT. We ran the algorithm with the frequency threshold $\sigma = 1(\%)$ on the data stream for the dataset *dblp* and measured the running time of the algorithm every two thousand nodes. The result is shown in Fig. 8. From this experiment, the proposed online algorithm StreamT runs almost linearly and thus continues to work efficiently for large semi-structured data streams.

6.3 Comparison of Algorithms

In the second experiment, we examined the effect of the candidate management policies and the choice of the online models discussed in Section 4 and Section 5, respectively. We implemented the three modifications, namely LB, LF, and EB,

[1] `http://dblp.uni-trier.de/xml/dblp.xml`
[2] Sosnoski Software Solution Inc., `http://www.sosnoski.com/opensrc/xmlbench/`

Table 1. The four algorithms compared in the experiments

Algorithms	Candidate management	Online strategy
LB	lazy	basic online model
LF	lazy	forgetting model
EB	eager	basic online model

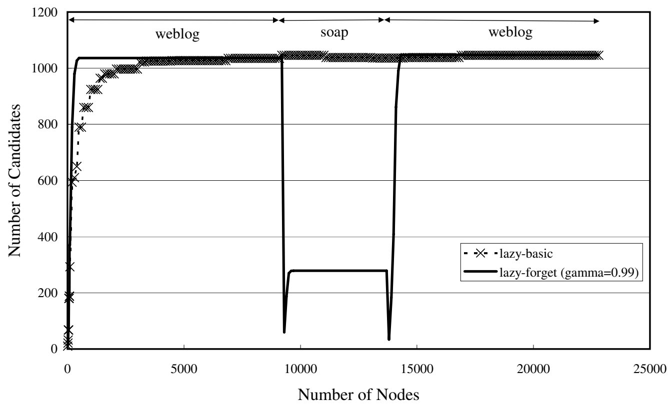


Fig. 9. The number of candidate patterns computed by *LB* and *LF*

of the basic StreamT algorithm shown in Table 1 and runs experiments on the dataset *web-soap* with a PC (PentiumIII 1.2GHz, 1GB RAM, g++ on cygwin, Windows2000).

Online strategies. Fig. 9 shows the number of candidate patterns computed by *LB* and *LF* with the forgetting factor $\gamma = 0.99$, with the same frequency threshold $\sigma = 5(\%)$. In the *web-soap* data, there are boundaries of two different datasets, i.e., weblog to soap or soap to weblog boundaries, at the data size of 9,000 and 14,000 (nodes). Also, the numbers of candidate patterns in these datasets significantly differs each other, i.e., they are about 1000 and 300 (patterns). In the plot for *LB*, the number of discovered patterns is constant even at these points. Thus, *LB* is insensible to the trend change. On the other hand, in the plot for *LF*, the number of discovered patterns quickly changed at these boundary points.

Forgetting factors. Then, we studied the performance of the algorithm StreamT in the forgetting model on a range of forgetting factors. Fig. 10 shows the number of candidate patterns computed at each stage by the algorithms *LF* with the different forgetting factors $\gamma = 0.99, 0.999, 0.9999$, and 0.99999, where the forgetting factor of 0.99 means that the ratio $0.01 = 1.00 - 0.99$ of the accumulated frequency is forgotten in a unit time.

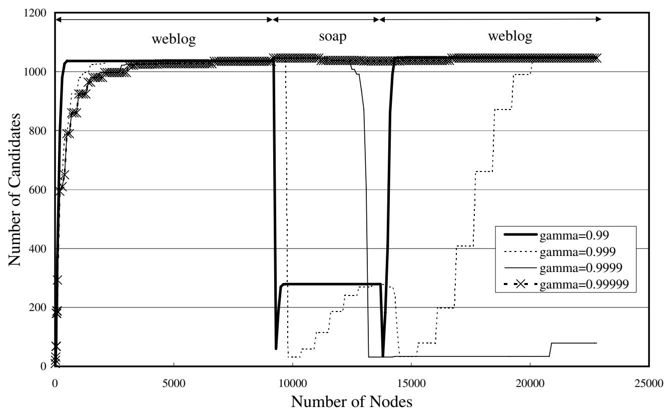


Fig. 10. The number of candidate patterns computed by LF, changing the forgetting factor $\gamma = 0.99, 0.999, 0.9999,$ and 0.99999

From the figure, we can see that algorithms with higher forgetting factors does not immediately compute current new patterns after a trend of the data stream changes. Furthermore, the performance of the algorithm with the forgetting factor $\gamma = 0.99999$ is similar to the performance of the algorithm in the basic online model shown in Fig. 9, since the value of forgetting factor is too high to decay the effect of past events.

Candidate management strategies. Finally, we compared the lazy and eager strategies in candidate management policies in Section 4.1. In Fig. 11, we show the number of candidate patterns and frequent patterns at each stage computed by LB and EB, with the frequency threshold $\sigma = 7(\%)$. From the figure, we

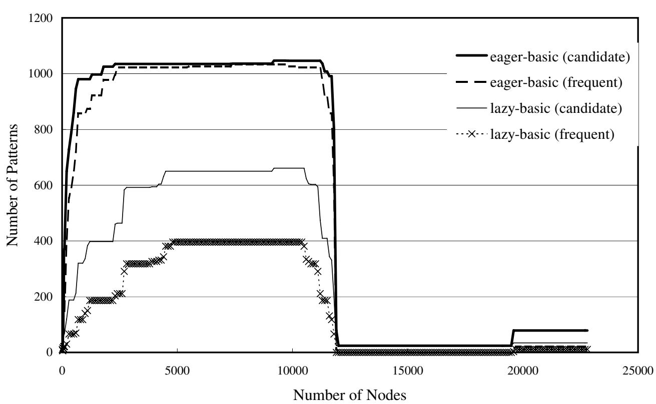


Fig. 11. The number of candidate and frequent patterns computed by LB and EB

can see that the algorithm with the eager management computes more patterns than the algorithm with the lazy management.

7 Conclusion

In this paper, we studied an online data mining problem from unbounded semi-structured data stream. We presented efficient online algorithms that are continuously working on an unbounded stream of semi-structured data with bounded resources, and find a set of frequent ordered tree patterns from the stream on request at any time. We modified our algorithm to other online models, and run experiments on real XML data to evaluate the proposed algorithms.

It is a future problem to examine the online property of the proposed algorithms using long and trend-changing semi-structured data streams in the real world.

Acknowledgments

Hiroki Arimura is grateful to Masaru Kitsuregawa for his direction of the author's interests to online data processing. The authors would like to thank Takeaki Uno, Ken Satoh, Masayuki Takeda, Ayumi Shinohara, Hiroshi Motoda, Takashi Washio, Takeshi Tokuyama, Akihiro Yamamoto, Yoshiharu Ishikawa, and Mohammed Zaki for fruitful discussions on data mining.

References

1. K. Abe, S. Kawasoe, T. Asai, H. Arimura, and S. Arikawa. Optimized Substructure Discovery for Semi-structured Data, In *Proc. PKDD'02*, 1–14, LNAI 2431, Springer, 2002.
2. S. Abiteboul, P. Buneman, D. Suciu, *Data on the Web*, Morgan Kaufmann, 2000.
3. A. V. Aho, J. E. Hopcroft, and J. D. Ullman, *Data Structures and Algorithms*, Addison-Wesley, 1983.
4. T. Asai, K. Abe, S. Kawasoe, H. Arimura, H. Sakamoto, and S. Arikawa, Efficient Substructure Discovery from Large Semi-structured Data. In *Proc. SIAM SDM'02*, 158–174, 2002.
5. T. Asai, H. Arimura, K. Abe, S. Kawasoe, and S. Arikawa, Online Algorithms for Mining Semi-structured Data Stream, In *Proc. IEEE ICDM'02*, 27–34, 2002.
6. T. Asai, H. Arimura, T. Uno, and S. Nakano, Discovering Frequent Substructures in Large Unordred Trees, In *Proc. DS'03*, LNAI 2843, 47–61, Springer-Verlag, 2003.
7. R. J. Bayardo Jr., Efficiently Mining Long Patterns from Databases, In *Proc. SIGMOD98*, 85–93, 1998.
8. M. de Berg, M. van Kreveld, M. Overmars, O. Schwarzkopf, Computational Geometry, Algorithms and Applications, Springer, 2000.
9. L. Dehaspe, H. Toivonen, R. D. King, Finding Frequent Substructures in Chemical Compounds, In *Proc. KDD-98*, 30–36, 1998.

10. P. B. Gibbons and Y. Matias, Synopsis Data Structures for Massive Data Sets, In *External Memory Algorithms*, DIMACS Series in Discr. Math. and Theor. Compt. Sci., Vol. 50, AMS, 39–70, 2000.
11. C. Hidber, Online Association Rule Mining, In *Proc. SIGMOD'99*, 145–156, 1999.
12. P. Kilpelainen, H. Mannila, Ordered and Unordered Tree Inclusion, SIAM J. Comput, 24(2), 340–356, 1995.
13. P. Laird, R. Saul, Discrete Sequence Prediction and Its Applications, Machine Learning, 15(1), 43–68, 1994.
14. H. Mannila, H. Toivonen, and A. I. Verkamo, Discovering Frequent Episode in Sequences, In *Proc. KDD-95*, 210–215, AAAI, 1995.
15. T. Matsuda, T. Horiuchi, H. Motoda, T. Washio, K. Kumazawa, N. Arai, Graph-Based Induction for General Graph Structured Data, In *Proc. DS'99*, LNAI 1721, 340–342, Springer-Verlag, 1999.
16. T. Miyahara, Y. Suzuki, T. Shoudai, T. Uchida, K. Takahashi, H. Ueda, Discovery of Frequent Tag Tree Patterns in Semistructured Web Documents. In *Proc. PAKDD'02*, LNAI 2336, 341–355, Springer-Verlag, 2002.
17. S. Nakano, Efficient Generation of Plane Trees, *Information Processing Letters*, 84, 167–172, Elsevier, 2002.
18. S. Parthasarathy, M. J. Zaki, M. Ogihara, S. Dwarkadas, Incremental and Interactive Sequence Mining, In *CIKM'99*, 251–258, 1999.
19. R. Rastogi, Single-Path Algorithms for Querying and Mining Data Streams, In *Proc. SDM'02 Workshop HDM'02*, 43–48, 2002.
20. W3C, Extensive Markup Language (XML) 1.0 (Second Edition), *W3C Recommendation*, 06 October 2000.
`http://www.w3.org/TR/REC-xml`
21. K. Wang, H. Liu, Discovering Structural Association of Semistructured Data, TKDE, 12(2), 353-371, IEEE Computer Society, 2000.
22. K. Yamanishi, J. Takeuchi, A Unifying Framework for Detecting Outliers and Change Points from Non-Stationary Time Series Data, In *Proc. SIGKDD-2002*, ACM, 2002.
23. M. J. Zaki. Efficiently mining frequent trees in a forest, In *Proc. SIGKDD-2002*, ACM, 2002.

Bus Information System Based on User Models and Dynamic Generation of VoiceXML Scripts

Shinichi Ueno, Fumihiro Adachi, Kazunori Komatani, Tatsuya Kawahara, and Hiroshi G. Okuno

Kyoto University, Kyoto 606-8501, Japan
{komatani,kawahara,okuno}@i.kyoto-u.ac.jp

Abstract. We have developed a telephone-based cooperative natural language dialogue system. Since natural language involves very various expressions, a large number of VoiceXML scripts need to be prepared to handle all possible input patterns. Thus, flexible dialogue management for various user utterances is realized by generating VoiceXML scripts dynamically. Moreover, we address the issue of appropriate user modeling to generate cooperative responses to users. Specifically, three dimensions of user models are set up: the *skill level* to the system, the *knowledge level* on the target domain and the degree of *hastiness*. The models are automatically derived by decision tree learning using real dialogue data collected by the system. Experimental evaluation showed that the cooperative responses adapted to individual users served as good guides for novices without increasing the duration of dialogue for skilled users.

1 Introduction

A Spoken dialogue system is one of the promising applications of the speech recognition and natural language understanding technologies. A typical task of spoken dialogue systems is database retrieval. Some IVR (interactive voice response) systems using the speech recognition technology are being put into practical use as its simplest manifestation. Due to the spread of cellular phones, spoken dialogue systems via the telephone enable us to obtain information from various places without the need for any special apparatus.

To realize user-friendly interaction, spoken dialogue systems should be able to (1) accept various user utterances to enable mixed-initiative dialogue and (2) generate cooperative responses. Currently, a lot of IVR systems via telephone operate by using VoiceXML, which is a script language to describe procedures of spoken dialogues. However, the procedures are basically designed as system-initiated one, where the system asked required items one by one, because only next behaviors corresponding to every input is described in VoiceXML scripts. The system should be able to accept various user-initiated utterances to realize mixed-initiative dialogue. By allowing acceptance of various user utterances, the combination of words contained in them accordingly gets enormous. It is

A. Sakurai et al. (Eds.): JSAI 2003/2004, LNAI 3609, pp. 46–60, 2007.

practically impossible to prepare all VoiceXML scripts that correspond to the enormous combinations in advance. It is also difficult to generate cooperative responses adaptively in the framework.

We propose a framework to generate VoiceXML scripts dynamically to realize the mixed-initiative dialogue, where the system is needed to accept various user utterances. This framework realizes flexible dialogue management without requiring preparation for a large number of VoiceXML scripts in advance. Furthermore, it enables the system to behave adaptively to various dialogue situations that can be described by the number of obtained query results, presence of alternatives, and so on.

Another problem to realize user-friendly interaction is how to generate cooperative responses. Dialogue strategies, which determine when to provide guidance and what the system should tell the user, are one of the essential factors in spoken dialogue systems. There are many studies in respect of the dialogue strategy such as confirmation management using confidence measures of speech recognition results [1,2], dynamic change of dialogue initiative [3,4,5], and adding cooperative contents to system responses [6].

Nevertheless, whether a particular response is cooperative or not depends on individual user's characteristic. To adapt the system's behavior to individual users, it is necessary to model user's patterns [7]. Most of conventional studies on user models have focused on user's knowledge. Others have tried to infer and utilize user's goals to generate responses that have been adapted to the user [8,9]. Elzer et al. [10] proposed a method of generating adaptive suggestions according to users' preferences. However, these studies depended on a thorough knowledge of the target domain, and therefore the user models need to be constructed manually to be applied to new domains. Moreover, they have assumed that the input is text only, which does not contain errors.

We propose more comprehensive user models to generate user-adapted responses in spoken dialogue systems taking available information into account specific to spoken dialogue. Spoken utterances include various types of information such as the interval between the utterances, barge-in (interruption of system prompts by users), which can be utilized to judge the user's character. These also have generality in spoken dialogue systems because they are not dependent on domain-specific knowledge. In [11], typical users' behaviors are defined so that spoken dialogue systems can be evaluated by simulation, and stereotypes are assumed such as patient, submissive and experienced. We did not introduce user models to define users' behaviors beforehand, but to detect users' patterns in real-time interactions.

We defined three dimensions in the user models, i.e., the *skill level* to the system, the *knowledge level* on the target domain, and the degree of *hastiness*. The user models were trained with decision tree learning using real data collected from the Kyoto city bus information system. We then implemented them and adaptive dialogue strategies on the system, and evaluated them using data collected with 20 novice users.

2 Bus Information System Based on Dynamic Generation of VoiceXML Scripts

VoiceXML[1] is a script language to describe procedures in spoken dialogues mainly on telephone. It is becoming to a standard language of IVR (interactive voice response) systems. The VoiceXML scripts consist of three parts: (1) specifications of system's prompts, (2) specifications of grammars to accept a user's utterance, and (3) description of the next behaviors.

However, most of existing services using the VoiceXML impose rigid interaction. User utterances are restricted by system-initiated prompts. Users are accordingly allowed to specify only requested items one by one. It is more user-friendly that users can freely convey their requests by natural language expressions.

We present a framework to realize flexible interaction by generating VoiceXML scripts dynamically [12,13]. The framework enables users to express their requests by natural language even in VoiceXML-based systems.

2.1 Dynamic Generation of VoiceXML Scripts

In VoiceXML scripts, acceptable keywords and corresponding next states must be explicitly specified. To enable the system to accept natural language expressions, enormous combinations of keywords must be specified in the scripts since the combinations arbitrarily appear in natural language utterances. It is practically impossible to prepare all VoiceXML scripts that correspond to the combinations. We then introduce the framework where VoiceXML scripts are generated dynamically to enable the system to accept natural language expressions.

Figure 1 shows the overview of the framework. The front end that operates based on VoiceXML scripts is separated from the dialogue management portion, which accepts speech recognition results and generates corresponding VoiceXML scripts. The user utterance is recognized based on grammar rules specified in VoiceXML scripts, and keywords extracted from a speech recognition result are passed to the CGI script. It retrieves corresponding information from the database on the Web, and generates VoiceXML scripts for next interaction. If sufficient information is not obtained from a user utterance, a script that prompts to fill remaining contents is generated. If there is ambiguity in a user utterance, a script that makes a disambiguating question is also generated.

The generation of VoiceXML scripts enables to accept natural language expressions without preparing a large number of scripts corresponding to various inputs beforehand. The framework also enables to generate cooperative responses adapted to the situation such as retrieval results without spoiling portability.

2.2 Kyoto City Bus Information System

We have developed the Kyoto city bus information system, which locates the bus a user wants to take, and tell him/her how long it will take before arriving. The

[1] VoiceXML Forum. `http://www.voicexml.org/`

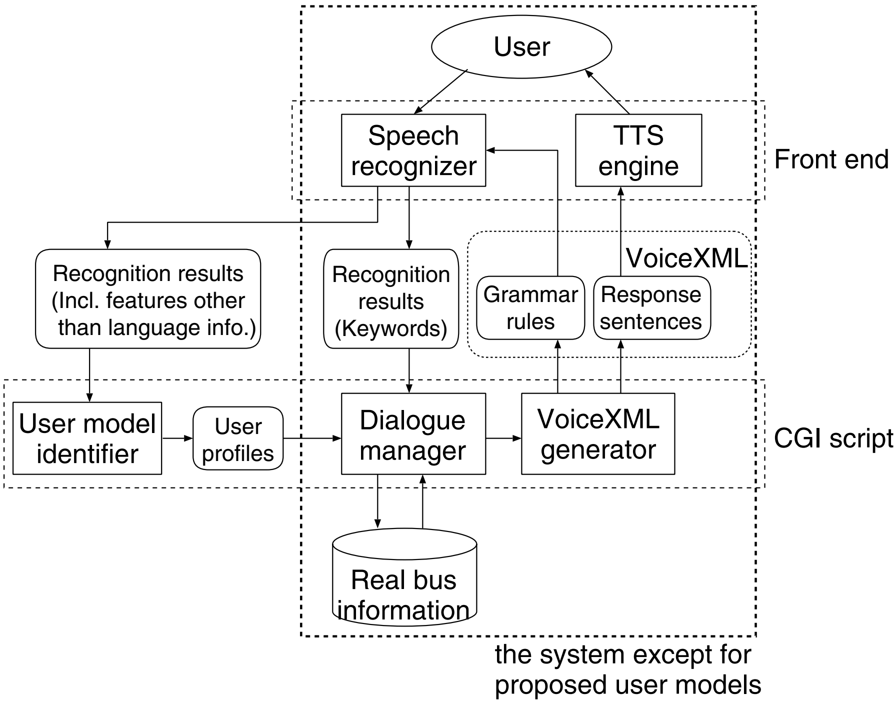


Fig. 1. Overview of bus system based on dynamic generation of VoiceXML scripts

system can be accessed via telephone including cellular phones[2] From any places, users can easily get the information on the bus that changes every minute. Users are requested to input the bus stop to get on, the destination, or the bus route number by speech, and consequently get the bus information. There are 652 bus stops and 85 bus routes in Kyoto city. The bus stops can also be specified by the name of famous places or public facilities nearby. Figure 2 shows a simple example of the dialogue.

Figure 1 also shows an overview of the Kyoto city bus information system with the user models. The system operates by generating VoiceXML scripts dynamically as described in Section 2.1. The real-time bus information database is provided on the Web, which can be accessed via Internet. Then, we explain the modules in the following.

Speech Recognizer
The speech recognizer decodes user utterances based on specified grammar rules and vocabulary, which are defined by VoiceXML at each dialogue state.

Dialogue Manager
The dialogue manager generates response sentences based on speech recognition results (bus stop names or a route number). If sufficient information

[2] +81-75-326-3116.

Sys:	What is your current bus stop, your destination or specific bus route number?
User:	Shijo-Kawaramachi.
Sys:	Will you take a bus from Shijo-Kawaramachi?
User:	Yes.
Sys:	Where will you get off the bus?
User:	Arashiyama.
Sys:	Will you go from Shijo-Kawaramachi to Arashiyama?
User:	Yes.
Sys:	Bus number 11 bound for Arashiyama has departed Sanjo-Keihanmae, two bus stops away.

Fig. 2. Example dialogue in bus system

to locate a bus is obtained, it retrieves the corresponding bus information on the Web.

VoiceXML Generator

This module dynamically generates VoiceXML scripts that contain response sentences and specifications of speech recognition grammars, which are given by the dialogue manager.

User model identifier

This module classifies user's characters based on the user models using features specific to spoken dialogue as well as semantic attributes. The obtained user profiles are sent to the dialogue manager, and are utilized in the dialogue management and response generation.

2.3 Cooperative Responses in Kyoto City Bus Information System

It is important to generate cooperative responses by guessing the user's intention in spoken dialogue systems. Several kinds of cooperative responses are listed in [6], such as adding information more than a user has requested, telling reason when there is no query result, presenting alternatives, and so on.

Our system can generate various responses according to situations such as query results, since we separated language-understanding modules from individual VoiceXML scripts and generate them dynamically. Examples of cooperative responses implemented in our system are listed below.

(1) Suggestive Answers

When query results are negative, the system can offer related information such as alternatives.

(Example)

User: From Hyakumanben to Kyoto Station by bus number 17.

System: Bus number 17 does not come for a while, *but bus number 206 bound for Kyoto Station has departed Takano, four bus stops away.*

(2) Corrective Answers
The reason is presented to users when there is no query result or when an utterance contains errors.
(Example)
User: From Kyoto Station to Arashiyama.
System: *There is no bus that directly goes from Kyoto Station to Arashiyama.*

(3) Intensional Answers
If several query results are obtained, the system generates intensional answers. This is more cooperative than listing all the results.
(Example)
User: From Shijo-Kawaramachi to Shijo-Karasuma.
System: *All buses* to Shijo-Karasuma do not come for a while.

(4) Completion Answers
The system can generate responses including information more than user's requests.
(Example)
User: From Kyoto Station by bus number 5.
System: Bus number 5 will depart at 17:34 *from platform 'A1'*.
(There are a lot of platforms in Kyoto Station.)

The responses listed above are not always cooperative for every user. For example, suggestive answers that provide alternatives may be verbose for hasty users, or completion answers are not informative for acquainted people. That is, whether such cooperative responses should be presented or not depends on individual users. We will describe user modeling to adapt system responses to users in the next section.

3 Response Generation with User Models

We propose comprehensive user models to generate user-adapted responses taking account of information specific to spoken dialogue. Whether a particular response is regarded as cooperative or not depends on individual user's characteristics. We therefore address appropriate user modeling to generate cooperative responses.

3.1 Classification of User Models

We defined three dimensions for the user models, which are listed below.

- Skill level to system
- Knowledge level on target domain
- Degree of hastiness

Skill Level to the System
Since spoken dialogue systems are not yet widespread, there is a difference in the skill levels of users in operating the systems. It is important for the system to change its behavior including response generation and initiative management according to the skill levels of users. In conventional systems, system-initiated guidance has been invoked on the spur of the moment either when the user remains silent or when speech recognition is not successful. In our framework, we find a radical solution for the unskilled users by modeling the skill level on the user's properties.

Knowledge Level on the Target Domain
There is also a difference in the knowledge level on the target domain between users. Thus, it is necessary for the system to change its information to present to individual users. For example, it is inappropriate to give too much detailed information to strangers. However, it is useful to omit obvious information and to output additive information for established residents. We therefore introduced a dimension that represents the knowledge level on the target domain.

Degree of Hastiness
It is more important in speech communication to present information promptly and concisely compared with the other communication modes such as browsing. Conciseness is preferred, especially in our bus system, because this information is urgently required by most users. Therefore, we also took the degree of hastiness of the user into account, and changed the system's responses in accordance with this.

3.2 Response Generation Strategy Using User Models

We will next describe the response generation strategies adapted to individual users based on the proposed user models: *skill level*, *knowledge level* and *hastiness*. The basic design of dialogue management is based on mixed-initiative dialogue, where the system asks follow-up questions and provides guidance if necessary while allowing the user to make utterances freely. Sadek investigated adding various content to system responses as cooperative responses [6]. Such additional information is usually cooperative, but some people may feel that such a response is redundant.

We thus introduced user models and controlled the generation of additive information. The system changes responses with the proposed user models through dialogue procedure and the content of responses.

Dialogue Procedure
The dialogue procedure is changed based on the *skill level* and *hastiness*. If a user is identified as having the high *skill level*, dialogue management is carried out in a user-initiated manner; i.e., the system generates only open-ended prompts. However, when user's *skill level* is detected as being low, the system takes the initiative and prompts for necessary items in order.

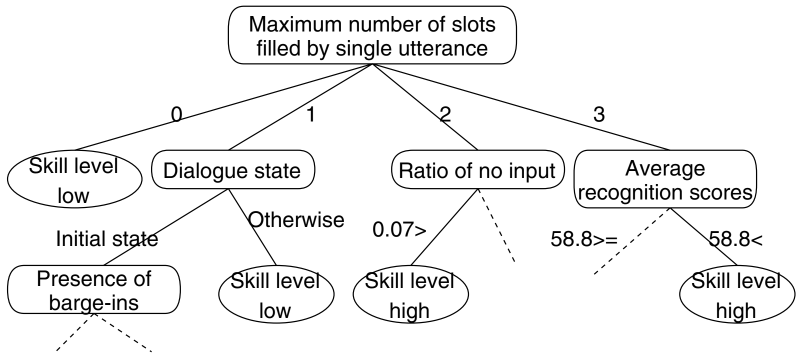


Fig. 3. Decision tree for *skill level*

When the degree of *hastiness* is low, the system confirms the input content. Conversely, when *hastiness* is high, such confirmation is omitted.

Contents of Responses
Information that should be included in the system response can be classified into the following:

1. Dialogue management information and
2. Domain-specific information.

Dialogue management information specifies how the dialogue is carried out including instructions concerning user expressions like "Reply with either yes or no." and explanations on the following dialogue procedure like "Let me confirm these in turn." This dialogue management information is determined by the user's *skill level* to the system, and is added to system responses when the *skill level* is considered low.

Domain-specific information is generated according to the user's *knowledge level* on the target domain. That is, for users unacquainted with the local information, the system adds explanations about the nearest bus stop, and omits complicated contents such as proposing another route.

The content described above is also controlled by *hastiness*. For users who are not in a hurry, the system generates additional contents as cooperative responses. For hasty users, however, the content is omitted to prevent the dialogue from becoming verbose.

3.3 Classification of Users Based on Decision Tree

We adopted a decision tree to implement the proposed user models as classifiers. It was constructed with decision tree learning algorithm C5.0 [14] with data collected with our dialogue system. Figure 3 shows the derived decision tree for *skill level*.

- features obtained from single utterance
  - dialogue state (defined by already filled slots)
  - presence of barge-in
  - elapsed time for current utterance
  - recognition results (something recognized / uncertain / no input)
  - score by speech recognizer
  - number of slots filled by current utterance
- features obtained from the session
  - number of utterances
  - dialogue state of previous utterance
  - elapsed time from beginning of session
  - number of the repetitions of same question
  - average number of repetitions of same question
  - ratio of total time of user utterances in whole of elapsed time
  - ratio of occurrence of barge-ins out of whole number of utterances
  - recognition results for previous utterance
  - ratio of something recognized
  - ratio of uncertain results
  - ratio of no input
  - number of barge-ins
  - number of something recognized
  - number of uncertain results
  - number of no input
  - average recognition scores
  - maximum number of slots filled by single utterance
- features obtained from session (used only in classification of *knowledge level*)
  - attribute of specified bus stops
  - way to specify bus stops (whether bus stop is specified by its correct name or not)
  - ratio of each attribute in specified bus stops
  - ratio of each way to specify bus stops

Fig. 4. Features to classify user models

We used the features listed in Figure 4. These include not only the semantic information contained in utterances but also information specific to spoken dialogue systems such as the duration of silence prior to utterances and barge-ins. Except for the last category in Figure 4 including "attribute of specified bus stops", most of the features are domain-independent.

Each dimension is classified for every user utterance except for the *knowledge level*. The user models can change during the dialogue. Features extracted from utterances are accumulated as history information during the session.

Figure 5 has an example of system behavior with the user models we propose. The *skill level* is classified as low by the decision tree, because the user's first utterance includes only one content word. The dialogue procedure is then changed to a system-initiated one. Similarly, *hastiness* is classified as low by the

case: [skill level: low] [knowledge level: low] [hastiness: low]

Sys: What is your current bus stop, your destination or specific bus route number?
User: Maruyama Park.
Sys: Let me confirm these in turn.
(← Explanation is added by [skill level: low].)
Will you get on at Maruyama Park?
(← Content is confirmed by [hastiness: low].)
Reply with yes or no.
(← Instruction is added by [skill level: low].)
User: Yes.
Sys: Where will you get off the bus?
Specify by bus stop name or a well-known landmark.
(← Explanation is added by [skill level: low].)
User: Tofukuji Temple.
Sys: Will you get off the bus at Tofukuji Temple?
Reply with yes or no.
User: Yes.
Sys: Following buses are located.
(← Explanation is added by [skill level: low].)
Nearest bus stop to Maruyama Park is Gion.
(← Domain information is added by [knowledge level: low].)
Bus number 202 bound for Nishioji-Kujo has departed Higashiyama-Sanjo, which is two stops away.
...

Fig. 5. Example dialogue with proposed user models

decision tree, and the system has the explanation for the dialogue procedure and instruction on expressing responses. They are omitted if *hastiness* is identified as high.

3.4 Decision Tree Learning for User Models

We trained and evaluated the decision tree for the user models using dialogue data collected with our system. The data was collected from December 10, 2001 to May 10, 2002. There were 215 sessions (telephone calls), and a total of 1492 utterances in the sessions. We annotated the subjective labels by hand. An annotator judged the user models for every utterance based on recorded speech data and logs. Labels were awarded to the three dimensions in Section 3.1 either 'high', 'undetermined' or 'low'. It is possible for the user's annotated model to change during the dialogue, especially from 'undetermined' to 'low' or 'high'. Table 1 has the number of labeled utterances.

We evaluated the classification accuracy of the user models with the labeled data. All the experiments were carried out with 10-fold cross validation. A

Table 1. Number of manually labeled items for decision tree learning

	Low	Undetermined	High	Total
Skill level	743	253	496	1492
Knowledge level	275	808	409	1492
Hastiness	421	932	139	1492

Table 2. Classification accuracy of proposed user models

Condition	#1	#2	#3
Skill level	80.8%	75.3%	85.6%
Knowledge level	73.9%	63.7%	78.2%
Hastiness	74.9%	73.7%	78.6%

process, where one tenth of all data was used as test data and the remainder was used as training data, was repeated ten times, and the average accuracy was computed. The results are listed in Table 2. Conditions #1, #2 and #3 in Table 2 can be described as follows.

#1: We did 10-fold cross validation per utterance.
#2: We did 10-fold cross validation per session (call).
#3: We calculate the accuracy under more realistic conditions. We also did 10-fold cross validation per session (call) as well as Condition #2. The accuracy was not calculated in three classes (high / undetermined / low) but the two that actually affected the dialogue strategies. For example, the accuracy for *skill level* was calculated for two classes: low and others. To classify the *knowledge level*, accuracy was calculated for dialogue sessions because features such as the attributes of specified bus stops were not obtained in every utterance. Moreover, to smooth out the unbalanced distribution of training data, we introduced a cost corresponding to the reciprocal ratio of the number of samples in each class. Through cost, the rate of chance of two classes becomes 50%.

The difference between Conditions #1 and #2 is that training was done in a speaker-closed or speaker-open manner. The former had better performance.

Under evaluation condition #3, the hypothesized *skill level* was accurate. The features that follow played an important part in the decision tree for the *skill level*. These were the number of slots filled by the current utterance, the presence of barge-in, and the ratio of no input. The recognition results (something recognized / uncertain / no input), ratio of no input, and the way of specifying bus stops (whether a bus stop is specified by its exact name or not) were effective for the *knowledge level*. *Hastiness* was classified mainly by the presence of barge-in, ratio of no input, and the elapsed time for the current utterance.

4 Experimental Evaluation of the System with User Models

We evaluated the system with the user models we proposed using 20 novice subjects who had not previously used it. The experiment was done in a laboratory under adequate controls. A headset microphone was used for speech input.

4.1 Experiment Procedure

We first explained the outline of the system to subjects and gave each a document that described the experimental conditions and scenarios. We prepared two sets of eight scenarios. Subjects were requested to acquire bus information using the system with/without user models. In the scenarios, neither concrete names for bus stops nor bus number were given. For example, one of the scenarios was as follows: "You are sightseeing in Kyoto. After visiting the Ginkakuji Temple, you go to Maruyama Park. In this scenario, obtain information about the bus." We also set constraints to vary subjects' hastiness such as "Be as quick as you can to conserve charge of your cellular phone."

Subjects were also told to peruse questionnaire items before the experiment, and answer these after using each system. This was aimed at reducing the subject's cognitive load and possible confusion due to switching systems [15]. The questionnaire consisted of eight items (see Table 4), e.g., "Did the system guide you intelligibly when the dialogue did not go well?" We established seven steps for evaluations on each item, and the subject selected one of these.

Furthermore, subjects were asked to write down the information obtained: the name of the bus stop to embark, the bus number and how long it would take before the bus arrived. Through this procedure, we planned to make the experimental conditions as realistic as possible.

The subjects were divided into two groups; half (Group 1) used the system in the order of "with user models → without user models", the other half (Group 2) used it in the reverse order.

Dialogue management in the system without user models was also based on mixed-initiative dialogue. The system generated follow-up questions and guidance where necessary, but behaves in a fixed manner. That is, additive cooperative content corresponding to the *skill level* described in Section 3.2 was not generated and the dialogue procedure was only changed after recognition errors occurred. The system behavior, without user models, was equivalent to the initial states of user models, i.e., *hastiness* was low, the *knowledge level* was low and *skill level* was high.

4.2 Results

All subjects successfully completed the given task, although they had been allowed to withdraw if the system did not work well. That is, the task success rate is 100%.

Table 3. Duration and the number of turns in dialogue

		duration (sec.)	# turn
Group 1	with UM	51.9	4.03
(with UM → w/o UM)	w/o UM	47.1	4.18
Group 2	w/o UM	85.4	8.23
(w/o UM → with UM)	with UM	46.7	4.08

UM: User Model

Table 4. Questionnaire items

1. Did you understand what to say in your turns?
2. Did the system guide you intelligibly when the dialogue did not go well?
3. Did not you feel uneasy because you did not understand the system's behaviors?
4. Did you get information quickly when in a hurry?
5. Was system's guidance concise and intelligible?
6. Did guidance contain required information in detail?
7. How was the user-friendliness of the system?
8. Do you want to use the system from now on?

Average dialogue duration and the number of turns in each case are listed in Table 3. Though users had not experienced the system at all, they got accustomed to it very rapidly. Therefore, as we can see in Table 3, both the duration and the number of turns decreased in the latter half of the experiment for both groups. However, in the initial half of the experiment, Group 1 completed it with significantly shorter dialogues than Group 2. This means that novice users got accustomed to the system more rapidly with user models, because they were instructed on use by cooperative responses that were generated when the *skill level* was low. The results demonstrate that cooperative responses that were generated according to the proposed user models could serve as good guides for novice users.

In the latter half of the experiment, the duration of dialogue and the number of turns were almost the same for the two groups. This means that the proposed models prevented the dialogue from becoming redundant for skilled users, although generating cooperative responses for all users generally made the dialogue verbose. This suggests that the proposed user models could appropriately control the generation of cooperative responses by detecting the characteristics of individual users.

Table 5 lists the results of the questionnaire. Subjects selected for eight items from seven scales, where a larger number denotes a better evaluation. The system with user models generally obtained better evaluation. There were statistically significant differences between the two groups in the initial half of the experiment for questionnaire items #1, #2, #7, and #8 with T-test ($p < 0.05$). This means cooperative responses that were generated according to the proposed user models were appropriate when users' skill levels were low. There were also statistically

Table 5. Results of questionnaire

Group 1 (with UM → w/o UM)		
	with UM	w/o UM
#1	4.5	4.4
#2	4.5	3.9
#3	3.9	2.5
#4	2.9	2.6
#5	4.2	4.0
#6	3.7	3.8
#7	3.7	3.3
#8	4.5	4.1

Group 2 (w/o UM → with UM)		
	w/o UM	with UM
#1	3.4	4.3
#2	2.6	3.9
#3	2.6	3.9
#4	2.1	3.6
#5	3.8	4.3
#6	4.5	4.3
#7	2.5	4.0
#8	2.5	3.5

significant differences between the initial and latter half of the experiment in Group 2 for items #2, #3, #4, #7, and #8. This suggests that responses of the system without the user models were not cooperative for novices. There was no significant difference for almost all items between the initial and latter half of the experiment in Group 1. This was because the behaviors of the system with/without the models became nearly equivalent after the users' skill levels became high.

5 Conclusions

We presented a framework to realize flexible interaction by dynamically generating VoiceXML scripts. This framework realizes mixed-initiative dialogues and cooperative responses in VoiceXML-based systems.

We also proposed and evaluated user models for adaptively generating cooperative responses to individual users. The proposed user models had three dimensions: *skill level* to the system, *knowledge level* on the target domain and the degree of *hastiness*. The user models were identified using features specific to spoken dialogue systems as well as semantic attributes. They were automatically derived by decision tree learning, and all features used for *skill level* and *hastiness* were independent of domain-specific knowledge. We therefore expect that user modeling can generally be applied to other domains.

The experimental evaluation with 20 novice subjects revealed that their skill levels were improved more rapidly by incorporating user models, and the duration of dialogue decreased almost immediately as a result. This was achieved by generating cooperative responses based on the user models we propose. The models also suppressed redundancy by changing the dialogue procedure and selecting contents of responses.

Thus, the framework generating VoiceXML scripts dynamically and the user models realized user-adaptive dialogue strategies. The cooperative responses that were generated based on the models served as good guides for novices without increasing the duration of dialogue for skilled users.

References

1. Komatani, K., Kawahara, T.: Flexible mixed-initiative dialogue management using concept-level confidence measures of speech recognizer output. In: Proc. Int'l Conf. Computational Linguistics (COLING). (2000) 467–473
2. Hazen, T.J., Burianek, T., Polifroni, J., Seneff, S.: Integrating recognition confidence scoring with language understanding and dialogue modeling. In: Proc. Int'l Conf. Spoken Language Processing (ICSLP). (2000)
3. Litman, D.J., Pan, S.: Predicting and adapting to poor speech recognition in a spoken dialogue system. In: Proc. of the 17th National Conference on Artificial Intelligence (AAAI2000). (2000)
4. Chu-Carroll, J.: MIMIC: An adaptive mixed initiative spoken dialogue system for information queries. In: Proc. of the 6th Conf. on Applied Natural Language Processing. (2000) 97–104
5. Lamel, L., Rosset, S., Gauvain, J.L., Bennacef, S.: The LIMSI ARISE system for train travel information. In: IEEE Int'l Conf. Acoust., Speech & Signal Processing (ICASSP). (1999)
6. Sadek, D.: Design considerations on dialogue systems: From theory to technology -the case of artimis-. In: Proc. ESCA workshop on Interactive Dialogue in Multi-Modal Systems. (1999)
7. Kass, R., Finin, T.: Modeling the user in natural language systems. Computational Linguistics **14** (1988) 5–22
8. van Beek, P.: A model for generating better explanations. In: Proc. of the 25th Annual Meeting of the Association for Computational Linguistics (ACL-87). (1987) 215–220
9. Paris, C.L.: Tailoring object descriptions to a user's level of expertise. Computational Linguistics **14** (1988) 64–78
10. Elzer, S., Chu-Carroll, J., Carberry, S.: Recognizing and utilizing user preferences in collaborative consultation dialogues. In: Proc. of the 4th Int'l Conf. on User Modeling. (1994) 19–24
11. Eckert, W., Levin, E., Pieraccini, R.: User modeling for spoken dialogue system evaluation. In: Proc. IEEE Workshop on Automatic Speech Recognition and Understanding. (1997) 80–87
12. Pargellis, A., Kuo, J., Lee, C.H.: Automatic dialogue generator creates user defined applications. In: Proc. European Conf. Speech Commun. & Tech. (EUROSPEECH). (1999)
13. Nyberg, E., Mitamura, T., Placeway, P., Duggan, M., Hataoka, N.: Dialogxml: Extending voicexml for dynamic dialog management. In: Proc. of Human Language Technology 2002 (HLT2002). (2002) 286–291
14. Quinlan, J.R.: C4.5: Programs for Machine Learning. Morgan Kaufmann, San Mateo, CA (1993) http://www.rulequest.com/see5-info.html.
15. Over, P.: Trec-7 interactive track report. In: Proc. of the 7th Text REtrieval Conference (TREC7). (1999)

Robotic Communication Terminals as a Ubiquitous System for Improving Human Mobility by Making Environment Virtually Barrier-Free

Ikuko Eguchi Yairi, Kentaro Kayama, and Seiji Igi

National Institute of Information and Communications Technology,
3–4 Hikarinooka, Yokosuka, Kanagawa 239–0847 Japan
yairi@nict.go.jp

Abstract. Mobility represents very basic and essential behavior for people: reaching a destination, strolling at will, and much more. Indispensable for independence, it also makes life more enjoyable. Yet moving from place to place may be difficult for disabled, elderly, or ill individuals affected by an impairment of sight, hearing, or lower-extremity motor function, which undermine abilities needed for mobility: recognizing things, controlling motor function, and accessing information. To offset this, countries and communities have been actively preparing systems and facilities in recent years to make routes barrier-free. But it would be unfeasible to make all routes barrier-free, and there continues to be a great need for mobility support with IT technology as an alternative means of assistance. We have been researching to put Robotic Communication Terminals (RCT) into practice, which supports the three elementary behaviors of recognition, actuation, and information access, targeting almost all the pedestrians including elderly and disabled people with various types, levels, and duration of disabilities. The RCT consists of three types of terminals: "environment-embedded terminal", "user-carried mobile terminal", and "user-carrying mobile terminal". These terminals communicate with one another to provide the users with a comfortable means of mobility. This paper introduces our recent research progress.

1 Introduction

Robotic Communication Terminals (RCT) which has been proposed by us since 1999 is a ubiquitous system for improving human mobility by making environment virtually barrier-free[1,2]. RCT was selected as a theme of Challenge for Realizing Early Profits (CREP) in 2000 - 2004 annual conference of the Japanese Society for Artificial Intelligence. The main purpose of the project is to develop elemental technologies for a ubiquitous system assisting independent and comfortable transportation for all pedestrians, especially targeting elderly and disabled people. Compared to those in 1999 when RCT was first planned

A. Sakurai et al. (Eds.): JSAI 2003/2004, LNAI 3609, pp. 61–75, 2007.

and proposed, the social and technological conditions have dramatically changed with mobile phones being used by general pedestrians for navigation system. Under these circumstances, RCT has been drawing a line against other human mobility support researches by setting an advanced target of "Developing human mobility support system to help recognition, actuation, and information access by actively gathering real world's information and by interacting with the users" and also by stressing the idea of "Pursuing the universal designs for human mobility support by targeting various users including elderly and disabled people".

Figure 1 shows the human mobility support by RCT. RCT consists of three types of terminals; "Environment-embedded terminals" installed at roads and stations, "user-carried mobile terminals" moving with the users, and "user-carrying mobile terminals" as intelligent vehicles. Environment-embedded terminals act like aerials and they have the functions of accumulating and providing dynamic information about obstacles and presence of people/cars in the pedestrian space by monitoring the surroundings of the places the terminals are set. User-carried mobile terminals are the advanced version of mobile phones and PDA. They have the interface using graphics, audio and touch sensors produced to fit the physical condition of the users and also the location detecting function, condition/intention detecting function, environment-information acquiring function and the function that provides information which is converted and reprocessed to an appropriate media to the user's physical condition. User-carrying mobile terminals are the advanced vehicles to carry elderly and disabled people. They are built to fit the users physical conditions. They have the driving assisting function enabling users to drive safely. These different types of terminals communicate each other to connect real world, computer network and users, and

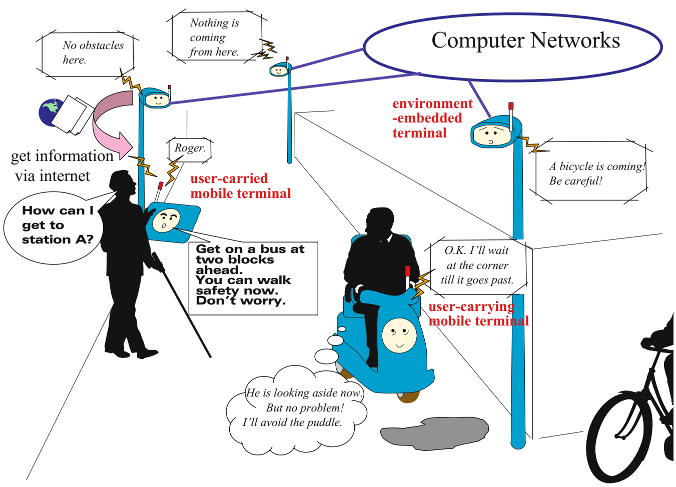


Fig. 1. RCT as a human mobility support system

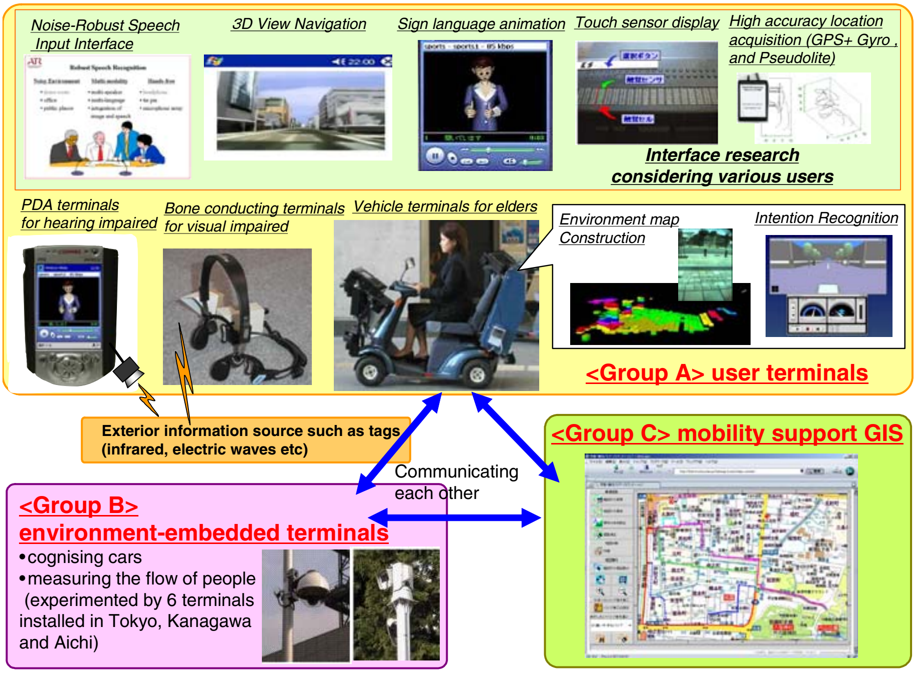


Fig. 2. Overview of RCT research

support the three elementary behaviors of recognizing things, controlling motor function, and accessing information, and assist the users to move around in the city.

In the project, we have mainly been focusing on the following two points; (1) to develop the elementary technology of the way of acquiring and providing real world information by considering various physical conditions of users through producing prototype of the terminals, (2) to propose universal design of the pedestrians' walking space contents, and aim for putting Mobility Support GIS (Geographic Information System) into practical use. This paper introduces recent research progress of our project. Figure 2 shows the overview of RCT researches we have done. Group A and B in Figure 2 are researches on the RCT terminals, which are presented in Section 2. Group C in Figure 2 is the research on Mobility Support GIS, which is described in Section 3. The trend of Japanese and oversea researches on pedestrians' mobility support are discussed in Section 4. Section 5 illustrate our future plan and the future of mobility support.

2 Researches on Terminals

This section introduces researches on developing terminals of RCT in Figure 2, group A and B. As the prototype of user-carried mobile terminal, two types were

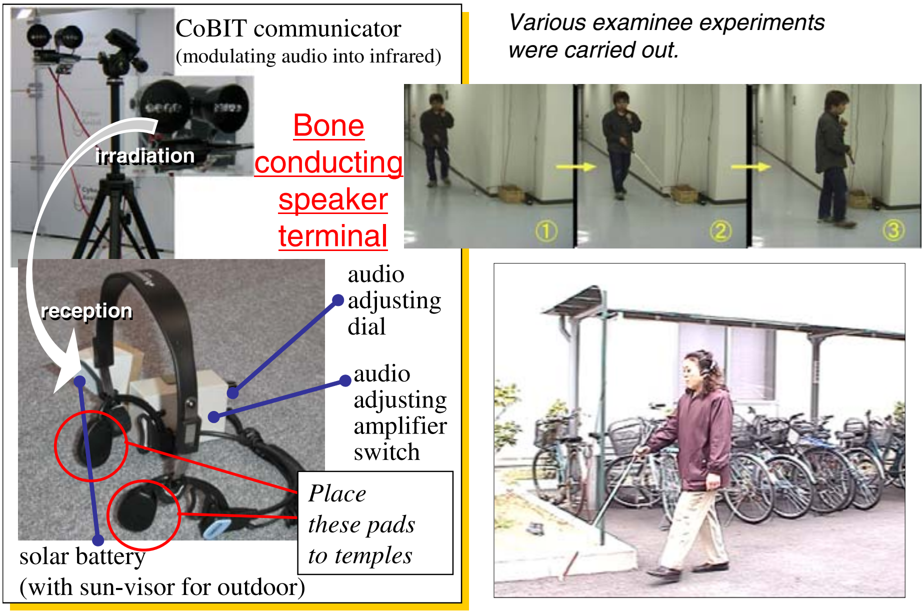


Fig. 3. Bone conducting headphone-type mobile terminals for blind people

made. One is the PDA terminal for deaf people which provides information by dactylology animations[3]. Another is the bone conducting headphone-type simple mobile terminal for blind people which provides information in audio received in infrared and AM radio waves as shown in Figure 3[4]. For the prototype of user-carrying mobile terminal, we developed a semi-autonomous scooter (named as ICW; Intelligent City Walker) targeting active elderly people with driving assisting functions such as detecting obstacles and automatic avoidance. Figure 4 shows the overview of ICW. At the same time, to use ICW as a data collecting system in pedestrian space, we are researching about producing environment maps using quaternion syntax and stereo image rows[5]. We are also carrying out a research on local and micro pattern learning modules from multidimensional time-series data for recognizing vehicle users' intentions, which are a very important function for semi-autonomous vehicles that help providing information and manoeuvring assistance[6]. Moreover, as an interface research for mobile terminals targeting various users, we have been researching on audio cognition under disturbing noises, navigation using 3-D view, producing dactylology animation, touch sensor display, and high accuracy location acquisition. For environment-embedded terminals, there are three prototype systems in NICT in Tokyo, one in Kanagawa and two in Chubu University in Aichi to develop the way to deal with various road conditions, weather and times in Figure 5[7].

We are checking the effectiveness by establishing network system assisting users transportation with linking environment and mobile terminals and by carrying out navigation experiments of blind people. We had been showing the

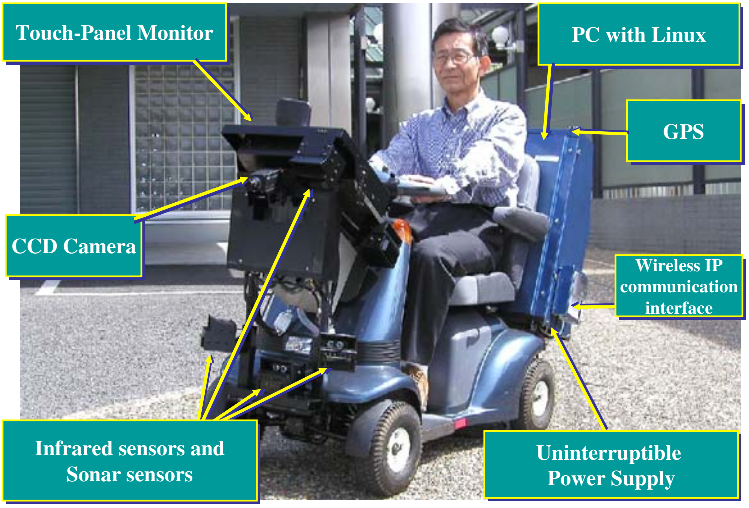


Fig. 4. User-carrying mobile terminal prototype ICW for elderly people

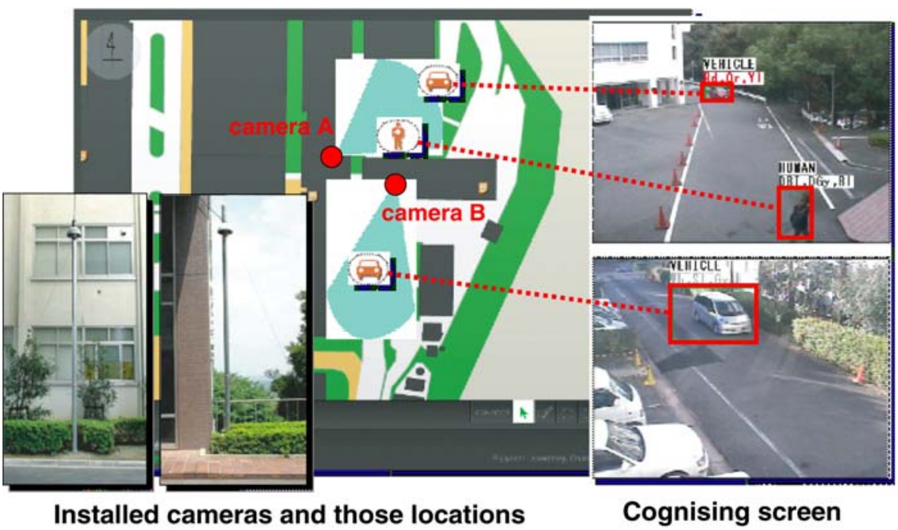


Fig. 5. Environment embedded terminal prototype in service (Chubu university campus)

results to the general users through conventions, open demonstrations and newspaper press. Figure 6 is the image of NHK's 19:00 news broadcasting the demonstration in March 2002, when ICW automatically stopped when it received the road information warning "A possibility of a car and the ICW crashing into each other at the blind intersection". At the same time, the independent terminal was

Fig. 6. Demonstrations of the interface between terminals (broadcasted in NHK news, March 2002) Caption: "Look out! Avoids automatically"

introduced during the demonstration of ICW broadcasted in NHK's live variety programme in May 2003, and was invited to the prime minister's office in January 2004 to demonstrate the simpler headphone-type terminals. We are going to join the human mobility support project of the Ministry of Land, Infrastructure and Transport (MLIT) using the knowledge of the elementary technology.

3 Research on Information Server for Pedestrians

We are going to introduce the Mobility Support GIS research in Figure 2; Group C.

3.1 Development of Mobility Support GIS

Mobility support GIS for pedestrians is information providing server for not only a mobile terminal proposed in RCT research. That is a system aiming to provide information on barrier/barrier-free in comprehensive pedestrian space related to choosing destinations and routes for all the pedestrians including elderly and disabled people who use the terminals such as mobile phones having the Internet features on. To be enacted and to become popular as a business, the most difficult aspect during the development was that we had to overcome the differences in the perspective of barrier and barrier-free which really depend on people's physical conditions and tastes, but still had to come up with a "universal design on data

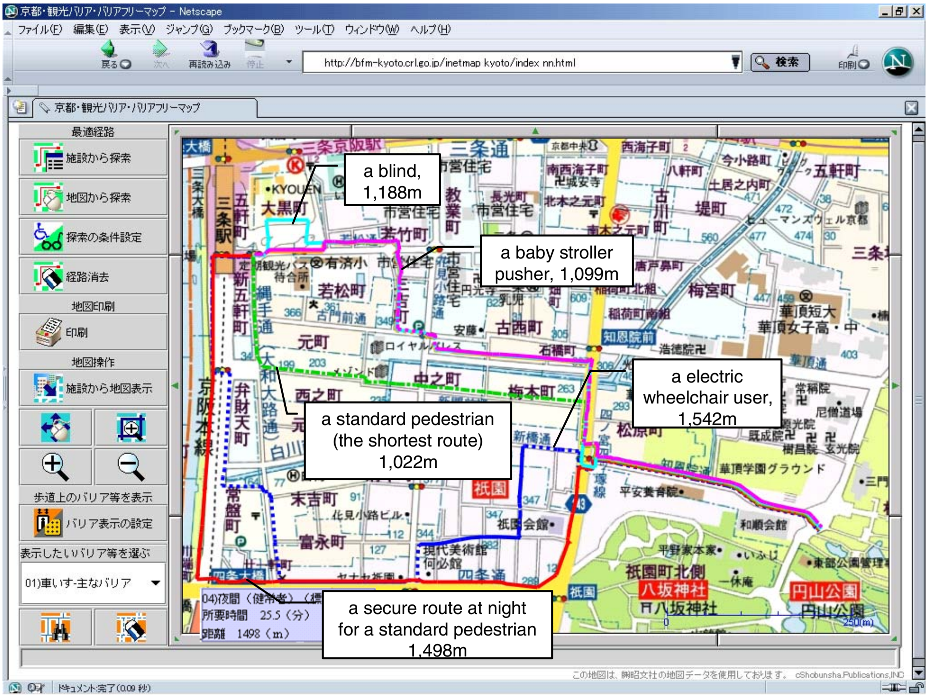


Fig. 7. Higashiyama, Kyoto's sightseeing area prototype, examples for the results of searching the most appropriate route (available on the Internet)

structure" that satisfies both the demands of pedestrians and the methods of researching and accumulating realistic data.

We have investigated the data structure and data accumulation methods by listening to the elderly and disabled people's opinions and creating GIS prototypes[8,9]. The first GIS prototype was developed by gathering data of an area as large as 12 km^2 including Koganei city in Tokyo and around the JR Kokubunji station's north exit (Kokubunji city), and was released as the "Koganei Barrier/Barrier-free map"on the Internet in May 2003[10]. However, the assessing experiment of the first prototype with 22 people working over 4 hours each, and the data research practice on the road, have brought us to yet another issue to solve, which we started reviewing the data structure and data accumulation methods. We came up with the second prototype, "Kyoto's Higashiyama area Sightseeing spots; Barrier/Barrier-free map (the Kyoto BFM)" which was released on the Internet in December 2003 to have a good press in newspapers[11]. The second prototype stores revised pedestrian network data in 2km^2 of Higashiyama area in Kyoto that includes popular sightseeing spots such as Kiyomizu-dera temple, Chion-in, Kohdai-ji, Gion, Shijyo and Shirakawa. Figure 7 shows the most appropriate route from the Keihan line's Sanjyo station to Chion-in which was searched by using the search parameter samples for

Fig. 8. Higashiyama, Kyoto sightseeing area prototype, search examples on Barrier-Barrier-free matters

various physical conditions such as "electronic wheelchair", "completely blind person", "baby stroller" and "physically unimpaired person at night time". Figure 8 is an example of the barrier/barrier-free objects search result with a pop-up window of the facility information of Yasaka-jinja Shrine. Kyoto BFM has become a research results to demonstrate a universal design of the data structure and system establishing methods.

3.2 Standardization and Practical Application Plans of the GIS

Mobility Support GIS's research result is valued highly as we are asked to participate in the JIS meetings to standardize it. The aspects that need standardizing are data structure, methods of data maintenance to assure the quality of accumulated data, and interface designs of searching. From now on, we need to review the discussions of consortium including disabled people's associations and private companies, and to maximize the opinions of both users and system producers to standardize it as a JIS and as a worldwide standard.

The adoption and operation of Mobility Support GIS can be divided into three steps of "Investigating the basic policy", "System establishment" and "Responding to the users' opinion". To become a real system to be used by real people, it is crucial to continuously correspond to the changes in cities and demands for added functions. Hence, we are thinking of putting the system into practice by local governments adopting it and local people to successively operate it. The process of local governments and people operating the pedestrian assisting GIS together can be described as "Community building by local people" which is from a brand new viewpoint of using the information-communication technology for local areas. The effectiveness of that is not only the direct ones like human mobility itself getting safer and more comfortable, but also (1) effectively distributing administrative resources to maintain pavement based on the information on barrier/barrier-free, (2) creating new jobs by disabled people participating in the scheme of receiving compensation for collecting data, renewing data and providing service, (3) creating markets for data maintenance, maintenance equipments, (4) mutual understanding between pedestrians with different physical conditions and the enlightenment of barrier/barrier-free, (5) maintaining the pedestrian space by observing every corner of the pedestrian space with people being more aware of them, which will lead to an improvement in scenery of the area and we can also expect its secondary effect of having less crime. In the future, we are aiming to propose the standardization of universal design of data structure for providing pedestrian space contents, and to simplify and promote efficiency of adopting and operating system by publishing the know-how guidelines of system adoption and operation and by developing and selling the tools of system adoption and operation.

4 Japan and Overseas Researches on Pedestrians' Mobility Support

4.1 National Project

RCT is an ITS (Intelligent Transport System) for pedestrians and it does not only target a portion of disabled people but also all the pedestrians and advocates the universal design of the service. The country's policy for pedestrian ITS is represented by following typical national projects; (1) the MLIT aiming to explain the concept of pedestrian ITS by demonstrating basic systems with private companies (from 2000)[12], (2) IT barrier-free project for disabled people by the Ministry of Economy, Trade and Industry aiming to demonstrate human mobility support on-site at the Aichi Expo (from 2003) with the trial of pedestrian ITS at city area of Aich[13], and (3) the autonomous human mobility support project by MLIT aiming to establish the system at Kobe (from 2004)[14], which we have always been participating through exchanging information and by being in the committee. The characteristics of these national projects are that appliance manufacturers and mobile phone companies bring their cutting-edge terminal technologies and do some demonstrations, which they seriously think about standardizing them and moreover to practice the service, terminals and ways of telecommunicating. On the other hand, the overseas' ITS related national projects focus on developing traffic lights for the safety of pedestrians and

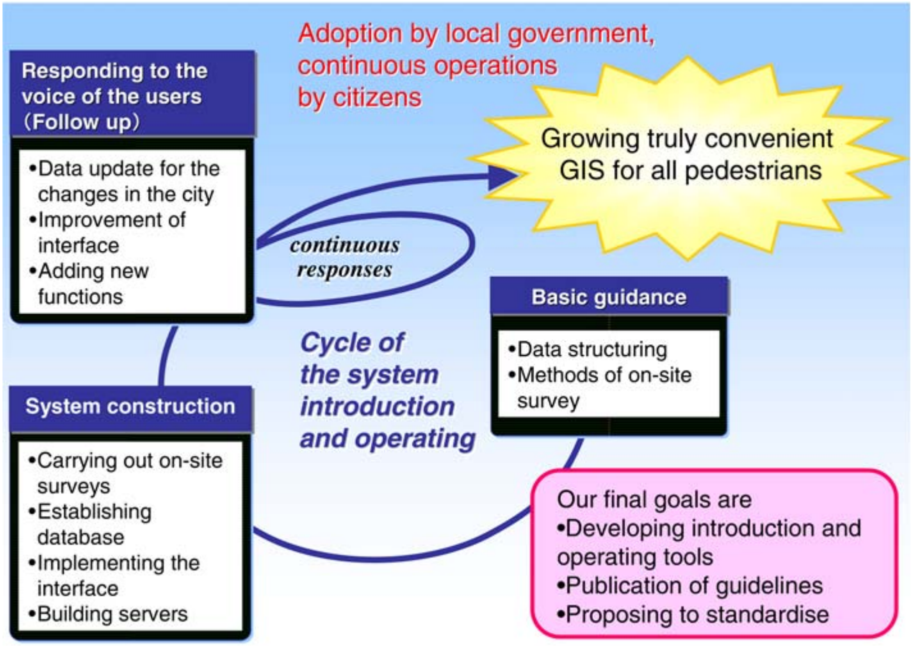


Fig. 9. Practical application of pedestrian-assisting GIS and the future issues

bicycles, or designing the roads which is more urban planning. We can be cynical by saying "Japan has abandoned urban planning and give over the future of foot-traffic to the high-tech field which they are good at". But basically, it is peculiar situation of Japan that the main transportation systems being only public transport and foot despite for metropolitan cities like Tokyo which has a huge urban area with overlapping satellite towns. In other words, Japanese metropolitan area is one of the very rare places that business can work out perfectly by pedestrians' transport assistance.

4.2 Academic Research

There are two academic research field related to the human mobility support; the terminal research and the contents research. The terminal research employs strategies of seeking opportunities where appliance manufacturers cannot reach, such as grand challenges of Human Interface, Robotics and Artificial Intelligence[15,16]. The characteristics of HI research is that they pursuit the navigation interface suited for advanced spatial perception mechanism such as sensory integration, cognition of spatial structure and processing kinetic systems[17]. Designing navigation interface for blind people who have limited visual expressions is still a very difficult task and we hope it develops in the future[18]. Robot researches are mostly developing intellectual interface of wheelchair and guiding robots for blind people[19,20]. For the background of the robot researches, there is this huge research theme of pursuing real time pattern recognizing technology which can become a substitution for environment recognition and manoeuvring. There still is the issue of safety before it gets practiced, due to people worried about something critical being substituted by robots, and that concern has to be overcome by having better interactions between users and robots and to have better coordination with implanted environment infrastructures such as maps and electric tags.

For contents research there are fields of GIS, VR, and image processing. GIS research is focused on developing GIS for providing information for pedestrians like us, and VR research is focused on developing 3-D contents technology for cities using animations and live-actions images[21]. These researches are being the important tasks for data maintenance and development of renewal method to be put into practice. At the same time, image-processing research had been focusing on pursuing image-processing algorism with having the application of road monitoring, but nowadays they are researching in a larger picture such as delivering movies and still images to users and accumulating information on maps and aiming to put them into practice[22].

5 The Future of Human Mobility Support

Human mobility support is a service based on the location of the users, and it is assumed the society with highly developed ubiquitous network and computer like the "network everywhere, computer everywhere" slogan. At the same time, it is appropriate as an integrated application for various technologies to actualize the ubiquitous technology society such as terminal technology, telecommunication technology, positional-information acquisition technology and contents providing technology. Nowadays, ubiquitous network and computer is one of the most active field and as being the important keyword for Japan's economic recovery and has been investing capitals and human resources nationwide, in private companies and in universities. While it is only starting up as a business, we should spread out the human mobility support without pressuring neither the private companies and the users by setting Dejours standards of universal design friendly to elderly and disabled people who are receiving less benefit from the advanced information society. For example of a JIS standards that are having this idea, there is the "Guidance of assisting system by audio assistance for walking and transportation of blind people" that is related to the information provision and there are four information especially set; warning information, location information, route information and territory information[23]. These types of information are also important for physically unimpaired people and other pedestrians. We hope that the society focuses on the perspective of the elderly and disabled people about the human mobility support's universal design and that becomes standardized and practiced.

6 Conclusion

RCT is one of the demonstration experiments to assist human mobility of the elderly and disabled people. For the background, it involves academic study fields such as recognizing motion picture images, artificial intelligence, human communication, computer graphics, welfare engineering and language processing. To become an admissible tool by the users, there are a lot of research tasks to overcome. It is a goldmine of research topics. The first thing to do is to dynamic renewal of barrier and barrier-free information which is hindering to use pedestrian assistance GIS. Dynamic renewal system is shown in Figure 10. Once this system gets put into practice, it enables to renew the ever-changing the live city's information such as landmarks on the given routes, changes in pedestrians space (roadworks, illegally parked bicycles), the assessments of the simplicity and enjoyment of passing the roads based on the image information from the mobile object in the city such as taxis or the information transmitted by users using RCT portable terminals or mobile phones.

The second issue is the research on navigation targeting children. The general idea is shown in Figure 11. This field has not been well researched in the past, however children possess physical and behavioral characteristics of having low/narrow viewpoints and poor sense of direction. We have to realize that there are some dangerous areas on the roads due to their viewpoints and sense of direction. Also there are still a lot of issues to be solved such as understand-

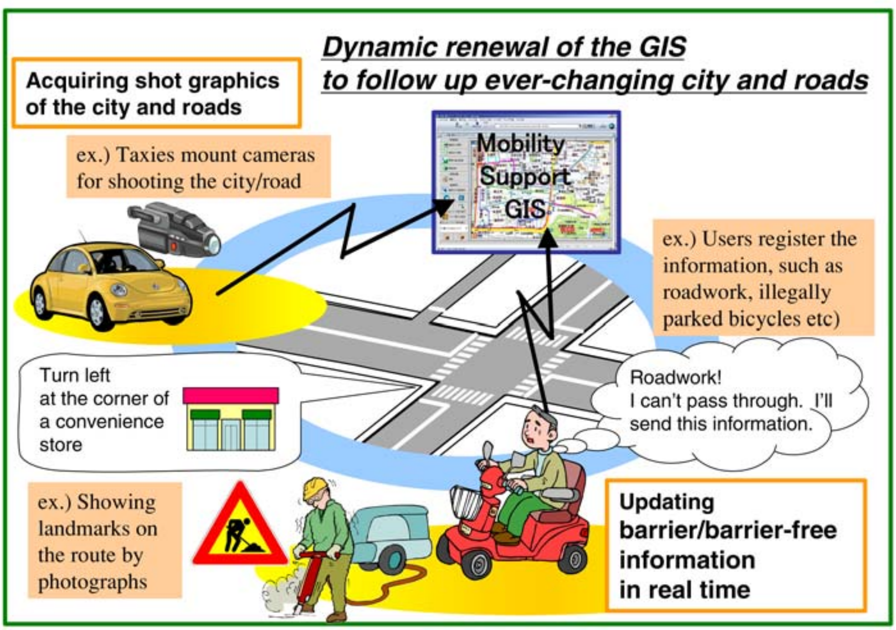


Fig. 10. Dynamic data renewal of GIS

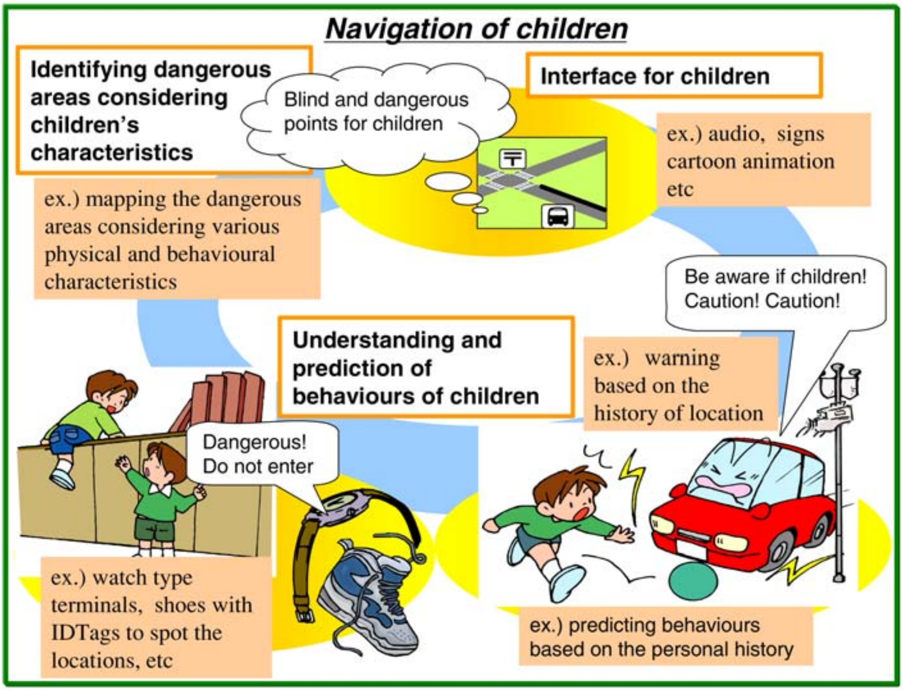


Fig. 11. Children navigation

ing/estimating children's behaviors and nonverbal interface. The problems with low viewpoint apply for the people on wheelchairs as well, and that with poor sense of direction also apply for adults.

Upon researching the RCT, we have been focusing especially on involving the elderly and disabled people into the development and ask them for their experience, ingenuity and knowledge related to the behavior based on the clear awareness and strong demand in everyday life. The researchers, who are categorized as healthy and young people, cannot have the perspective for the inner nature of the transport assistance. We are aiming to develop a useful transport assistance technology not only for elderly and disabled people but also for all people. We are going to continue examining and experimenting on examinees to improve the RCT.

References

1. I.E. Yairi and S. Igi,"A Mobility Support System Assisting the Elderly and the Disabled by Connecting the Real-world, Computer Networks, and Users," in proceedings of the 7th International Workshop on Mobile Multimedia Communications, 3B-3, pp. 1-6, 2000.

2. I.E. Yairi and S. Igi,"Robotic Communication Terminals as a Mobility Support System for Elderly and Disabled People(3)", Trans. of the Japanese Society for Artificial Intelligence, vol. 18, no. 1, pp.29-35, 2003(in Japanese).
3. Seiji Igi, Mitsue Tamaru, Yasunori Yamamoto, Megumu Ujitani and Shigeharu Sugita, "Sign Language Synthesis for the Mobile Environments," in proceedings of WSCG 2003, Plzen, Czech Republic, February, 2003.
4. Shinya Oyama, Ikuko Eguchi Yairi, Seiji Igi and Takuichi Nishimura, "Walking Support Using Compact Terminals with Infrared Audio Communication," in proceedings of ICCHP 2004, pp. 468-475, Paris, France, July, 2004.
5. Kayama, K., Yairi, I.E., and Igi, S., "Construction of Elevation Map for User-Carried Outdoor Mobile Robot using Stereo Vision," in proceeding of 2003 IEEE International Conference on Systems, Man and Cybernetics, pp. 4631–4636, Washington D.C., USA, Oct, 2003.
6. Ikuko Eguchi Yairi, Takehisa Yairi, Seiji Igi: Intention Recognition for Vehicle Driving by Sensing of User and Environment. Applied Informatics 2003: pp. 166-171, The 21st IASTED International Multi-Conference on Applied Informatics (AI 2003), February 10-13, 2003, Innsbruck, Austria.
7. Kayama, K., Yairi, I.E. and Igi, S., "Road Observation System for Robotic Communication Terminals supporting Pedestrians," in proceedings of the IEEE 5th International Conference on Intelligent Transportation Systems, Vol.1, No.O6-4, pp.635–640 (2002).
8. I.E. YAIRI, H. Yoshioka, et al., "Mobility Support GIS with Route Accessibility Information for Pedestrians," Trans. of Human Interface Society, vol.5, no.4, pp.413-420, 2003(in Japanese).
9. I.E. Yairi and S. Igi,"Accessible Route and Area Information Support GIS for All Pedestrians Including the Elderly and the Disabled," in proceedings of 11th World Congress on ITS October 18-24, 2004, Nagoya, Japan.
10. http://bfms.nict.go.jp/koganei/index.html(in Japanese)
11. http://bfms.nict.go.jp/kyoto/index.html(in Japanese)
12. http://www.mlit.go.jp/road/ITS/
13. http://www.meti.go.jp/information/data/c30319aj.html (in Japanese)
14. http://www.jiritsu-project.jp/(in Japanese)
15. R.G. Golledge, R.L. Klatzky, et al., "A geographical information system for a GPS based personal guidance system," International Journal of Geographical Information Science, Vol. 12, No. 7, pp.727-749, 1998.
16. T. Hatakeyama, F. Hagiwara, et al., "Remote Infrared Audible Signage System," International Journal of Human-computer interaction, vol.17, no.1, pp.61–70, 2004.
17. van Veen, H.A.H.C., H.K. Distler, et al.,"Navigating through a virtual city: Using virtual reality technology to study human action and perception," Future Generation Computer Systems 14, pp.231-242 (1998).
18. Rieser, J.J., Hill, E.W., Talor, C.R., Bradfield, A., and Rosen, S. (1992). Visual experience, visual field size, and the development of nonvisual sensitivity to the spatial structure of outdoor neighborhoods explored by walking. Journal of Experimental Psychology: General, 121(2), 210-221.
19. E. Prassler, J. Scholz and Paolo Fiorini, "A Robotic Wheelchair for Crowded Public Environments," IEEE Robotics and Automation Magazine, Vol. 8, No. 1, pp.38-45(2001).
20. S. Shoval, I. Ulrich and Johann Borenstein, "NavBelt and the GuideCane," IEEE Robotics and Automation Magazine, Vol. 10, No. 1, pp. 9-20(2003).

21. Ogawa, T. and Tsukamoto, M., “Tools for Constructing Pseudo-3D Space on the WWW using Images,”New Generation Computing, Vol. 18, No. 4, pp. 391-407, 2000.
22. A. Argyros, P. Georgiadis, et al., “Semi-autonomous Navigation of a Robotic Wheelchair,” J. of Intelligent and Robotic Systems, Vol. 34, No. 3, pp. 315-329, July, 2002.
23. http://www.meti.go.jp/information/downloadfiles/c30319b04j.pdf(in Japanese)

Workshop on Agent-Based Modeling

After the success of the first workshop of Agent-based approach in Economic and Social Complex Systems (AESCS) in Shimane, 2001, we have held a similar workshop in Niigata, 2003. The title is International Workshop on Agent-Based Modeling. The aims and scope are summarized as follows: Agent-Based Modeling has become one of major techniques to design and analyze complex adaptive systems including societies, economics, organizations, business management, Web applications, and the other engineering fields. The objective of the Workshop is to continue the efforts to foster the formation of an active multi-disciplinary community on multiagent, computational economics, organizational science, social dynamics, and complex adaptive systems, in conjunction with the 17th annual conference of JSAI, the largest AI related annual conference in the Pacific Asia region.

At the workshop, we have had twelve oral presentations. They were pearly reviewed after the workshop, and then we have decided to accept the following five papers for the inclusion of these post-proceedings: PjS. Tomita, A. Namatame: Bilateral Trading with and without Strategic Thinking QjK. Miyanishi, K. Suzuki: Cooperative Behavior with Common Information Controller in Minority Game RjK. Izumi: Analysis of Efficiency and Accuracy of Learning in Minority Game SjK. Yuta et al.: Sectioned Random-Network Agent Model Demonstrates Behavior of Sectionalism TjM. Kunigami, T. Terano: Interaction and Control in Learning Agent System

The selected papers cover the wide area of agent-based modeling and have shown the effectiveness of the approach to both theoretical and practical issues of the domains.

Following to the workshop, we are continue to have the AESCS workshop series in 2002, 2004 and 2005, for Pacific Asian regions. Furthermore, in August 2006, integrating European, American, and Pacific-Asian academic associations, we held the First World Congress on Social Simulation in Kyoto. The post-proceedings will be also published from Springer Verlag, very soon in 2007. We believe that agent-based modeling for social complex systems is quite promising for the future Artificial Intelligence researches.

September 2006 Takao Terano

A. Sakurai et al. (Eds.): JSAI 2003/2004, LNAI 3609, p. 79, 2007.

Mechanism Design for Environmental Issues

Shinji Tomita and Akira Namatame

Dept. of computer Science,
National Defense Academy, Yokosuka, 239-8686, Japan
{g41044,nama}@nda.ac.jp

Abstract. In this paper, we consider the problem of the mechanism design for the multi-agents system. We develop the social learning model for the mechanism design for creating the collective action with an efficient cost sharing rule. We consider the situation in which self-interest agents have incentives to cooperate each other for jointly acquiring the environmental level with sharing the necessary cost. We obtain the optimal level of the environment to be acquired and the cost allocation rule so that their individual rationality is satisfied, and at the same time the social rationality is also satisfied. We show that the factors such as the value (worth) of the environmental level perceived by each agent and the cost affect the level the collective action. A social rule of allocating the common cost among agents is developed with decentralized transaction mechanism. We formulate and analyze the problem of cooperating multiple agents under uncertainty. We show that when agents cooperate in order to encounter uncertainty when acting alone, their benefits would not be as attractive, and hence cooperate to share the risk. As a specific example, we consider the model of obtaining the environmental level by sharing cost. We propose the negotiation mechanism for sharing cost among agents. With that mechanism, they can learn and obtain the unbiased and fare cost distribution rule.

1 Introduction

Like a human individual or an organization, an agent is a theoretical concept constructed to exhibit the property of an autonomous entity, which seeks to achieve a self-interest goal. They may form an organization because of the joint interest for efficient resource acquisition or allocation. They may form an organization for sharing the necessary cost of the common infrastructure [2][3][4][6].

In this paper, we consider the collective action problem of multi-agents. When acting alone without cooperation, their standalone behavior would not be as attractive, and hence the collective action can be attractive. Collective action for sharing cost is promoted here by firstly satisfying each agent's individual rationality with the appropriate distribution of the cost that could not be accepted by each single agent incurred. Agents may benefit from the collective action of sharing the cost.

We formulate the optimal problem of acquiring the environmental level with sharing the cost. The very basic question is then stated what level of the environment or infrastructure should be acquired as the collective action and how agents negotiate to share the common cost. We show that the factors such as the worth and the sharing

A. Sakurai et al. (Eds.): JSAI 2003/2004, LNAI 3609, pp. 80 – 94, 2007.

cost of the environmental level perceived by each agent may affect the formation of the collective action. We obtain the mechanism of an efficient and unbiased sharing rule of the common cost. We show such a social rule varies with the changing incentive of each agent. Therefore in order to form the optimal collective action, it may require that their interest or incentive is not jeopardized, and the conditions of both the individual rationality and the social rationality should be satisfied. However, conflicts may occur when agents behave in such that the cost accordingly to the declared value of each agent.

We can also consider a society of agents each of them behaves with the common goal. The key element that distinguishes such a common goal from an agent's individual goal is that it requires a collective action. These collective action problems also pose difficult quandaries for a society. However, each member of the society who wants to attain a common goal, may be tempted to benefit from goal without contributing to the common goal. There is a long history of interest in such social dilemmas, known as free riding problems of this type in many fields [3]. It may require same enforcement in the form of social rule in order to achieve a common goal [1]. The problem of incentive compatible may occur when these agents behave in such that the cost accordingly to each value possessed.

In these cases, an efficient and unbiased social rule is thus required to evaluate the best sharing rule among the agents to even out the conflicts. We will discuss the collective action problem with basing the mechanism design. With this mechanism, the optimal level of the environment can be acquired. The mechanism also provides an efficient and unbiased social rule for sharing the common cost necessary for acquiring the environmental level at the optimal level. We design the model of social learning in which an efficient and unbiased cost sharing as a social rule is obtained through the decentralized adjustment process.

2 Environmental Issues of the Collective Actions

When agents deal with one another, they often bring to the encounter differing goals, and the interaction process takes conflict into account. Each agent pursues its own goals through encountering with other agents; arrangements should be made so that each individual's goal can be satisfied. Agents also may promise, threaten, and find compromises together that will satisfy all agents.

The very basic question is then stated as why and how do agents should cooperate than to act independently. There can be many answers to this multi-edged question, but generally, it can be deduced that agents cooperate in order to share the common benefit or the certain load which each individual cannot fulfill alone. The agents benefit from the cooperative behaviors of sharing the cost or a load where they cannot fulfill alone [5][7].

In this paper, we consider the cooperating problem of multiple agents under uncertain situations. As a specific example, we consider the problem of cooperating agents in the environmental maintenance model. The price is not set on the loss by environmental pollution and when the polluter produces and consumes without considering the loss, environmental pollution as the external diseconomy is generated. For many sorts of pollution, particularly that of the atmosphere or sea, it is fairly accurate to say that a polluter cannot choose to pollute one group of agents rather than another, that is, pollution can be said as a pure public bad and hence pollution reduction as a public good [8]. In this situation, two or more agents cooperate to share the cost of obtaining certain environmental level from environmental maintenance activity. The agents benefit from the cooperative behavior in maximizing their utility. When acting alone by themselves, their benefits would not be as attractive, and hence they cooperate by sharing the cost.

We say, there is uncertainty involved here and agents may acquire certain environmental level together to share their risk or to lower their possible cost. Cooperative behavior under uncertain circumstances can be made possible by considering their utilities while acquiring certain environmental level. Cooperative behaviors under uncertainty is promoted here by, firstly satisfying each agent's individual rationality by appropriately distributing the worth and cost that could not be accepted by both or single agent incurred from acquiring the environmental level be accepted by all; and also by satisfying the agents' social rationality. The following factors like the value of environmental level possessed by each agent, the utility through acquiring environmental level should be considered throughout the negotiation. Similarly, conflicts occur when agents behave in such that the cost accordingly to each value possessed. In this case, an efficient and unbiased sharing rule of the common cost among the agents that evens out the conflicts should be designed.

Fig.1 illustrates the concept of the collective action to maintain an environment or the common infrastructure for multi-agents.

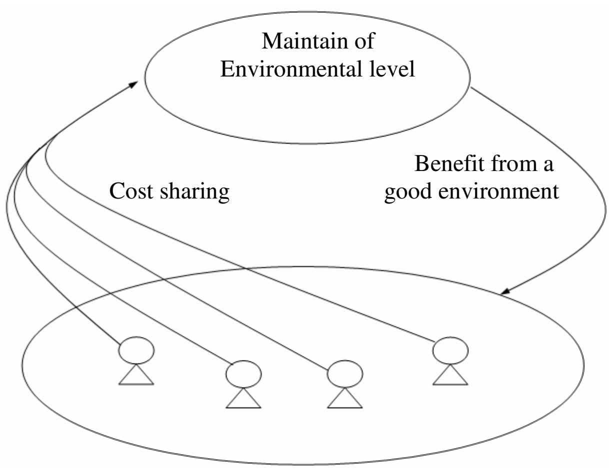


Fig. 1. The collective action problem of obtaining the environmental level

The very basic question is stated as why and to whom do agents cooperate than to act independently. How agents with the different worth of the environmental level attempt to share the common cost? The following factors like the value of the environmental level possessed by each agent, the utility through acquiring the environmental level should be considered. Each agent attains a certain worth from the environmental level, however, they are required to pay or share for every cost associated with the environmental maintenance activity. In this instance, each agent considers about his worth and cost of the environmental level, and finds out that by cooperating with other agents to share the cost so that his utility will be improved. In this case, the agent is said to be benefiting from the collective action with sharing the common cost.

3 Environmental Issues of the Collective Actions

In this section, we obtain the optimal level of the environmental level to be acquired by a society and investigate the properties of the collective action. Each agent assesses all the results of his behaviors, and this assessment is measured in terms of his utility. To participate in the collective action through negotiation, each agent first examines if his utility is improved by obtaining such the environmental level. In other words, individual rationality must first be satisfied. At the same time social rationality must be satisfied. This condition requires that when an agreement is reached, neither of the agent's utility will be increased without declining the other agents' utilities.

Condition 1: Individual Rationality
Each agent's utility is denoted as, $u_i(\Omega_i - c_i, Y)$, where $\Omega_i, i = 1,2,...,n$ represents the initial private resource, $c_i, i = 1,2,...,n$ represents the cost shared for environmental maintenance, and Y represents the environmental level. The condition of the individual rationality is then given as :

$$u_i(\Omega_i - c_i, Y) \geq u_i(\Omega_i, 0) \tag{3.1}$$

As a special example, we consider the following semi-linear utility function of each agent.

$$u_i(\Omega_i - c_i, Y) = \Omega_i - c_i + 2\alpha_i \sqrt{Y} \tag{3.2}$$

where α_i represents the benefit coefficient received from the environment. This is private information. If one agent has big value of α_i, he can receive big benefit from environment. Contrary, if another agent has small value of α_i, he receives small benefit from environment. Then the condition of the individual rationality is given as

$$u_i(\Omega_i - c_i, Y) - u_i(\Omega_i 0) = 2\alpha_i \sqrt{Y} - c_i \geq 0 \tag{3.3}$$

Therefore, each agent agrees to share the environmental level if the following condition is satisfied.

$$\sqrt{Y} \geq c_i / 2\alpha_i \tag{3.4}$$

Condition 2: Social Rationality
The condition of the social rationality requires that when an agreement is reached, no agent's utility will decline, and no one will be improved by any means. We specify the optimal solution that maximizes the summation of each individual utility in (3.3) as a solution satisfying the social rationality.

$$\sum_{i=1}^{n} u_i(x_i, Y) \geq \sum_{i=1}^{n} u_i(\Omega_i, 0) \tag{3.5}$$

where $x_i = \Omega_i - c_i$ represents the private resource. We can obtain the solution satisfying the condition of the social rationality by solving the following optimal problem:

$$\underset{Y}{Max} \sum_{i=1}^{n} u_i(x_i, Y)$$

$$s.t.\ F(X,Y) = 0, \sum_{i=1}^{n} x_i = X \tag{3.6}$$

where $F(X,Y)$ represents the kind of the production function of the environmental level Y in terms of the private resource $\sum_{i=1}^{n} x_i = X$. The optimal solution of (3.6) is given by solving the following function,

$$\sum_{i=1}^{n} (\partial u_i / \partial Y) / (\partial u_i / \partial x_i) = F_Y / F_X \tag{3.7}$$

As a special case, we consider the following quasi-linear utility function of each agent i, $i = 1,2,...,n$

$$u_i(x_i, Y) = x_i + v_i(Y) \tag{3.8}$$

We also define the exchange function of the environmental level to private resource by

$$g(Y) \equiv \sum_{i=1}^{n} \Omega_i - \sum_{i=1}^{n} x_i \tag{3.9}$$

The production function of the environmental level in terms of the private resource appeared as the constraint in (3.8) is defined using the exchange function $g(Y)$

$$F(X,Y) = X + g(Y) - \sum_{i=1}^{n} \Omega_i$$

$$= \sum_{i=1}^{n} x_i + g(Y) - \sum_{i=1}^{n} \Omega_i = 0 \tag{3.10}$$

Therefore the substitute rate of the environmental level to the private resource is given as

$$F_Y / F_X = g'(Y) \tag{3.11}$$

Similarly, the substitute rate of the environmental level Y to the private resource of agent i is given as follows:

$$(\partial u_i / \partial Y) / (\partial u_i / \partial x_i) = v_i'(Y) \tag{3.12}$$

Therefore, the optimal level of the environmental level Y^* to be acquired by the society is given as the solution that satisfies

$$\sum_{i=1}^{n} v_i'(Y^*) = g'(Y^*) \tag{3.13}$$

The sum of each agent's utility after obtaining the environmental level at the optimal level Y^* is then given as

$$\sum_{i=1}^{n} u_i(x_i, Y^*) = \sum_{i=1}^{n} x_i + \sum_{i=1}^{n} v_i(Y^*) \tag{3.14}$$

The sum of each agent's utility before acquiring the environmental level is given as

$$\sum_{i=1}^{n} u_i(x_i^0, 0) = \sum_{i=1}^{n} \Omega_i \tag{3.15}$$

Therefore if the following condition is satisfied

$$\begin{aligned} &\sum_{i=1}^{n} u_i(x_i, Y^*) - \sum_{i=1}^{n} u_i(x_i^0, 0) \\ &\quad = \sum_{i=1}^{n} x_i + \sum_{i=1}^{n} v_i(Y^*) - \sum_{i=1}^{n} \Omega_i \\ &= \sum_{i=1}^{n} v_i(Y^*) - g(Y^*) > 0 \end{aligned} \tag{3.16}$$

The agents should participate in the collective action and jointly acquires the environmental level. The effectiveness of such a collective action is given as

$$h(Y^*) = \sum_{i=1}^{n} v_i(Y^*) - g(Y^*) \tag{3.17}$$

As a special example, we consider the following case where

$$u_i(x_i, Y) = x_i + 2\alpha_i \sqrt{Y},$$
$$g(Y) = kY$$

$$F(X, Y) = X + kY - \sum_{i=1}^{n} \Omega_i \tag{3.18}$$

The optimal level of the environmental level Y^* to be acquired is given as the solution of

$$\begin{aligned} h'(Y^*) &= \sum_{i=1}^{n} v_i'(Y^*) - k \\ &= \sum_{i=1}^{n} \alpha_i Y^{-1/2} - k = 0 \end{aligned} \tag{3.19}$$

Therefore the optimal level of the environmental level Y^* is given as

$$Y^* = \left(\sum_{i=1}^{n} \alpha_i / k \right)^2 \quad (3.20)$$

$$h(Y^*) = \sum_{i=1}^{n} v_i(Y^*) - g(Y^*)$$

$$= \sum_{i=1}^{n} (\alpha^2) / k \quad (3.21)$$

Condition 3: free riding problem
The cost shared by each agent to obtain the optimal environmental level Y^* is given as

$$\sum_{i=1}^{n} c_i = kY^* = \sum_{i=1}^{n} (\alpha_i^2) / k \quad (3.22)$$

Distribution rule of the cost is undecided. In this case, it seems that each agent bears the cost according to his benefit coefficient received from the environment α_i is the most unbiased rule. Then the cost shared by each agent is given as

$$c_i = \left(\alpha_i / \sum_{i=1}^{n} \alpha_i \right) kY^* \quad (3.23)$$

However, because the cost depend on private information α_i and no one can know it, the problem is caused that private information is not honestly declared.

For instance, if agent *i* declares his private information as $\overline{\alpha_i}(\overline{\alpha_i} = \alpha_i - \beta)$ where β represents the amount of manipulation, the cost shared by him is given as

$$\overline{c_i} - c_i = -k\alpha_i\beta / \left(\sum_{i=1}^{n} \alpha_i - \beta \right) \left(\sum_{i=1}^{n} \alpha_i \right) < 0 \quad (3.24)$$

He can decrease the cost with false. From this, someone comes to manipulate his private information. Clark-Groves mechanism is known well as a method of solving this problem. Though as for this mechanism, the surplus is caused in formulation of the rule side. So the cost shared by each agent increases [9].

4 An Evolutional Approach for Designing Cost Sharing Rules

We now discuss about how the process of mutual learning is carried out in order to satisfy the individual rationalities of all agents. The problem of incentive is defined as the issue of declaring private information of each agent without false. Then the question is how to design the cooperation protocol compatible with incentive of each self-interest agent. We propose such a mechanism shown in Fig. 2 as follows:

Step1: Each individual agent i declares his utility for the environmental level $v_i(Y), i = 1,2,...,n$.

Step2: The socially optimal level of the environmental level is determined by solving

$$\sum_{i=1}^{n} v_i'(Y^*) = g'(Y^*) \tag{4.1}$$

Step3: The cost shared by each agent $c_i(Y^*), i = 1,2,...,n$, is determined as follows:

$$c_i(Y^*) = \left(v_i'(Y^*) / \sum_{i=1}^{n} v_i'(Y^*) \right) C(Y^*) \tag{4.2}$$

As a member of the society, the agent autonomic behaves in such as to improve his utility with the social competence to behave together with other agents. This is the basic foundation for the social learning.

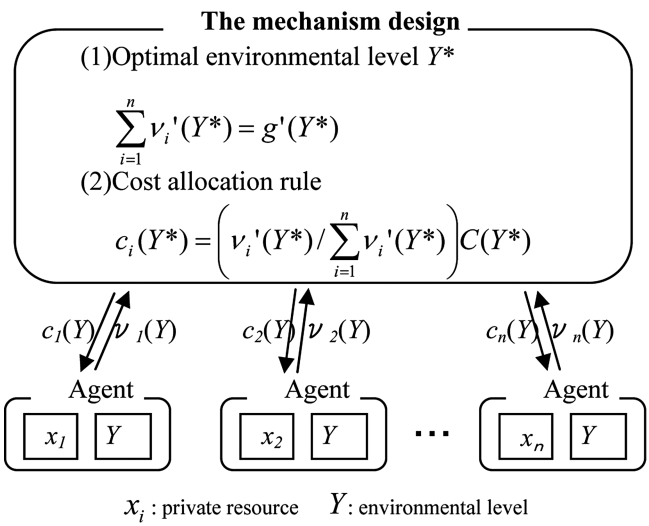


Fig. 2. The mechanism design for the common cost

The problem of incentive is defined as the issue of declaring private information of each agent without false. Then the question is how to design the cooperation protocol that is compatible with incentives of each selfish agent. In this section, we develop the mutual adjustment process among agents that compatible with each individual agent and it may reveal the true cognitive states. The preceding sections explained mainly about the rationality of the agents in cooperating with each other, and the cost distribution rules that governs the cooperative behaviors.

4.1 Proposed Model 1

In this section, we will discuss about how negotiation is carried out to satisfy the rationalities of the agents during cooperative behaviors. In our proposed model, there exists a negotiation manager within the same community of the participating agents as illustrated in Fig. 3. We show that Pareto-efficient can be achieved if there is no manipulator.

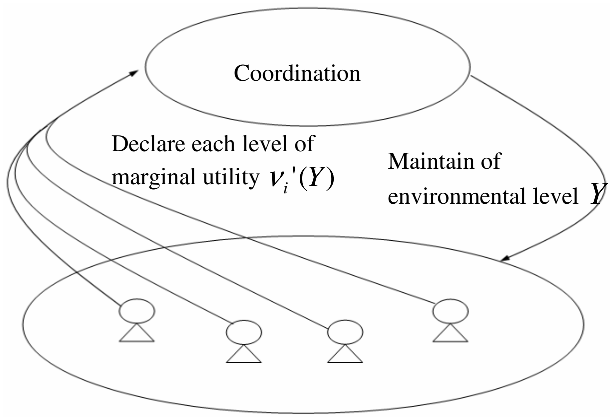


Fig. 3. The coordination of the cooperation protocol 1

Step1: Initially, every agent will receive environmental level Y from the negotiation manager upon agreement to negotiate.
Step2: Each agent calculates his marginal utility based on the benefit coefficient received from the environment α_i and declares it.
Step3: The negotiation manager proposes new environmental level Y_{t+1} based on the marginal utility declared by each agent according to the following rule.

$$\text{if } \sum_{i=1}^{n} \alpha_i Y_t^{-1/2} - k > 0 \text{ then } Y_{t+1} := Y_t + \delta (\delta > 0)$$

$$\text{if } \sum_{i=1}^{n} \alpha_i Y_t^{-1/2} - k < 0 \text{ then } Y_{t+1} := Y_t - \delta \tag{4.3}$$

where δ represents the adjustment speed and the cost shared by each agent is provided by the next expression.

$$c_i = \left(v_i'(Y_t) / \sum_{i=1}^{n} v_i'(Y_t) \right) k(Y_{t+1}) \tag{4.4}$$

Step4: Each agent calculates his marginal utility to new environmental level and declares it.
Step5: Repeat Step1 to Step4.

The point of this negotiation mechanism is that the cost shared by each agent is decided according to his marginal utility.

A simulation applying the above suggested protocol was performed to examine the behaviors of the agents during the negotiation process. In this simulation, when we sat the parameters of population $n = 9$, initial private resource $\Omega_i = 1000$, benefit coefficient received from the environment $\alpha_i = 6 \sim 14$ (rectangular distribution) and $k = 1$. We examine the validity of the model based on these optimal values. Fig. 4 represents the environmental level over time and Fig. 5 represents the utility over time.

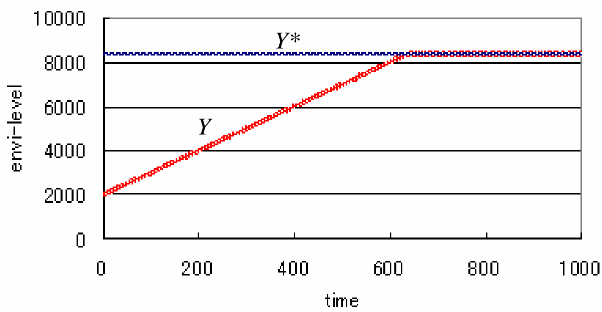


Fig. 4. Environmental level over time

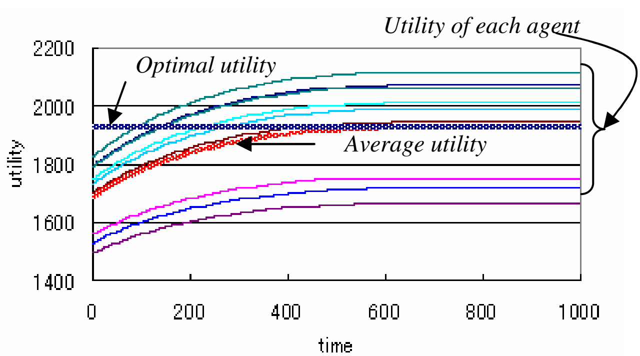


Fig. 5. Utility over time

From Fig 4,5 it is understood that environmental level Y and average utility are settled to optimal environmental level Y^* and optimal utility respectively. This result means that this model can achieve the Pareto-efficient, even if no one knows the optimal environmental level Y^*. However, this model has a potential problem that the free ride on the other's loads is possible by the manipulation of own benefit coefficient received from the environment α_i.

To show this potential problem, we put the manipulator in this simulation. The manipulator decreases his cost given as (3.24). In this simulation, we sat the amount

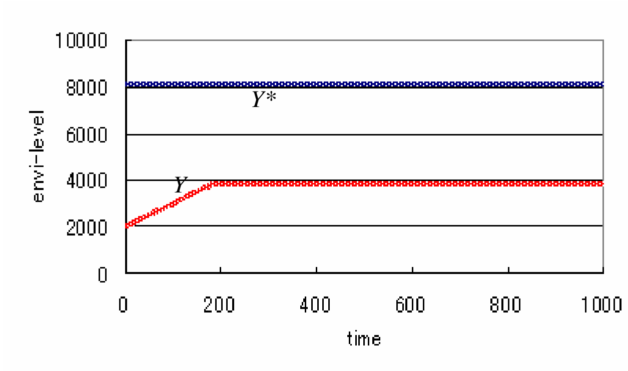


Fig. 6. Environmental level over time (The ratio of manipulator: 80%)

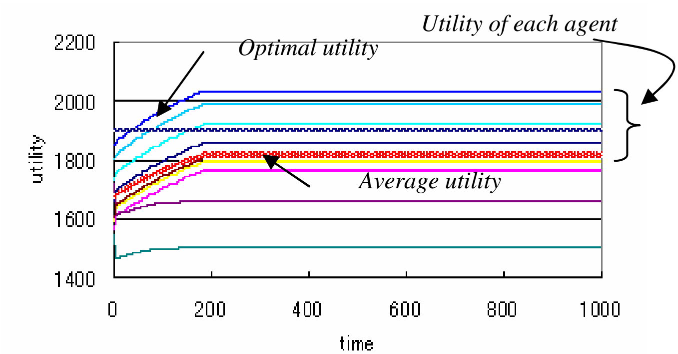


Fig. 7. Utility over time (The ratio of manipulator: 80%)

of manipulation $\beta = 4$. Fig. 6 represents the environmental level over time and Fig. 7 represents the average utility, when there are 80% manipulators.

From Fig.6,7 we can see that environmental level Y and average utility are not settle to optimal environmental level Y^* and optimal utility respectively. When a lot of people try to get a free ride, it becomes a society that most people lose.

4.2 Proposed Model 2

The point of problem is that we can not know each agent's private information. We can not find the manipulator and give them penalty individually. Then, we added the third party to model 1 as illustrated in Fig. 8.

We bring in the mechanism of all-or-nothing. The mechanism is that we added two steps after Step2 of Model 1.

Step2.1: if the amount of cost shared is lower than the proposed environmental level of third party, negotiation manager does not maintain the environmental level. The amount of cost is accumulated in negotiation manager while the environmental level is not maintained.

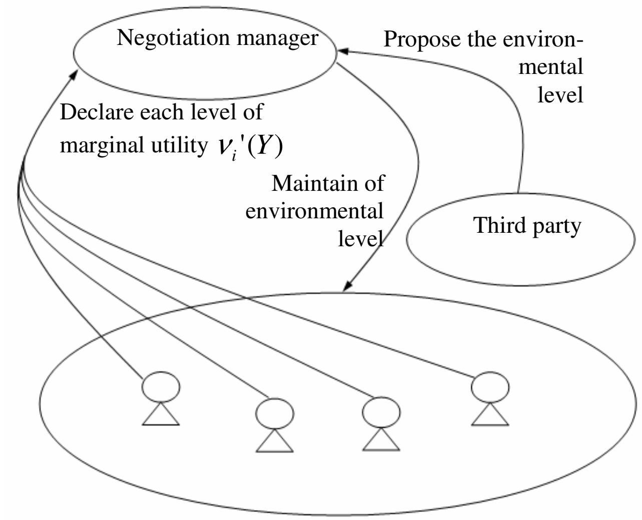


Fig. 8. The coordination of the cooperation protocol 2

Step2.2: if the utility of manipulator in certain time is lower than previous time, he improves his benefit coefficient received from the environmental level $\overline{\alpha_i}$ in ε. ε represents the amount of improvement.

If negotiation manager does not maintain environmental level, that manipulation is meaningless. Because manipulator can not obtain the benefit from an environmental level, his utility is lower than initial private resource.

We show the result of the simulation when we sat the parameter of the proposed environmental level of third party =6000, the amount of manipulation $\beta = 4$, rate of manipulator is 80% and the amount of improvement $\varepsilon = 0.05$. Fig. 9 represents the environmental level over time and Fig. 10 represents the average utility, when there are 80% manipulators.

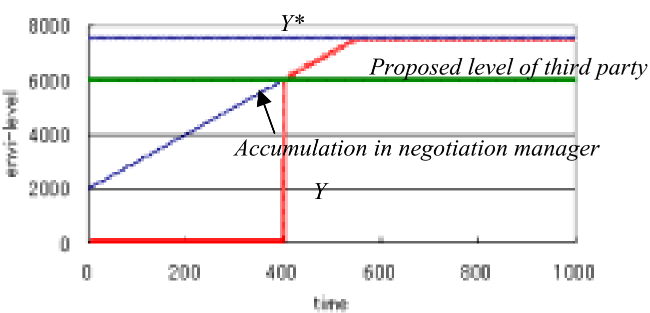


Fig. 9. Environmental level over time (The ratio of manipulator: 80%)

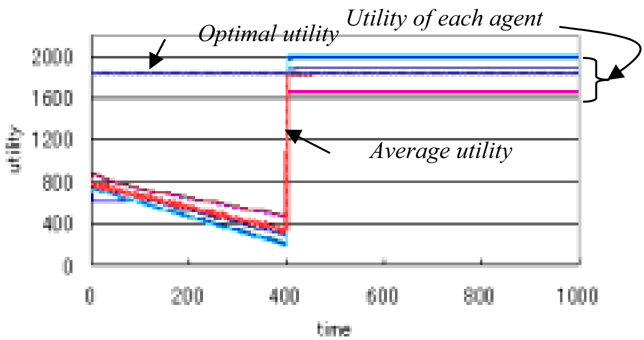


Fig. 10. Utility over time (The ratio of manipulator: 80%)

From Fig.9,10 we can see that environmental level *Y* and average utility are settled to optimal environmental level *Y** and optimal utility respectively. However the social meaning of the accumulated cost to becomes a problem. This should be small.

Next, we examined how an increase of the population and difference of the proposed environmental level by third party influences on the simulation. We set the parameter of population $n = 100$ and compare the proposed level 60000 and 600000. Fig 11,12 and Fig 13,14 are the results when the proposed level is 60000 and 600000 respectively.

The first understanding from these results, optimal levels have risen compared with the case of small population. This is a natural result. The point that should be paid attention, the big difference is caused by the difference of the proposed environmental level of third party. If the third party proposes the low level 60000, environmental level *Y* and average utility rise early. But they don't reach the optimal level. Oppositely, if the third party proposes the high level 600000, environmental level *Y* and average utility rise late. Especially, there is a problem of minus the utility. But they reach the optimal level.

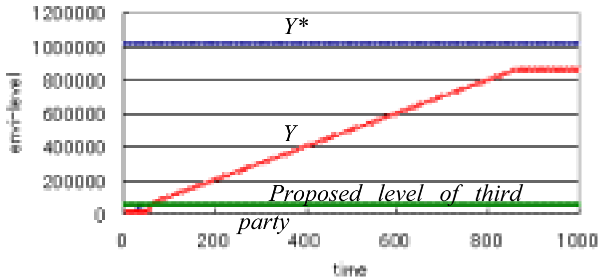


Fig. 11. Environmental level over time (The ratio of manipulator: 80%)

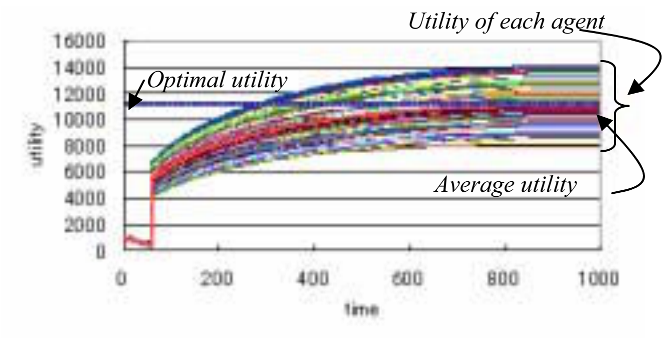


Fig. 12. Utility over time (The ratio of manipulator: 80%)

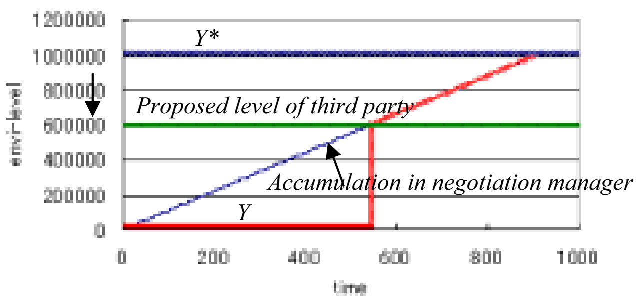


Fig. 13. Environmental level over time (The ratio of manipulator: 80%)

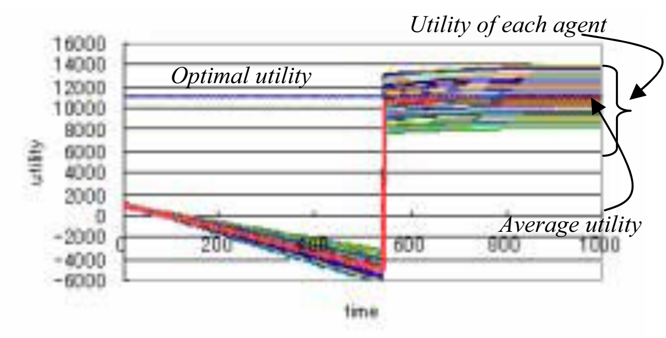


Fig. 14. Utility over time (The ratio of manipulator: 80%)

5 Conclusions

We started out with examining the nature of the collective action of agents, that is, why they cooperate, and then further discussed about how they cooperate in order to share the common cost. Next, we propose the negotiation mechanism for sharing cost among agents. We showed that this mechanism can obtain the optimal level of the environment according to the private information of each agent. We must consider the social meaning of third party. We think that non-governmental organization like NPO and NGO can perform the role of third party. However the problem still remains which way we should adopt the proposal, high level or low level. It is a trade-off relation. It also can be said that this is a social dilemma.

References

1. Campbell, Donald E, "Incentives", Cambridge University Press, 1995
2. Detlof von Winterfeldt, Ward Edwards, "Decision Analysis And Behavioral Research", Cambridge University Press, 1986
3. Fudenberg Drew "Game Theory", The MIT Press, 1991
4. Rosenschein,J. Zlotkin,G "Rules of Encounter, Designing Conventions for Automated Negotiation among Computers", MIT Press 1994
5. Gasser, L "Computational Organization Research", ICMAS'95 Proceedings, pp414-415, 1995
6. Zlotkin, G, Rosenschein, J .S. "Mechanism design for automated negotiation" Artificial Intelligence, Vol.86, pp195-244, 1996
7. Kumon, S " Information Civilization Theory " NTT Press, 1979
8. Sandeep Baliga, Erick Maskin "Mechanism Design for the Environment", 2002
9. Masahiro, O and Koutaro, S "Micro-economics Ⅱ" Iwanami press, 1988

Cooperative Behavior with Common Information Controller in Minority Game

Keiji Miyanishi and Keiji Suzuki

Future University - Hakodate
Kameda-Nakano 116-2, Hakodate City 041-8655, Japan

Abstract. In this paper the Minority Game is applied to examine the effects of the common information controller for generating the cooperative behavior among the agents. If the controller modifies the information accurately, it is expected that social profit can be raised even under the situation where coopration is difficult. Through the simulation the effects and the properties of the controller are examined.

1 Introduction

The research on an agent-based simulation is attracting attention in recent years. An agent based simulation is the method of analysis and prediction of the group behavior which is caused by the interaction among the agent based on its individual decision making[1]. It is used effectively in research of a social system, for example to clarify and analyze the phenomenon like cooperation or competition, or to indicate the system design for an organization improvement, or to investigate the mechanism which produces harmony and so on[2].

In this paper, the minority game, which is one of the game models, is taken up. In the game, agents compete for limited resources based on common information. It is known that even if the agents behave selfish the agents as a whole can get high rewards by using appropriate length of information. But if the length becomes shorter or longer, agent can't get high rewards. We are interested in how cooperation can be caused in a situation where cooperation is difficult. One solution is to design agents to behave altruistic, but here we assume that we can't change the selfish behavior of the agents. We try to solve it by introducing some rules from the outside. We consider what rules can bring out a better performance from the society. We introduce the information controller in this game. The controller modifies the information to increase the total rewards among the agents. If controller modifies the information adequately, it is expected that total rewards can be increased. We propose the information transmission method that adapts to agents behavior. Through the simulation the effects of the controller are examined and its factor is analyzed by comparing with feature of an original game.

A. Sakurai et al. (Eds.): JSAI 2003/2004, LNAI 3609, pp. 95–102, 2007.

Table 1. An example of a strategy table ($M = 3$)

history	prediction
A A A	B
A A B	B
A B A	A
A B B	A
B A A	B
B A B	A
B B A	B
B B B	B

2 Minority Game

2.1 Rule

Minority Game[3] is the repeated game which models the competition. This is the abstraction of the famous El-Farol's bar program[4]. At each time step N(odd) players have to choose whether to be in group A or group B. Those who belong to minority side win and get a reward. The common information called histories is shared among the agents. Histories are the last M times wining side. Agents possess the S set of the strategies. The strategy consists of every history patterns and corresponding prediction. Table.1 is an example of a strategy table in the case of $M = 3$. At the beginning of the game S strategies are randomly assigned to each agent from 2^{2^M} possible strategies. Agents use the strategy that would have predicted the winning side the most successful.

2.2 Feature

We show the basic result of simulation in fig.1. The setup of the game is $N = 101$, $S = 2$, $M = 1, 2...10$. The horizontal line in the graph is the results when agents decide the action randomly. It is knows that it is divided into two phases bordering on M_c (M_c is the length of the history that causes the highest profit in the fixed N game)[5,6].

- symmetric phase ($M < M_c$)
  As M became short agents competes and the total rewards became low.
- asymmetric phase($M > M_c$)
  As M became long the total reward approaches the random case.

3 Extended Game with Information Controller

3.1 Assumption of an Extended Game

In real economic field, the economical history is interpreted and modified by the commentators. The participants will be affected by such information.

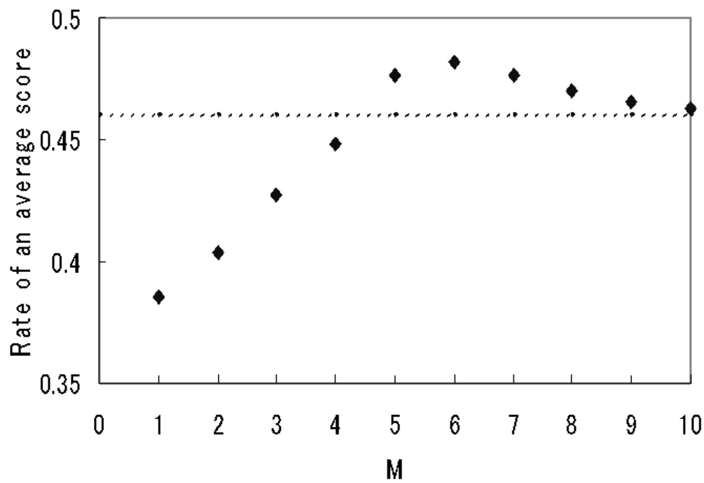


Fig. 1. Rate of an average score as a function of M (N=101,S=2,repeaeted 6000times)

Furthermore, the results of the behaviors of the affected participants will change the opinions of the commentators. To reflect such situations, the common information controller is introduced in the minority game.

3.2 Information Controller

The controller judges the fluctuation of the total rewards among the agents. In order to increase the rewards it changes the common information by masking (formula (1)). The last M_S length of the real histories are used as before and the remaining parts are masked with A (fig:2). h^1 is the latest history.

$$h_s^j = \begin{cases} h^j & (\,if\; j \leq M_s) \\ A & (\,if\; j > M_s) \end{cases} \quad (j = 1, 2, \cdots, M) \tag{1}$$

M_S is adapted with the following.

$$M_s(t) = M_s(t-1) + V(t), \tag{2}$$
$$where\;\; V(t) = \begin{cases} V(t-1) & (\,if\; R(t) \geq R(t-1)) \\ -V(t-1) & (\,if\; R(t) < R(t-1)) \end{cases}$$

$R(t)$ is the total rewards among the agents for each 2000 time steps. $V(t)$ is the length to which M_S is changed.

4 Simulation

4.1 Effects of the Controller

To examine the effects of the controller, we did a simulation.

(a) With controller
(b) With no controller
(c) Control is stopped on the way

The setup of the game is the same as chapter 2.2 and initial $M_s = 10$, initial $V = 10$. M_S is updated every 2000 time steps. The game is repeated 60000 times. We show the results in fig.3. In the case of (a) the rewards increased compared with (b). While if control is stopped on the way rewards decreased at the same level as (b). The effect of the controller is confirmed.

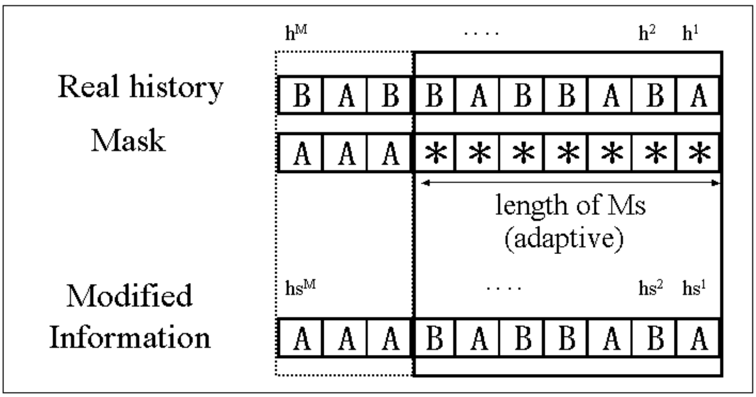


Fig. 2. Masking the information

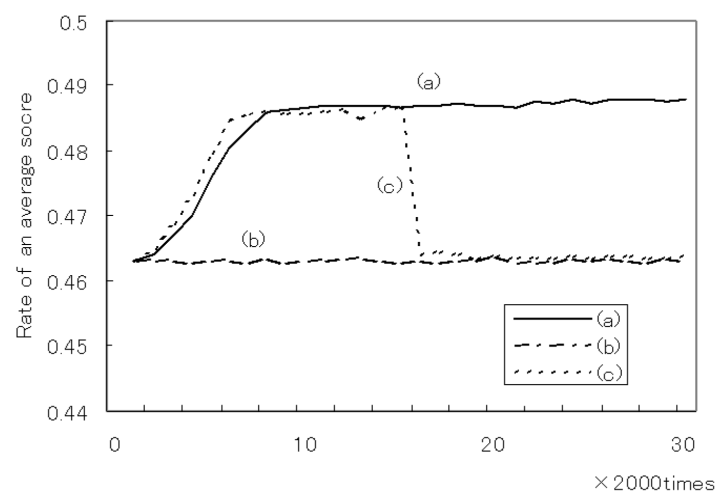


Fig. 3. The effect of the controller, M=10, average of 10 samples

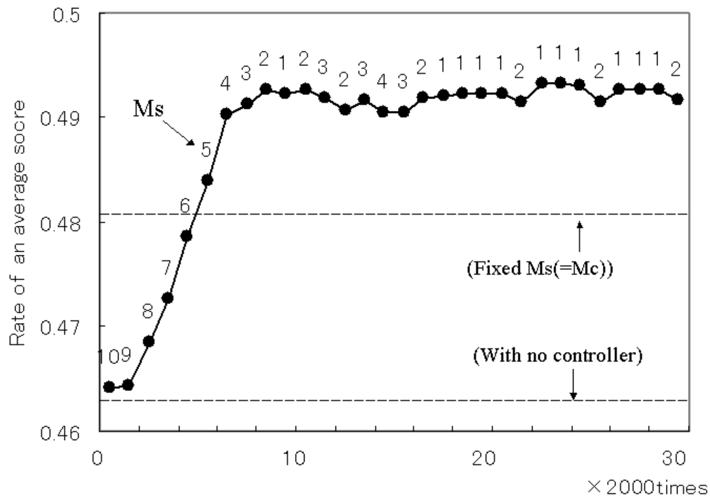


Fig. 4. Average points and Ms

4.2 Relation Between M_s and Total Rewards

Here we examine the relation between M_s and total rewards. In fig.4 we show the characteristic pattern of M_s transition. The numerals on the graph are the length of M_s which was adjusted by the controller in every 2000 times. The horizontal lines of the graphs are the case of no-controller and the case of fixed $Ms(Ms = Mc)$, respectively from the bottom. The results shows that total rewards increased than original case $(M = Mc)$ by making M_S short gradually.

5 Discussion

Here we analyze the factor of the profit improvement by the controller. In the original game when the common information is short $(M < M_c)$ the social profit is lower than random game as shown in fig.1. On the other hand in the game with controller when $M_s < M_c$ agents can get higher rewards. It is still higher than the highest profit in an original game when M=M_c. It is thought to be caused through the process which the game shifted from asymmetric phase to symmetric phase by adjusting M_s adaptive.

5.1 The Ratio of ϕ Group

In the original game especially in the setup that agents can get high rewards, there arises the difference of relative merits of strategy tables in each agent. As a result a finite fraction of the agents ends up using only one strategy [6]. Here we call such agent ϕ agent. In fig.5 we show the ratio of ϕ agents to all agents as a function of M in the original game. In symmetric phase ϕ group is not generated

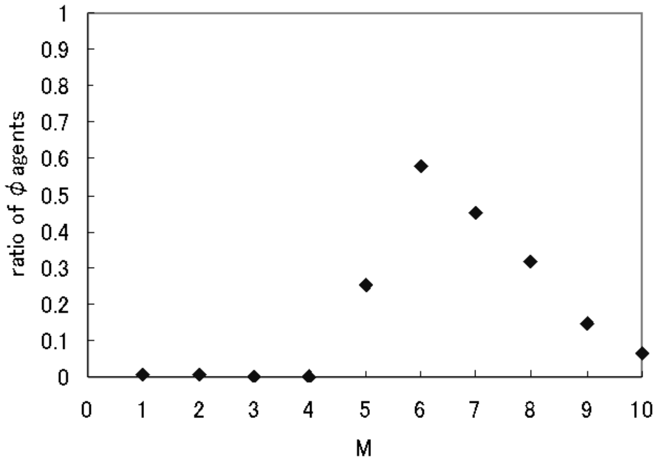


Fig. 5. Rate of number of ϕ (original game, average of 10 samples)

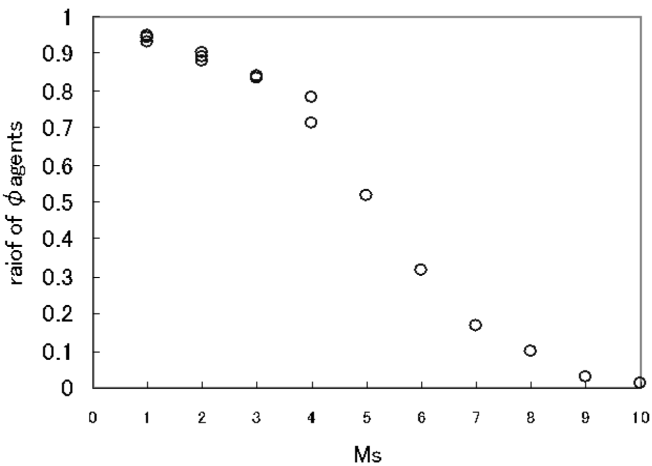


Fig. 6. Rate of number of ϕ (with controller)

but in the asymmetric phase ϕ group has generated. The ratio of ϕ agent is the maximum when $M = 6(M = M_C)$. In fig.6 we show the ratio of ϕ agents to all agents in the game with controller. When we compare the graphs by contrasting M with M_S, ϕ agent is generated at $M_s < 4$, that can't be seen in the original game. Additionally the ratio of ϕ became high as M_s became short. ϕ agent is the factor of the high rewards in the original game. These points are considered to have brought about high rewards at short M_s in the extended game.

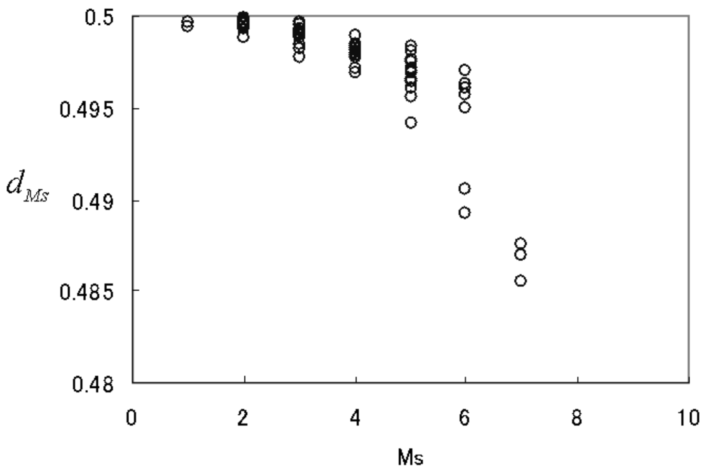


Fig. 7. Average distance among ϕ versus M_s

5.2 Strategy Table - Hamming Distance Among ϕ Agents

In this game suppose agents use the strategy table similar to others, such agent can't get high rewards. Because the agent using the strategy similar to others will do the same action as majority agents, therefore can't belong to the minority group. So it is important to have and use the strategy that is dissimilar to others in order to get high rewards.

$$d_{M_s}(i) = \frac{1}{2^{M_s}} \sum_{j}^{2^{M_s}} |S_i(h_j) - \mathrm{T}(h_j)|, \tag{3}$$

$$where \;\; T(h_j) = \frac{1}{n(\phi)} \sum_{i}^{n(\phi)} S_i(h_j)$$

In formula (3) $S_i(h_j) \in \{0, 1\}$ is the i-th agent's strategy corresponding to the history h_j, $T(h_j)$ is the average of the strategy of ϕ group corresponding to the history h_j, $n(\phi)$ is the number of ϕ agent (The strategies A and B in Table.1 were made to correspond to 0 and 1, respectively).

Higher value of d_{M_s} indicates the lower similarity with the whole. We show the average d_{M_s} among ϕ agents in fig.7. The average d_{M_s} became high as M_s became short. Namely it is thought that the possibility of the improvement of total rewards increased by making M_s short.

6 Conclusion

In this paper we introduced the common information controller in the Minority Game. The purpose of the controller is to raise social performance. Through the simulation it is confirmed that the controller can increase the total rewards among the agents by adjusting real history (M_S length) adaptively. Through the process of the game when M_S became short we found the followings. (a)There are more ϕ agents (the agents which uses the same strategy) than the original game. (b)The strategy table used among the agents can be expected to gain high profit.

References

1. Eric Bonabeau, "Predicting the Unpredictable, *Harvard Business Review*, March 2002(in Japanese, translated by N.Nishi)
2. T.Terano, H.Deguchi, "Perspective on Agent-Based Simulation Study in the Social Science," *Technical Reports of IEICE*, Artificial Intelligence and KnowledgeCJanuary 2002 (in Japanese)Computing (AI2001-59)Cpp25-32
3. D.Challet and Y.-C.Zhang, "Emergence of Cooperation and Organization in an Evolutionary Game", *Physica A* , Vol. 246, pp. 407-418, 1997
4. Brian W. Arthur, "Inductive reasoning and bounded rationality (The EL Farol Problem)", *American Economic Review*, vol. 84, pp. 406, 1994
5. Radu Manuca, Yi Li, Rick Riolo and Robert Savit, "The Structure of Adaptive Competition in Minority Games", *PSCS Technical Reports*, PSCS-98-11-001, 1998
6. D.Challet and M.Marsili "Phase Transition and Symmetry Breaking in the Minority Game", *Phys. Rev. E 60* , R6271, 1999

Analysis of Efficiency and Accuracy of Learning in Minority Games

Kiyoshi Izumi

DHRC, AIST, 2-41-6 Aomi, Koto-ku, Tokyo 135-0064, Japan
kiyoshi@ni.aist.go.jp
http://staff.aist.go.jp/kiyoshi.izumi/

Abstract. In this paper, we constructed three types of agents, which are different in efficiency and accuracy of learning. They were compared using acquired payoff in a game-theoretic situation that is called Minority game. As a result, different types of learning methods got the highest payoff according to the complexity of environmental change and learning speed.

1 Introduction

1.1 Why Do Humans Have Intelligence?

The question "why humans have intelligence" is an ultimate question. It can not be answered simply, even if there is an answer. Then, by this research, I will change it and consider it the form of a problem as follows.

Why do human beings have an intelligence? ⇒ "Intelligence" is knowledge and capability to reason using knowledge. To "have" is to be assumed to have. ⇒ What conditions do make it better to assume that the others have knowledge and capability to reason using knowledge in creating others' models?

That is, the purpose of this paper is to examine the situation when it is advantageous to create others' model with representation, thinking, and learning. If it can be explained well, we can understand the reason why human beings are assumed to be have intelligence. And it will be the first step to an elucidation of the emergence of intelligence.

1.2 Frame of an Idea

What are the conditions that make one to assume that others has capability to process information, to update an internal state to others, and to determine action? In this research, it is thought that the following two conditions are key points.

- **Time pressure to reasoning and learning**. Asynchronous nature of decision-making.
- **Complexity** produced from the internal interaction.

A. Sakurai et al. (Eds.): JSAI 2003/2004, LNAI 3609, pp. 103–113, 2007.

If there are the following three conditions when an others model is constituted, constituting the model with reasoning capability is advantageous.

1. Environment continues to change. → Necessity for learning. If a player does not change its mental model, it cannot be adapted.
2. Players are heterogeneous. → Necessity of taking in others' information to learning.
3. Time pressure to reasoning and learning exists. → Necessity for effective learning. The learning with an assumption that the others have intelligence will be effective.

Especially the last two conditions have not been considered well in conventional economic models.

1.3 Using Information Is Not Free

In the classic economic model, a player can spend infinite time for reasoning the other players' action. That is, while changing time from t to $t+1$, all players can spend time infinitely until they finish to reason the economic structure and the others' behavior at the next period $t+1$. The efficiency of learning using information was not almost taken into consideration at all. Therefore, with economic rationality, it was assumed that a player can acquire the best learning results by using all the information that one can get and spending time thoroughly. However, when there is time pressure to learning and reasoning, all players will have to concern the efficiency of learning method.

Then, what can be considered to be the efficient learning method? In the field of game theory, the strategy called tit-for-tat was adaptive in the situation called a prisoner's dilemma game [1]. The tit-for-tat strategy is the strategy that a player imitates action of others at the previous time step, and he also performs the same action as the next time step. In other words, it is the parasitism to others' action. This is the learning method with efficiency to some extent. However, when the environmental change is more complicated, this simple learning method, imitation of others' *action*, will not work. The learning method which imitates others' *strategy* may succeed in such conditions. That is, it is the parasitism to others' strategy.

In order to perform parasitism to others' strategy, others' strategy have to be inferred. To infer others' strategy, it is necessary to assume that others have capability to process information, to update their internal state to others, and to determine action,

1.4 Hypothesis

The hypothesis of this research is as follows.

> Assuming that there is capability to process information to others, to update an internal state, and to determine action occurred from two conditions; (a) the complexity of environmental change, and (b) the time pressure to learning.

In this research, as a preliminary experiment for investigating this hypothesis, the computer simulation was performed on the topic of the game-theory-situation called minority game.

2 Framework of Preliminary Experiments

2.1 Minority Game as a Model of Financial Markets

A minority game is a repetition game in which N (odd number) players must choose one of two alternatives at each step. A payoff is given to minority group, players who chose the alternatives which fewer people chose between two alternatives. Arthur, who is one of the proposal persons of complexity systems economy, proposed the idea of a bar problem that people try to drink at the bar which fewer person chose between two bars [2]. Then, many researchers pointed out and analysis the feature as a nonlinear phenomenon and made various extensions [3].

Since the mechanism that a minority group win is seen also in an actual financial market, it can be considered that a minority game is a simple model of financial [4]. In this research, standard minority game [5, 6] was developed and it considered as the framework of a preliminary experiment of the following framework interpreted as a model of a financial market. The agent of N (odd number) participates in a game, and time progresses dispersedly. One period consists of four steps; (1) Determination of action, (2) Price determination, (3) Calculation of payoffs, and (4) Learning.

(1) Determination of Action: Each agent i determines dealing action $h^i(t)$ of financial capital at t using knowledge called *memory*, $\mathbf{P}^m(t-1)$, the time series data of price change of the financial price at the past m time steps.

$$\mathbf{P}^m(t-1) = \{P(t-1), P(t-2), \cdots, P(t-m)\} \tag{1}$$

$P(\tau) = \{+1, -1\}$ expresses price change of the financial price at τ. When financial price rose (dropped) at τ, $P(\tau) = +1(-1)$. Each agent i has the rule that determines its behavior $h^i(t)$ (buy or sell) according to each pattern of price change $\mathbf{P}^m(t-1)$. This rule is called strategy $S^i(t)$ (see Figure 1).

(2) Price Determination: The supply and demand of all N agents are accumulated. When there are more agents to buy, the price rises. When there are more agents to sell, the price falls. That is, t price change $P(t)$ is decided by the form like majority.

$$P(t) = \begin{cases} +1 \text{ up} & (\sum_{i=1}^{N} h^i(t) > 0) \\ -1 \text{ down} & (\sum_{i=1}^{N} h^i(t) < 0) \end{cases}$$

This simplifies the relation between financial prices and supply and demand in a market.

$$
\begin{array}{ccc}
 & & \text{Dealing behavior } h^i(t) \\
\text{Price change } \mathbf{P}^m(t-1) & & \text{Buy or sell} \\
\{-1,-1,-1,\cdots,-1,-1\} & \rightarrow & \left(\begin{array}{c} +1 \text{ or } -1 \\ +1 \text{ or } -1 \\ \vdots \\ +1 \text{ or } -1 \end{array}\right) \\
\{-1,-1,-1,\cdots,-1,+1\} & \rightarrow & \\
\vdots & \vdots & \\
\{+1,+1,+1,\cdots,+1,+1\} & \rightarrow & \\
 & & \uparrow \\
 & & \text{Strategy } S^i(t)
\end{array}
$$

Fig. 1. Each agent's strategy

(3) Calculation of Payoff: Payoff$^i(t)$ is calculated to each agent i from price change $P(t)$ and dealing behavior $h^i(t)$.

$$\text{payoff}^i(t) = -P(t) \cdot h^i(t) \tag{2}$$

By this equation, at the time of a price rise ($P(t) = +1$), the agent of sell behavior ($h^i(t) = -1$) belongs to a minority group, and a minority group's agent acquires a positive gain. On the other hand, at the time of a price decline ($P(t) = -1$), he buys it, the agent of behavior ($h^i(t) = +1$) belongs to a minority group, and a minority group's agent acquires a positive gain similarly. Therefore, it is a minority game.

Economically, the equation 2 assumes that the financial price returns to the average value. That is, when the price rose (fell), it will fall (rise) to the previous level in the future of the infinite point. And the equation 2 also assumes that the payoff is calculated based on the final value of the financial capital in terms of the price in the future of the infinite point. Therefore, the payoff is positive when a player bought (sold) the financial capital in drop (rise) of the price. The payoff is negative when a player sold (bought) the financial capital in drop (rise) of the price. Such a regression-assumption is the entity which considers that the finance market is a minority game.

(4) Learning: The learning in a standard minority game is simple; selection of strategy. Each player has s strategies generated by random at the beginning of a game, and continue to have those strategy without modifying them. And each strategy has a specific value called a virtual value. It is a number of times that the dealing behavior derived from a strategy could acquire a positive payoff. Each agent chooses one strategy with the highest value from s strategies, and uses it at the time of the behavior decision at the next step.

In standard studies of the minority game, only the very simple learning was assumed and they analyzed in many cases about the relationship between memory length m and the price fluctuation. However, in a previous study [7], it is suggested that it has essential significance that all agents are homogeneous, that is, all agents share the same information and the same learning algorithm. It is shown that the same results was obtained in both cases that all agents used information about price movement and that all agents used random data. In our study, an agent is heterogeneous and the information and the learning to use differ among agents.

2.2 Agents

First, **Chartist (Ch)** was prepared as a standard agent. This agent determines behavior based on the time series of a past price change (chart information). It is extended from the player in the standard minority game, described in section 2.1.

In this study, we prepared three kinds of agent; **Hand imitator (HI)**, **Strategy imitator (SI)**, and **Perfect predictor (PP)**. They are different in terms of kinds of information that they use (see Table 1). Each agent took part in the minority game against Chartists, and its payoff was compared with each other. Hand imitator performs simple learning of imitating the behavior of other

Table 1. Each agent type's characteristics

Efficiency	HI	>	SI	>	PP
Accuracy	HI	<	SI	<	PP
Information	Payoff		Payoff		Payoff
	Behavior		Behavior		Behavior
			Price		Price
					Game structure

(HI: Hand imitator, SI: Strategy imitator, PP: Perfect predictor)

agents with high payoff. Since it uses the least information, the speed of learning is quick and the efficiency is good, but the learning results will be inaccurate. On the other hand, Perfect predictor infers both other agents' strategy and the game structure (payoff matrix) using all kinds of information. Since it uses much information, the obtained result is exact, but the speed of learning is slow and the efficiency is bad. Strategy imitator is in the middle of these two types. It performs only a inference of other agents' strategy, and imitates the strategy of other agents with high payoff. The efficiency and accuracy of its learning are in the middle of HI and PP.

Chartist(Ch). Chartist's behavior decision is the same as stated in section 2.1. It is extended about learning. Since a agent in standard minority games continues to have its strategies given first without changing, as described in section2.1. Thus, it can not search for all of solution spaces. Then, we extended Chartist's learning method as follows, to enable it to search for all solution spaces.

Decision of Behavior. Chartist has one strategy described in figure 1. The pattern matching of the price change of the past m steps, $\mathbf{P}^m(t-1)$, to the strategy $S^i(t)$ determines Chartists' behavior $h^i(t)$.

Learning. When Chartists acquires a positive payoff at t, it does not change its strategy. When it got a negative payoff, its strategy $S^i(t)$ is updated at a certain probability α (learning rate). The bit of the behavior rule (buy or sell) to this price change is reversed.

Hand Imitator (HI). Hand imitator performs simple learning of imitating the behavior which other agents with a high payoff.

Decision of Behavior. According to a certain probability p_{buy}, Hand imitator buys the financial capital. Probability to sell p_{sell} is $1 - p_{buy}$.

Learning. The probability of dealing behavior of other agents with a high payoff is copied.

1. *Inference of others' dealing probability*: About other agent js of each other than itself, an estimated probability to buy $\tilde{p}^j_{buy}(t)$ is updated by the following equation.

$$\tilde{p}^j_{buy}(t) = (1-\beta) \cdot \tilde{p}^j_{buy}(t-1) + \beta \cdot \text{action}^j(t) \tag{3}$$

action$^j(t)$ is probability in t which he buys agent j this term.

$$\text{action}^j(t) = \begin{cases} 1 & (\text{Agent } j \text{ bought at } t) \\ 0 & (\text{Agent } j \text{ sold at } t) \end{cases}$$

Parameter $0 \leq \beta \leq 1$ expresses the rate which updates the estimated value of the probability of dealing behavior of other agents, and means the learning speed of others' models.

2. *Accumulation of payoff*: About all agents j including itself, the accumulation value $R^j(t)$ of payoff is updated by the following equation.

$$R^j(t) = (1-\gamma) \cdot R^j(t-1) + \gamma \cdot \text{payoff}^j(t) \tag{4}$$

payoff$^j(t)$ is a parameter showing the payoff of agent j this term. $0 \leq \gamma \leq 1$ expressed the update rate of the accumulation value of a payoff, and fixed it to 0.5 in this study.

3. *Copy of the behavior according to the payoff*: Copy probability which he buys with a certain probability α (learning rate). One agent j^* is chosen by the probability proportional to accumulation value $R^j(t)$ of each payoff from all the agents that he also contains first. And probability $p^{j^*}_{buy}(t)$ which the agent buys is copied to the probability which he buys.

Strategy Imitator (SI). Strategy imitator performs only an estimation of other agents' strategy, and imitates the strategy of the high agent of a payoff.

Decision of Behavior. The pattern matching of the price change of the past m steps, $\mathbf{P}^m(t-1)$, to the strategy $S^i(t)$ determines dealing behavior $h^i(t)$. It is the same as that of Chartist.

Learning. Strategy imitator estimatesothers' strategies and imitates the other's strategy with high payoff.

1. *Estimation of others' strategies*: Strategy imitator estimates whether the other agent js to buy or sell from the pattern matching of the price change $\mathbf{P}^m(t-1)$ to estimated others' strategies $\tilde{S}^j(t)$. If the estimated behavior is different from the actual behavior which agent j performed, by a certain probability $0 \leq \beta \leq 1$ (learning speed of others' models), the estimated strategy $\tilde{S}^j(t)$ will be updated. The bit of the agent j's behavior corresponding to the price change in $\tilde{S}^j(t)$ will be inverted.
2. *Accumulation of a payoff*: About all agents j including itself, the accumulation value $R^j(t)$ of payoff is updated by the equation 4. It is the same as that of Hand imitator.
3. *Imitation of strategy according to the payoff*: Strategy is copied by a certain probability α (learning rate). One agent j^* is chosen by the probability proportional to accumulation value $R^j(t)$ of each payoff from all the agents. And the agent's strategy $\tilde{S}^{j^*}(t)$ is copied to its strategy.

Perfect Predictor (PP). Perfect predictor estimates both other agents' strategies and the game structure (the payoff matrix) using all information.

Strategy. Perfect predictor estimates others' behavior and decide its own behavior according to the estimated payoff matrix.

1. *Estimation of others' behavior*: Perfect predictor estimates whether the other agent js to buy or sell from the pattern matching of the price change $\mathbf{P}^m(t-1)$ to estimated others' strategies $\tilde{S}^j(t)$.
2. *Decision of behavior*: Perfect predictor has its own strategy that represents estimated game structure (the payoff matrix). That is, the strategy shows which behavior (buy or sell) can acquire a positive payoff corresponding to the others' behavior. According to the strategy and the other agent j's estimated behavior, Perfect predictor decides its own behavior.

Learning. Perfect predictor learns both others' strategies and the game structure.

1. *Learning of others' model*: The estimated strategies $S^j(t)$ about other agent js are updated. Learning method is the same as that of Strategy imitator.
2. *Learning of the game structure*: When its payoff is lower than the average of other agents' payoff, the bit of behavior corresponding to the estimated others' behavior is reversed by a certain probability α (learning rate). This means renewal of the knowledge about the payoff structure.

Table 2. Setting of the simulation

Number of agents	$N = 5$
Agents' combination	(Ch 4 and HI 1), (Ch 4 and SI 1), and (Ch 4 and PP 1)
Memory length	$m = \{1, 2, 3, 4\}$
Learning speed	$\alpha = 0.8$ (fixed) s
Learning speed of others' model	$\beta = \{02, 05, 08\}$
Update rate of payoff	$\gamma = 0.5$ (fixed) s
Number of simulation runs	10 times every parameter combination
Comparison method	Improvement rate of the payoff of HI, SI, and PP when the average of the payoff of four Chs is set to 100.

(Ch: Chartist, HI : Hand imitator, SI: Strategy imitator, and PP : Perfect predictor)

3 Simulation Results

3.1 Setting of the Simulation

Setting of the simulation of the preliminary experiment in this study is shown in Table 2.

Five agents participate in each trial of the minority game. Four are Chartists as standard agents in all trials. The remaining one agent is Hand imitator, Strategy imitator, or Perfect predictor.

Learning speed α of their strategy and the update-rate of a payoff γ were fixed. Memory length m and the learning speed of others' models β of HI, SI, and PP were changed as shown in Table 2. When Chartists had longer memory, the behavior of the financial price became more complicated. Thus, the memory length is related to the complexity of market dynamics.

The simulation was performed 10 times every parameter set; {agent's combination (HI, SI, or PP) $\times$ Memory length $\times$ Learning speed of others' models}. The agents (HI, SI, and PP) were compared by the improvement rate of payoff when the average of payoff of the four standard agents (Chartists) wad set to 100.

3.2 Simulation Results

Simulation results are shown in Figure 2.

The results show that when memory length is short (the complexity of a market place is low) and the learning speed of others' models is slow, Hand-imitator's payoff is high. And when the learning speed of others' models is fast, Perfect-predictor's payoff is high. Strategy-imitator's payoff becomes high as memory length becomes long (as the complexity of a market place becomes high). These results are summarized in Figure 3.

Consideration of these results shows the following thing. When the behavior of the whole system is comparatively simple and the learning speed of the

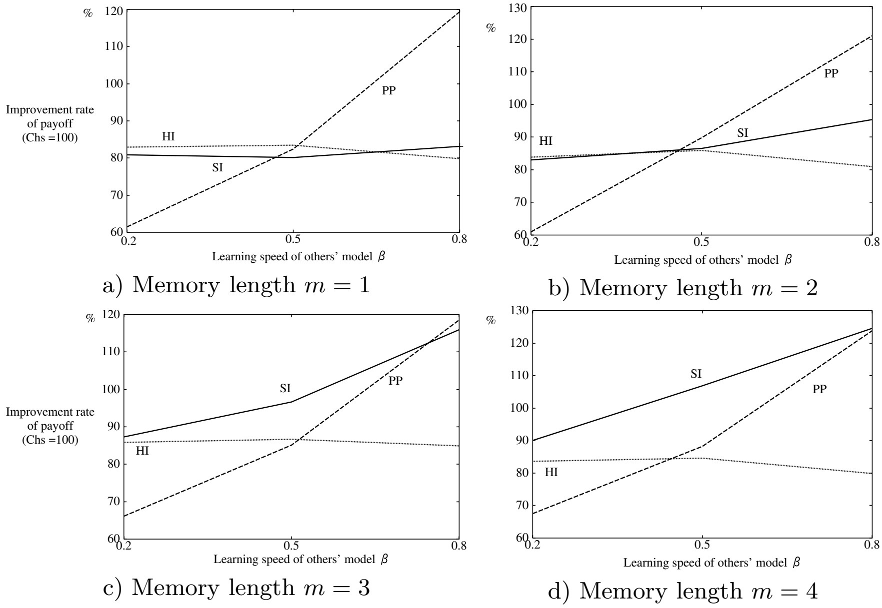


Fig. 2. Simulation results: In figures (a-b), when memory length is short and the learning speed of an others model is slow, the payoff of HI (Hand imitator) is high. When the learning speed of an others model is fast, the payoff of PP (Perfect predictor) is high. In figures (c-d), The payoff of SI (Strategy imitator) becomes high as memory length becomes long.

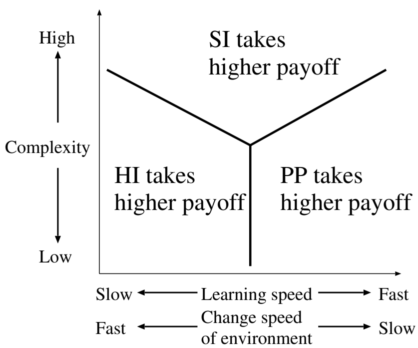
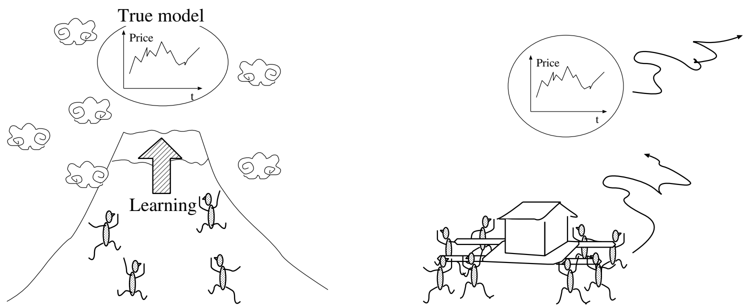


Fig. 3. Summary of results (HI: Hand imitator, SI: Strategy imitator, and PP : Perfect-predictor)

others' model is slow, simple learning methods like HI can correspond to change of others' strategy immediately and such methods are advantageous. Therefore, Hand-imitator payoff was high. When the behavior of the whole system is comparatively simple and the learning speed of the others' model is fast, learning methods using all information like PP can obtain exact learning results. There-



(a) Classic economic models (b) Agent-based models

Fig. 4. Classic economic models and agent-based models

fore, Perfect-predictor payoff is high. However, when other agents' strategy is complicated, simple learning methods can only acquire inaccurate learning results. And learning methods using all information spend too much time to get learning results at a limited learning speed. Therefore, payoff of Hand imitator and Perfect predictor had fallen when the behavior of the whole system is comparatively complicated. Then, Strategy imitator, that is in the middle of these two learning types, got high payoff in this area.

4 Discussion

In conventional economic models, when the participant of an economic system was not using all information, only some negative causality had been considered. For example, information cannot be acquired (in-accessibility). Information has many noises and it cannot be trusted (low reliability). Many costs are required to acquire information (high cost).

These models have an implicit assumption, "a forecast would become exact so much if many information is used." Although its attention had been directed to some extent by the cost which acquires information, probably, about the cost which learns using the acquired information, it had not almost been taken into consideration at all. It is because the condition that a participant's learning speed assumes infinitely the ideal condition of being fast, in the economic model with this classic, therefore the speed of change of the environment which is a learning target can be relatively regarded to a zero is considered. This is shown in figure 3 as an ultimate situation which is in the right at an infinite distant place. The algorithm which performs an exact learning in this condition using all information, without considering the efficiency of a learning is the most advantageous.

The image of the situation of such a classic economic model is drawn in figure 4a. The "true model" showing the behavior of the target economic phenomenon exists solemnly on a cloud as a given thing. And the participants who have an infinite learning speed using all information are learning toward a true model.

On the other hand, in the model by the agent simulation which was shown by this study, the behavior of an economic phenomenon is not being fixed and it has generated internally from the interaction between participants (figure 4b) Each participant thinks that he will forecast well and is performing various learnings. The learning behavior carries out an interaction and the behavior pattern of the whole economic system changes. While one mikoshi is shouldered all together and a motion of everybody collides, it is moving on the whole. The target to pursue also escapes or approaches according to a motion of a mikoshi.

As for a learning in the environment where the target of a learning changes behavior in response to reaction from the learning behavior of oneself, like a society and an economic system, learning using all information is not necessarily good. At the point, a classic economic model differs from the condition of having assumed implicitly. Then, what does actual human being do? Probably, paying attention to what information, heuristics is used about how it learns. Heuristics is the learning method discovered experientially. Although it cannot necessarily obtain an exact learning result, it can obtain efficiently the result which can be satisfied moderately. Strategy imitator is also a kind of heuristics of imitating the strategy of others who succeeded in fact.

It is also interesting to analyze the heuristics of a learning of actual human being found by the cognitive psychology and some learning methods in artificial intelligence in the viewpoint of the efficiency and accuracy of a learning, and to evaluate as a model of the learning in a society and an economic system. And finally it will become one approach to the elucidation of the function of an intelligence in social / economical situation.

Acknowledgments

This research was partially supported by the Ministry of Education, Science, Sports and Culture, Grant-in-Aid for Scientific Research (B), no. 16300047.

References

1. Axelrod, R.: The Evolution of Cooperation. Basic Books (1984)
2. Arthur, W.B.: Inductive reasoning and bounded rationality (the el farol problem). American Economic Review **84** (1994) 406
3. Minority Game's web page: (`http://www.unifr.ch/econophysics/`)
4. Zhang, Y.C.: Modeling market mechanism with evolutionary games. Europhys. News **29** (1998) 51–54
5. Challet, D., Zhang, Y.C.: Emergence of cooperation and organization in an evolutionary game. Physica A **246** (1997) 407–418
6. Marsili, M.: Market mechanism and expectations in minority and majority games. Physica A **299** (2001) 93–103
7. Cavagna, A.: Irrelevance of memory in the minority game. PHYSICAL REVIEW E **59** (1999) R3783–R3786

A Partitioned Random Network Agent Model for Organizational Sectionalism Studies

Kikuo Yuta[1,2], Yoshi Fujiwara[1], Wataru Souma[1],
Keiki Takadama[1,3], Katsunori Shimohara[1,2], and Osamu Katai[2]

[1] ATR Network Informatics Laboratories
Hikaridai, "Keihanna Science City," Kyoto 619-0288, Japan
{kikuo,yfujiwar,souma,katsu}@atr.co.jp
[2] Kyoto University, Graduate School of Informatics,
Yoshida-Honmachi, Sakyo-ku, Kyoto 606-8501 Japan
katai@i.kyoto-u.ac.jp
[3] Tokyo Institute of Technology,
Interdisciplinary Graduate School of Science and Engineering,
4259 Nagatsuta-cho, Midori-ku, Yokohama, Kanagawa 226-8502 Japan
keiki@dis.titech.ac.jp

Abstract. This paper presents a new organization model that addresses the effects of networks on the sectionalism phenomenon, defined as excessive concern that members of a section have for the interests of their own section. No studies tackled the relationship between human communication networks and sectionalism. The points of our model design are: network distributed agents with a sense of values, extended random network structures, and a new index to monitor sectionalism. A homogeneous effect of communication networks and a heterogeneous effect of sectional specialization were also introduced into the model. Empirical results showed that sectionalism behavior and the performance of the proposed index were superior to conventional indices when capturing sectional structures. Finally, we showed one example of the availability of such a multi-agent network approach. Simulation results clearly illustrated the effect of cross-sectional links on sectionalism reduction by following a so-called "power law."

1 Introduction

When sectionalism functions well, the progress of local rules creates more efficiency. On the other hand, excess sectionalism causes a value gap between sections. In this case, that the term "sectionalism" implies a sectional "wall" that functions as a barrier or safety shield to the members of the section and a blocking layer to outsiders. A prime example of a blocking layer can be found in official bureaucracies. Japanese newspapers or magazines use "sectionalism" to indicate the problem of overemphasis on self sections that prevails in large companies and bureaucracies. At the same time in the United States, the "section" part in "sectionalism" mostly means an area in the country. Although there is related conventional research such as the research field known as "conflict management,"

A. Sakurai et al. (Eds.): JSAI 2003/2004, LNAI 3609, pp. 114–125, 2007.

which emphasizes the final result of sectionalism that denotes "conflict," not the progress of the specialized multi polarization of the section itself, so that it is not enough. Inherently, the section's function is to raise performance by specialization. In innovation research attracting attention as a new organizational design, such as *cross-functional team (CFT)*, the importance of specialists and the cooperation between sections is noted by McDonough [1]. So, the function of section is needed but it must be kept at adequate levels. The division and specialization of functions causes the both merit and demerit. Although the question what causes sectionalism is important, it is not enough. What drives sectionalism? How can it be monitored? Those are much more important questions.

To investigate the impetus of the sectionalism phenomenon, we employed a social simulation approach by *multi-agent model (MAM)*, well reviewed by Weiß[2]. Social simulations, especially organizational simulations through MAM are well reviewed by Prietula, Carley, and Gasser [3]. Previous research shows that a multi-agent model is a valid method to generate macro-behavior from a micro-agent design and to examine the macro-level between the macro- and micro-levels. However, even if all factors are included in a model make it theoretically, to meaningful results are not necessarily obtained. What is the important factor to generate sectionalism? Such causes as organization structure, goal setting, chain of command, and personnel evaluation systems, can be considered. In this study, we regarded a communication network as important and installed it at the core of our model. Bases on empirical cases organizational surveys, and projects which the author personally exprienced as a human-resources management consultant, a company with plenty of cross-sectional communication rarely sinks into sectionalism.

There is much previous research on the importance of a communications network in the organizations. For example, Torenvlied's case study [4] shows that an informal communication network influences feelings of resistance toward change in an organization. Although the study did not focus on the section level, it is still important evidence that people exchanged opinions through an informal communication network. Considering the causes of sectionalism on micro-agent levels, it is a problem that priority is usually given to the rules of its own section, without understanding the rules of other sections, such as technical terms and senses of values. Bar-Hillel and Carnap discussed "semantic noise" [5]. In their paper, feelings of resistance produced from misinterpretation or misapprehension of information from other sections were called "semantic noise." While local language or coding schemes smoothly circulate information inside a section, it may prevent acquiring or interpreting the information from outside the section. By a psychological approach, Ross [6] discovered that people are easily influenced by the opinion of others and those related, so it became impossible to recognize the state of the entire public consensus. This phenomenon was called "false consensus effect." These findings indicate that sectional communication deeply affects a member's sense of values.

Furthermore, when seeing a member in a section in an actual organization from the view of a communication network structure, we realize that most communication links inside of the section are dense, and cross-sectional communication links are sparse. If there is a propagation of a sense of values by communication, people in the same section will tend to share the same values, producing a gap between different sections. This gap is "sectionalism." When we dissect a sectionalism model, the most important submodels are the following: 1) Different sense of values for every section. 2) Network structure classified in the section. 3) Propagation mechanism. In this study we install the three submodels into a multi-agent model and generate a sectionalism phenomenon. To check the generation of sectionalism, an evaluation system is needed. Therefore, we introduce a new index to evaluate gaps between sections. Ultimately, the target was to make a platform for the experimental research of sectionalism by modeling the emergence sectionalism and its evaluation.

This paper is organized as follows. Section 2 describes the model and the methods of generating a partitioned random network and index. Section 3 shows the detailed designs of the simulation experiment by installing parameters. Section 4 indicates the experimental results. Section 5 discusses the results and indicates future works. Finally, conclusions are given in Section 6.

2 Model Design for Sectionalism Studies

2.1 Model Outline

Briefly, the model consists of distribution agents which have a sense of value and interact on a network. A sense of values is designed as a vector of parameters in this model. Each parameter denotes independent organizational factors of culture, priority, history, and communication. Organizational sections have their specific pattern of this parameter that a effect agents close to it. An agent's vector is affected by a sectional vector and by the vectors of other agents when interactions happen. In total, each sectional sense of value propagation through agents' networks causes cross-sectional heterogeneity and sectional homogeneity. The sense of value gap between agent and organizational average is the individual deviation. Adequately gathering of this individual deviation from a section indicates sectional deviation. This model employs sectional deviation as an index to monitor sectionalism. To apply it to various sectionalism studies, the model was constructed with an object-oriented approach. Such dynamic parts as methods of interaction were designed to exchange with the various submodels. If all variables were dynamic, it would be hard to comprehend what has occurred. So, any variables were set up static in the simulation, which are explained in the following section. In this section, we explain the inherent property of the model.

2.2 Sectionalism Phenomenon Model Design

Organization Design. An organization was designed that consisted of '`N`' number of agents and '`S`' number of sections. The model can change the number of agents and sections to reflect such events as recruitment, retirement, company mergers, and divestiture.

Organization Factors into Section Design. The number of parameters for the sense of values is '$\mathtt{Q}$'. A sectional vector of the section of number '$\mathtt{k}$' is expressed as $\boldsymbol{S_k}(k = 1, 2, \ldots, \mathtt{S})$. We express each parameter of the organization factor 'i' of section $\mathtt{k}$ as 'sf_{ki},' and the sectional effect vector is expressed as $\boldsymbol{S_k} \equiv (sf_{k1}, sf_{k2}, \ldots, sf_{k\mathtt{Q}})$. In the simplest organization of only three sections, Sales, Operations, and Manufacturing, $\mathtt{Q}$ sets to 3 and $\boldsymbol{S_1} = [100, 0, 0]$ as Sales, $\boldsymbol{S_2} = [0, 100, 0]$ as Operations, and $\boldsymbol{S_3} = [0, 0, 100]$ as Manufacturing. A big firm has thousand of sections, and the number of sections does not denote the number of organization factors. In this case, $\mathtt{Q}$ is smaller than $\mathtt{S}(\mathtt{Q}<\mathtt{S})$.

Agent Data Structure. Agents have three types of informational data: 1) The list of the number of other agents to whom an agent has links, 2) The section number which the agent is affiliated, 3) The vector of sense of values that denotes organization factors of agent $\mathtt{i}$, expressed as $\boldsymbol{A_i} \equiv (af_{i1}, af_{i2}, \ldots, af_{iQ})$. These sense of values initialized all the same initial values, which were set to $\mathtt{Val}$. $(af_{i1} = af_{i2} = \ldots = af_{iQ} = \mathtt{Val})$

Network Structure by a Partitioned Random Network (PRN). For generalization of an organizational section structure, we propose a *partitioned random network (PRN)* which is an extension of a random network design proposed by Erdös and Rényi [7]. Partitioned random networks have two types of link probability: own-partition ($\mathtt{p_{self}}$) and cross-partition. When the total number of partitions is q, the link probability of an agent in partition i toward an agent in partition j is expressed as $p_{ij}(i, j = 1, 2, \ldots, q)[0, 1]$, calculated as pick up probability from partition j. In a simple case number between sections and partitions is same ($\mathtt{S}=q$). Symmetric section designs are powerful enough to sectionalism phenomenon studies. First, $\mathtt{p_{self}}$ is fixed and uses this link probability for all partitions. Next, all cross-partition link probability $p_{ij}(i \neq j)$ are calculated as $(\mathtt{1} - \mathtt{p_{self}})/(\mathtt{S} - \mathtt{1})$. When $\mathtt{p_{self}}$ is the $\mathtt{1/S}$, network is uniform. When $\mathtt{p_{self}}$ is 100%, the network is isolated. Such uniformity is important in partitioned networks so that we introduce a new coefficient as follows:

$$\text{uniformity coefficient } :\mathtt{UC} \equiv \frac{(\mathtt{1} - \mathtt{p_{self}})}{\mathtt{S}(\mathtt{S} - \mathtt{1})}. \tag{1}$$

The adjacency matrix of symmetric PRN ($\mathtt{N}$=30,$\mathtt{S}$=3,$\mathtt{P_{self}}$=33.3%, 80%, and 100%) is shown in Fig. 1. The horizontal and vertical axes denote the agents' number. The cross point $[i, j]$ denotes the links of agent i to agent j.

Sectional Effects on Agents. In this paper, we propose a simple submodel of sectional effect. Here, we set two parameters to carry out this submodel. The sectional effect frequency ($\mathtt{SE_{freq}}[\mathtt{0}, \mathtt{1}]$) denotes probability of occurrence. The sectional effect size ($\mathtt{SE_{size}}[\mathtt{0}, \mathtt{1}]$) denotes the ratio effect vector size to sectional vector size. Agent i in section k comes under the sectional effect in $\mathtt{SE_{freq}}$ of occurrence probability per unit period. After the effect, the vector of agent will be $\boldsymbol{A_i}$ + $\mathtt{SE_{size}} \times \boldsymbol{S_k}$. Because vectors denote the agents' sense of value, incrementing the sum of the factors is inappropriate. Then, the vector is normalized

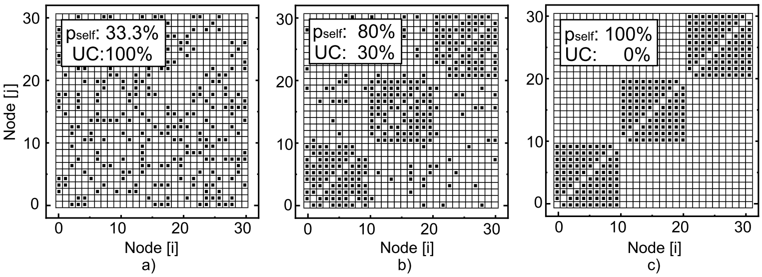


Fig. 1. Adjacency matrix of S-PRN: symmetric partitioned random network. (N=30, S=3, UC=0%, 30%, 100%).

so that the sum of the factors might maintain a value equal to the sum of the initial setting.

Interaction Effects on Agents. Previous research denotes that communication shares sense. Allen described a gatekeeper as a bridge between different companies [8]. We propose a shared sense of value model as a submodel of interaction. When agent b that has vector $\boldsymbol{A_b}$ interacts with agent c that has vector $\boldsymbol{A_c}$, the agents b and c share the difference of vectors. Adding new parameter $\mathtt{p_{share}}$ [0,50%], $\boldsymbol{A_b}$ is set to $\boldsymbol{A_b}$ - $\mathtt{p_{share}}\times$ ($\boldsymbol{A_b}$ - $\boldsymbol{A_c}$) and $\boldsymbol{A_c}$ is set to $\boldsymbol{A_c}$ + $\mathtt{p_{share}}\times$ ($\boldsymbol{A_b}$ - $\boldsymbol{A_c}$). The $\mathtt{p_{share}}$=50% indicates perfect sharing that create $\boldsymbol{A_b} = \boldsymbol{A_c}$ after the interaction. The $\mathtt{p_{share}}$=0% indicates perfect nonsharing that never makes any change through interactions.

2.3 Sectionalism Monitoring Index

Conventional Network Indices. Recently, as Barabási showed in his enlightening book [9], much research about network science has been done in such fields as bioscience, the internet, and economic networks. The importance of social networks has also been pointed out in sociology. An excellent book of practical review for social network analysis was published by Wasserman [10]. There are many network indices. We use the two fundamental indices: the average path length `L` and the cluster coefficient `C`, proposed by Watts and Strogatz [11]. These two evaluate designed network and new index for sectionalism. In explanations of network indices, the term,‘node’ and ‘edge’ are generally used, but to deepen our understanding of a network agents model, we use ‘agent’ and ‘link’.

Average path length L is the group average of the shortest course lengths from agents `i` to `j` that expresses d_{ij} for each combination. Therefore, when an individual average is expressed as L_i, `L` is calculated by as follows:

$$\mathtt{L} \equiv \frac{1}{N}\sum_{i=1}^{N} L_i, \qquad L_i \equiv \frac{1}{(N-1)}\sum_{j\neq i} d_{ij}. \tag{2}$$

Cluster coefficient $\mathtt{C}$ quantifies the gathering conditions of the agents that constitutes the network. It is called a cluster when there are three agents whose linkage forms a triangular structure. When agent $i$has k_i links, we can calculate the number of actual clusters as E_i and the number of potential clusters as R_i, and the individual cluster coefficient is C_i. The average of C_i is $\mathtt{C}$, which is calculated by the following equations:

$$\mathtt{C} \equiv \frac{1}{N}\sum_{i=1}^{N} C_i, \qquad C_i \equiv \frac{E_i}{R_i} = \frac{2E_i}{K_i(K_i-1)}. \tag{3}$$

New Index of Sectional Deviation (SecD). For practical research and applications, we introduce a new index to monitor sectionalism thet gathers the deviation of agents in group $\boldsymbol{G}$ with an independent group size. The average sense of value vectors for all agents in organization $\boldsymbol{O}$ is set as follows:

$$\overline{\boldsymbol{af}} \equiv \frac{1}{N}\left(\sum_{i=1}^{N} af_{i1}, \sum_{i=1}^{N} af_{i2}, \ldots, \sum_{i=1}^{N} af_{iQ}\right) \equiv (\overline{afo_1}, \overline{afo_2}, \ldots, \overline{afo_Q}). \tag{4}$$

When group $\boldsymbol{G}$ included in $\boldsymbol{O}$ has g agents: $\boldsymbol{G} \equiv (\boldsymbol{A_1}, \boldsymbol{A_2}, \ldots, \boldsymbol{A_g})$, we can calculate the variance of organization factors in $\boldsymbol{G}$ as V_G and variance per agent in $\boldsymbol{G}$ as $\overline{V_G}$. The proposed index of sectional deviation $\mathtt{secD}$ is calculated as:

$$\mathtt{SecD} \equiv \sqrt{\frac{\overline{V}_G}{Q}}, \quad \overline{V}_G \equiv \frac{V_G}{g}, \quad V_G \equiv \sum_{i=1}^{g}\sum_{q=1}^{Q}(\mathtt{af_{iq}} - \overline{\mathtt{afo_q}})^2. \tag{5}$$

3 Simulation Experiments

Simulation Outlines. First simulation consists of initialization, and then batch processes are counted in turns. Total turns are expressed as $\mathtt{T}$ in this paper. Initialization consists of the creation of networks and agents' data input. A batch process consists of the procedures of agent interaction and sectional effects. Each batch process treats all agents individually, picking them randomly.

Experiment of Model Behavior. Parameters are set as N=300, S=Q=3, and Val=100. A *symmetric partitioned random network* was designed *S-PRN* (UC=0%, 6%, 100%). Simplicity avoids confusion. Since the network considers interaction between sections, at least three sections are needed so that influence may extend indirectly. The same is true of $\mathtt{Q}$. $\mathtt{N}$ is related to the capability of getting to know each other. When $\mathtt{UC}$ is 0%, all section members must communicate in this model simulation. So, $\mathtt{N}$ is set to 300 whose sections have 100 agents each. The $\mathtt{Val}$ is just a standardization number, so 100 has no meaning.

Next are the remaining parameters: $\mathtt{SE_{freq}}, \mathtt{SE_{size}}, \mathtt{p_{share}}$ and $\mathtt{T}$. This experiment has no change in the number of agents, no learning, and no evolution. It focuses on the effects of network structures on sectionalism situations. So, this experiment has to converge into one final status all of the vector's values related to a single network design. We tuned those parameters, which reflect each other, with several constrains.

In our simulation, one turn is correspond to one week as setting image. Constrains include: 150 turns (about 3 years) at an isolated design is enough for the convergence of sectionalism, and enough uniformity design to keep secD at zero, starting from no deviation of sense of value, which denotes that secD is zero. Finally, those parameters are set as $\mathtt{SE_{freq}}$=10% ,$\mathtt{SE_{size}}$=30%, $\mathtt{p_{share}}$=25%, and T=1000. T of 500 is enough for the convergence so that an average between 501 to 1000 is set as the final organization factor of each agent.

Experiment of Indices Comparison. To investigate the performance of our proposed index on conventional indices, four *asymmetric partitioned random networks (A-PRN)* weere introduced as shown in Fig. 2. The basic parameters were set as N=360, S=Q=3, Val=100, $\mathtt{SE_{freq}}$=10%, $\mathtt{SE_{size}}$=30%, $\mathtt{p_{share}}$=25% and T=1000. The middle group of figures shows four organization designs that have three sections of the same number of agents. Section A has 60, Section B has 180, and Section C has 120. In all organizations there are 720 interacting links between Sections A and B, 360 between Sections B and C, and no links between A and C. The general structures between these sections are invariant. Furthermore, there are internal links in each section set between subdivisions. For example, Org. 2 divides Section B into two smaller classes: 60+120. Org. 3 divides Section B into three even smaller classes: 60+60+60. Org. 3 has separated subsections of B from both A and C. Org. 4 contains additional divisions to Org. 3, and Section C is divided into three subsections. In the figure, the small letter of a, b, c, b1, b2, b3, c1, c2, and c3 indicate section and subsection names respectively. The link probability form section x to section y ($\mathtt{P_x y}$) is expressied by the string 'x'-'y' inside figure.

The $A - PRN$ is designed as that agents keep their link amounts in all four organizations. Agents at Section A in each four organizations keeps individual 60 links, agents in Section B keeps 42, and agents in Section C keeps 33. Subsections keep this design role, too. Comparison focuses more on the sectional structure. The bottom of Fig. 2 shows an adjacency matrix of those organizations. The x- and y-axes, and the cross point of x-y have the same indications of Fig. 1.

Experiment of Cross-Sectional Link Effects. This experiment investigated the relationship between cross-sectional links and sectionalism. Basic parameters were set as N=300, S=Q=3, Val=100, $\mathtt{SE_{freq}}$=10%, $\mathtt{SE_{size}}$=30%, $\mathtt{p_{share}}$=25% and T=1000. There are two gradual experimental scale of UC change. One is from 0% to 10% and the other is from 10%, in each case was divided into 8 scales. The investigation of SecD in each UC on $S - PRN$ shows the relationship.

4 Experimental Results

Sectionalism Behavior. The results of a simple sectionalism model simulation of $S - PRN$ is shown in Fig. 3. These three graphs are the results of simulations

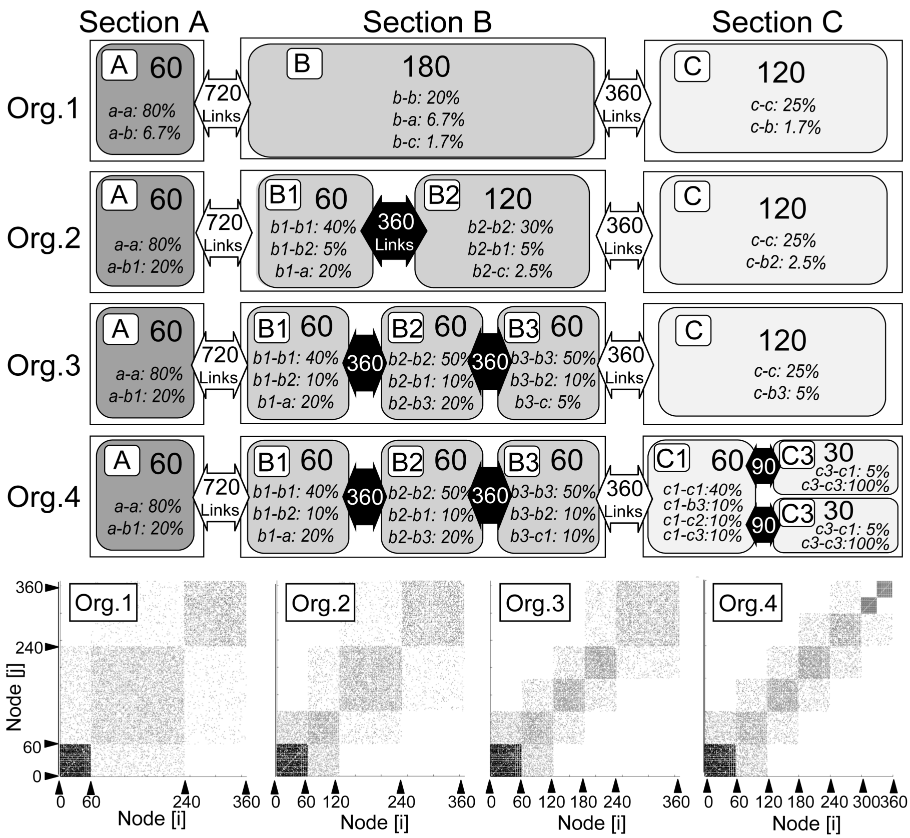


Fig. 2. Design and adjacency matrix of A-PRN: asymmetric partitioned random network. (N=360, S=3).

based on three different UC (0%: Isolation, 6%: sectional structure, and 100%: uniformity). In those graphs, the vertical axis denotes the average of organizational factors from all agents in section S_1, and the horizontal axis denotes simulation turns. '$org.f.\mathbf{S_1}$' indicates an organization factor of S_1. This behavior of section S_1 means that $org.f.\mathbf{S_1}$ is apt to increase and the others are apt to decrease. Those three graphs show parameters that keep constrains. For uniformity there was no sectionalism behavior which became a deviation of the factors. For isolation, 300 turns was enough to eliminate other factors.

Higher Performance of New Index (SecD). The results of the model simulation that compared of network indices on A-PRN are shown in Fig. 4. The horizontal axis denotes each organization designed as shown in Fig. 2. The vertical axis denotes the ratio to Org. 1 in each index. Such horizontal order from left to right means that organizational structures are being deeply divided into subsections. Consequently, all SecD values show higher than conventional indices of average path length L and cluster coefficient C.

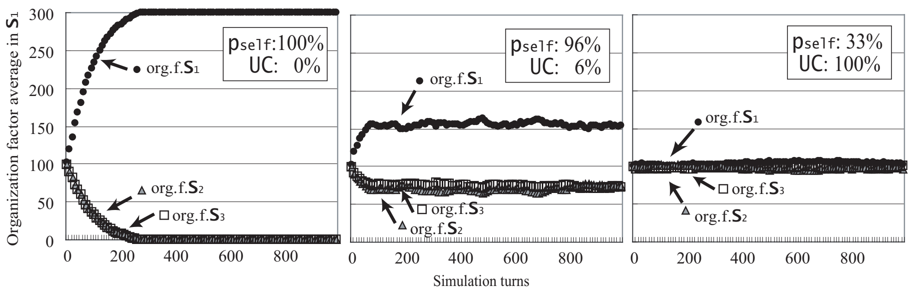


Fig. 3. Experimental results on S-PRNMAM: relationship between simulation turns and sectional average of three organizational factors. (N=300, S=3, T=1000).

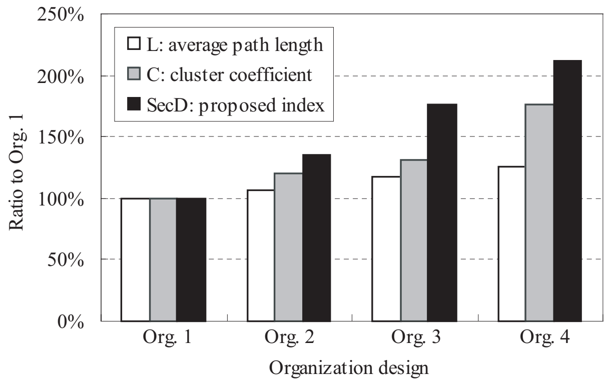


Fig. 4. Comparison of sensitivity performance to different organizational structures between conventional network indices and experimental results of secD on A-PRN

Power Law of Cross-Sectional Link Effectiveness. The experimental results of the relationship between `SecD` and `UC` on this $S - PRN$ is shown as Fig. 5. The horizontal axis denotes `UC` (%), and the vertical axis denotes SecD in both graphs. The graph on the left is drawn in normal scales, and the graph on the right is drawn in logarithmic scales. At `UC` of 10% or less, `SecD` rose rapidly, and at `UC` over 20% `SecD` decreased slowly. The graph on the right shows a straight line called power law, as which is under the statistics,a result of multiple regression analysis, as drawn in the figure.

5 Discussion and Future works

Through simulations, the generation and evaluation of sectionalism phenomenon were performed, which verified that the model functioned well. For example, an investigation of the network effects on cross-sectional job rotation, which was designed by changes in the section of agents' affiliation, and the others of

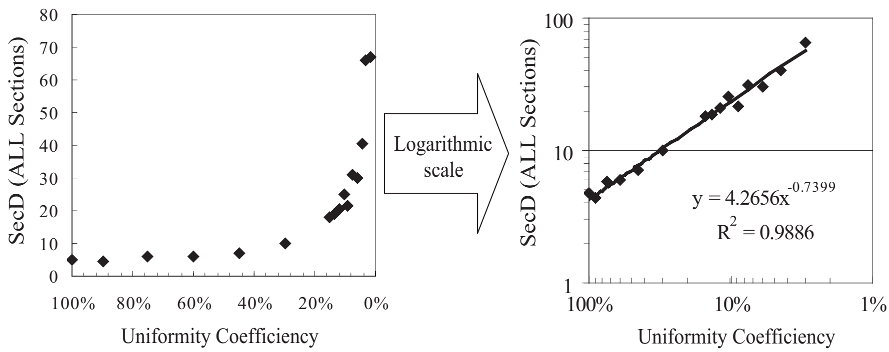


Fig. 5. Experimental results on S-PRN: relationship between UC and secD.(N=300,S=3,T=1000, the scale of graphs are normal and logarithmic.)

model were the same. Various experiments can be flexibly conducted simply by changing the agent and interaction designs.

Why did `SecD` show better sensitivity than cluster coefficient or average path length? Sectional organization is always highly clustered with shorter average path lengths. Section structure itself is set for clustered, and only one cross-sectional link will connect these two clusters so that all members between these two sections are connected by relaying 3-links. This is exactly the situation of the so-called "Small World Network"(see [9] and [11]). Furthermore, `SecD` has basically the same tendency as cluster coefficients because it is influenced by the linked agent situation. An advance of `SecD` is a sectional effect on an asymmetrical partitioned network. `SecD` gets more deviation of organizational factors' propagation relayed by the mediate section of B. We notice that most organization structure must be handled with sectional structure in its actual existence.

Incremental cross-sectional links effectively reduce sectionalism tendencies in this model, as seems case in real companies too. Why did the uniformity coefficient carry out the power law to `SecSD`? There was no explicit setup in which an exponential effect was derived. The power law is believed to influence the effect of section structure on a network. Since the members of a section greatly influence each other in a high dense network, the influence of a single agent is shared quickly. Where there is little connection between cross-over sections, one new link of a cross-over section will influence all links of the target. The power of the second link apparently serves as an exponential influence.

This study is the first fundamental step of an informatics approach toward the sectionalism phenomenon, upgraded by much research field knowledge. To sophisticate this organization informatics approach, achievements of multi-agent modeling is set to the core. This is deeply related to sociology and organizational theory and the field of innovation, conflict management, and studies of the role of a 'gatekeeper' as a bridge between companies and sections. For the details of modeling, especially for submodel of interaction and sectional effects, much attention has to be paid to studies of 'false consensus effect' or 'semantic noise' that originated from psychology and social psychology. To clarify network effect, knowledge from mathematic sociology or econophysics, network centrality, assortative mixing, power law distribution, and affiliation network are also powerful concepts to that can enrich the comprehension of the sectionalism phenomenon. It is high priority in future works to apply our model to empirical research by investigating human network structures and organization factors in actual company, public entities, or academic organizations.

6 Conclusion

To study the effect of communication network structures on the sectionalism phenomenon, we designed an applicable organizational model, based on a multiagent model. There are four new approaches in this model study. 1) An agent has distributed network information and a sense of values. 2) By extending a conventional random network, a partitioned random network is introduced. 3) Sectional and interaction effects were incorporated as changes of agents' sense of values. 4) An index for sectionalism was designed independently by gathering groups of agents.

Our proposed index reflects the heterogeneity of sections so that its sensitivity is about 1.3 time's higher than cluster coefficients as a conventional index in experimental simulations of handmade designed organizations. One example study clearly illustrated the effects of cross-sectional links on sectionalism reduction by following the so-called "power law." We noted a lot of future works in the discussion, since we believe in the importance of network effects on inter-heterogeneity and intra-homogeneity structures such as actual organizations.

Acknowledgments

This research was funded as part of "Research on Human Communication" by the National Institute of Information and Communications Technology. It was also supported in part by a Grant-in-Aid for Scientific Research, Japan (encouragement of young scientists (B) 15700122) and by the "Foundation for Fusion of Science and Technology (FOST)."

References

1. McDonough, III, E.F.: Investigations of factors contributing to the success of cross functional teams. Journal of Product Innovation Management **17**(3) (2000) 221-235
2. Weiß, G.: Multiagent Systems -Modern approach to distributed artificial intelligence. The MIT Press (1999)
3. Prietula, M. J., Carley, K. M.,and Gasser, L. (Ed): Simulating Organizations – Computational models of institutions and groups. The MIT Press (1998)

4. Torenvlied, R.,and Velner, G.: Informal networks and resistance to organizational change –The Introduction of quality standards in a transport company. Computational & Mathematical Organization Theory **4**(2) (1998) 165-188
5. Bar-Hillel, Y.,and Carnap, R.: “Semantic Information.” British Journal of the Philosophy of Science **4** (1953) 47-157
6. Ross, L., Greene, D.,and House, P.: The “false consensus effect” –An egocentric bias in social perception and attribution processes. Journal of Experimental Social Psychology **13** (1977) 279-301
7. Erdös, P.,and Rényi, A.: On random graphs. Publ. Math. Debrecen **6** (1959) 290-297
8. Allen, T.: Managing the Flow of Technology. The MIT Press (1977)
9. Barabási, A-L.: Linked –The new science of networks. Perseus Publishing (2002)
10. Wasserman, S.,and Faust, K.: Social Network Analysis –Methods and applications. Cambridge University Press (1997)
11. Watts, D.J.,and Strogatz, S.H.: Collective dynamics of “small-world” networks. Nature **393** (1998) 440-442

Analyzing and Taming Collective Learning of a Multiagent System with Connected Replicator Dynamics

Masaaki Kunigami[1] and Takao Terano[2]

[1] Graduate School of Business Science, Tsukuba University
3-29-1, Otsuka, Bunkyo-ku, Tokyo 112-0012, Japan
{kunigami, terano}@gssm.otsuka.tsukuba.ac.jp
[2] Computational Intelligence and Systems Science,
Tokyo Institute of Technology
4259 Nagatsuda-Cho, Midori-ku, Yokohama 226-8502, Japan
terano@dis.titech.ac.jp

Abstract. This paper analyzes complex collective behaviors of a multiagent system, which consists of interacting agents with evolutionary learning capabilities. The interaction and learning of the agents are modeled by the concept of Connected Replicator Dynamics expanded from evolutionary Game Theory.The dynamic learning system we analyze shows various behavioral and decision changes including bifurcation of chaos in the sense of physical sciences.The main contributions of the paper are summarized as follows: (1) In amultiagent system, the emergence of chaotic behaviors is general and essential, even if each agent does not have chaotic properties; and (2) However,asimple controlling agent with the Keep-It-Simple-Stupid (KISS) principle, or a sheep-dog agent, is able to domesticate or tame the complex behaviors.

1 Introduction

In multi-agent system for social and economic study, we often observe the emergence of spontaneous cooperation and self-organization phenomena. On the other hand, in a complex adaptive system, chaotic and unpredictable behaviors in the sense of physical sciences are also very common. Therefore, to understand the collective behaviors of a multiagent system, in addition to the equilibrium of the system, unstable states which seem to be succeeding permanently also must be investigated.

Expressing the agent functions by the expanded Replicator Dynamics, this paper describes that such a state instability will emerge among the interacting agents. Moreover, to tame such complex and unpredictable states of a multi-agent system, a simple mechanism, or a sheep dog agent is designed based on a bounded rationality with the Keep-It-Simple-Stupid (KISS) principle described in Axelrod [1].

In a multi-agent system, each of the agents does not face to a static environment independently. If an environment changes independently (or quasi-statically), a rational agent can acquire the optimal behavior easily. However, since a multi-agent system includes the other agents, which also learn and interact the environment and the landscape inside each agent change simultaneously. For this reason, a way of interaction between agents and reaction speed over environment are considered sometimes more important than rationality or optimality.

A. Sakurai et al. (Eds.): JSAI 2003/2004, LNAI 3609, pp. 126 – 134, 2007.

There are static and dynamic approaches to study the behavior of such agents that interact each other. Static approaches analytically derive the equilibrium under the given rationality of agents, but cannot cope with permanent states except the equilibrium such as a hetero-critical cycle or a strange attractor. Dynamic approaches observe the change of the states with explicitly given decision mechanisms. While it is difficult for analytical handling, it enables us to describe a macro attitude change based on bounded rational micro agents, and the path dependency in multiple equilibriums, and the existence and properties of long-lived transients (Axtell [2]). This paper focuses on the features of the multi-agent system as dynamics and shows the emergence of state instability or chaotic behaviors as a non-equilibrium state in the above learning agent systems, moreover examines the possibility of the adaptive control. In precedent researches, Hogg and Hurberman [3], Ushio, Imanori, & Yamasagi, [4] studied on the chaos and its stabilization in a multi-agent system in which agents use multiple resources competitively with time delay and incomplete information. Since in the Hogg-Hurberman system, the uncertainty and the time delay are represented explicitly with a nonlinear function, their agent system is not enough universal, and is hard to become a general illustration of the emergence of state instability in an interacting agents system.

Kunigami and Terano [5] have shown that the chaotic motion emerges from a mutually learning agents system by connecting two Replicator Dynamics. They have suggested that a simple chaos control is effective for multi-agent system. Independently, Sato and Crutchfield [6] have studied on Coupled Replicator Dynamics. They have investigated the relation of the game structure and bifurcation to the chaos.

This paper will demonstrate the following three propositions. 1) In the multi-agent system, emergence of state instability is common phenomenon; 2) The interaction between agents is essential in state instability; and 3) There exists a standard chaos control technique which can be applicable and has preferable natures as control mechanism for a multi-agent system.

2 Agent's Decisions and Their Interactions

Based on the concept of replicator dynamics in evolutionary game theory or theoretical biology, we formulate the internal state of the agent as follows:

$$\frac{1}{x_i}\frac{dx_i}{dt} = c \cdot \left((\mathbf{Ax})_i - \mathbf{x} \cdot \mathbf{Ax} \right), \ \sum_{i=1}^{N} x_i = 1, \qquad (1)$$
$$\mathbf{x} = \left(x_1, x_2, \cdots, x_N \right), \ \mathbf{A} \in M_{N \times N}.$$

In the evolutionary game theory, the state variable x_i means the population ratio of the individual who chooses the i-th alternative of N kinds of pure strategies. The matrix A= $\{a_{ij}\}$ represents a pay-off, when an individual taking the i-th strategy plays a game against another taking the j-th one. At each moment, some individuals are picked randomly, play games, and take payoff. The RD (1) describes that a population ratio of i-th strategy player.

In this paper, we reinterpret (1) as inner state description of an agent. Where, x_i means the agent's *i*-th decision value, *A* is a weight matrix among decisions to represent the internal model of the agent, and c is a constant value to determine the speed of the decision changes. The agent makes decisions based on the inner rationality or an inner belief A. Under our interpretation, the decision value x_i means the ratio of the strategy i chosen by the agent. We assume that the evaluation (i.e. a_i $(x_i|\mathbf{x})$) to the choice i of decision $\mathbf{x}$ is given with an linear function (i.e. a_i $(x_i|\mathbf{x}) = (\mathbf{Ax})_i$) and that the changes of the decisions are described from the inner state driven by the difference (satisfaction or regret) from the expected values (i.e. $\mathbf{x}\cdot\mathbf{Ax}$).

Furthermore to express the interaction between two or more agents, we assume that the state of the observed external world is evaluated by linear function $\gamma A^0 y$. Thus, dynamics of interacting agents represented by the equation (2) is derived, which expresses that an agent behaves based on an inner model for evaluation of the observed environment and its own decisions.

$$\frac{1}{x_i}\frac{dx_i}{dt} = c_a \cdot \left((\mathbf{Ax})_i + \gamma_a \cdot (\mathbf{A}_0\mathbf{y})_i - \mathbf{x}\bullet\mathbf{Ax} - \gamma_a \cdot \mathbf{x}\bullet\mathbf{A}_0\mathbf{y} \right),$$
$$\frac{1}{y_i}\frac{dy_i}{dt} = c_b \cdot \left((\mathbf{By})_i + \gamma_b \cdot (\mathbf{B}_0\mathbf{x})_i - \mathbf{y}\bullet\mathbf{By} - \gamma_b \cdot \mathbf{y}\bullet\mathbf{B}_0\mathbf{x} \right), \quad (2)$$
$$\sum_{i=1}^{N} x_i = \sum_{i=1}^{N} y_i = 1.$$

The interaction of the agents is illustrated in the Figure 1. Please note that (1) the formulation is a very simple extension of conventional Replicator Dynamics; (2) For the first and second propositions in Section 1, agents in the model have the minimum internal states and interaction functionalities according to Axelrod's KISS principle.[1] and have no individual chaotic property nor bifurcation parameters to cause the state instability; and (3) For the third proposition, all the parameters required for control are adaptively determined from the observation of the macro behaviors of the agents.

However, using the formulation, when we set the parameters γ_a and γ_b gradually larger from 0, the system will show from simple, then periodic, finally to the chaotic behaviors in the sense of the changes of decisions. This means that very complex collective learning phenomena will emerge in our multiagent system.

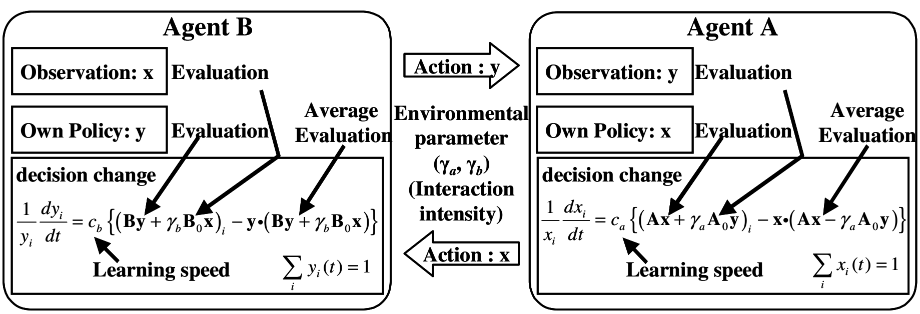


Fig. 1. Formulation of the Interaction between two Learning Agents

Such an agent's interaction can be found out in the consumption game of goods with the nature which flaunts own riches. Each agent evaluates the utility from its consumption of goods by the inner model represented by matrices $\gamma_a\mathbf{A}_0$ and $\gamma_b\mathbf{B}_0$, in which γ_a and γ_b respectively represent the sensitivity of the agent A and B. Therefore, this dynamics is expected to express the bandwagon effect (utility from following the others) and the snob effect (the utility from conspicuousness) in consumption behaviors.

3 Emergence of Chaos

On the "connected Replicator Dynamics" described in the previous section, the emergence of chaos is numerically observed. Typical evaluation matrices ($\mathbf{A}$, $\mathbf{A}_0$, $\mathbf{B}$, $\mathbf{B}_0$) and time constants of learning (c_a, c_b) are as follows:

$$\mathbf{A}=\begin{pmatrix}1 & 3 & -0.4\\ -1 & 1 & 6\\ 3 & -2 & 1.3\end{pmatrix},\ \mathbf{A}_0=\begin{pmatrix}2 & 3.33 & 1.6\\ 0.67 & 2 & 6.67\\ 4 & 1.33 & 2.67\end{pmatrix},\ \mathbf{B}=\begin{pmatrix}1 & -2 & 1.8\\ 4.5 & 1 & -0.9\\ 0.4 & 3.7 & 1\end{pmatrix},\ \mathbf{B}_0=\begin{pmatrix}2 & 0.26 & 2.67\\ 4.67 & 2 & 1.33\\ 1.73 & 4.4 & 2\end{pmatrix},\ \begin{pmatrix}c_a=0.5\\ c_b=0.5\end{pmatrix} \tag{3}$$

$$\mathbf{A}=\begin{pmatrix}-1 & 1 & -2.5\\ -3 & -1 & 4\\ 1.5 & -4 & -1\end{pmatrix},\ \mathbf{A}_0=\begin{pmatrix}0.5 & 1 & -2\\ -1 & 0.5 & 3\\ 1.5 & -1 & 0.5\end{pmatrix},\ \mathbf{B}=\begin{pmatrix}-1 & -3.5 & 1.5\\ 3 & -1 & -3\\ -2 & 2 & -1\end{pmatrix},\ \mathbf{B}_0=\begin{pmatrix}-0.5 & -1 & 1.5\\ 2 & -0.5 & -1.5\\ -1.5 & 3 & -0.5\end{pmatrix},\ \begin{pmatrix}c_a=0.5\\ c_b=0.5\end{pmatrix} \tag{4}$$

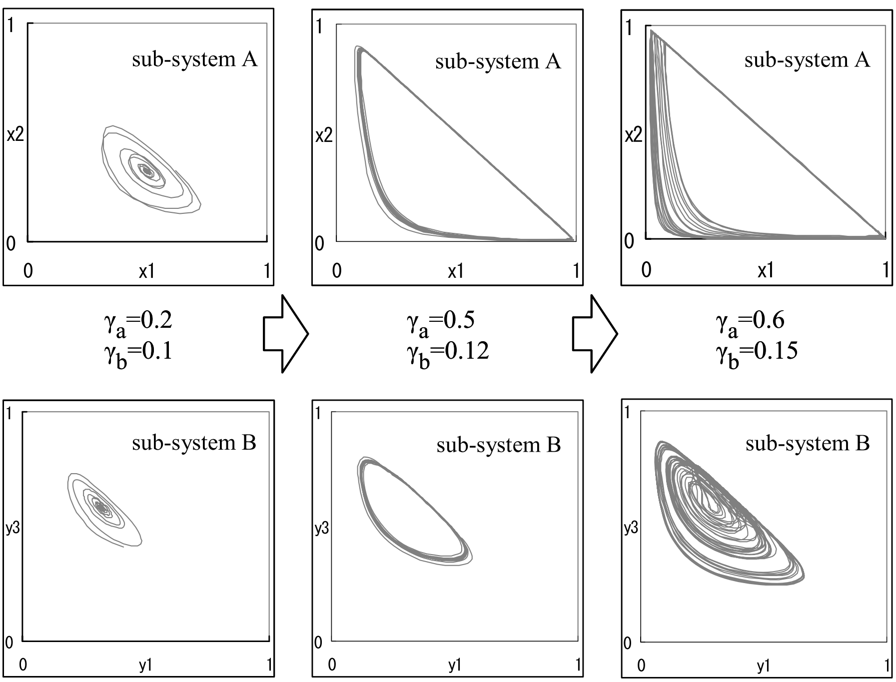


Fig. 2. Bifurcation to chaos, with equation (3)

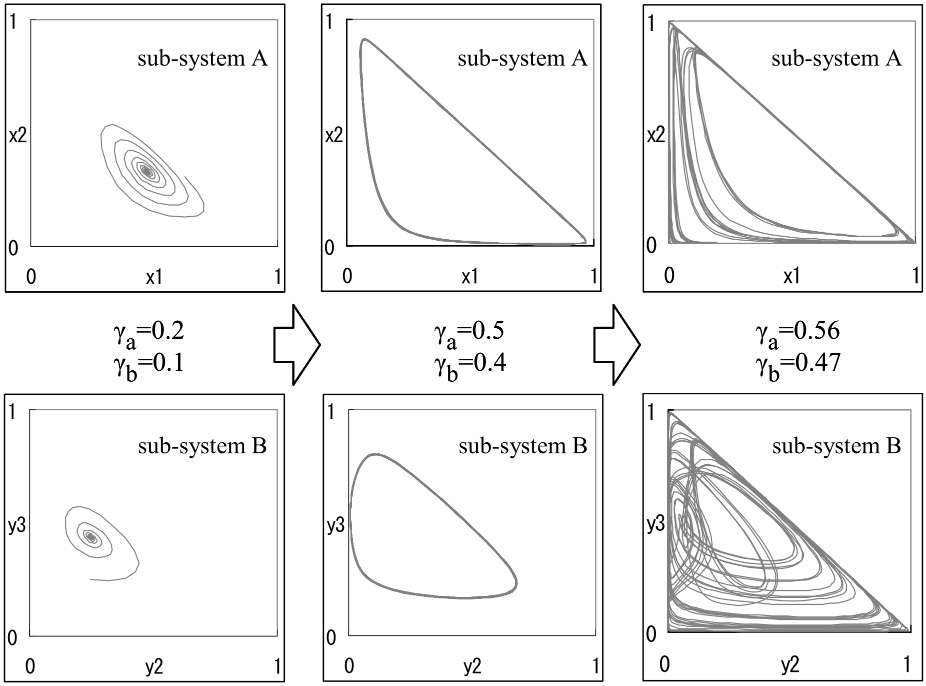


Fig. 3. Bifurcation to chaos, with equation (4)

The interaction coefficients γ_a and γ_b are changed as the bifurcation parameters.

At first, in the no-interaction case (i.e. $\gamma_a = \gamma_b = 0$), both agents in the system converge to steady state independently. As gammas grow large, the system shows complex behaviors after periodic / quasi-periodic motions. (Fig.2 and Fig.3). The positive maximum Lyapunov exponents that observed along these attractors prove the chaos.

The system consists of interacting two 3-dimensional continuous dynamics. Since each dynamics has one constraint, the independent degrees of freedom are 2 each. Thus, from the Poincaré-Bendixson's theorem, any agents do not generate chaos independently (gamma= 0).

For this reason, the chaos of this multi-agent system is essentially caused by the mutual interaction of agents who are not chaotic.

In this minimally constructed example, the bifurcation to chaos is caused by the changes of the interaction intensity parameter (gamma). Such interaction intensity between agents reflects to the spatial distance or the social relation. It suggests the importance of control and/or the stabilization of a chaotic behavior that even in the individually stable agents, some changes of their interaction intensity or reaction speed will cause chaotic behavior in a whole system.

4 Sheep Dog Agent to Tame the Chaotic Behaviors

In the multi-agent system literature, the concept of the institutional design by the micro-macro link is proposed, in which the equilibria are rearranged by changing the macroscopic system parameters. In the view of chaos control, this corresponds to tuning of the bifurcation parameter (OGY method; Ott, Grebogi, & Yorke [7]) based on the knowledge or identification about dynamics.

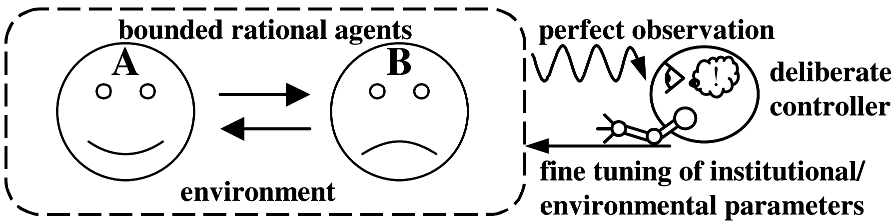


Fig. 4. Controlling Chaotic Behaviors often requires omniscience

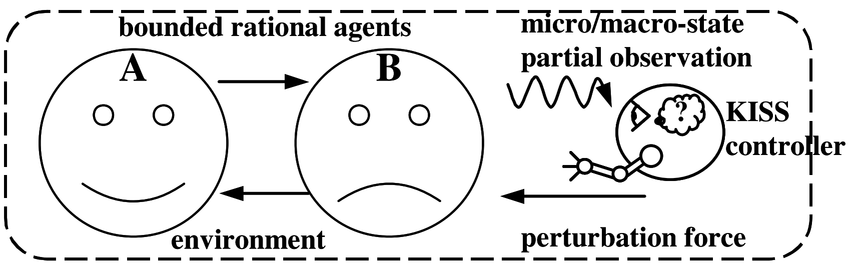
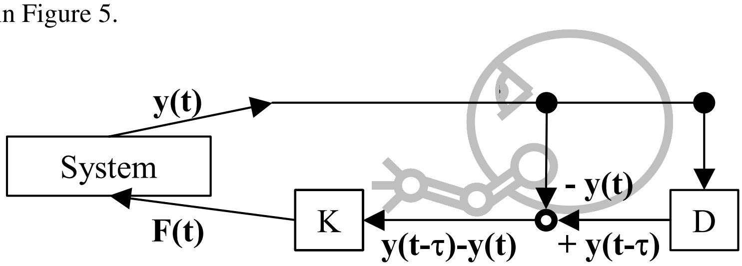


Fig. 5. Taming Chaotic Behaviors of Agents via Sheep Dog Agent

In such deliberative control concepts, omniscience and omnipotence (like God) are required for a controller. Even if we could design an indirect system control mechanism, it would not be realistic that perfect knowledge is assumed over individual agent's inner norm (the utility and preferences) or over the system equation. For agent-based stabilization of an autonomous distributed system, therefore, we believe that a KISS principle based control agent like a "sheepdog" is adequate to the system shown in Figure 5.



$$\left(\mathbf{K}\left(\mathbf{y}(t-\tau)-\mathbf{y}(t)\right)\right)_i \equiv \min\left[k_0 \cdot \left(y_{ii}(t-\tau)-y\ (t)\right),\ k_{\max}\right]$$

$$\tau \equiv \Delta t_{\text{peek}}^{(n)} \equiv t_{\text{peek}}^{(n)} - t_{\text{peek}}^{(n-1)}$$

Fig. 6. Outline of the Pyragas' Chaos Control Method

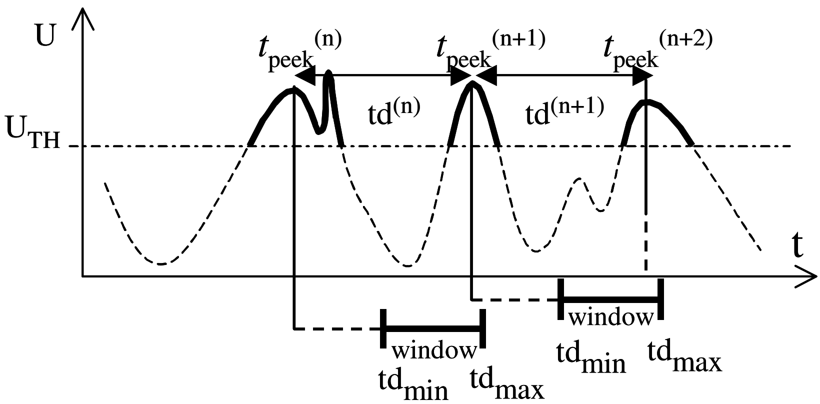


Fig. 7. The delay-time window selects admissible peak time. U_{TH} is the threshold of macro-state variable (e.g. social utility).

Although various methods were proposed about chaos control, in this research we have implemented improved Pyragas' method [8], since it is a continuous system control, and the simple control which does not require the prior knowledge on the system. Figure 6 shows the outline of Pyragas' method.

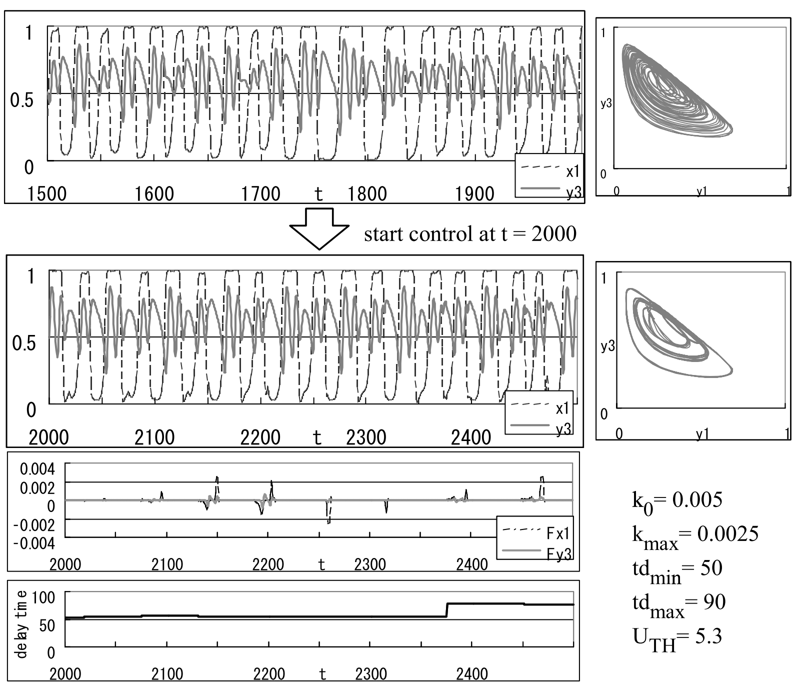


Fig. 8. Numeric example: Taming the chaotic motion with the equation (3) to periodic motion

Pyragas' method determines the feedback force by the difference between delayed system output and current one, then stabilizes a chaotic orbit into a periodic orbit. On the other hand, Pyragas' method cannot stabilize any system. In addition, the gain matrix and the stabilized period cannot be determined analytically.

As a method of determining delay-time autonomously from observed macro-system variable, Kittel-Pyragas [9] has proposed self-adapting delay-time control. However, this method is weak to noise, since small delay time around peak easily makes the system be unstable. Kunigami-Terano [5] improved stability with low-pass filtering. This research introduces a delay-time window that consists of minimum delay-time (td_{min}) and maximum one (td_{max}). In addition, we also use a threshold of social utility. Figure 7 illustrates how the self-adaptive delay time reflects the controller's preferences through the delay-time window and the utility threshold.

Figure 8 and 9 show numeric examples of the control corresponding with the equation (3) and (4).

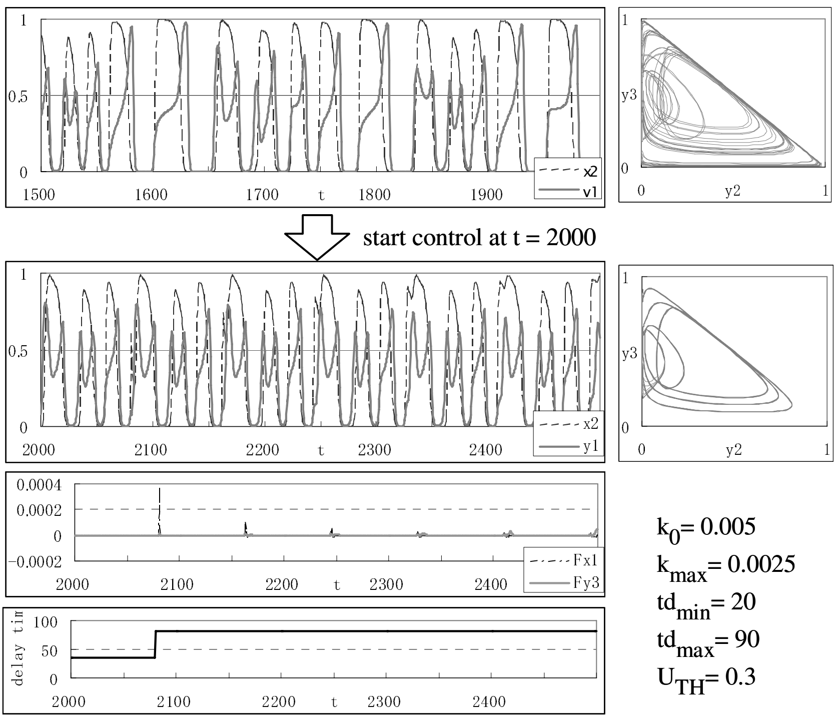


Fig. 9. Numeric example: Taming the chaotic motion with the equation (4) to periodic motion

5 Concluding Remarks

This paper has described the basic idea and configuration of interactive learning agents with Connected Replicator Dynamics using the two agent examples. Also we have shown the experimental results, the system shows that (1) The chaotic behaviors will emerge, even if each agent does not have chaotic properties; and (2) However, a simple controlling agent with the Keep-It-Simple-Stupid (KISS) principle, or a sheep-dog agent, is able to domesticate or tame the complex behaviors.

Furthermore, we have observed that, using three or more agents' interaction models with various communication network topologies, our learning agent framework also shows state instabilities. In the present stage, a comprehensive framework has not been established about various interactions of a multi-agent system. However, for example, it is possible to analyze Connected Replicator Dynamics or other decision-making dynamics about the universality and peculiarity of the dynamics reflecting the social structures.

References

1. Axelrod, R.: *The Complexity of Cooperation*. Princeton University Press (1997).
2. Axtell, R.L.: Why Agents? On The Varied Motivations for Agent Computing in the Social Sciences. Working paper No.17, *Center on Social and Economic Dynamics, The Brookings Inst*. (2000), http://www.brook.edu/es/dynamics.
3. Hogg, T., Huberman, B.: Controlling Chaos in Distributed Systems. *IEEE Transactions on Systems, Man, and Cybernetics*, Vol.21 No.6 (1991) 1325-1332.
4. Ushio, T., Imamori, T., Yamasagi, T.: Controlling Chaos in Discrete-Time Computational Ecosystem. In: Chen (ed.): *Controlling Chaos and Bifurcations in Engineering Systems*. CRC Press (2000) 625-644.
5. Kunigami, M., Terano, T.: Connected Replicator Dynamics and Their Control in a Learning Multi-Agent System. In: Liu,J., Cheung,Y., Yin,H.(eds.): Intelligent Data Engineering and Automated Learning. Lecture Notes in Computer Science, Vol. 2690. Springer-Verlag, Berlin Heidelberg New York (2003) 18-26.
6. Sato, Y., Crutchfield, J.P.: Coupled Replicator Equations for the Dynamics of Learning in Multiagent Systems, working paper of Santa Fe Institute, April (2002), http://www.santafe.edu/sfi/publications/Working-Papers/02-04-017.pdf.
7. Ott, E., Grebogi, C., Yorke, J.A.: Controlling Chaos. Physical Review Letters, vol.64 (1990) 1196-1199.
8. Pyragas, K.: Continuous Control of Chaos by Self-Controlling Feedback. *Physics Letters A*, Vol.170, No.6 (1992) 421-428.
9. Kittel, A., Parisi, J.: Pyragas, K., "Delayed Feedback Control of Chaos by Self-Adapting Delay Time. *Physics Letters A*, Vol.198 (1995) 433-436.

Semantic Authoring and Semantic Computing

Kôiti Hasida

ITRI, AIST
10F. Dai-Building, 1-18-3, Sotokanda, Chiyoda-ku, Tokyo 101-0021, Japan
hasida.k@aist.go.jp

Abstract. Semantic Computing is to design and operate information systems based on meaning and vocabulary shared by people and computers. It aims at closing the semantic gap, thus enabling closer cooperation between people and information systems, and thereby semantically enriching our life-world. Ontologies and constraints are major technologies to let people and information systems share common meaning. Semantic authoring is to compose information content together with explicit semantic structure based on ontologies. This not only reduces the cost of content composition but also improves the quality of the resulting content, by both freeing the author from worries about the order of presentation and providing her a perspicuous view of the logical content structure. Social interactions are much more generally modelled in terms of constraints than in terms of workflows or procedures. CBTO (compositional business-task organization) is a constraint-based framework to concisely describe uniformities of social interactions and thus provides a semantic-level scheme for coordinating various, possibly interactive, services with each other.

1 Introduction

Semantic computing is a technology to compose information content (including software) based on meaning and vocabulary shared by people and computers and thereby to design and operate information systems (i.e., artificial computing systems). Its goal is to plug the **semantic gap** through this common ground, to let people and computers cooperate more closely, to ground information systems on people's life world, and thereby to enrich the meaning and value of the entire life world.

Various technologies are necessary to allow people and computers to share meaning. For instance, so-called ubiquitous computing is to provide various information services based on the grounding of information systems on physical aspects of our life world by networks of sensors and other information artifacts ubiquitous in our physical environment. The rest of the present paper discusses the roles which **ontologies** and **constraints** play for people and computers to share meaning.

A. Sakurai et al. (Eds.): JSAI 2003/2004, LNAI 3609, pp. 137–149, 2007.

2 Semantic Authoring

Ontologies help people and computers share meaning because the graph-type representation is easy for people to understand and its syntax and semantics are both relatively simple. Although as discussed later not all graph-type representations are easy to understand, the simplicity of syntax and semantics contributes not only to the usability of information systems but also to the intuitive understandability for people. This section mainly discusses the role of ontologies in the creation of information content.

Semantic authoring is to compose intelligent content (content with explicit semantic structure) based on an ontology [1]. Authoring of content together with semantic structure may be considered more costly than usual authoring. However, the reality is the other way around. With an ontology and a user interface appropriate for the sort of content to compose, semantic authoring is easier than traditional composition of content and the resulting content is of higher quality. Below we consider such ontologies and user interfaces.

2.1 Discourse Semantic Authoring

Let us first consider text[2] as a type of content to compose. As depicted in Figure 1, traditional text composition has mainly two drawbacks. First, a lot of

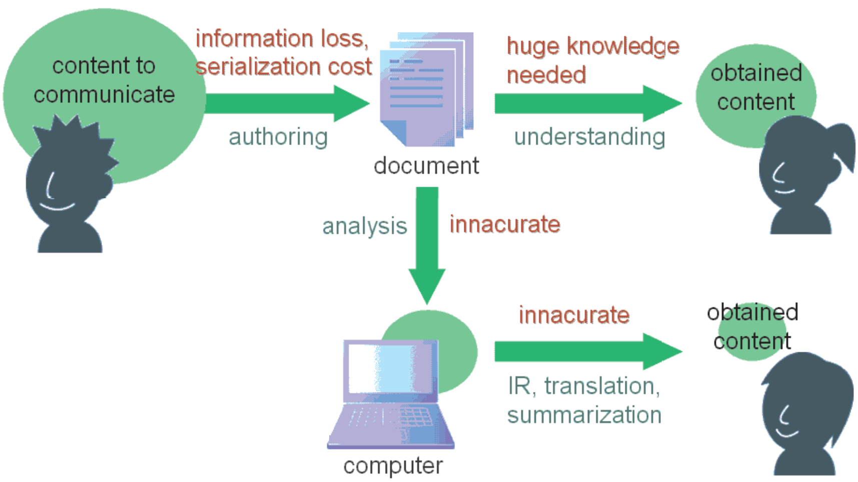


Fig. 1. Traditional Authoring

[1] This term 'semantic authoring' was probably coined by Handschuh and Staab [2].

[2] Here we consider documents whose content is basically logical, such as technical papers, contracts, specifications, and laws. Poems, novels, and so forth are out of our current scope.

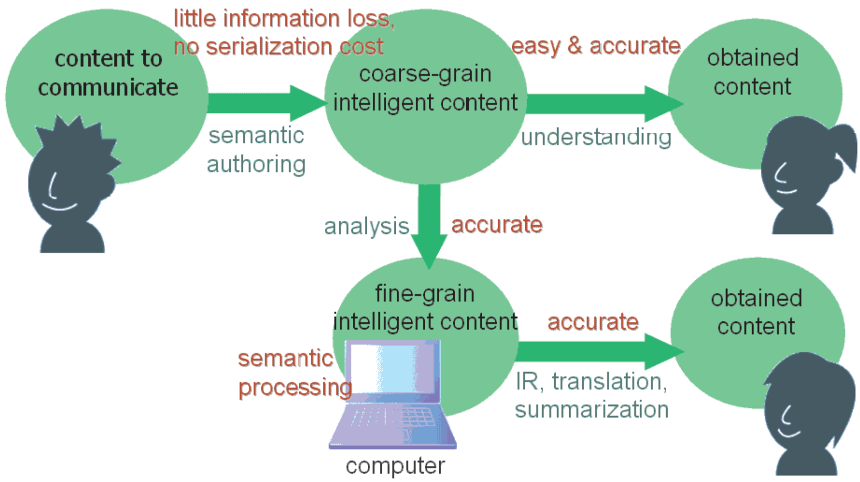


Fig. 2. Semantic Authoring

information is lost when people express their thoughts in terms of text. In fact, ambiguities arise as to the referents of missing subjects, the meaning of empty conjunctions, and so forth. This causes misunderstandings and other miscommunications. Also, a huge amount of knowledge is necessary in order to recover the missing information. Since computers do not have such knowledge, however, they cannot accurately understand texts. The quality of retrieval, summarization, translation, and so on based on such an inaccurate understanding is low. The second drawback of text is in its serial form with a fixed order of presentation. Text composition entails an extra cost of serialization, because the content that people have in mind lacks such an order most of the case. This serialization requires a complicated bookkeeping during the combinatorial problem solving.

Unlike the traditional text composition, on the other hand, the authoring method depicted in Figure 2 is much easier, both because it allows people to straightforwardly express the semantic structure of what they want to communicate, and because it entails little serialization for linear presentations. The resulting coarse-grain intelligent content encodes semantic structures more explicitly than text does, so that it is more comprehensible and less misleading to both the author and the other people. Also, information services such as retrieval and summarization are of much higher quality because computers can analyze such content very precisely.

Coarse-grain intelligent content is a graph such as in Figure 3. Below is a textual equivalent.

(*) A huge amount of content is necessary in order to realize ubiquitous information service. So the authoring of content must be easy. Semantic annotation is hence necessary. Retrieval must be quick and easy also in order to realize ubiquitous information service, and so again semantic annotation is necessary.

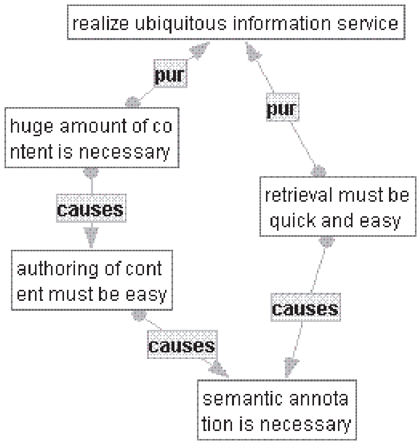


Fig. 3. Coarse-grain intelligent content

The content of each node in this graph is basically a simple sentence. The links are labelled with discourse relations such as 'causes' and 'pur(pose)' (and possibly dialogue acts as well, as discussed later), representing semantic relations among nodes. A node need not be a simple sentence, but as an endpoint of a link it must be an argument of the semantic relation represented by the links connected with it. Of course a node need not be a text, but may be a moving picture or audio data as far as its content can be an argument of the relation encoded by the link.

Given such short texts (plus annotations of coreferences, though we do not enter details here), computers can obtain a minute semantic structure as shown in Figure 4 with a high accuracy, and thus can provide high-quality services such as information retrieval, summarization, translation, and so on.

Both Figure 3 and Figure 4 straightforwardly depict instances of some ontologies as entity-relationship graphs. The ontolgy underlying Figure 3 addresses classes of eventualities (i.e., what sentences means, which are events, processes, and states), and properties of semantic relations among them. In addition to these classes and properties, the ontology used in Figure 4 defines classes of objects represented by noun phrases, and properties encoding **semantic roles** (their relation with eventualities) such as 'obj(ect)' and other semantic relations such as 'msr' (for 'measure').

Not all graph-based representations are easy for people to understand. Coarse-grain graphs such as in Figure 3 are more comprehensible than ordinary text, but fine-grain graphs such as in Figure 4 are rather hard to understand. This may be because the comprehension of a simple sentene is an almost unconscious and cognitively light process (simple sentences hence tend to be chunks) whereas the comprehension of a longer text accompanies a larger cognitve cost (long texts thus cannot be chunks). A piece of content corresponding to an ordinary text should hence be treated as a graph whose nodes are basically simple sentences

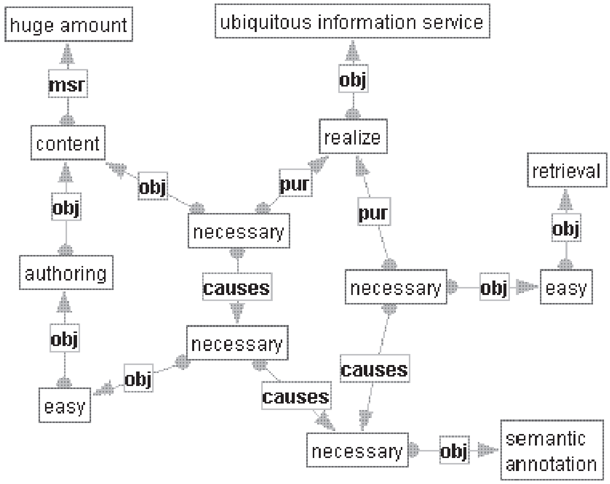


Fig. 4. Fine-grain intelligent content

and whose links are semantic relations among them. [3] We will refer to such a task of authoring as **discourse semantic authoring** (DSA).

Graph representations at an appropriate granularity, as mentioned above, are known to promote people's understanding and idea generation by providing explicit access to the logical structure of texts. Graph representations have been used in so-called idea processors. Idea processors assist idea generation by lowering the cost of thinking about the semantic content itself. Munermori, Yagishita, and Sudo [4] have shown that a text composition following a graph composition using an idea processor results in a better text than without a graph composition, where a text is good when it covers more points relevant to the theme (i.e., its scope is wide) and contains long inference chains (i.e., its thoughts are deep).

Idea processors have so far spread very little, however, compared with word processors and other authoring support tools. This is probably because the meaning of the graphs made by idea processors is rather unclear because the meaning of the links therein is not clearly standardized. Ordinary texts often contain explicit connectives such as 'because,' 'but,' 'for instance,' and so forth, but graphs have so far not clarified the meaning of links in a standard way. This is why graphs have not been able to convey complex semantic content to other people.

This problem is solved by standardizing the relations represented by links and socially sharing them. Classifying natural-language connectives (such as 'so that' and 'for the sake of') and compiling the set of the semantic relations (such as 'causes' and 'purpose') they encode, we can use those relations to express

[3] The semantic authoring editor we are developing at AIST supports graphs containing hypernodes (nodes containing graphs). A large graph can be treated as a reusable unit of content by storing it into a hypernode.

the logical content of arbitrary text in terms of a graph. The author convenes ISO/TC37/SC4/TDG3 to compile standard registries of such semantic relations and concepts. DSA will use about fifty relations including discourse relations and some dialogue acts out of the result of this standardization activity. Users of DSA need not memorize all those relations, however, because relation are never used uniformly. Much less than twenty relations are constantly used, and users will find it easy to memorize and utilize them.

Given this standardized set of semantic relations, ordinary texts can be (semi-) automatically generated from graphs. A text like (*), for instance, can be generated from the graph in Figure 3 by serializing the nodes while putting them together via connectives such as 'so' and 'hence' corresponding to the discourse relations represented by the links. (Of course pronominalization and other paraphrasing are also necessary, but we do not go into details here.) Another weakness of traditional idea processors is that composing both a graph and then a text after that looks like doing almost the same thing twice and appears to be more costly than the traditional text composition [4], which is dissolved by this functionality of automatic text generation.

2.2 Variation

DSA is appropriate for authoring the content of ordinary text, [5] but other types of ontologies and user interfaces are better for dealing with other types of content. To edit PERT diagrams and Gantt charts, for instance, you need an ontology to define the inclusion and dependency relations among events and the traditional diagrammatic representation for such content will be more comprehensible and easier to handle than a graph representation. For another example, a tabular (spreadsheet) interface is considered more convenient when treating many pieces of content of the same type. Note that they are variants of semantic authoring in the sense of authoring on the basis of some ontology.

It is important that those different sorts of content coordinate with each other via an ontology-based infrastructure even though they are operated on through different user interfaces with different types of views. For instance, you can use the DSA interface to comment on tasks in a Gantt chart and link these comments with each other, so as to coordinate DSA and project management across multiple tasks. Combinations of multiple user interfaces such as a PERT diagram and a graph will be necessary for dealing with such hybrid content based on a compound ontology.

[4] On the contrary, the entire cost of authoring should be smaller when authoring a rather complex content.

[5] Of course DSA is yet to be improved in several respects. During DSA, for instance, you probably want to label nodes with not only sentences but also noun-phrases. When you create noun-phrase nodes, you will need links connecting them with sentence nodes, but it is very cumbersome to assign semantic roles to those links. More details shall be discussed elsewhere.

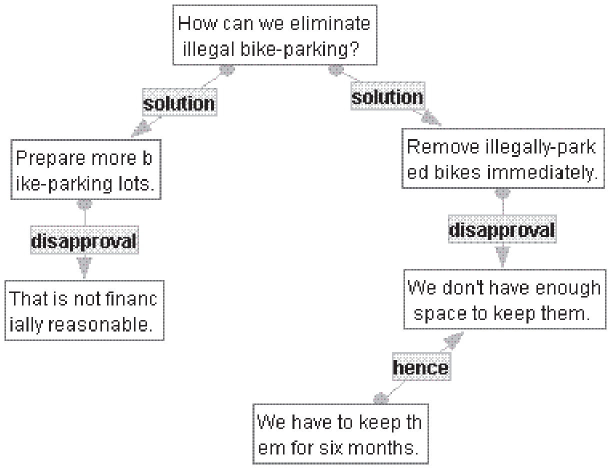


Fig. 5. Structure of dicusssion

2.3 Collaborative Semantic Authoring

We have so far discussed semantic authoring by individual users. All this discussion can of course be extended to group works. In fact, multiple users can enjoy better functionality than that of traditional groupware by sharing with one another the graph-type content they compose through semantic authoring (in particular DSA).

Discussion-supporing groupware such as gIBIS [1] allows you to structure discussions as graphs like Figure 5 and share them among the discussion participants. This will almost eliminate redundant repetitions of the same argument, allow simultaneous utterances, decrease oversights and deepen discussions.

Despite these merits, such groupware has not spread very much. This is because this type of groupware can be used for group works only, and those group works are remote from ordinary individual works.

The structure of discussion developing in such groupware and the graph content composed by DSA are essentially of the same type, except that the links in the latter have more concrete semantic types. So DSA serves as discussion-supporting groupware just by sharing the composed graphs among the discussion participants, without changing environments between individual works and group works. The group work on this basis allows you to share contexts more easily and to advance discussions more efficiently than by using mailing lists, blogs, and former discussion-supporting groupware, because in DSA explicit semantic (or pragmatic) structures such as in Figure 5 emerge and are shared as discussions develop. The quality of such a discussion will be improved further by high-quality retrieval, summarization, translation, and so on. All this holds true for other types of group works such as project management, too.

You can view response relations among e-mails, but that is insufficient to deal with discussions in general, which are often not trees. Trackbacks of blogs may form a graph like Figure 3. Like traditional idea processors, however, traditional blogs utilize no explicit standard of the meaning of links, which makes it hard for both people and computers to recognize the structure of discussions. In this connection, each message in traditional blogs has to be some substantial amount of text (something like a paragraph), also because standardized semantic relations are not utilized there. Also lacking a rich ontology for intersentential relations, traditional groupware for discussion is limited in addressing semantic structures of discussions.

DSA may be regarded as a type of blog in which the meaning of trackbacks are made explicit based on a standard ontology of semantic relations and each message may hence be as short as a simple sentence. This allows each user to contribute just a simple-sentence node or a link between nodes, rather than a paragraph, which should make it easier for people to join the discussion. Due to the explicit logical structure, the progress of discussion becomes more perspicuous and the discussion hence becomes more efficient, not just assisting existing participants but also allowing newcomers to quickly catch up and start contributing. It is also possible to automatically evaluate the quality of discussion and the contribution of each participant [3]. In addition to retrieval and summarization, this will also be useful for supporting collaborative creation of knowledge.

3 Constraint-Based Model of Society

In 2 we have discussed sharing meaning between people and computers concerning static types of content. Here we consider dynamic content, or programs. These programs are not necessarily localized within information systems, but in general may provide a computational model of the society involving both people and information sytems.

People and information systems do not share meaning of programs because programs are hard for people to understand. [6] Programs are hard to understand because they are complex. This complexity mainly arises due to descriptions of procedures. [7] The complexity of programs drastically decreases by abstracting procedures away. A program without procedural descriptions is called a constraint. A constraint program statically describes conditions on a certain object, but does not stipulate the procedure to meet those conditions. Compared with procedural programs, constraint programs account for more complex computational processes by simpler specification.

Constraints of course cannot account for all the computational tasks. However, constraint-based computational models can probably explain cognitive processes

[6] It is controversial whether computers understand the meaning of programs, but we do not consider this issue here.

[7] Another cause of the complexity is intermediate values which are hard to understand based on daily intuitions. Since procedures tend to need a working memory and a working memory tends to contain such intermediate values, however, the description of procedures is the main cause of the difficulty, after all.

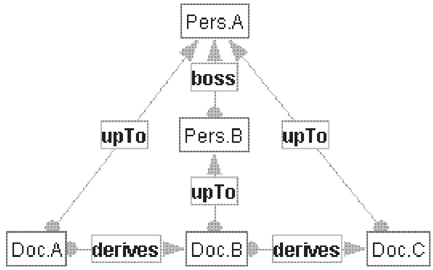


Fig. 6. A constraint-based workflow description

in general. Along the same line, social interactions may be modeled in terms of constraints, too.

Let us consider workflows as an example of social interaction. Figure 6 is a constraint program representing a workflow involving three tasks: the composition of document A, B, and C. Person A is responsible for document A and C, and person B for document B. When document A is finished, the composition of document B is requested, and when document B is done, the composition of document C is requested. Binary relation `derives` is a constraint to the effect that its second argument is newer than its first argument. Suppose for instance that document A has just been composed. Then it is newer than document B, which violates this constraint, so that a renewal of document B is requested. This request is implemented by notifying person B that she should work on document B.

The constraint program in Figure 6 does not necessarily mean that document A, B, and C should be worked on in this fixed order. For example, person A may touch upon document A while working on document C after document A and B have been composed. This violates again the constraint that document B should be newer than document A, and hence a request will be issued to revise document B. Here person A is regarded as having pulled her task back to document A or having passed the work back to document B. Thus Figure 6 accounts for very complex processes involving lots of pullbacks and pushbacks.

Figure 6 as a whole may be regarded as one compound document consisting of the three documents and relations among them. Many (perhaps practically all) workflows can be seen as such compound documents. Collective decision-making processes and other processes with greater degree of freedom concerning the order of tasks can of course be accounted for by such constrained compound documents.

We are currently developing **CBTO** (Compostitional Business-Task Organization), a constraint-satisfaction system to provide such a general framework for computational models of society. In CBTO, the execution of a business task is viewed as the composition of a proof tree, and services are regarded as providing parts to this proof tree. Constraints are described as Horn-clause programs. Each Horn clause is unfolded either interactively following the human user's operation or automatically as part of a program transformation.

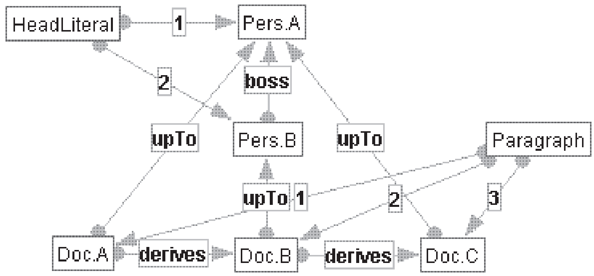


Fig. 7. A Horn clause generating the workflow in Figure 6

The workflow in Figure 6 can be instantiated by unfolding the Horn clause in Figure 7. Here the predicate defined by this Horn clause is `workflow`; Predicates are classes in the ontology and atomic formulae (literals) are their instances, in the current formulation of CBTO. We have omitted the `paragraph` literal in Figure 6. The link labelled with integer i connects a literal to its i-th argument. A symbolic representation of this Horn clause is shown in Figure 8. This clause

```
workflow(Pers_A,Pers_B) :- boss(Pers_B,Pers_A),
  upTo(Doc_A,Pers_A), upTo(Doc_B,Pers_B), upTo(Doc_C,Pers_A),
  derives(Doc_A,Doc_B), derives(Doc_B,Doc_C),
  paragraph(Doc_A,Doc_B,Doc_C).
```

Fig. 8. Symbolic Representation of the Horn clause in Figure 7

is unfolded with two people corresponding to person A and person B specified as arguments to the head literal.

Figure 9 depicts the only definition clause of predicate `paragraph`. The body of this clause is assumed to instantiate never or arbitrarily many times. (This formulation, rather than recursive repetition, is to allow deletions in an order different from that of instantiations.) Figure 10 shows the result of two instantiations. Note here that this allows pullbacks and pushbacks not only concerning an entire document but also concerning each paragraph.

Built-in predicates and demons are used to implement constraints which are hard or impossible to address by Horn clauses. For instance, the `derives` property discussed above is a built-in predicate. The computation to obtain eigenvalues of matrices should be implemented as a built-in predicate, too. The constraint that one same person cannot participate in two different events at one time can be realized as a demon triggered when somebody is involved in two different events on, say, the same day. Implemented as another demon is the constraint to the effect that two events at different places involving the same person must accompany an event of her transportation in between.

Many built-in predicates restrict the input/output directions of their arguments. For instance, it is unrealistic to reverse the I/O of the computation of

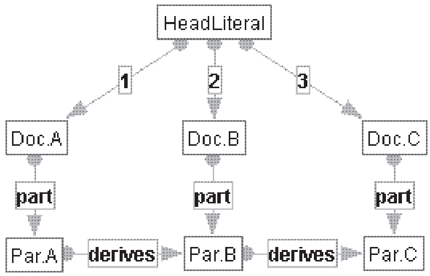


Fig. 9. A Horn clause to add paragraphs

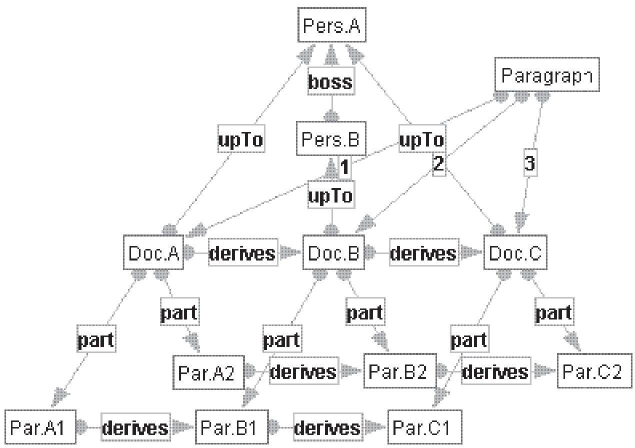


Fig. 10. A workflow with two paragraphs in each document

matrix eigenvalues. Predicates defined by Horn clauses may also be directed if some of their definition clauses contain directed predicates. To guarantee that no deadlock occurs in CBTO programs, we can check that each Horn clause contains no fixed I/O cycle, assuming that each argument of each predicate has a fixed I/O pattern; i.e., an argument is either fixed for input only, for output only, or for both.

Introducing external Web services as atomic formulae (literals) with exter nally-defined predicates, CBTO provides a light-weight and simple infrastructure for semantic Web service to coordinate various services based on their meaning. In this connection, various types of groupware including project management systems and enterprise systems can be implemented by combining CBTO with GUI generated by ontology-based stylesheets, though we do not discuss further details here.

Several frameworks of semantic Web services have been proposed. OWL-S [5] is close to traditional procedural programming language. WSMO [7] is a framework based on the planning technology in AI. WSMO may be regarded as a constraint-based method, but WSMO programs tend to be more complicated than CBTO programs because in WSMO you must specify preconditions and postconditions of each action. Rule language WRL [6], which is presupposed by OWL-S and WSMO, involves Horn clauses. So CBTO is a subset of WRL and thus both its specification and its programs are much simpler than in OWL-S, WSMO, and even WRL.

Another feature of CBTO is that its constraint-based coordination can accommodate not only one-shot services but also interactive ones, including both ordinary client-server interactions and constraint-satisfaction interactions. For instance, the business interaction involving pullbacks and pushbacks discussed above is of a constraint-satisfaction type. OWL-S, WSMO, and other frameworks of Web services, semantic or not, fail to directly address such interactions as far as they assume rather straightforward grounding on WSDL.[8] Unfortunately, OWL-S and WSMO are fundamentally incompatible with interactive services due to their procedural and/or plan-based formulations.

In this connection, the functionality required for CBTO which most of the constraint-satisfaction systems lack is the treatment of assumptions and contradictions for the sake of interactions with the environment involving human users. As discussed regarding Figure 6, for example, there may arise contradictions during constraint satisfaction. The assumptions containing such contradictions must be maintained rather than rejected as long as these contradictions may be dissolved through some interactions with the environment.

4 Concluding Remarks

We have discussed the role of ontologies and constraints in establishing a common semantic ground between people and computers as a basis of semantic computing. Something like CBTO, which is an integration of various functionalities for fundamental computation and coordination based on ontologies and constraints, is considered as a next-generation operating system. Through semantic Web service and ubiquitous computing, however, this new OS will be something like a computational model grounded on the entire society, which is very different from the current operating systems. The potential impacts of this new technology are yet to investigate.

References

1. Jeff Conklin and Michael L. Begeman. gibis: A hypertext tool for team design deliberation. In *Proceeding of the ACM conference on Hypertext*, pages 247–251, 1987.

[8] Of course, the grounding of CBTO is more complex in order to address interactive services in terms of multiple calls for one-shot services via WSDL.

2. Siegfried Handschuh and Steffen Staab. Authoring and annotation of web pages in cream. In *Proceedings of WWW2002 (ACM 1-58113-449-5/02/0005)*, 2002. Institute AIFB, University of Karlsruhe, 76128 Karlsruhe, Germany.
3. Naoki Kamimaeda, Noriaki Izumi, and Kôiti Hasida. Discovery of key persons in knowledge creation based on semantic authoring. In *Proceedings of KMAP 2005*, 2005.
4. Jun Munemori, Kazuyo Yagishita, and Masaru Sudo. Evaluation of idea generation method and its supporting groupware. In *Proceedings of the Third International Conference on Knowledge-Based Intelligent Information Engineering Systems*, pages 54–57, 1999.
5. W3C. OWL-S: Semantic markup for web services, 2004. http://www.w3.org/Submission/OWL-S/.
6. W3C. Web rule language (WRL), 2005. http://www.w3.org/Submission/WRL/.
7. WSMO.ORG. Web service modeling ontology, 2006. http://www.wsmo.org/.

Social Summarization for Semantic Society

Yasuhiro Katagiri[1], Toru Takahashi[1], and Noriko Arai[2]

[1] ATR Media Information Science Laboratories
{katagiri,toru}@atr.jp
[2] National Institute of Informatics
arai@nii.ac.jp

Abstract. We propose the concept of social summarization as a new technology for semantic computing. Social summarization focuses on human evaluative acts toward information, and provides an alternative to the content-based methods employed in the conventional information summarization technologies. We describe the idea of social summarization and its implementation in the community system TelMeA2003, which is being developed to investigate its effectiveness in supporting collaborative activities in online communities. We also report on the preliminary analysis of TelMeA2003 based on our experience obtained in a distance learning community e-Kyositu.

1 Introduction

Online communities provide a means to exchange ideas and opinions among people and facilitate new ideas and mutual understandings. Information contents are not always guaranteed to be presented in precise forms. Ideas are proposed, countered, evaluated and modified before they get finally agreed upon. Information filtering [1] is obviously lacking in its scope to capture and support this social process of information processing. Any form of information summarization is invaluable to drive this social process forward.

We propose the concept of social summarization, a new method to support the structuring and summarizing of information being created and exchanged in online communities.

2 Social Summarization

Text summarization techniques developed in the Natural Language Processing community have been focusing on the textual contents and forms. Lexical/encyclopedic knowledge, discourse models and statistics are employed to assess the importance of information pieces expressed in the text.

The idea of social summarization is to shift the focus from text itself to people participating in discussions. People express and negotiate for their assessment on pieces of information. More people would engage in assessment for more important ideas, and most important ones should be shared among people. Evaluative

A. Sakurai et al. (Eds.): JSAI 2003/2004, LNAI 3609, pp. 150–157, 2007.

attitudes are often indicated by the choice of speech acts or even expressed with non-verbal means. Social summarization aims to capture the core information parts from evaluative assessment expressed by discussion participants.

The basic ideas to implement social summarization in online communities are the following:

Act-based evaluation. Categorize participants' utterances and other behaviors into two types of acts; information acts and evaluative acts. Information acts are to establish informational relationships between messages, whereas evaluative acts are to perform evaluative assessments on messages and participants. Examples of information acts and evaluative acts are shown in Table 1.

Evaluation network construction. Build a network structure out of conversation log of exchanges in a community based on the categorization of acts. The network consists of nodes of messages and links of either informational or evaluative types. Types of information acts, such as questioning, answering, agreeing, determine types of informational links, whereas types of evaluative acts, such as agreeing, acting respectfully, exhibiting happy emotion, determine types of evaluative links and target objects. Performative verbs are mostly directed at messages, whereas interpersonal attitudes and affective expressions are for participants.

Summarization from network. Once the network structure is extracted, various network analysis techniques, such as are described in [2], can be applied to assign importance measures to messages, from which summarization can then be constructed.

This scheme of social summarization has the following benefits:

(1) Information collection is automatic, and no extra steps, such as questionnaires, are needed for collecting evaluative assessment from participants.
(2) Social summarization itself does not require text content analysis, although, of course, Natural Language Processing techniques can be employed to obtain better summarization results.

Table 1. Examples of information/evaluative acts

Information act	Evaluative act		
ask	agree	act respectfully	happy
query	appreciate	act friendly	joyful
answer	criticize	act emotionally	surprise
agree	contradict	act frowningly	anxious

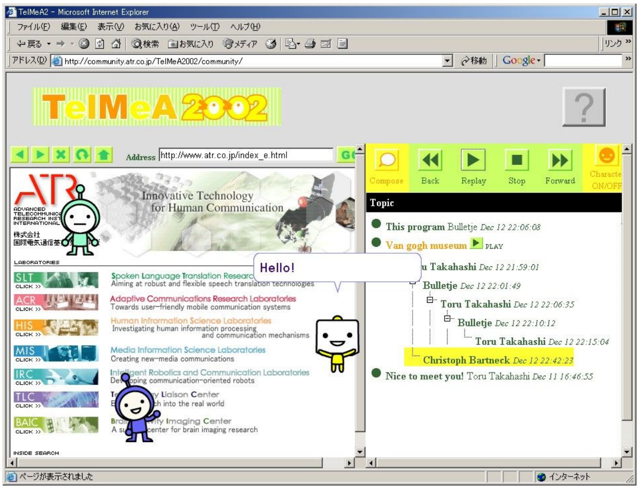


Fig. 1. Asynchronous community system with personified media

(3) The general scheme proposed above is open to wide range of summarization algorithms to accommodate different types of communities.

3 TelMeA: Asynchronous Online Community System with Personified Media

In order to further investigate the idea of social summarization, we are developing an asynchronous online community system TelMeA2003. TelMeA2003 employs personified media [3], a set of animated screen characters for community participants to engage in conversations. The design of TelMeA2003 was motivated by the following strategies for the development of social summarization technologies:

Personified media: Employ a set of animated characters, each of which is controlled by a participant and functions as an agent for her. Animated characters can provide participants with a rich repertoire of expressing themselves with facial expressions and gestures, which enable them to openly and freely disclose their evaluative attitudes toward information as well as toward people.

Non-verbal cues: Exploit non-verbal cues employed by the participants to obtain evaluative assessment information which can then be utilized in summarization.

Script language: Design a script language for both media control and participant behavior analysis. Community participants specify their own characters' behaviors. With a suitable GUI, the specifications can be made in terms

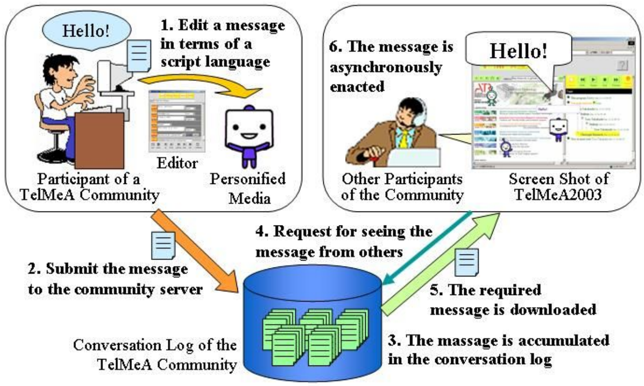


Fig. 2. Conversation process in TelMeA2003

of a script language. They will be accumulated as a conversation log, which are then subjected to summarization.

Figure 1 shows a screen shot of TelMeA2003, and how the conversational interactions proceed in TelMeA2003 is schematically shown in Figure 2. Different from conventional text-based bulletin board systems (BBSs), TelMeA participants specify the behaviors, both verbal and non-verbal, of their personified media by means of the TelMeA editor (see Figure 3), and share multimodal behavior representations of their personified media.

The Editor provides its users with five types of action representations for composing messages; 1) speech, 2) affective expression, 3) interpersonal attitude, 4) document reference, and 5) comment on document. A button-based GUI is provided and the users can easily compose complex messages by selecting and sequencing these action representations.

Figure 3 shows a screen shot of the message editor of TelMeA2003. Components lined up at the middle of the editor mean representation sequence for personified media. By using each five buttons arranged at the top of the editor, users can add a new representation into the sequence. These buttons are associated with the five types of representations.

Figure 4 shows an exemplar step of message composition. The user selected the speech representation in the Editor, typed in the text to be spoken, and is now selecting a performative verb for the type of gestures and facial expressions. These selections are integrated and determine how this portion of the message delivery is performed by her personified media. We have prepared a fine-grained set of performative verbs, as well as of affective expressions and interpersonal attitudes, to exploit the rich expressive potentials of non-verbal cues. The verb "agree," for example, is further sub-instantiated to represent the entire range of agreeing, e.g., from smiling to nodding to thumbing up. We have currently 35 types for performative verbs, 48 types of affective expressions and 13 types of interpersonal attitudes.

Fig. 3. Message posting with verbal and non-verbal expressions

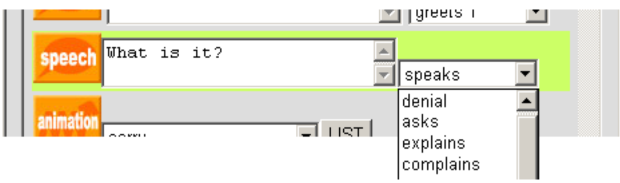


Fig. 4. Selection of performative verbs

```
<speech
  performance="deny"
  animation="shrug"
  utterance="What is it?" />
<speech
  performance="ask"
  animation="spread_arms"
  utterance="What is it?" />
```

Fig. 5. Script representation of character acts

Messages are translated into a script language format [4] before they are submitted to the community server. Script language examples are shown in Figure 5. Even when two character actions share the same sentence for their speech, they can be distinguished by the different choices of their non-verbal acts.

Once the messages are submitted to the community server, they become shared by all the participants in a community. Other participants freely reproduce the message and add their messages by using their own personified media. The exchange of messages are accumulated in the form of conversation log, on which social summarization is to be applied. We focus on non-verbal cues for extracting summarization information, and do not look into textual contents. Non-verbal cues are explicitly specified in the action representations of personified media in the form of the choice of performative verbs, affective expressions and interpersonal attitudes.

4 Practice and Analysis

TelMeA2003 has been introduced in the e-Kyositu (e-classroom) community for distance learning in Japan. After usability testing and system adjustment, initial data collection has been started. Investigations are under way to establish an algorithm to reliably extract evaluative information from users' choices of non-verbal cues.

e-Kyositu has about 250 student participants ranging from elementary school to senior high school students. Most of the e-Kyositu classroom activities are conducted using the conventional style of text-based bulletin board system, and the TelMeA classroom is the only exception. The TelMeA classroom was opened in January 2003, and currently has about 20 active participants.

We've conducted a preliminary analysis on the conversation log of the TelMeA classroom. At the time of analysis, TelMeA classroom produced a total of 251 messages in 17 threads. We first classified each of the animations employed in the participant messages according to the two scales of valence and arousal [5]. The valence classification was meant to capture the polarity, either positive or negative, of the evaluative attitude depicted in the animated action of the character, whereas the arousal classification was intended to capture the intensity of emotion expressed by the character actions. Each message was then rated based on the types of animated actions employed in the messages directed to it.

Figure 6 shows the result of analysis in the form of an evaluation score map for all the messages. From the figure, we can see that many of the messages are located in the central area. These messages should be of little significance for social summarization since they didn't receive much attention from other participants. Some of the messages, on the other hand, are clearly placed far from the central area. The positive valence region, particularly, has a number of messages that obtained high evaluation scores. These messages located in the periphery region in the evaluation map are the prime candidates for constructing social summarization.

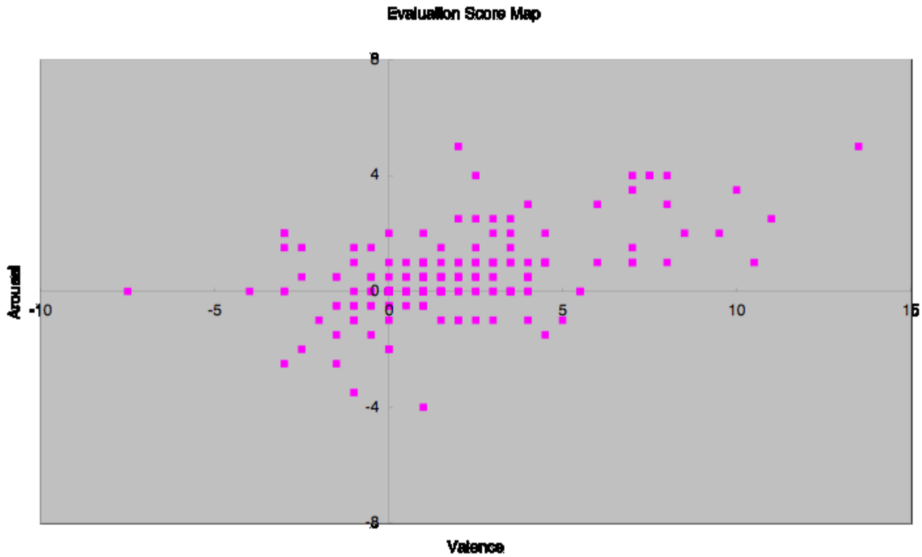


Fig. 6. Evaluation score map for messages produced by the e-Kyositu participants

From the observation of the messages scored in the periphery region in the map, we found that:

– Many high score messages were the messages that started new threads, though not all thread beginning messages were scored high.
– Reporting of good events, e.g., a success in school entrance examination, often makes high scores with high positive valence values.
– Introducing a controversial topic, e.g., Iraq war, can make high scores with high negative valence values.

5 Implications and Future Works

In knowledge intensive semantic societies, an increasing emphasis is going to be placed on bottom up collaborative activities, in business, in education, and in other facets of life, rather than on fixed top-down information delivery, for fostering and enhancing creative innovations. Social summarization has a huge potential in providing a support technology for them. The strong point of the technique is that it is free from any demand on extra steps for data collection, which often makes a bottleneck for annotation-based semantic computing technologies. Another point of the technique is its independence from textual content information. As long as evaluative attitudes of participants toward pieces of information presented is available, the method can be applied in domains where it is difficult to apply conventional content-based text analysis techniques.

The present paper demonstrates the possibility of social summarization. Obviously, the technique needs further development both in the area of evaluation methods and in the techniques of summarization constructions. Integration with text analysis techniques and incorporation of ideas with corpus-based study of multimodal dialogues are also promising directions for future works.

Acknowledgments

The authors wish to thank people participated in e-Kyositu community for their creative, inspiring and jovial interactions. This research was supported in part by the National Institute of Information and Communications Technology of Japan.

References

1. Balabanovi, M., Sholam, Y.: Content-based collaborative recommendation. Communications of ACM **3** (1997) 66–72
2. Knoke, D., Kuklinsiki, J.H.: Network Analysis. SAGE Publications (1991)
3. Takahashi, T., Takeda, H.: Telmea: An asynchronous community system with avatar-like agents. In: Proceedings of INTERACT2001, Tokyo Japan, IOS Press (2001) 480–487
4. Takahashi, T., Takeda, H.: Proposal of a script language for embodied conversational agents as asynchronous conversational media. In: Proceedings of AAMAS2002. (2002)
5. Lang, P.J.: The emotion probe: Studies of motivation and attention. Americal Psychologist **50** (1995) 372–385

Discussion Mining: Knowledge Discovery from Semantically Annotated Discussion Content

Katashi Nagao

EcoTopia Science Institute
Nagoya University
Furo-cho, Chikusa-ku, Nagoya 464-8603, Japan
nagao@nuie.nagoya-u.ac.jp

Abstract. We present *discussion mining* as a preliminary study of knowledge discovery from discussion content of offline meetings. Our system generates minutes for such meetings semi-automatically and links them with audio-visual data of discussion scenes. Then, not only retrieval of the discussion content, but also we are pursuing the method of searching for a similar discussion to an ongoing discussion from the past ones, and the method of generation of an answer to a certain question based on the accumulated discussion content. In terms of mailing lists and online discussion systems such as bulletin board systems, various studies have been done. However, what we think is greatly different from the previous works is that ours includes face-to-face offline meetings. We analyze meetings from diversified perspectives using audio and visual information. We also developed a tool for semantic annotation on discussion content. We consider this research not just data mining but a kind of real-world human activity mining.

1 Introduction

It will be very helpful that we can proceed to a discussion smoothly in a meeting or in a place for debating, and that we can retrieve and reuse the content such as minutes of these discussions. Generally speaking, however, unlike online discussions on bulletin boards on the Web, on offline ones at face-to-face meetings, making minutes reusable is difficult.

In this research, we propose a system that generates structured data on the discussion content semi-automatically and that displays them being visualized in real time. In addition, the data are registered in an XML database and in a relational database so that retrieval and summarization will be available. To apply this function, we developed an experimental system for supporting discussions.

One of the purposes of our research is to acquire knowledge from read world human activities such as conversations and discussions. Also, we are researching methods to annotate human activity records like meeting minutes with some semantic structures and to reuse them for supporting ongoing activities.

We have been developing some technologies which can record the sounds and the visions in meetings and generate multimedia minutes semi-automatically.

A. Sakurai et al. (Eds.): JSAI 2003/2004, LNAI 3609, pp. 158–168, 2007.

Our technologies also include search and summarization methods for finding and browsing relevant and useful information.

In this paper, we focus on two technologies. One creates discussion content not just meeting minutes but hypertext documents including participants' statements linked with video and audio. The discussion content also includes discussion graphs that visualize semantic/pragmatic structures of discussions. The other allows the users to annotate the discussion content with additional information about the structure of the discussion. The semantically-annotated discussion content is easily converted into more appropriate formats to be reused.

2 Creation of Discussion Content

There is a method mainly using video (image and sound) to make a discussion reusable. What we employed is a method that generates data on minutes primarily by text input through the forms in a Web document. The forms are subdivided and not only the content of the discussion but the information such as presenter, date, title and participants are inputted.

We especially focused on information about presentation slides in the trial system. In recent years people tend to use slides made by Microsoft PowerPoint for meetings. We employed JACOB (A Java-COM Bridge) to handle COM (Component Object Model) with Java. Specified with Microsoft PowerPoint from the ones made in advance, the slide is converted to GIF format and is displayed on an input page of minutes. Simultaneously the information of characters in the slide is obtained and is added to the minutes. Then the minutes are structuralized by relating with the slide and a statement.

Furthermore the information of the statement is input after being categorized into three types, question, answer, and comment. We limited the types of statement to the three in order to avoid troubles when inputting its content in plain text. We assumed that the information of a discussion consisted of three kinds of statement, question, answer, and comment. So when a subject changes to another, we realized a system that enabled to show a new form for input.

Figure 1 shows a display screen of inputting minutes.

gIBIS [2] seems to represent groupware by structural approach. This can display the structure of a discussion graphically to turn into facilitating the grasp of the content of minutes and encouraging effective statements.

In this system, we visualized the structure of minutes by creating graphical display and edit mode of statements with the use of SVG (Scalable Vector Graphics) [6]. The graph was semi-automatically structuralized with the pertinent information and keyword on the statements and slide as shown in Figure 2. In result, it allows users to edit.

Minutes are mainly created from the inputted text and the presented slide. The format of minutes is XML (eXtensible Markup Language) and they are stored in an XML database and a relational database at a server machine with Java Servlet. On this study, we used Xindice [1] as the XML database and PostgreSQL as the relational database.

Fig. 1. Input of Minute

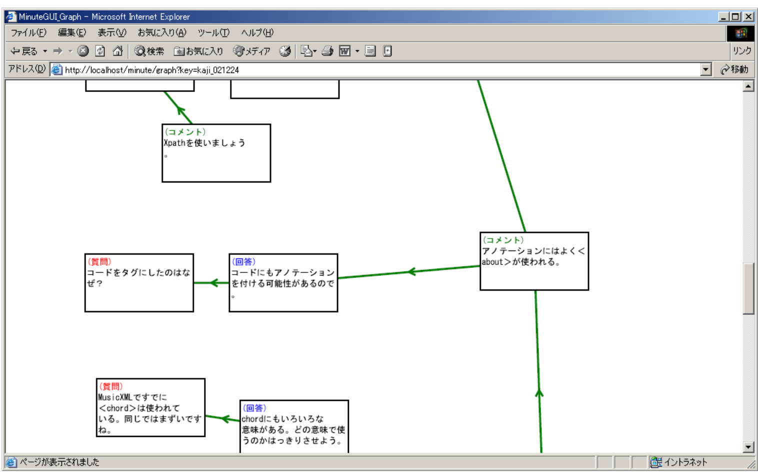


Fig. 2. Graph View of Minute

3 Reuse of Discussion Content

Some stylesheets on DOM (Document Object Model) object in Xindice allow real-time view of the minutes registered in the database through conventional Web browsers.

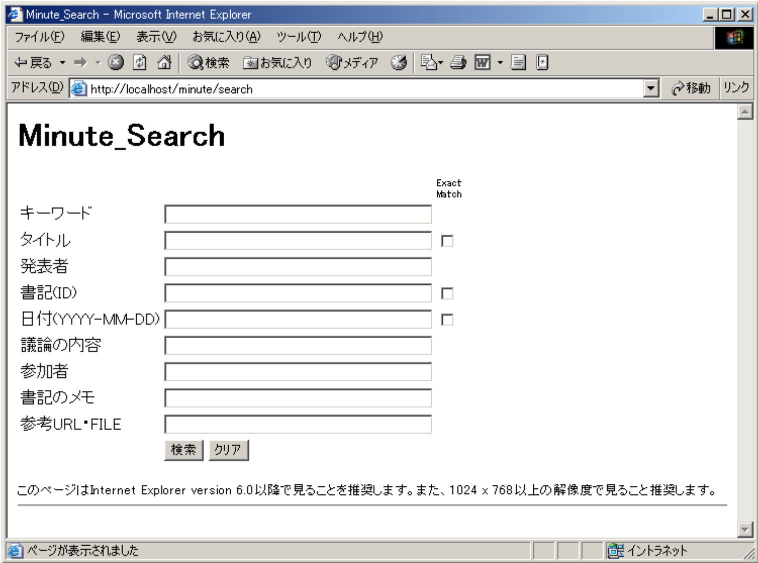


Fig. 3. Minute Search

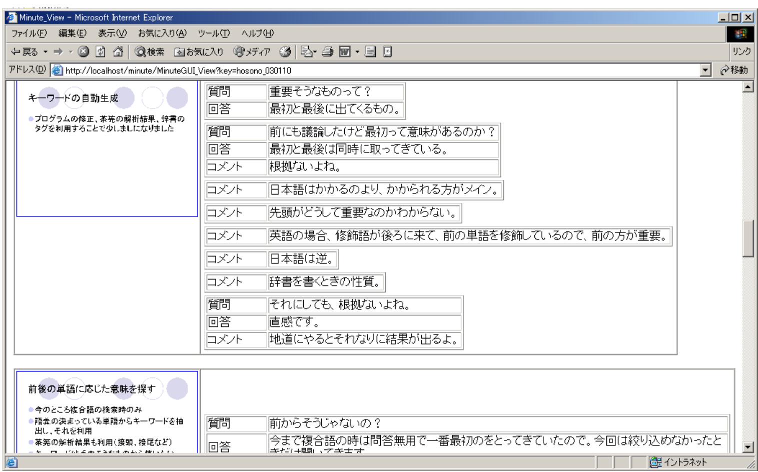


Fig. 4. Summarization of Minute (before)

Additionally, the users can access and retrieve only required minutes from the search form on the Web as shown in Figure 3. As the retrieval field is subdivided into several fields for title, date, a presenter and so on, we can easily access, retrieve and view the minutes that we want. As far as the retrieval, high-speed access will be available thanks to PostgreSQL.

The users can reedit and add the minutes, and correct their inputting errors through the Web browsers. When the users need to see the main points of

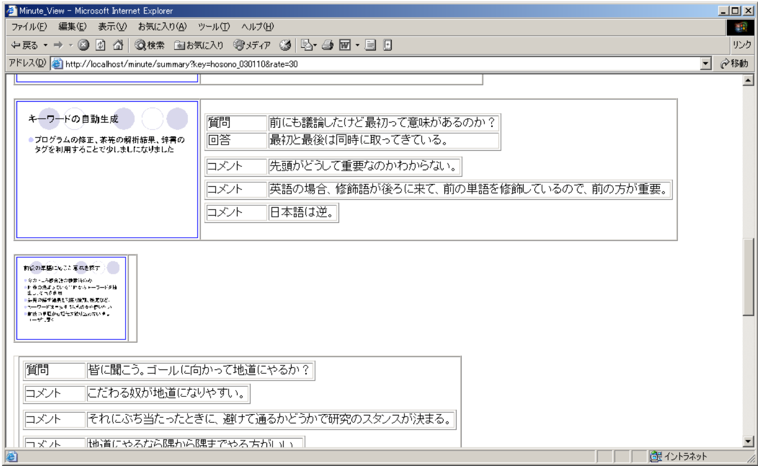


Fig. 5. Summarization of Minute (after)

the minutes, they can also see the summarized versions of the minutes in the order of importance. We utilize the structure of the statements at the time of inputting, the keyword of the statements, and the earned points for grades added to manually that are described later.

Figure 4 shows an original version of a minute and its summarized version is shown in Figure 5.

4 Semantic Annotation on Discussion Content

To utilize minutes more effectively, we add some semantic features to minute data such as linguistic structures, reference information to external content, and exact meanings of ambiguous words. We developed an authoring tool to create semantic annotation on minutes by linguistic analysis of text such as sentence parsing, anaphora resolution, and word sense disambiguation [5].

By using semantic information associated with mimutes, we can extract some important points in meetings and create summaries that include short history of discussions on the important topics.

Although annotation-based intelligent content processing is one of the major topics in some international research activities such as MPEG-7 [3] or Semantic Web [7], it has not obtained a definite result. We have already developed the system for interactive summarization and translation of Web content including multimedia data based on semi-automatically-created semantic annotation. Our annotation-based architecture is called "Semantic Transcoding" [5].

The annotation here includes the detailed language structure of the sentences in the document and the semantic segments of multimedia content which are generated semi-automatically. Each piece of the annotation is correctly associated with some part of the content.

The annotation enables Web audiences use the content more easily by using the semantic relationships among some media. For example, searching for videos using a keyword and summarization of videos connected with the summaries of voice transcript texts became possible because of the annotation. However, generation of the annotation costs very much. We need some methods to make this task more cost effective without making the annotation generation more complicated and inaccurate. One of our solutions captures some explicit hints or suggestions which are given by human who can easily understand the meaning of content and infers the semantic structure underlining the content. Our new user tools and software will make these processes simple and easy.

5 Discussion Mining

Discussion mining is a new research domain that aims at digitization of discussion content and discovery of knowledge from the content. Our developed system records audio-visual scenes of discussions in face-to-face meetings and visualizes discussion content in the form of Web documents linked with videos captured by several cameras and microphones.

We have done some preliminary studies and evaluations of the prototype systems for discussion capture and visualization. We have also implemented a minute editing tool that is capable of semantic annotation.

The meeting participants support the automatic generation of the minutes by transmitting their user IDs and types of their statements at the meeting via InfraRed signals by using a special device called "the discussion tag". Each participant has three types of the discussion tag colored green, yellow, and red, respectively. The green one is indicated to all participants when the participant wants to make a comment. The red one is used for asking a question. The yellow one is for answering it. All tags are also used for evaluation of whole discussion. Our system detects IR signals from each participant's discussion tag and generates a graph structure consist of each statement (some keywords included in the utterance are manually inputted by a meeting secretariat) as a node and each relation between statements (question-its answer, referred comment-referrer comment, etc.) as an arc. The nodes are also linked with the corresponding scenes in the recorded video and audio.

Figure 6 shows a scene of the discussion using the system, and Figure 7 shows a screen shot of a structured minute generated semi-automatically. The minute is visualized as a graphical structure of the discussion, and allows the users to retrieve the video related to specific scenes. Figure 8 illustrates the whole system configuration, and Figure 9 illustrates the discussion tag system. How to use the discussion tag is very simple and easy. A potential speaker will hold up the discussion tag just before beginning of a talk at the meeting. The tag automatically transmits the data to the system.

As mentioned earlier, the content of the minutes is represented in an XML data format and stored in the XML database. The audio-visual content is

Fig. 6. Scene of Discussion

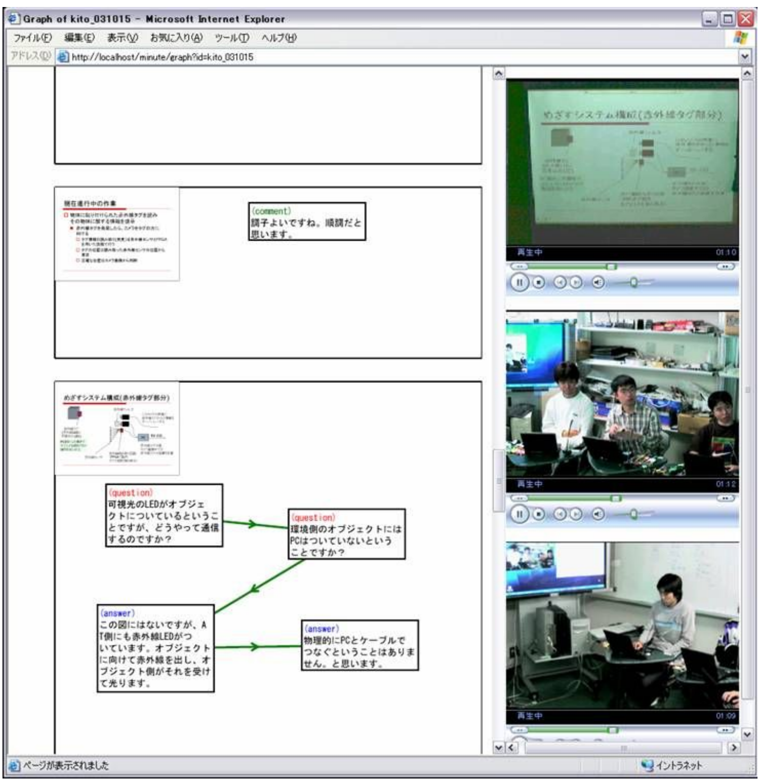


Fig. 7. Screen Shot of Structured Minute with Videos

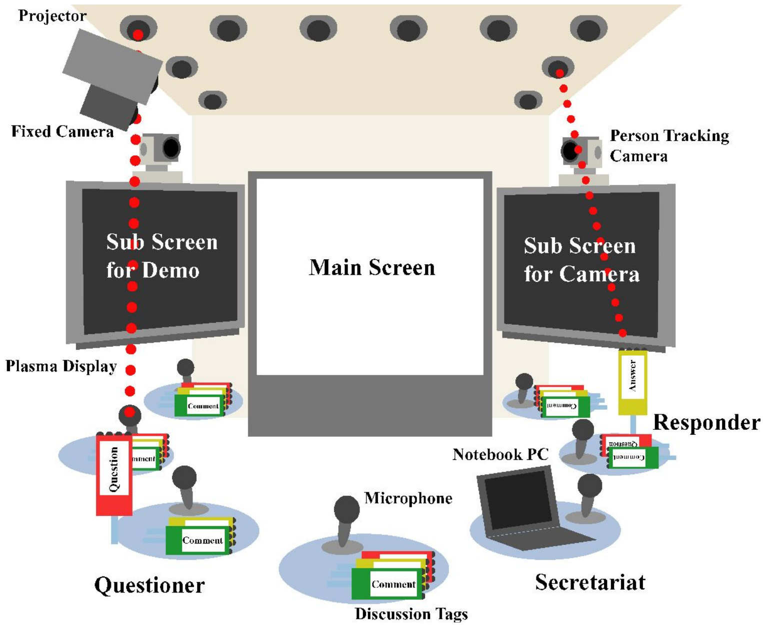


Fig. 8. Configuration of Discussion Mining

accumulated in the multimedia database. These databases are connected with the network and the XML data include the pointers to the multimedia data.

The minutes in the XML format enable keyword-based retrieval and summarization. The summarization method is based on importance values calculated by node importance (more linked modes are more important) and word importance (more salient words are more important). We evaluate each discussion by using discussion tags (the green, yellow, and red tag indicate "good", "soso", and "not good", respectively). We also evaluate each statement at a meeting by using a Web-based discussion annotation tool. These evaluations are also considered to calculate the importance of the discussion.

Not only visualization, retrieval and summarization of single minute content, we have also developed a mechanism of visualization of multiple minutes. Our system performs grouping of the accumulated minutes and linking between related minutes. A group of the minutes includes similar minutes and their correspondences. The groups of the minutes are visualized as a "minute map" which shows distances among the minutes as shown in Figure 10. The map supports the users to observe the whole discussions deeply and to discover knowledge. In order to make the map more informative, we have to create the annotation that indicates semantic relationships among the minutes. By using such annotation, we can discover more detail about discussion structures across several minutes such as some background information of an argument and its derivations.

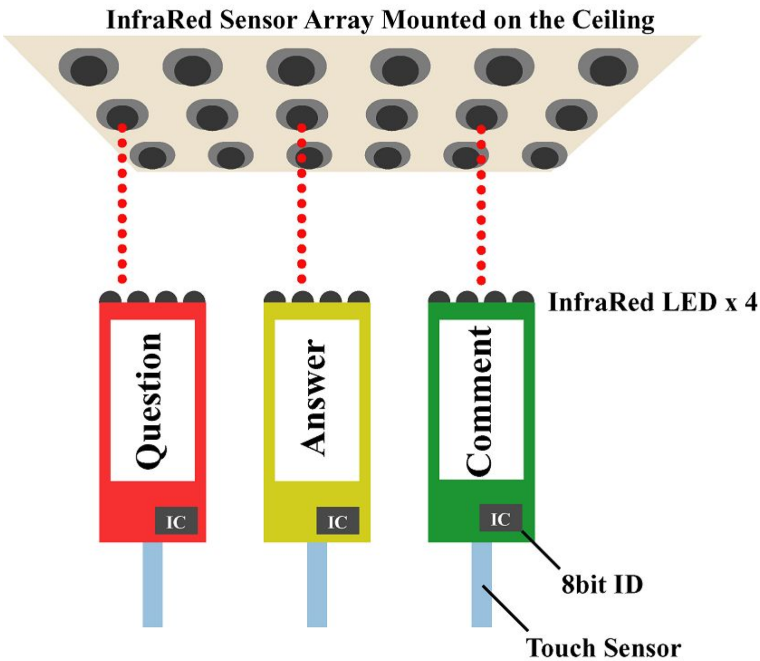


Fig. 9. Discussion Tag System

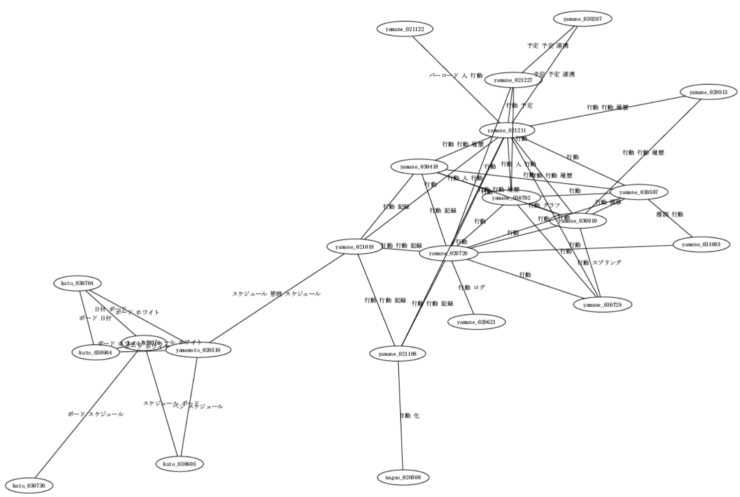


Fig. 10. Minute Map

From the minute map, we can also find the most important minute, the most significant statement and the most active speaker and questioner. Furthermore, we are analyzing the role of the sounds and the visions in meetings more deeply.

We are also considering a new way of linking between multiple media related to human real-world activities such as conversations and discussions.

We should pursue some methods for not just capturing real world activities but also utilizing them for some meaningful tasks such as remembrance assistance and knowledge discovery.

6 Concluding Remarks

In this paper, we described the system of registering the minutes of meetings in database and reusing them. The following is the issues that we are planning to examine and to improve the system:

1. Annotation over Multiple Discussions
   Minutes generated by this system are not related to the past ones, for they only refer to a certain meeting. The system should be improved to examine relevance to the past minutes semi-automatically from a keyword and to summarize covering multiple minutes. It will help to create annotations across multiple minutes.
2. Use of Voice Transcripts
   The use of voice will enable us to refer afterward to the information that cannot be contained in the minutes. Since the voice data is difficult to be searched as it is, using annotation can make it possible to largely reduce the problems like retrieval if speech recognition creates transcripts.
3. Use of Video Annotation
   Since video data is available, the system will be able to utilize hard-to-get data from other information like mood of meetings and gestures and facial expressions of the participants. We have also developed a video annotation editor to annotate visual objects in video frames with some notes and attributes and to edit scene descriptions and voice transcripts [4].

In this research, we have shown one example of the knowledge discovery from the information in the real world, and the multiple media integration based on semantic annotation.

Our annotation system will be a strong tool to associate some elements in the media data with the other media elements. For example, the image and the motion of the speakers' reference expressions such as finger pointings will be connected with the particular linguistic expressions such as deixis in the voice and the text data. The relationships among the multiple media will be pointed out clearly by using the semantic annotation, and contribute to finding out the rule of them.

Furthermore, the integration of other knowledge sources with the semantic annotation on the minutes makes it possible to search for and add to the information about speakers who made good arguments.

Semantic annotation is supposed to cost much in generally. However, involving human intervention effectively into the annotation task when the original content or the material of content is created improves cost performance very much.

Needless to say, improving the quality of automatic analysis of the content such as vision and speech recognitions is able to facilitate the efforts for creation of the semantic annotation.

We also need more user-friendly tools that reduce the human labor as much as possible. For example, in addition to the discussion tags, a mechanical button on the desk will be useful during meetings. Just pressing it by each participant lets the system know which comment is important in the discussion and which action is meaningful in the physical event. Such simple mechanism will work for automatic annotation to the audio-visual data.

Acknowledgments

The author would like to thank Katsuhiko Kaji, Daisuke Yamamoto, Toshiyuki Shimizu, Hironori Tomobe, and other members of Nagao Laboratory at Nagoya University for their cooperation in implementation and experiments of the discussion mining system.

References

1. The Apache XML Project. Apache Xindice. http://xml.apache.org/xindice/. 2001.
2. Conklin, J. and Begeman, M.L. gIBIS: A Hypertext Tool for Exploratory Policy Discussion. Proc. of CSCW '88. pp. 140–152, 1988.
3. MPEG. MPEG-7 Overview. http://www.chiariglione.org/mpeg/standards/mpeg-7/mpeg-7.htm. 2002.
4. Nagao, K., Ohira, S., and Yoneoka, M. Annotation-Based Multimedia Annotation and Transcoding. in *Proceedings of the Nineteenth International Conference on Computational Linguistics (COLING 2002)*. pp. 702–708, 2002.
5. Nagao, K. *Digital Content Annotation and Transcoding*. Artech House Publishers. 2003.
6. W3C. Scalable Vector Graphics (SVG) 1.0 Specification. http://www.w3.org/TR/SVG/. 2001.
7. W3C. The Semantic Web Community Portal. http://www.semanticweb.org/. 2002.

Semantic Computing with Conversations and Stories

Toyoaki Nishida

Department of Information and Communication Engineering
Graduate School of Information Science and Technology
The University of Tokyo
7-3-1 Hongo, Bunkyo-ku, Tokyo 113-8656, Japan
nishida@kc.t.u-tokyo.ac.jp

Abstract. Semantics is grounded on intellectual activities in various communities in the human society. Conversations and stories are primary media for substantiating the semantic processes in a community. In this paper, I model a semantic process as a coevolutionary spiral of conversations and stories in a community, and present a simple but practical computational model featuring knowledge cards that can serve as a semantic component, lifecycle support of knowledge cards, and strategic control of information streams. Then, I generalize the approach as a technique called conversation quantization, which is based on the idea of approximating a continuous flow of conversation by a series of conversation quanta that represent points of the discourse. I show some implemented systems to show how these ideas are implemented.

1 Community as a Ground of the Semantic Process

Semantics is dynamic in the sense it continually evolves by incorporating new thoughts arising in the history. This nature of semantics necessitates the effective support for semantic processes of meaning evolution in order for semantic computing to function effectively. The ground of a semantic process is a community which is a collection of people who build and share a common context. Communities of practice, a group of people who share common work practice, develop informal networks of relationships cutting across organizational boundaries in search for better cooperation [1].

We aim at building a suite of information and communication technologies for facilitating the semantic process in a community. The semantics process is social in nature. In order for our technology to be effective, we need to take into account the following aspects.

(1) Reality Sharing. Knowledge and information contain not only explicit aspects but also tacit dimensions; our knowledge and information cannot be completely described in the sense that they contain much incompleteness and inconsistency and are very hard to articulate. After all, it appears impossible to completely encode the tacit dimension and we have to compromise with a

A. Sakurai et al. (Eds.): JSAI 2003/2004, LNAI 3609, pp. 169–184, 2007.

weaker sense of communicating tacit information by employing an approximate means that sheds light on a certain limited aspects. In addition, knowledge and information need to be substantiated in the user's daily living space so that the user can understand the meaning from her/his own viewpoints. A virtual reality using a large immersive display will help the user understand the reality. It should noted, however, that perceptual information alone is not enough to communicate tacit information; it should to be combined with verbal information to give the user a clear and structured message about the subject.

(2) Social Awareness. Awareness is critical for collective intelligence to emerge. Awareness is information that enables people to act. Social awareness provides with information about other agents' activities and knowledge. Awareness is partial information that enables people to monitor critical events in the environment while shifting primary attention away from the subject. Information and computational support (ICS) for social awareness is necessary to support network communities whose members are geographically distributed.

(3) Context and Knowledge Sharing. In order for a collection of people to be a community, they need to share a common context consisting of social rules and principles, the discourse of universe, human relations, the background, and so on. Those propositions are often tacit, and hardly spoken explicitly. As a result, new-comers are often puzzled, and even old-timers may occasionally fall into misunderstanding. ICS should be able to help community members build, explicate, and maintain the shared context.

(4) Collaboration. Even though the primary function of a community is to develop and maintain human relations, collaborative work is necessary from time to time. Traditionally, CSCW (computer-assisted collaborative work) is the area for designing a number of useful tools that help people work together [2]. Seamless transition between the community oriented support and group oriented support are needed.

(5) Knowledge Cycle. Today's knowledge is dynamic in the sense that it changes quickly and soon becomes obsolete. Knowledge should accompany a proper process of keeping it fresh by incorporating new pieces of knowledge. In other words, knowledge is not like durable goods that sustain their value once they are built, rather it becomes more like consummatory goods that quickly loose their value after they are put to use. Accordingly, the knowledge cycle in the community need to be sustained properly.

The more knowledge is applied, the more opportunities arise concerning knowledge creation. Through exercises of knowledge applications, people build an expertise about when and where to apply the knowledge, what are the expected benefits and underlying risks, and what are the limitations of the knowledge. Through applications, the learners raise questions or comments, encouraging instructors to reply. ICS should integrate the knowledge creation, application, and maintenance phases.

(6) Dispute Resolution. It is quite probable that one community member's intention may conflict with another member in the community. It is desirable that the community has a means for predicting, preventing, and resolving disputes in the community without resorting to the external authority. In particular, ICS should facilitate mutual understanding by circulating enough information to tame ignorance and remove sources of misunderstanding.

(7) Large Scale Arguments. The potential of a community can be fully exploited by enabling the community-wide discussion. The goal is quite a challenging problem when a community is as large as a city. ICS should permit the user to be aware of arguments that are relevant to her/his own problems, understand its structure, raise questions, submit proposals, and take a leadership when appropriate.

2 Conversation-Story Spiral

Conversations and stories are primary media for substantiating the semantic process in a community. I model the semantic process in a community as a co-evolutionary spiral of conversations and stories. In order to exploit the benefits of conversation and compensate for its limitations, I propose a model called the conversation-story spiral. The model shown in Fig. 1 suggests that we use two modes of knowledge representation. The story archive is a structure-oriented representation of knowledge. It is suitable for representing the static aspects of knowledge, such as the structure or relation. In contrast, the conversation process is the interaction-oriented representation. It helps one understand and discover novel interactions among knowledge pieces. These two modes of knowledge representation are related to each other by a couple of conversion processes: conversation making and story making. Conversation making is a process for adapting story pieces into utterances that are suitable for each scene of a conversation. Various types of transformation are needed to tailor each piece of information so that it can conform to what is expected in the context.

Story making is a process of composing well-structured stories from a series of utterances. It involves seeking for an optimal story structure that can incor-

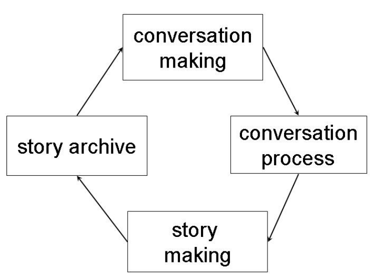


Fig. 1. The conversation-story spiral

porate conversation fragments, changing the tacit conversational knowledge into explicit representation, transforming components of information representation into those with a uniform grain size, and so on.

3 The Knowledge Channel Approach to the Conversation-Story Spiral

I present a simple but practical computational model based on several novel ideas. The model employs knowledge card as a component for representing conversation. A knowledge card is a data structure that encapsulates a narrative and its references for representing the content of a conversational unit. Each knowledge card is a relatively self-contained package of tacit and explicit knowledge, enabling one to implement a rather robust conversational system without introducing a complex discourse comprehension and generation algorithm. In order to support the lifecycle of knowledge cards, people are encouraged to create, edit, share, and utilize knowledge cards. It will bring about a evolutionary system of knowledge cards where a population of knowledge cards evolve by absorbing knowledge and information possessed by community members under selection and mutation. The knowledge channel mechanism is introduced to allow information providers to describe a policy for long-term interactions by strategically controlling the flow of knowledge cards.

3.1 The Knowledge Channel Approach

The knowledge channel approach uses the data structure called knowledge card as a representation of the conversation/story unit. A knowledge card is a data structure that encapsulates a narrative and its references for representing the content of a conversational unit (Fig. 2). Another data structure called a story is used to represent a discourse structure consisting of one or more knowledge card. A knowledge card server maintains an archive of knowledge cards. It delivers knowledge cards to clients on demand. Embodied conversational agents take knowledge cards to generate utterances in conversations. Knowledge channel is a conduit connecting one or more server and client. One can use one or more knowledge channel control policy to control the flow of knowledge cards in a knowledge channel. For example, one might write a knowledge channel policy for a server and/or a client. When more than one knowledge channel policy exist for a knowledge channel, negotioation will be made to determine the actual knowledge card flow. Dynamic program table spatially represents an overview of all knowledge flows.

This approach can approximate the conversation-story spiral even with a simple mechanism, and better with a more sophisticated dialogue engine. Tacit knowledge can be embedded into knowledge cards such as visual images or sound, even though it cannot be encoded into knowledge representation. The current natural language processing techniques such as document classification or summarization may be applied to turn the collection of knowledge cards into a structured knowledge.

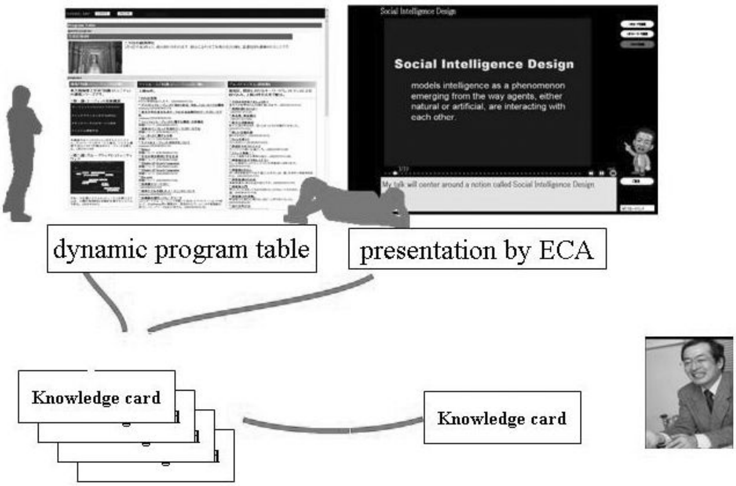


Fig. 2. A general framework of the knowledge channel approach

I illustrate how the conversation-story spiral is implemented in (a) EgoChat which is a conversational agent system based on the talking-virtualized-egos metaphor, and (b) S-POC which is an integrated communication support system.

3.2 EgoChat

EgoChat is a system for enabling an elaborate asynchronous communication among community members. It is based on the talking-virtualized-egos metaphor. A virtualized ego is a conversational agent that talks on behalf of the user. Each virtualized ego stores and maintains the user's personal memory as a collection of knowledge cards and presents them on behalf of the user at appropriate situations. It not only helps the user develop her/his personal memory but also better understand other members' interest, belief, opinion, knowledge, and way of thinking, which is valuable for building mutual understanding. We use a powerful dialogue engine that permits a virtualized ego to answer questions by searching for the best match from a potentially large list of question-answering pairs prepared in advance (Fig. 3).

The EgoChat system provides a couple of unique features. First, it integrates personal and interpersonal knowledge life cycles [3]. At earlier stages of the lifecycle when knowledge is not well captured, the user might want to describe her/his idea as one or more knowledge card and have the virtualized ego to present them for personal review. After a while when the knowledge becomes clear and evident to the user, s/he might want to delegate her/his virtualized ego to her/his colleagues to present the idea and ask for critiques. The automatic question answering facility of EgoChat will encourage the audience to ask questions or give comments. The virtualized ego can reproduce the question-answering session with a communication partner so that the owner can review

Fig. 3. EgoChat: normal presentation mode by a single agent (left), Question mode where the user's question is read by an avatar (right)

the interactions. It will highly motivate the owner to supply more knowledge or improve existing knowledge for better competence.

Second, EgoChat allows for describing a channel policy that is used to define the control strategies of the sender and the receiver. Four types of strategies are identified depending on whether the strategy is about the order of programs in a single program stream or about the way multiple program streams are mixed, and whether the program scheduling is static or dynamic, as shown in Table 1. The skeleton of the actual flow structure of knowledge cards for a given pair of the sender and receiver is determined by resolving constraints of their channel policies. It can be visually presented to the user by the dynamic program table, as shown in Fig. 4 [4].

Table 1. Four strategies in the knowledge channel policy

	Static scheduling	Dynamic scheduling
Intra-channel arranging	**Order strategy** Giving the outline of the content (messages and channels) in a channel ・course <[*message* | *channel* | (*a set of content,* sorting [*by title* | *category* | *length* | *rating* | *activity* | *author* | *date* | *keyword* |...], order [*ascendant* | *descendant*])], ... >	**Access strategy** Changing the stream of the content depending on the interaction and the situation ・browsing [{(behavior [*play* | *skip* | ...], condition [*always* | *in events* |...], permission [*all* | *members* | *owner* | ...]), ...} | *null*] ・contribution [{(action [*post* | *query* | *rate* | *create* | ...], condition [*always* | *in events* | ...], response [*agent replies* | *content changes* | *never* | ...], modification [*add the message* | *change the rating* | *never* | ...], permission [*all* | *members* | *owner* | ...]), ...} | *null*]
Inter-channel arranging	**Link strategy** Giving a connection with other channels ・loopback [*loopback* | *null*] ・channels [{(*channel,* elements {*channel name,* ...}), ...} | *null*]	**Interchange strategy** Navigating channels ・navigation [{(*target channel,* behavior [*branch* | *merge* | *parallel* | ...], condition [*always* | *in events* | ...], permission [*all* | *members* | *owner* | ...]), ...} | *null*]

<x_1, x_2, ... > represents ordered elements, {x, y, ...} represents unordered elements, [x | y | ...] represents x, y or other elements and (x, y , ...) represents a set of elements, where elements x and y are defined contextually.

Fig. 4. Example of a dynamic program table

3.3 S-POC

S-POC was designed to serve as an integrated communication environment for a multi-faceted community activities such as risk communication. S-POC supports reality sharing based on visual images, knowledge sharing, and community-wide discussion for decision-making (Fig. 5).

The S-POC uses the POC system [5,6] as a knowledge sharing engine and extends it in several respects. First, S-POC allows for sophisticated presentation mechanism using an embodied conversational agent. Second, it supports video editing and presentation. Third, it allows a community discussion support system to be plugged in [7,8].

(1) The Presentation Mechanism Using ECA. The RISA-CAST component allows for automatically generating from plain text as a presentation featured with an animated agent. It selects and generates appropriate gestures and facial expressions for a humanoid agent according to linguistic information in the text. An agent animation system called RISA can produce and animate the visual image of an embodied conversational agent (ECA) on web-based applications. The ECA subsystem called CAST (The Conversational Agent System for neTwork applications) generates agent's nonverbal behavior automatically based on theories of human communication (Fig. 6).

(2) Video Editing and Presentation Tools. The S-POC system allows for video presentation, using streamed video clips and camera works, such as zoom

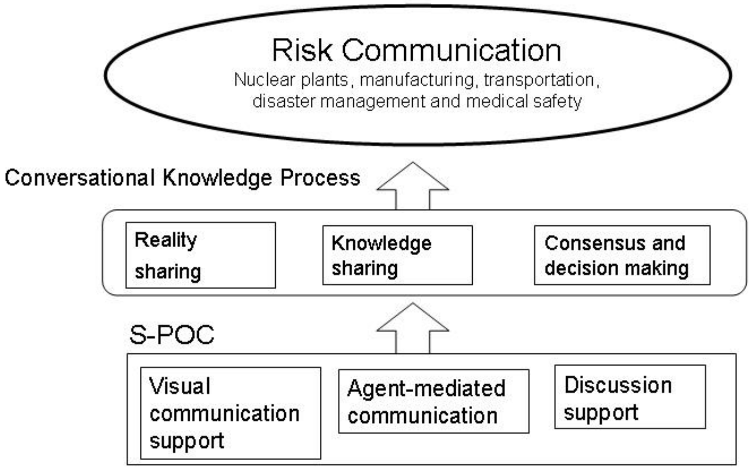


Fig. 5. The framework of S-POC

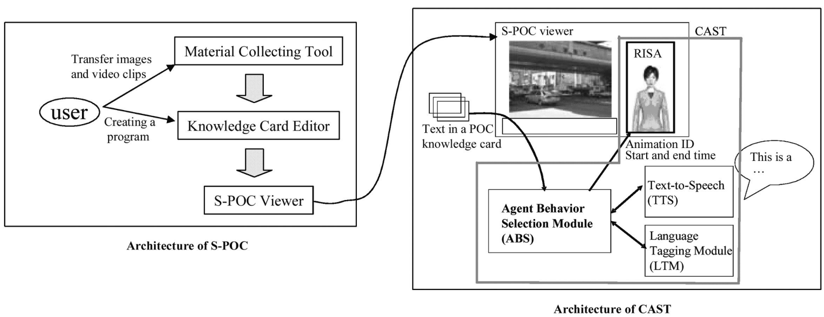


Fig. 6. RISA-CAST subsystem for generating ECA presentation

and pan, applied to graphic images. In order to help the casual user create and review the video contents, the following tools are incorporated.

– tools for collecting and accumulating materials (e.g. pictures, graphics and videos)
– tools for creating video contents by assembling components of various kinds (Fig. 7)
– a viewer for converting a knowledge card into a synchronized presentation with the ECA and multimedia materials.

(3) Discussion Support System. S-POC enables CRANES (Coordinator for Rational Arguments through Nested Substantiation) to be plugged in, which supports discussions in the community for consensus making [9]. CRANES uses a probabilistic approach to selectively identify “stirring arguments” that have

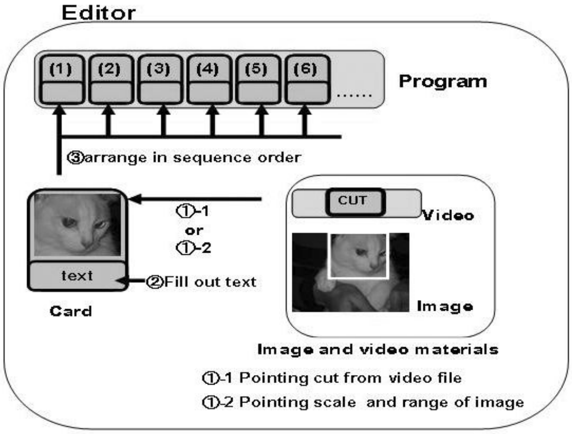


Fig. 7. The overview of the video editor

stimulated the entire group processes. Two random-tree models are developed for the purpose of defining an indicator of noteworthiness. In the perfect random tree, the probability for a new comment to be associated with a certain existing node is assumed to be the same among all existing nodes. A biased random tree model is a more elaborate model that inflates the probability of the last node's producing a new child compared to other existing nodes. It reflects an intuition that when readers of a BBS consider which comment they respond to, they are more likely to choose the last node in the whole thread or those with some specific features. CRANES was applied to a community conflict over a regeneration scheme in London, having resulted in an augmented social construction of the conflict with multiple dimensions.

4 Conversation Quantization

Conversation Quantization is a technique of articulating a continuous flow of conversation by a series of objects called conversation quanta each of which represents a point of the discourse. Conceptually, it consists of extraction, accumulation, processing, and application of conversation quanta (Fig. 8).

The extraction of conversation quantum is called "materialization" or "capture." It is a process of identifying a conversation quantum in a conversational situation in the real or virtual world, and encoding it in an appropriate information medium. Materialization permits to simulate the behavior of a conversation quantum as an object, and store it in a server, and apply appropriate conversations such as merging several conversation quanta into one and reducing it in the size. A collection of materialized conversation quanta can be spatially configured so that the their mutual relation can be visually grasped and even sensed tangible objects.

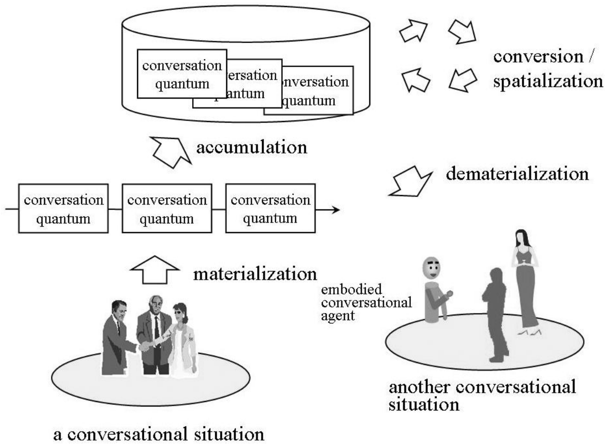


Fig. 8. The Framework of Conversation quantization

Alternatively, a conversation quantum can be applied to ongoing conversations by creating appropriate conversational interactions based on the information embedded in it. The process is called "dematerialization" or "release" of a conversation quantum.

Conversation quantization allows for implementing a conversation system by reusing a collection of conversation quanta gathered from real/hypothetical conversation situations. Given a conversational situation, a conversation quantum that best matches it will be sought from the collection of conversation quanta, and one role of the participants of the retrieved conversation quantum can be replayed by an embodied conversational agent, and other roles will be mapped to the participants in the given conversational situation. Such an algorithm is relatively easy to implement and rather robust in nature.

The granularity and size of conversation quanta essentially depend on the context and background knowledge of the observer. Although the detailed investigation of the nature of conversation quantization is left for future, we conjecture, based on experiments made so far, that each conversation quantum roughly corresponds to a small talk often identified in the discourse of daily conversations.

The implementation of conversation quantization depends on the data structure for representing conversation quanta. One may use plain video clips as representation but its utility in retrieving and processing will be limited and a certain amount of cost will be required in retrieving, editing, and applying conversation quanta. Alternatively, a deep semantic representation using logical formulas or case frames would not be ideal due to the expense and their limited capability of representing nonverbal information. A reasonable implementation appears to use annotated videos and images to represent a conversation quantum. The knowledge card approach can be characterized as an implementation in this vein.

5 Communicative Intelligence

A five-year research project "Intelligent Media Technology for Supporting Natural Communication between People" sponsored by Japan Society for the Promotion of Science (JSPS) addresses communicative intelligence as a generalization of semantic computing with conversations and stories. The term "communicative intelligence" reflects the idea of communicative intelligence implies a view that communication plays a critical role in both individual and collective intelligence. The project encompasses the environmental medium technology for embedding computational services in the environment, the embodied conversational agent technology for having computers interact with people in a social context, and the communication model of conversations that serves as a theoret-

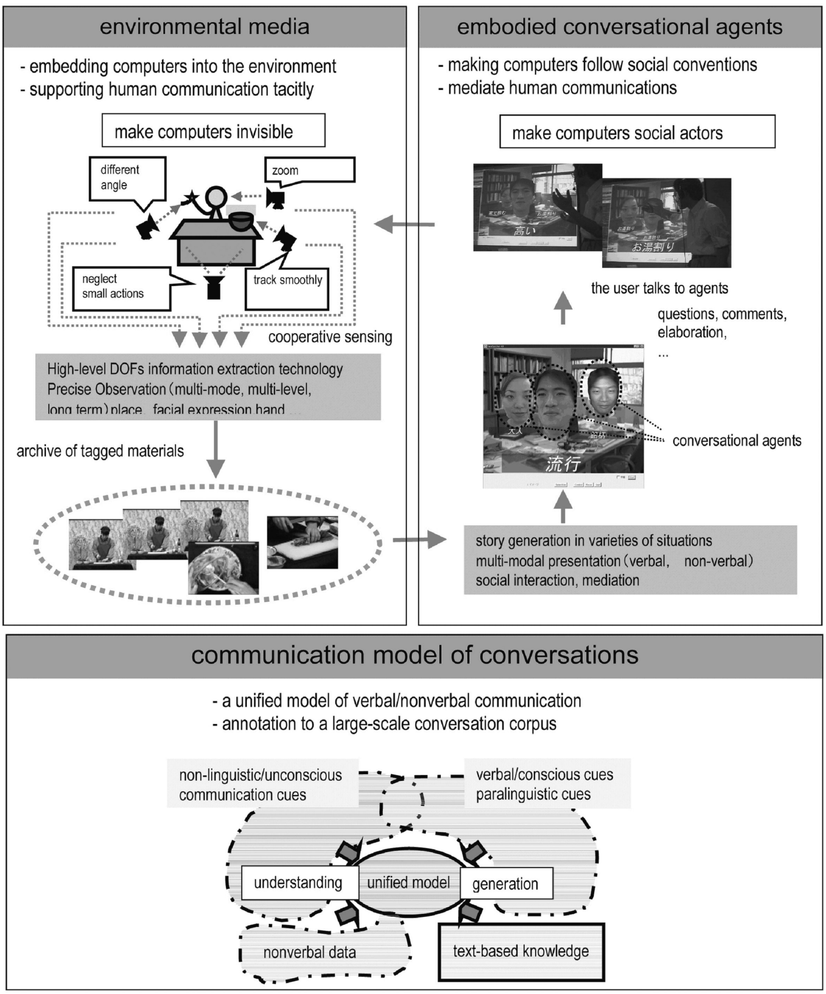


Fig. 9. The conceptual framework of research into conversational content

ical basis of the intelligent media research (Fig. 9). It places particular emphasis on *conversational content*, which is a package of information created and consumed in conversational contexts.

The environmental medium technology is based on the "making-computers-invisible" approach, aiming at embedding computers into the everyday environment so that they can assist people in pursuit for their goals without enforcing them to pay special attention to computer operations. We attempt at developing cooperative intelligent activity sensors, automatic analysis of nonverbal communication with high-resolution scene analysis, personalization of environment medium, and intelligent editing of audio-visual streams.

The embodied conversational agent technology is based on the "making-computers-social-actors" approach, aiming at developing embodied conversational agents that can make conversation with people by following various social conventions. The embodied conversational agent is characterized as an interface between people and conversational content. We attempt at developing virtualized egos that can talk on behalf of the user, socially intelligent robots that facilitate conversations among people, and agents that can create stories from their experiences.

In research into the communication model of conversations, we aim at establishing a theoretical foundation of conversational content. We focus on formulating a unified model of verbal and nonverbal communication, developing a language-centric technology for editing and summarization of conversational contents and building a large-scale corpus of conversations that accumulates conversation records with annotations to verbal and nonverbal events.

The entire conceptual framework of research on communicative intelligence is shown in (Fig. 10). At the bottom, language processing, speech processing, image processing, and human computer interaction make up the foundation of communicative intelligence. At the second lowest level are environmental media and embodied conversational agents technologies. The third level is about conversational contents. The topmost level is the application layer including community support systems, e-learning, risk communication, and so on.

The major research issues are:

1. Conversation measurement and analysis: aims to establish a method of capturing various aspects of conversations for analysis, ranging from the local features such as prosody to holistic aspects such as conversation atmosphere.
2. Conversation environment design: aims to bring about an augmented environment that can sense and support conversational behaviors.
3. Conversational artifacts: aims to develop artifacts that possess an ability of communicating their experiences with humans and other communicative agents.
4. Conversational contents: aims to shed light on the content that is shared and communicated in the conversation.
5. Applied conversation: aims to develop a task-specific conversation technologies suitable for applications such as e-learning or knowledge management.

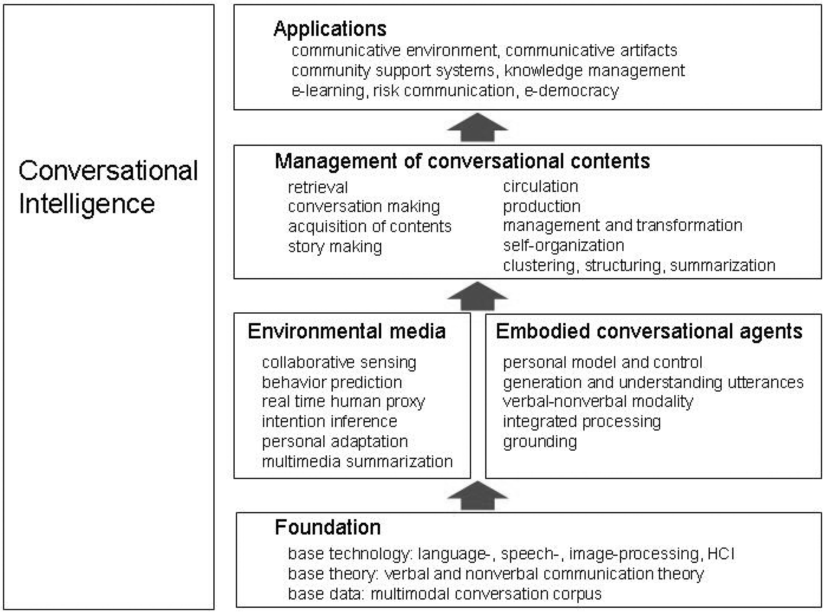


Fig. 10. The conceptual framework of research on communicative intelligence

6 Social Intelligence Design

The evolution of media technology supported by the progress of ICT may have both positive and negative sides. On the positive side, an enormous number of opportunities has arisen. In fact, an enormous number of new applications and businesses is springing up every day in collaborative environment, e-learning, knowledge management, community support systems, digital democracy, and so on. On the negative side, the advent of a new technology brings about new social problems that did not occur with the old technology. In addition to new kinds of crimes and ethical problems, some effects of ICT are very subtle, yet fairly destructive to our spiritual world.

In social psychology, it is known that some fundamental mechanisms of a group hinder constructive discussions and rational decision-making. Notorious examples include the "flaming war" (i.e., a barrage of postings containing abusive personal attacks, insults, or chastising replies to other people) and the "spiral of silence" (i.e., the social pressure discouraging people from expressing their views when they think they are in the minority). These phenomena might be amplified in the network society and bring about unexpected negative impacts on human society.

Social Intelligence Design is a new research area aimed at designing new computer-mediated communication systems based on in-depth understanding of intelligence as a social phenomenon [10,11]. Social Intelligence Design centers on understanding and designing social intelligence. As a personal attribute, social intelligence is contrasted with other kinds of intelligence such as problem-solving intelligence (ability to solve logically complex problems) or emotional intelligence (ability to monitor one's own and others' emotions and to use the information to guide one's thinking and actions). Alternatively, social intelligence might be attributed to a collection of agents and defined as an ability to manage complexity and learn from experiences as a function of the design of social structure. This view emphasizes the role of social conventions that constrain the way individual agents interact with each other. A good social culture will allow the members of the community to learn from each other.

Social Intelligence Design is an interdisciplinary field. On the one hand, it involves engineering approaches concerning the design and implementation of systems, ranging from group/team-oriented collaboration support systems that facilitate intimate goal-oriented interaction among participants to community support systems that support large-scale online discussion. On the other hand, it involves scientific approaches addressing cognitive and social psychological understanding of social intelligence, and provides a means for predicting and evaluating the effect of a given communication medium on the nature of discussions and conclusions. In addition, it encompasses pragmatic considerations from economics, sociology, ethics, and many other disciplines, for social intelligence design has a direct relation with the society.

The central idea underlying the evaluation is to define a standardized measurement of social intelligence of people. Should such a measurement be introduced, it is also applied to measure the effect of a given information system with respect to the degree of improvement of social intelligence of people. Social Intelligence Quantity is a framework of standardized quantitative measurement of social intelligence. Our approach combines qualitative evaluation consisting of questionnaire and protocol analysis and quantitative evaluation consisting of network log analysis, factorial experiment, and standardized psychological scale, as shown in Fig. 11 [12,13].

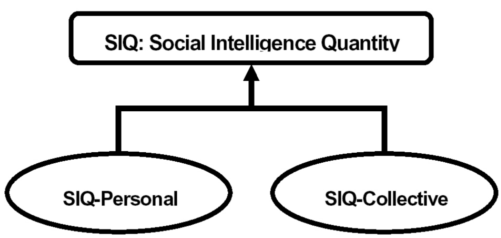


Fig. 11. The framework of Social Intelligence Quantity

SIQ can be defined for an individual as an SIQ-personal, representing a standardized psychological measurement that specifies the individual's social information processing activities such as information acquisition and publication activities, interpersonal relationship building desire, and monopolized information possession desire. We made an experiment of partially measuring

SIQ-personal by using a suite of questionnaires to measure the strength of each individual's desire of acquiring and publishing information.

SIQ can also be defined for a community as an SIQ-collective, representing the community's status of information and knowledge sharing or knowledge creation. SIQ-collective specifies the degree of community activities. SIQ-collective might be measured by observing individuals' behavior. In early communities where community maintenance functions are more dominant than community performance functions to increase the identity of communities, information acquisition desire and interpersonal relationship building desire might be more intensively associated with SIQ-collective. In contrast, information publication desire might be more associated with SIQ-collective, for community performance function becomes more prominent.

7 Concluding Remarks

In this paper, I have pointed out that semantics is grounded on intellectual activities in various communities in the human society. Conversations and stories are primary media for substantiating the semantic process in communities. I have proposed a framework of the conversation-story spiral for supporting communities, and presented the knowledge channel approach featuring knowledge cards for representing conversational units, the knowledge lifecycle support, and the strategic control of information stream. I have shown a couple of implemented systems – EgoChat and S-POC – to show how these ideas are implemented. Future work involves development of sophisticated technology for story-making and conversation-making in real world conversational settings.

References

1. Etienne Wenger. Communities of Practice – Learning Meaning, and Identity, Cambridge University Press, 1998.
2. S. Kelly and M. Jones, "Groupware and the Social Infrastructure of Communication," Comm. ACM, Dec. 2001, pp. 77-79.
3. Hidekazu Kubota, Koji Yamashita, Toyoaki Nishida, Conversational Contents Making a Comment Automatically, Best PhD paper award, In: E. Damiani, R. J. Howlett, L. C. Jain, and N. Ichalkaranje (eds.): Proc. KES 2002, pp. 1326-1330, KES'2002 Sixth International Conference on Knowledge-Based Intelligent Information & Engineering Systems, 16, 17 & 18 September 2002, Crema, Italy, 2002.
4. H. Kubota and T. Nishida. Channel Design for Strategic Knowledge Interaction. to be presented at: KES'2003 Seventh International Conference on Knowledge-Based Intelligent Information & Engineering Systems, University of Oxford, United Kingdom 3, 4 & 5 September 2003.
5. T. Nishida, ed., Dynamic Knowledge Interaction, CRC Press, Boca Raton, Fla., 2000.
6. T. Fukuhara, M. Chikama, and T. Nishida. A Platform for Investigating a Knowledge Creating Community: Community Analysis and Maintenance Functions in the Public Opinion Channel. presented at Social Intelligence Design SID 2003, 6th - 8th July 2003, Royal Holloway University of London, Egham, Surrey U.K.

7. Y. Nakano, T. Murayama, D. Kawahara, S. Kurohashi, and T. Nishida. Embodied Conversational Agents for Presenting Intellectual Multimedia Contents. to be presented at: KES'2003 Seventh International Conference on Knowledge-Based Intelligent Information & Engineering Systems, University of Oxford, United Kingdom 3, 4 & 5 September 2003.
8. T. Murayama, Y. Nakano, and T. Nishida. Participatory Broadcasting System Using Interface Agent and Multimedia, to be presented at Social Intelligence Design SID 2003, 6-th - 8th July 2003, Royal Holloway University of London, Egham, Surrey U.K.
9. M. Horita and N. Iwahashi. On Discovery of Stirring Arguments: A Random-Tree Approach to Collaborative Argumentation Support, to be presented at Social Intelligence Design SID 2003, 6th - 8th July 2003, Royal Holloway University of London, Egham, Surrey U.K.
10. T. Nishida. Social Intelligence Design for Web Intelligence, Special Issue on Web Intelligence, IEEE Computer, Vol. 35, No. 11, pp. 37-41, November, 2002.
11. Toyoaki Nishida. Social Intelligence Design – An Overview, in: Takao Terano, Toyoaki Nishida, Akira Namatame, Yukio Ohsawa, Shusaku Tsumoto, and Takashi Washio (eds): Exploring New Frontiers on Artificial Intelligence – Selected Papers from the First International Workshops of Japanese Society of Artificial Intelligence –, Lecture Notes on Artificial Intelligence LNAI2253, Springer Verlag, December 2001.
12. Koji Yamashita and Toyoaki Nishida: SIQ (Social Intelligence Quantity): Evaluation Package for Network Communication Tools, APCHI 2002 – 5th Asia Pacific Conference on Computer Human Interaction – Beijing, China, 1-4 November 2002.
13. Ken'ichi Matsumura. The factors to activate communication in the network community – New comers or Messages? –, to be presented at Social Intelligence Design SID 2003, 6th - 8th July 2003, Royal Holloway University of London, Egham, Surrey U.K.

Award-Winning Papers (Overview)

Kôiti Hasida

Information Technology Research Institute
National Institute of Advanced Industrial Science and Technology (AIST)
10F. Dai-Building, 1-18-3, Sotokanda, Chiyoda-ku, Tokyo 101-0021, Japan
hasida.k@aist.go.jp

This chapter features seven out of nine awarded papers, selected from JSAI 2004 — the 18th annual conference of Japan Society for Artificial Intelligence. These awarded papers are truly excellent, as they were chosen out of over 290 papers, with the selection rate just about three per cent, and the selection involved approximately seventy reviewers. In artificial intelligence and other related fields, empirical methods have recently been rather prevalent in comparison with theoretical approaches. However, all these awarded papers are both theoretically sound and practically feasibile. Synopses of the seven papers follow, with short comments for the award.

Ide et al. consider the issue of online anomaly detection from a time series of directional data (normalized vectors) in high dimensional systems. In spite of its practical importance, little is known about anomaly detection methods for directional data. Devising a concept of effective dimensions of such systems, they successfully establish an anomaly detection method exempt from the "curse of dimensionality."

Murata et al. report a new method to retrieve relevant documents by utilizing non-relevant documents. Conventional relevance feedback requires both relevant documents and non-relevant ones, but in reality an initial set of documents evaluated by the user may include no relevant documents. The proposed method, *non-relevance feedback*, selects and exploits documents classified as "not non-relevant" and close to the boundary defined by the discriminant function obtained from one-class SVM.

Takeda et al. describe a method to add tags to design documents in order to extract knowledge from texts for intelligent design support. Knowledge description form that they propose is used for capturing knowledge from texts and for inference for the sake of creative design, and design knowledge document contains both human-readable texts and machine-readable knowledge such as propositions and rules.

Hamasaki et al. build a scheduling support system for academic conferences, which assists information exchange among the participants and information discovery by visualizing the participants' interpersonal network. The system was tested at JSAI2003 by 276 active users and proven to promote information exchange among them by visualizing presence of a person to her aquaintances and by providing recommendations in that connection.

A. Sakurai et al. (Eds.): JSAI 2003/2004, LNAI 3609, pp. 187–188, 2007.

Sato et al. study the dynamic nature of social structures by adopting cognitive agents having internal dynamics to changes the agents' internal states governing their behavior. In a simulation where agents are implemented by a simple recurrent network with self-influential connections and play a minority game, they observe a macroscopic structure itinerating among various dynamical states.

Kobayashi et al. introduce CEA (Commands Embedded in Actions) into an interaction model of a human-robot cooperative task. Applying it to a cooperative sweeping task by a human and a small mobile robot, they show that CEA reduces the human cognitive load in the task better than conventional direct commanding methods.

Makio et al. address a problem of discovering skills from human motion data. They extract dependency relations among body parts and frequent patterns (motifs) of motion under the MDL principle. They compare the motions of skilled tennis players and beginners based on the proposed method, and find why skilled players can better serve.

On behalf of the JSAI 2004 program committee, I would like to thank all the chairpersons, discussants, and attentive audience who contributed to selecting these exciting papers, and of course the authors who contributed these papers.

Effective Dimension in Anomaly Detection: Its Application to Computer Systems

Tsuyoshi Idé and Hisashi Kashima

IBM Research, Tokyo Research Laboratory, 1623-14 Shimotsuruma, Yamato-shi, Kanagawa 242-8502, Japan
{goodidea, hkashima}@jp.ibm.com

Abstract. We consider the issue of online anomaly detection from a time sequence of directional data (normalized vectors) in high dimensional systems. In spite of the practical importance, little is known about anomaly detection methods for directional data. Using a novel concept of the effective dimension of the system, we successfully formulated an anomaly detection method which is free from the "curse of dimensionality." In our method, we derive a probability distribution function (pdf) for an anomaly metric, and use a novel update algorithm for the parameters in the pdf, where the effective dimension is included as a fitting parameter. For directional data from a computer system, we demonstrate the utility of our algorithm in anomaly detection.

1 Introduction

A general approach in anomaly detection from vector sequences is to introduce a probability distribution function (pdf) of the collection of negative (or normal) examples. Using a pdf, a threshold value to identify positive (or anomalous) examples can be calculated in a consistent fashion. However, such probabilistic methods often fail for systems with higher dimensions. One major reason is that some of the degrees of freedom in such systems are inactive (almost constant) in many practical applications. A typical example can be found in online text classification, where the dimensions of the document vectors are often on the order of one million, but the dimensions that are effective in classification are known to be on the order of several hundreds [4].

In this paper, we formulate an online anomaly detection algorithm for directional data (normalized vectors) based on the von Mises-Fisher (vMF) distribution [10]. Although directional data often appears in many practical situations [10,2,8], little is known in the context of anomaly detection.

Explicitly, the vMF distribution is given by,

$$p(\boldsymbol{u}|\kappa, \boldsymbol{\mu}) = \frac{\kappa^{\frac{N}{2}-1}}{(2\pi)^{N/2} I_{\frac{N}{2}-1}(\kappa)} \exp\left(\kappa \boldsymbol{\mu}^T \boldsymbol{u}\right), \tag{1}$$

where $\boldsymbol{\mu}$ is a mean direction and $I_l(\cdot)$ represents the modified Bessel function of order l. The value $1/\kappa > 0$ is a constant parameter called the angular variance.

A. Sakurai et al. (Eds.): JSAI 2003/2004, LNAI 3609, pp. 189–204, 2007.

The intrinsic dimension of the directional data $\boldsymbol{u}$ is denoted as N. Intuitively, the vMF distribution describes fluctuations of $\boldsymbol{u}$ around the mean direction. The vMF distribution is the most natural distribution for directional data in that it can be derived using the maximum entropy principle under the conditions that (1) the total probability is unity and (2) that the average over $\boldsymbol{u}$ on the unit sphere is $\boldsymbol{\mu}$. Since the normal distribution is derived if the second condition is replaced with that of the average in the whole N-dimensional space, the vMF distribution can be regarded as the "normal" distribution for directional data.

To detect anomalies in an online fashion, we need to update the pdf in accordance with the data just given at the current time, t. One possible way is to perform maximum likelihood estimation (MLE) continuously for the new data. Banerjee *et al.* [2] employed a mixture of vMF distributions, and derived an approximated version of the MLE procedure. However, parameter estimation in this case is quite difficult due to the modified Bessel function. Especially, the parameter N often degrades the accuracy of the approximations of $I_l(\cdot)$ in many application areas if N is relatively large [1].

In this paper, we introduce a novel concept, effective dimensions, to overcome the curse of dimensionality. Starting from the vMF distribution, we derive a pdf for an anomaly metric based on the Fisher kernel [9]. The pdf contains two parameters that the angular variance and the effective dimension, n, instead of the intrinsic dimension of the directional data, N. Then we introduce a new online algorithm to update the pdf at each time step. To the best of our knowledge, this is the first attempt to overcome the curse of dimensionality using the notion of the effective dimension. Note that existing formulation using Gaussian mixtures [15,14] are not appropriate in this case because of the degeneracies of the distributions due to the normalization condition and the existence of inactive variables.

We will experimentally show that the effective dimension is actually much less than the nominal dimension N for feature vectors extracted from a computer system. Also, we demonstrate that anomalies can be detected by comparing with a given critical probability that is independent of the details of the system.

The rest of this paper is organized as follows: In Section 2, we define the dependency matrix in computer systems and recapitulate our method of feature extraction [8]. In Section 3, we define an anomaly metric. In Section 4, a generative model for the anomaly metric is derived from the vMF distribution, and introduce the concept of effective dimensions. In Section 5, a novel incremental algorithm is proposed to estimate the parameter in the model. In Section 6, we report on experimental results in a benchmark system. In the final section, we summarize the major results in this paper.

2 Modeling Computer Systems

2.1 Dependency Matrix

To model the behavior of Web-based computer systems at the application layer, where the interaction between servers is essential, we define a *service* as a quartet of

$$(I_s, I_d, P, Q),$$

where I_s and I_d represent source and destination IP (Internet Protocol) addresses, respectively, and P denotes the port number of the destination application. We also use an attribute called the transaction type Q. Figure 1 illustrates a benchmark system. There are four server boxes in this system, and two server processes with port numbers p_1 and p_2 are installed on each of the boxes at i_2 and i_3.

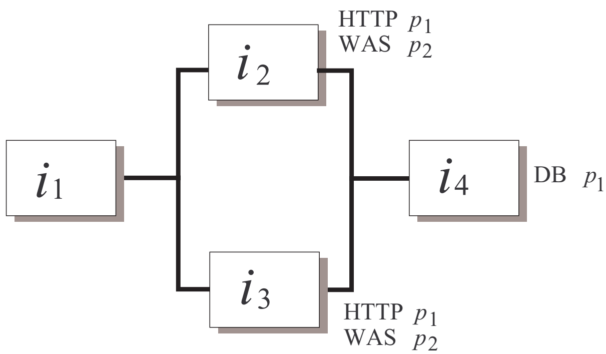


Fig. 1. Configuration of benchmark system. IP addresses and port numbers are denoted by i_k $(k = 1, .., 4)$ and p_j $(j = 1, 2, 3)$, respectively.

Consider a system with N different services, and imagine a graph each of whose nodes is one of the services. For the edge weights, we employ the following quantity [8]:

$$\mathsf{D}_{i,j} = [f(d_{i,j}) + f(d_{j,i})]\,(1 - \delta_{i,j}) + \alpha_i \delta_{i,j}, \tag{2}$$

where $\delta_{i,j}$ is Kroneker's delta function and the α_is are constants introduced to stabilize the numerical calculations. Considering the bursty nature of Web traffic, we use $f(\cdot) = \ln(1 + \cdot)$. In principle, the quantity $d_{i,j}$ can be measured through server logs or some estimation algorithm [13,6]. By definition, D is a square nonnegative matrix. Hereafter, we use a sans serif font to indicate matrices and use bold italic to indicate vectors. The norm of vectors is defined as the L_2-norm.

Figure 2 shows a subgraph of the dependency graph expected in the system depicted in Fig. 1. We drew links if $I_s = I_d$ holds between two services, and services involving only q_1 and q_2 are shown there. Generally, the dependency graph of a Web-based system is quite complicated even if the corresponding IP network is simple. For an instance of services, see Section 6.

2.2 Definition of Feature Vector

Let us assume that the data for the dependency matrix D is sequentially obtained at each time step t=1,2,... with a fixed interval, and that the dependency

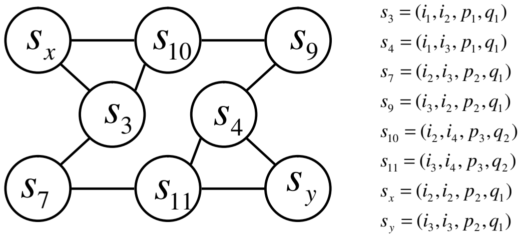


Fig. 2. A part of the dependency graph for the system in Fig. 1. Only services which have $Q = q_1$ or q_2 are shown. Graph edges are drawn if $I_s = I_d$ holds between two vertices.

graph has a single connected component. We define the feature vector $\boldsymbol{u}$ of D as

$$\boldsymbol{u}(t) \equiv \arg\max_{\tilde{\boldsymbol{u}}} \left\{ \tilde{\boldsymbol{u}}^T \mathsf{D}(t) \tilde{\boldsymbol{u}} \right\} \tag{3}$$

subject to $\tilde{\boldsymbol{u}}^T \tilde{\boldsymbol{u}} = 1$, where T denotes transpose. Since D is a non-negative matrix, one can see that the maximum value is attained if the weight of $\boldsymbol{u}(t)$ is larger for services where $\mathsf{D}_{ij}(t)$ is larger. If a service i actively calls other services, $\boldsymbol{u}(t)$ has a large weight for the i-th element. Following this interpretation, we call this feature vector an *activity vector*.

By introducing a Lagrange multiplier λ, Eq. (3) can be rewritten as

$$\frac{d}{d\tilde{\boldsymbol{u}}} \left[\tilde{\boldsymbol{u}}^T \mathsf{D}(t) \tilde{\boldsymbol{u}} - \lambda \tilde{\boldsymbol{u}}^T \tilde{\boldsymbol{u}} \right] = 0, \tag{4}$$

so that

$$\mathsf{D}(t) \tilde{\boldsymbol{u}} = \lambda \tilde{\boldsymbol{u}}. \tag{5}$$

While this equation holds for any of the eigenvectors of $\mathsf{D}(t)$, the feature vector corresponding to Eq. (3) is defined as the principal eigenvector (the eigenvector whose eigenvalue is the largest). Since Eq. (5) is homogeneous in $\tilde{\boldsymbol{u}}$, the direction of the activity vector is invariant with respect to $\mathsf{D}(t) \to \eta \mathsf{D}(t)$ for any nonzero real number η. Thereby we can exclude overall traffic changes from analysis. It is the eigenvalue that is proportional to the global traffic volume. This is important to abstract a hidden structure from D.

To understand the meaning of $\boldsymbol{u}$ further, one can relate $\boldsymbol{u}$ with a stationary state of a discrete-time linear dynamical system whose equation of motion is given by

$$\boldsymbol{x}(\tau + 1) = \mathsf{D}(t) \boldsymbol{x}(\tau), \tag{6}$$

where τ denotes a virtual time being independent of the actual time t, and $\boldsymbol{x}$ is associated with $\boldsymbol{u}$ by $\boldsymbol{u}{=}\boldsymbol{x}/||\boldsymbol{x}||$. Since $\mathsf{D}(t)$ is symmetric and of full-rank at least for $\alpha > 0$, all eigenvalues are real. Using the eigenvalues, $\boldsymbol{x}(0)$ can be expressed as a linear combination of the eigenvectors, so that

$$\boldsymbol{x}(\infty) = \lim_{n \to \infty} \left[\mathsf{D}(t) \right]^n \boldsymbol{x}(0) = \lim_{n \to \infty} \sum_{i=1}^{N} [\lambda_i(t)]^n c_i(t) \boldsymbol{u}_i(t),$$

where the eigenvalues and the normalized eigenvectors are denoted by $\lambda_i(t)$ and $\boldsymbol{u}_i(t)$ for i=1, 2, ..., N, respectively, and $c_i(t)$'s are coefficients of the linear combination. Evidently, the term of the maximum eigenvalue becomes dominant as $n \to \infty$. Thus, we have

$$\boldsymbol{u}(t) = \boldsymbol{x}(\infty)/||\boldsymbol{x}(\infty)||.$$

Specifically, the state vector approaches $\boldsymbol{u}$ after an infinite number of transitions. For computer systems, the stationary state can be interpreted as the distribution of the probability amplitude that a service is holding the control token of the system at a virtual time point of τ.

2.3 Activity Vectors in Disconnected Systems

In real computer systems, the dependency graph is often disconnected. For such systems, a permutation matrix P exists such that

$$\mathsf{P}^T \mathsf{D} \mathsf{P} = \begin{bmatrix} \mathsf{D}_1 & & 0 \\ & \mathsf{D}_2 & \\ 0 & & \ddots \end{bmatrix},$$

where D_1, D_2, ... are square submatrices. To be concrete, consider the system shown in Fig. 3. Using

$$\mathsf{P} = \begin{bmatrix} 1\,0\,0\,0\,0\,0 \\ 0\,0\,0\,0\,1\,0 \\ 0\,0\,1\,0\,0\,0 \\ 0\,0\,0\,0\,0\,1 \\ 0\,1\,0\,0\,0\,0 \\ 0\,0\,0\,1\,0\,0 \end{bmatrix},$$

the whole dependency matrix is decomposed into two square submatrices:

$$\mathsf{D}_1 = \begin{bmatrix} 0 & a_{15} & a_{13} & 0 \\ a_{15} & 0 & 0 & a_{56} \\ a_{13} & 0 & 0 & a_{36} \\ 0 & a_{56} & a_{36} & 0 \end{bmatrix}, \ \mathsf{D}_2 = \begin{bmatrix} 0 & a_{24} \\ a_{24} & 0 \end{bmatrix}. \tag{7}$$

Evidently, each submatrix corresponds to a single connected subgraph. Since the eigenvalue equation is invariant with respect to orthogonal transformations, the whole eigenvalue equation is written as

$$0 = \det \left| \mathsf{D}_1 - \lambda \mathsf{E}_{(4)} \right| \cdot \det \left| \mathsf{D}_2 - \lambda \mathsf{E}_{(2)} \right|, \tag{8}$$

where $\mathsf{E}_{(n)}$ represents the n-dimensional identity matrix. Consequently, the solution of the whole system can be obtained as the union of the solutions of each connected component. This fact allows us to analyze each subgraph separately.

For each connected component, the Perron-Frobenius theorem [3], which holds for non-negative irreducible matrices, guarantees that the principal eigenvector is positive, where an eigenvector is said to be positive if all the components of $\boldsymbol{u}$ or $-\boldsymbol{u}$ are positive and the corresponding eigenvalue is positive. This naturally supports the interpretation of the principal eigenvector as the activity vector, since the magnitude of the activities should be positive. In addition, the Perron-Frobenius theorem also guarantees that the principal eigenvalue is real [1] and has no degeneracy. From this, we understand that the activity vector is free from subtle problems due to level crossings of the eigenstates within a single connected component in the normal state of the system. If a level crossing easily occurs due to small fluctuations, the transition from one eigenstate to another eigenstate may be recognized as an outlier, resulting in a false alert. For more discussion on the stability of activity vectors, see [8].

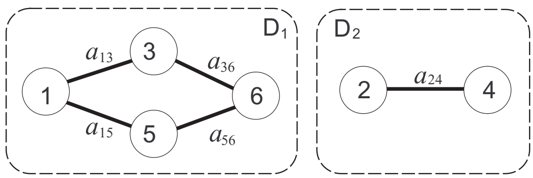


Fig. 3. Example of a disconnected graph

2.4 Remark

Our feature extraction technique provides a natural way to summarize the information contained in D. The eigencluster decomposition allows us to analyze each single eigencluster separately, and the activity vector extraction technique allows us to further reduce the degrees of freedom. When the set of all services is unknown, it is practically possible to find the activity vectors by choosing positive vectors from a set of eigenvectors [11]. Thus, we expect that the degrees of freedom of each of subproblems are still moderate even when the whole degrees of freedom are very large. In addition, the feature vector has a clear interpretation that is comprehensible to system administrators. Understanding what is happening is as essential as detection itself in practical situations. These are advantages over naive approaches such as defining a feature vector simply by connecting all of the column vectors, where the scalability cannot be achieved and interpretation of results is often unclear.

For the numerical calculations, an extremely fast and simple algorithm called the power method [12] is known to find the principal eigenvector. While the activity vector must be calculated online whenever D is updated in the given time interval Δt, typically on the order of a few tens of seconds, our experience shows that the time to convergence is far less than Δt even for N on the order of 10^3.

[1] In this case, all of eigenvalues are real since Eq. (2) makes D real and symmetric.

3 Anomaly Metric

3.1 Definition

Now we consider how to detect anomalous changes from the sequence of activity vectors $\{\boldsymbol{u}^{(t)}\}$ for $t = 1, 2, ...$ Since $\boldsymbol{u}^{(t)}$ is normalized, this is a time sequence of *directional data*.

To define the anomaly measure, recall the fact that the Fisher kernel function [9] defines a natural affinity between observables in terms of Fisher's information matrix. For the vMF distribution, the Fisher kernel function is given as

$$K(\boldsymbol{u}_i, \boldsymbol{u}_j) = \kappa^{-2} \left[\frac{\partial \ln p(\boldsymbol{u}_i | \kappa, \boldsymbol{\mu})}{\partial \boldsymbol{\mu}} \right]^T \frac{\partial \ln p(\boldsymbol{u}_j | \kappa, \boldsymbol{\mu})}{\partial \boldsymbol{\mu}} = \boldsymbol{u}_i^T \boldsymbol{u}_j.$$

This is nothing but the cosine similarity. Since it takes a value within $[0, 1]$, we define the anomaly (or dissimilarity) measure $z(t)$ as

$$z(t) \equiv 1 - K\left(\boldsymbol{r}(t), \boldsymbol{u}(t)\right), \tag{9}$$

where $\boldsymbol{r}(t)$ denotes the past typical pattern defined at t. The value of $z(t)$ is unity if the present activity vector is orthogonal to the typical pattern, and zero if the present activity vector is identical to the typical pattern. In the present context, if $z(t)$ is greater than a given threshold, we infer that an anomalous situation is occurring in the system.

3.2 Extraction of Typical Activity Pattern

We define a matrix $\mathsf{U}(t)$ by

$$\mathsf{U}(t) = \left[\boldsymbol{u}(t-1), \boldsymbol{u}(t-2), \cdots, \boldsymbol{u}(t-W)\right], \tag{10}$$

where W is a window size. Clearly, $\mathsf{U}(t)$ is an $N \times W$ matrix. We suppose that the typical pattern is a linear combination of the column vectors:

$$\boldsymbol{r}(t) = c \sum_{i=1}^{W} v_i \boldsymbol{u}(t-i), \tag{11}$$

where c is the normalization constant to satisfy $\boldsymbol{r}^T\boldsymbol{r} = 1$ under the condition of $\sum_{i=1}^{W} v_i^2 = 1$. The easiest way to obtain $\boldsymbol{r}(t)$ is to assume that the v_is are independent of i. In that case, $\boldsymbol{r}(t)$ is parallel to the mean vector, $\overline{\boldsymbol{r}}(t)$. Practically, a good way to reduce the unwanted effects of noisy fluctuations is to optimize the coefficients $\boldsymbol{v}^T = (v_1, v_2, ..., v_W)$ based on

$$\boldsymbol{v}(t) \equiv \arg\max_{\tilde{\boldsymbol{v}}} \left\| \sum_{i=1}^{W} \tilde{v}_i \boldsymbol{u}(t-i) \right\|^2 = \arg\max_{\tilde{\boldsymbol{v}}} \left\{ \tilde{\boldsymbol{v}}^T U(t)^T U(t) \tilde{\boldsymbol{v}} \right\} \tag{12}$$

subject to $\tilde{\boldsymbol{v}}^T\tilde{\boldsymbol{v}} = 1$. It is well-known in the field of pattern recognition that the solution of this equation is given by the Karhunen-Loève decomposition [5]. Specifically, $\boldsymbol{v}(t)$ is a right singular vector of $\mathsf{U}(t)$, and c is the inverse of the corresponding singular value. So, we conclude that $\boldsymbol{r}(t)$ is the principal left singular vector of $\mathsf{U}(t)$, where a singular vector is said to be principal if it corresponds to the largest singular value. Again, the power method [12] is a good way to perform the singular value decomposition (SVD).

4 Generative Model for Anomaly Metric

4.1 Marginal Distribution over z

A conventional method to detect anomalies in a time sequence of multivariate vectors is to find outliers using a generative model that describes the distribution of the multivariate vectors [15,14,2]. However, as discussed in Introduction, such approaches have difficulties for high-dimensional data. Instead, we consider a pdf of the anomaly measure itself, assuming that the distribution of $\boldsymbol{u}$ basically obeys the vMF distribution given in Eq. (1).

Before plunging into the detail, we summarize our anomaly detection procedure in Fig. 4, where we denote the angle between $\boldsymbol{r}(t)$ and $\boldsymbol{u}(t)$ as θ. As shown, the basic procedure is to extract a typical pattern from the past activity vectors, and to calculate the dissimilarity of the present activity vector from this typical one. We believe that this is reasonable approach if the typical pattern is relatively stable, and it is the case at the application layer of Web-based computer systems.

Since θ has a one-to-one correspondence to z as $z = 1 - \cos\theta$, one can derive the pdf over z through the marginalized distribution with respect to θ, starting from Eq. (1). We perform a transformation of the variables from $\boldsymbol{u}$ to angular variables $\{\theta, \theta_2, ..., \theta_{N-1}\}$ of the N-dimensional spherical coordinates. By using

$$d^{N-1}\Omega = d\theta d\theta_2 \cdots d\theta_{N-1} \sin^{N-2}\theta \sin^{N-3}\theta_2 \cdots \sin\theta_{N-2},$$

where $d^{N-1}\Omega$ is the area element on the unit sphere in an N-dimensional Euclidean space, the marginalized distribution for θ is written as

$$\int d\theta_2 d\theta_3 \cdots d\theta_{N-1} \left[p(\boldsymbol{u}|\kappa, \boldsymbol{\mu}) \sin^{N-2}\theta \sin^{N-3}\theta_2 \cdots \sin\theta_{N-2} \right]. \tag{13}$$

Since

$$z(t) \simeq \frac{\theta^2}{2}, \ \cos\theta \simeq 1 - \frac{\theta^2}{2}, \ \sin\theta \simeq \theta$$

hold for $|\theta| \ll 1$, we see that the distribution for z is given by

$$q(z) \propto \exp\left[-\frac{z}{2\Sigma} \right] z^{\frac{N-1}{2}-1}, \tag{14}$$

where we used $\theta d\theta = dz$ and set $1/2\kappa$ to be Σ. Apart from a prefactor and the scaling factor Σ, this is the same as the χ^2-distribution with $N - 1$ degrees of freedom.

We have derived this generative model from the vMF distribution of $\boldsymbol{u}$, which is the most natural assumption as long as the fluctuations around the mean

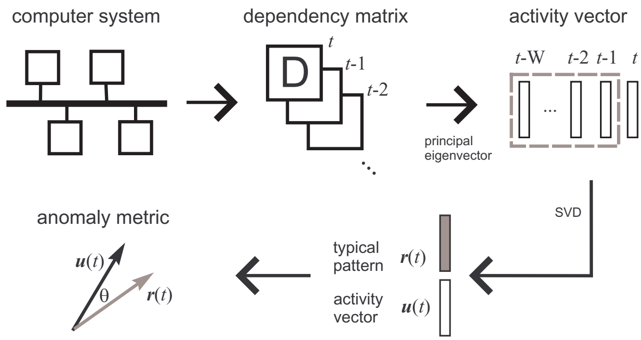


Fig. 4. Summary of our anomaly detection procedure

direction is relatively small. However, our empirical study shows that the above model is not consistent with the experimental distribution at all. One reason can be found in the fact that some of the degrees of freedom happen to be inactive over some duration of time. In the derivation of the vMF distribution, an implicit assumption is that all of the degrees of freedom are equally active. These observations lead us to the concept of the effective dimension.

4.2 Effective Dimension

One of the most important steps in our formulation is to replace N in Eq. (14) with a parameter n, and regard it as a fitting parameter. We call n the *effective dimension* of the system. If properly estimated, the effective dimension represents the active degrees of freedom of the system. We expect that n is much smaller than N in many application domains. For Web-based systems, the activities of some of services are much lower than those of others, so that n is much smaller than N, as shown in Section 6.

Since the function $q(z)$ rapidly decreases as $z \to \infty$ for a moderate value of the degrees of freedom, the normalization constant can be evaluated by integrating over $[0, \infty)$. Using the definition of the gamma function $\Gamma(\cdot)$, we have

$$q(z|n, \Sigma) = \frac{1}{(2\Sigma)^{\frac{n-1}{2}} \Gamma(\frac{n-1}{2})} e^{-z/(2\Sigma)} z^{\frac{n-1}{2}-1}, \tag{15}$$

where we use the notation of $q(z|n, \Sigma)$ instead of $q(z)$ to emphasize the effective dimension as a fitting parameter. We see that the distribution of $z \in [0, 1]$ can be written as the χ^2-distribution with $n - 1$ degrees of freedom.

5 Online Estimation of Parameters

5.1 Moment-Based Estimation Scheme

To estimate the two parameters n and Σ in Eq. (15), we employ a moment-based estimation scheme. Note that MLE for the χ^2-distribution is intractable due to the gamma function. Fortunately, analytic formulas about the first (m_1) and second (m_2) moments are available for the χ^2-distribution. By using the definition of the gamma function, it is easy to show

$$m_1 \equiv \int_0^\infty dz\ q(z|n, \Sigma) z = (n-1)\Sigma$$

and

$$m_2 \equiv \int_0^\infty dz\ q(z|n, \Sigma) z^2 = 2(n^2-1)\Sigma^2,$$

where we again extended the domain of integration to $[0, \infty)$. This can be also justified within the approximation made above. Solving these equation with respect to n and Σ, we have

$$n = 1 + \frac{2{m_1}^2}{m_2 - {m_1}^2} \quad \text{and} \quad \Sigma = \frac{m_2 - {m_1}^2}{2m_1}. \tag{16}$$

These relations provide a way to evaluate the parameters n and Σ experimentally. For example, if we have T samples $\{z(t)|t = 1, 2, ..., T\}$, the experimental first and second moments can be calculated as $\hat{m}_1 = (1/T)\sum_{t=1}^{T} z(t)$ and $\hat{m}_2 = (1/T)\sum_{t=1}^{T} z(t)^2$, respectively.

5.2 Incremental Algorithm

These formulas can be extended to their incremental versions. Using an identity

$$\frac{1}{t}\sum_{i=1}^{t} z(t) = \left(1 - \frac{1}{t}\right)\frac{1}{t-1}\sum_{i=1}^{t-1} z(i) + \frac{1}{t} z(t).$$

and setting $1/t$ as β, we have

$$\hat{m}_1(t) = (1-\beta)\cdot \hat{m}_1(t-1) + \beta z(t). \tag{17}$$

Similarly for the second moment, we have

$$\hat{m}_2(t) = (1-\beta)\cdot \hat{m}_2(t-1) + \beta z(t)^2. \tag{18}$$

Naturally, β satisfies $0 < \beta < 1$ and is called the discounting factor.

Since $1/\beta$ can be associated with the number of data points, a rough estimate of β may be $\beta \sim \Delta t/L$, where L denotes the time scale we are interested in. Similarly, W can be estimated as $W \sim L/\Delta t$. In our benchmark system, we empirically take L on the order of 10 minutes.

Based on the above discussion, we have an online algorithm to calculate a threshold value to judge whether it is anomalous or not:

1. Give a critical boundary $0 < p_c < 1$.
2. Calculate $\hat{m}_1$ and $\hat{m}_2$ at t using Eqs. (17) and (18).
3. Calculate n and Σ using Eq. (16).

4. Find z_{th} numerically such that $\int_0^{z_{\text{th}}} dz\, q(z|n, \Sigma) = p_c$.
5. Emit an alert if $z(t) > z_{\text{th}}$.

The above algorithm includes three parameters, p_c, β, and W. Since β and W can be easily estimated with L and Δt, the only parameter we must specify is substantially p_c, which is totally independent of the details of the system.

6 Experiment

6.1 Experimental Settings

The configuration of our benchmark system is illustrated in Fig. 1. As shown, the HTTP servers and WASs are doubly redundant. On the WASs, two applications, "Trade" and "Plants," are running. Trade is a standard benchmark application called Trade 3 [7], and Plants is a sample application bundled with IBM WebSphere Application Server V5.0 and simulates an online store dealing with plants and gardening tools. For both, the number of clients was fixed to be 16 and the think time was randomly chosen from 0 to 4 seconds.

We generated a matrix D every 20 seconds using a method that evaluates $d_{i,j}$ from captured IP packets. Loopback packets were ignored in the experiments, so that the services s_x and s_y in Fig. 2 are not observed for $i_1 = 192.168.0.53$ and $i_2 = 192.168.0.54$. The principal eigencluster is defined in Table 1, and small perturbations affecting it were ignored. In Table 1, the zeroth service was introduced to describe the situation where an optimal pair between callee and caller could not be identified. For example, services triggered by those outside the intranet will be associated with the zeroth service.

Apart from these, there are other service types, "DB2" and "JMS," in Table 1. DB2 denotes a request for the DB server, and JMS is for communications related to the Java Messaging Service.

Table 1. Services appearing in the principal eigencluster

Index	I_s	I_d	P	Q
0	0.0.0.0	0.0.0.0	0	(none)
1	192.168.0.19	192.168.0.53	80	Plants
2	192.168.0.19	192.168.0.54	80	Plants
3	192.168.0.19	192.168.0.53	80	Trade
4	192.168.0.19	192.168.0.54	80	Trade
5	192.168.0.54	192.168.0.53	5558	JMS
6	192.168.0.53	192.168.0.54	9081	Plants
7	192.168.0.53	192.168.0.54	9081	Trade
8	192.168.0.54	192.168.0.53	9081	Plants
9	192.168.0.54	192.168.0.53	9081	Trade
10	192.168.0.53	192.168.0.52	50000	DB2
11	192.168.0.54	192.168.0.52	50000	DB2

6.2 Statistical Properties in the Normal State

We calculated $\boldsymbol{u}$ and z online over a period when the system exhibited no failures. The dependency matrix was generated over 52.7 minutes, so we had 158 matrices. The α_i values were taken as small random numbers on the order of 0.01. To see the fluctuation in D, we show in Fig. 5 the time dependence of $d_{9,11}$ as an example. We see that there are approximately 500 calls within 20 seconds under these experimental conditions and that the amplitude of fluctuation of $d_{9,11}$ is almost of the same order as the average. Hence, it makes little sense to place a threshold value on an isolated $d_{i,j}$.

To experimentally validate the pdf of z, we plotted the frequency distribution of z in Fig. 6 (a), where the χ^2 pdf is also shown. The parameters of the χ^2 pdf were calculated using all of the 158 data points with no discounting. The result was

$$n = 4.62 \quad \text{and} \quad \Sigma = 6.79 \times 10^{-5}.$$

It is noteworthy that the calculated effective dimension is much less than $N = 12$. In spite of the limitation of the number of data points, the frequency distribution is a good fit to the χ^2 pdf. We also drew a quantile-quantile plot in Fig. 6 (b). As shown, the experimental data is well placed on the 45 degree line. These results clearly support our formulation.[2]

6.3 Detection of an Application Fault

Next, we performed a more realistic experiment: A bug in one of the applications ("Plants") only on 192.168.0.54 causes a malfunction of the service of 11 at a time point. The server process itself continues running, so the network communication is normal at the IP layer or below. Since two Web servers are working on the system, a client may feel no change in response time as long as the overall traffic is sufficiently small. Although this defect occurs within a single service, it can cause a massive change in D. In fact, the dependencies of the services directly related with the service 11 will be considerably changed. What we would like to detect is a transition of this kind.

Figure 7 shows the generated time-series of the activity vector. We see that a sudden change in activities is observed at t_A and t_B, which correspond to the malfunction of the service 11 and its recovery. From the figure, the activities of the services 2, 6, and 11 are clearly decreased during this period. This result demonstrates that the service activity vector actually expresses the activity of services, and suggests a way to visualize the whole system.

To detect this fault automatically, we calculated z and its threshold value, following the algorithm explained in Section 5. In Fig. 8, we depicted the z values with vertical bars and the threshold values with thick gray curves for W =5, 25, and 50. The discounting factor and the critical boundary were taken

[2] We rounded the $n-1$ value to be 4 to fit the χ^2 pdf because of the limitation of the numerical library we used. This is the main reason the deviation of the χ^2 pdf from the experimental frequency.

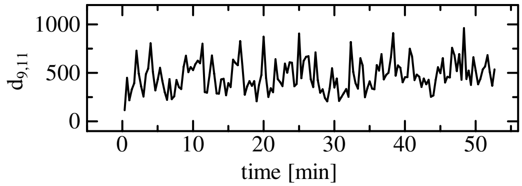


Fig. 5. Time dependence of $d_{9,11}$

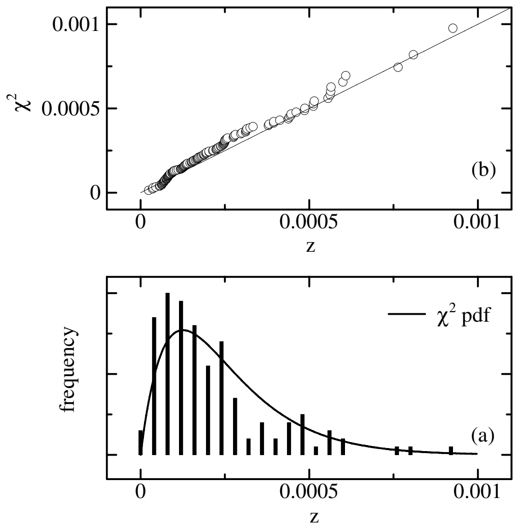


Fig. 6. Statistics of z in the normal state. (a) Comparison of the experimental frequency and the χ^2 pdf. (b) The quantile-quantile plot.

as $\beta = 0.005$ and $p_c = 0.5\%$, respectively. While the result is considerably affected by the choice of W, we observe clear features at t =35.0 and 45.7 minutes, which correspond to t_A and t_B in Fig. 7, respectively. These time points are highlighted with dashed vertical lines in Fig. 8. Note that the feature at t_B (recovery from the malfunction) demonstrates the learnability for gradual changes of the environment. The dependence on W is an inevitable consequence of the choice of the applications. Since the benchmark applications simulate human behavior, they must have a characteristic time scale. Comparing Fig. 7 with Fig. 8, we conclude that an appropriate value of W is about 25 (8.3 minutes). We see that this value of W allows us to pinpoint the time points t_A and t_B.

The curves plotted with thin lines ("not SVD") in Fig. 8 represent the result using the simple mean vector $\overline{\boldsymbol{r}}$ instead of $\boldsymbol{r}$. The trend of z is similar to that of the SVD-based method, but is blurred out by the noise. This result demonstrates the effectiveness of the SVD-based pattern extraction technique.

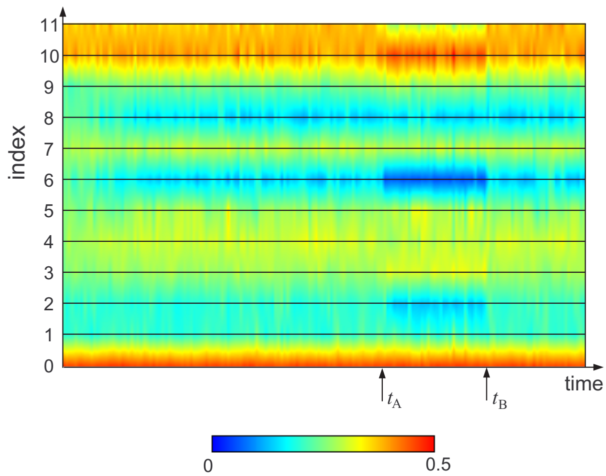


Fig. 7. Time dependence of the activity vector. The failure duration starts at t_A and ends at t_B, as shown by arrows. The definition of the service indices are shown in Table 1.

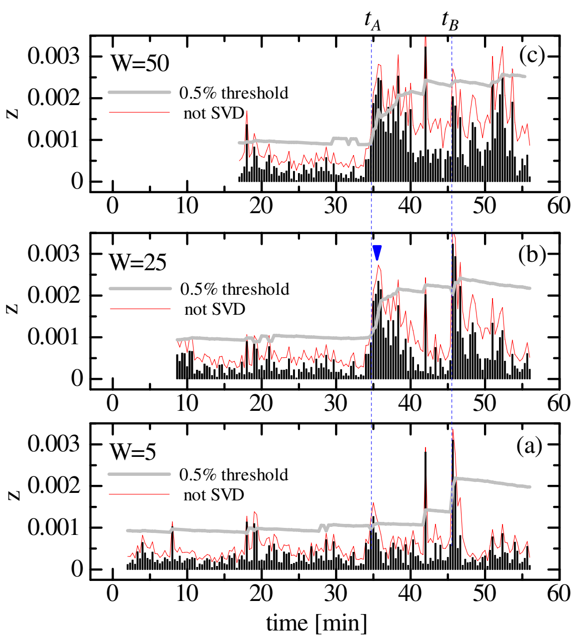


Fig. 8. The dependence of z for W = (a) 5, (b) 25, and (c) 50. The 0.5% threshold is denoted by gray curves.

For the limitations of our approach, first, the probability of false alarms will be finite even if W is set to be the optimal value. As understood from Fig. 8, there is small finite probability of having outliers beyond a threshold value. Second, since the basic assumption of our approach is the stability of the direction of the activity vector, our approach is not appropriate for anomaly detection of rarely invoked services. Finally, there is much room for improvement in the calculations of the threshold values since the numerical library we used handles only integer degrees of freedom in the χ^2 pdf.

7 Summary

We have proposed a new framework of statistical anomaly detection for a time-sequence of directional data. First, we defined an anomaly metric, z, based on the Fisher kernel function of the von Mises-Fisher distribution, and derived its probability distribution as the χ^2 distribution in an approximated manner. Second, we proposed a new concept of the effective dimension, n, and gave its online estimation algorithm based on the method of moments. Our generative model of z is the χ^2 distribution with $n-1$ degrees of freedom. Third, we derived an online algorithm to calculate threshold values of z. Only a value of the critical probability p_c is needed to determine the threshold. Finally, we demonstrated the utility of our method in a fault detection task in a benchmark computer system.

References

1. A. Banerjee, I. Dhillon, J. Ghosh, and S. Sra. Expectation maximization for clustering on hyperspheres. *Technical Report*, TR-03-07, 2003. Department of Computer Sciences, University of Texas at Austin.
2. A. Banerjee, I. Dhillon, J. Ghosh, and S. Sra. Generative model-based clustering of directional data. In *Proceedings of the Ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 19–28, 2003.
3. A. Berman and R. J. Plemmons. *Nonnegative Matrices in the Mathematical Sciences*, volume 9 of *Classics in applied mathematics*. SIAM, 1994.
4. S. C. Deerwester, S. T. Dumais, T. K. Landauer, G. W. Furnas, and R. A. Harshman. Indexing by latent semantic analysis. *Journal of the American Society of Information Science*, 41(6):391–407, 1990.
5. R. O. Duda, P. E. Hart, and D. G. Stork. *Pattern Classification*. Wiley-Interscience, 2nd edition, 2000.
6. M. Gupta, A. Neogi, M. K. Agarwal, and G. Kar. Discovering dynamic dependencies in enterprise environments for problem determination. In *Proceedings of 14th IFIP/IEEE Workshop on Distributed Systems: Operations and Management*, pages 221–233, 2003.
7. IBM. Trade3; `http://www-306.ibm.com/software/webservers/appserv/benchmark3.html`.
8. T. Idé and H. Kashima. Eigenspace-based anomaly detection in computer systems. In *Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2004.

9. T. Jaakkola and D. Haussler. Exploiting generative models in discriminative classifiers. In *Advances in Neural Information Processing Systems*, volume 11, pages 487–493, 1999.
10. K. Mardia. *Multivariate Analysis*. Academic Press, 1980.
11. S. Sarkar and K. Boyer. Quantitative measures for change based on feature organization: Eigenvalues and eigenvectors. *Computer Vision and Image Understanding*, 71:110–136, 1998.
12. G. Strang. *Linear Algebra and its Applications*. Academic Press, 1976.
13. The Open Group. Application response measurement — ARM; `http://www.opengroup.org/tech/management/arm/`.
14. K. Yamanishi and J. Takeuchi. A unifying framework for detecting outliers and change points from non-stationary time series data. In *Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 676–681, 2002.
15. K. Yamanishi, J. Takeuchi, G. Williams, and P. Milne. On-line unsupervised outlier detection using finite mixtures with discounting learning algorithms. In *Proceedings of the Sixth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 320–324, 2000.

Document Retrieval Using Feedback of Non-relevant Documents

Hiroshi Murata[1], Takashi Onoda[1], and Seiji Yamada[2]

[1] Central Research Institute of Electric Power Industry, Tokyo, Japan
{murata,onoda}@criepi.denken.or.jp
[2] National Institute of Informatics, Tokyo, Japan
seiji@nii.ac.jp

Abstract. This paper reports a new document retrieval method using non-relevant documents. Suppose, we need to find documents interesting to the user in as few iterations of human intervention as possible. In each iteration, a relatively small set of documents is evaluated in terms of the relevance to the user's interest. Ordinary relevance feedback needs both relevant and non-relevant documents, but the initial set of documents checked by the user may often not include relevant documents. Accordingly we propose a new feedback method using non-relevant documents only. This *"non-relevance feedback"* selects documents classified as "not non-relevant" and close to the boundary defined by the discriminant function obtained from one-class SVM. Experiments show that this method can efficiently retrieve a relevant documents.

1 Introduction

With the continued progression of Internet technology, the amount of information accessible by end users is increasing explosively. In this situation, it is now possible to access a huge document database through the web. However, it is difficult for a user to retrieve relevant documents from which he/she can obtain useful information, and therefore, many studies have been done on information retrieval, particularly document retrieval [11]. Various studies on such document retrieval have been reported in TREC (Text Retrieval Conference) [10] for English documents, and IREX (Information Retrieval and Extraction Exercise) [2] and NTCIR (NII-NACSIS Test Collection for Information Retrieval System) [3] for Japanese documents.

In most frameworks for information retrieval, a vector space model is used in which a document is described with a high-dimensional vector [7]. An information retrieval system using a vector space model computes the degree of similarity between a query vector and document vectors by using the cosine of the two vectors, and then indicates to the user a list of retrieved documents.

In general, since a user rarely describes a query precisely in the first trial, an interactive approach to modifying the query vector on the basis of an evaluation by the user of documents in a list of retrieved documents, has been proposed. This method is called *relevance feedback* [6] and is used widely in information

A. Sakurai et al. (Eds.): JSAI 2003/2004, LNAI 3609, pp. 205–215, 2007.

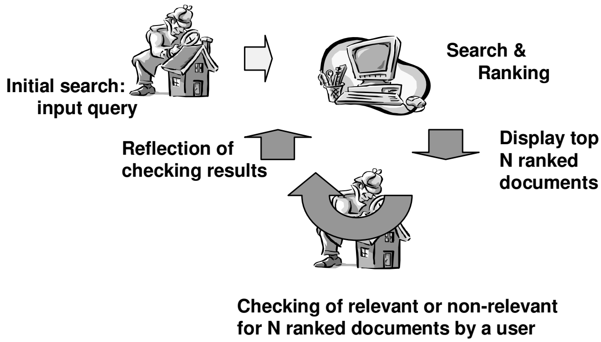


Fig. 1. Relevance Feedback

retrieval systems. In this method, a user directly evaluates whether a document in a list of retrieved documents is relevant or non-relevant, and the system modifies the query vector on the basis of the user's evaluation. A conventional way to modify a query vector is through a simple learning rule which reduces the difference between the query vector and the documents evaluated as relevant by a user. A conceptual diagram of relevance feedback is shown in Figure 1.

Another approach has been proposed in which classification learning treats relevant and non-relevant document vectors as positive and negative examples for a target concept [4]. Some studies have proposed that a Support Vector Machine (SVM) with excellent ability to classify examples into two classes be applied to the classification learning of relevance feedback [1].

We have proposed a relevance feedback framework with an SVM for active learning. In contrast to a conventional relevance feedback system which indicates a list of the most relevant documents to a user, our system provides a list of the most relevant documents which are difficult for the SVM to classify [5].

In relevance feedback, however, the user evaluates many documents until a relevant document is obtained. If all documents are judged by the user as non-relevant, classification learning cannot be applied to relevance feedback.

Classification learning which deals with one class has been developed recently. Accordingly, we propose a framework for relevance feedback based on such classification learning, using only information on non-relevant documents. We call the feedback method which uses only the non-relevant documents, *non-relevance feedback*. We use a one-class SVM [9] in our approach to classification learning.

In the remainder of this paper, we explain the one-class SVM algorithm in the second section, and our document retrieval method with a one-class SVM for non-relevance feedback in the third section. As discussed in the fourth section, in order to evaluate the effectiveness of our approach, we carried out experiments using a TREC data set obtained from the Los Angeles Times, and we present the experimental results. Finally, we conclude our work and discuss the remaining problems in the fifth section.

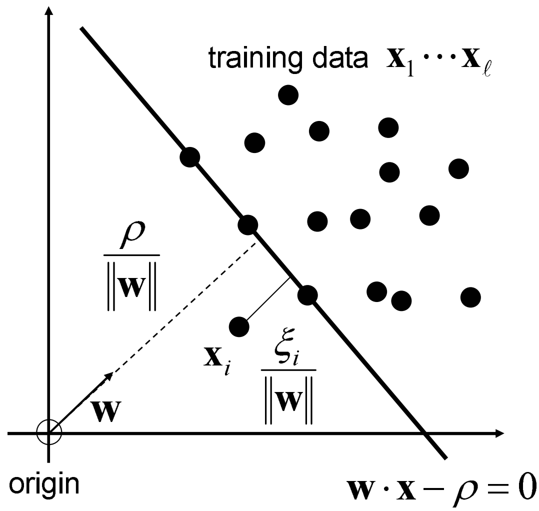


Fig. 2. One-class SVM

2 One-Class SVM

Let the training data be $\mathbf{x}_1, \ldots, \mathbf{x}_\ell$, $\mathbf{x} \in \mathbf{R}$. A one-class SVM [9] returns a function f that takes the value $+1$ in a small region that captures most of the training data points, and -1 elsewhere. This strategy is to separate the data from the origin with a maximum margin. To separate the data set from the origin, we solve the following quadratic program:

$$\begin{aligned} \min \quad & \frac{1}{2}\|\mathbf{w}\|^2 + \frac{1}{\nu\ell}\sum_i \xi_i - \rho \\ \text{subject to} \quad & (\mathbf{w} \cdot \mathbf{x}_i) \geq \rho - \xi_i, \\ & \xi_i \geq 0. \end{aligned} \tag{1}$$

Here, $\nu \in (0, 1)$ is a parameter whose meaning is the fraction of the outliers.

A conceptual diagram of the one-class SVM is shown in Figure 2.

Since the nonzero slack variables ξ_i are penalized in the objective function, we can expect that if $\mathbf{w}$ and ρ solve this problem, then the decision function

$$f(\mathbf{x}) = \operatorname{sgn}((\mathbf{w} \cdot \mathbf{x}) - \rho) \tag{2}$$

will be positive for most examples of $\mathbf{x}_i$ contained in the training set. For a new point $\mathbf{x}$, the value $f(\mathbf{x})$ is determined by evaluating which side of the hyperplane it falls on.

Using multipliers $\alpha_i, \beta_i \geq 0$, we introduce a Lagrangian

$$L(\mathbf{w}, \boldsymbol{\xi}, \rho, \boldsymbol{\alpha}, \boldsymbol{\beta}) =$$

$$\frac{1}{2}\|\mathbf{w}\|^2 + \frac{1}{\nu\ell}\sum_i \xi_i - \rho - \sum_i \alpha_i((\mathbf{w}\cdot\mathbf{x}_i) - \rho + \xi_i) - \sum_i \beta_i\xi_i \quad (3)$$

and set the derivatives with respect to the primal variables $\mathbf{w}, \xi_i$ and ρ to be equal to zero, yielding

$$\mathbf{w} = \sum_i \alpha_i \mathbf{x}_i, \quad (4)$$

$$\alpha_i = \frac{1}{\nu\ell} - \beta_i \leq \frac{1}{\nu\ell}, \quad \sum_i \alpha_i = 1. \quad (5)$$

In (4), all patterns $\{\mathbf{x}_i : i \in [\ell], \alpha_i > 0\}$ are called support vectors. The support vector expansion transforms the decision function (2)

$$f(\mathbf{x}) = \mathrm{sgn}\left(\sum_i \alpha_i \mathbf{x}_i \cdot \mathbf{x} - \rho\right). \quad (6)$$

Substituting (4) and (5) into (3), we obtain the dual problem:

$$\min_{\boldsymbol{\alpha}} \ \frac{1}{2}\sum_{i,j} \alpha_i\alpha_j \mathbf{x}_i \cdot \mathbf{x}_j \quad (7)$$

$$\text{subject to } 0 \leq \alpha_i \leq \frac{1}{\nu\ell}, \quad \sum_i \alpha_i = 1. \quad (8)$$

One can show that at the optimum condition, the two inequality constraints (1) become equalities if α_i and β_i are nonzero, namely, if $0 < \alpha \leq 1/(\nu\ell)$. Therefore, we can recover ρ by exploiting the fact that for any such α_i, the corresponding pattern $\mathbf{x}_i$ satisfies

$$\rho = (\mathbf{w}\cdot\mathbf{x}_i) = \sum_j \alpha_j \mathbf{x}_j \cdot \mathbf{x}_i. \quad (9)$$

3 Non-relevance Feedback

In this section, we describe a method of document retrieval which uses a one-class SVM for non-relevance feedback.

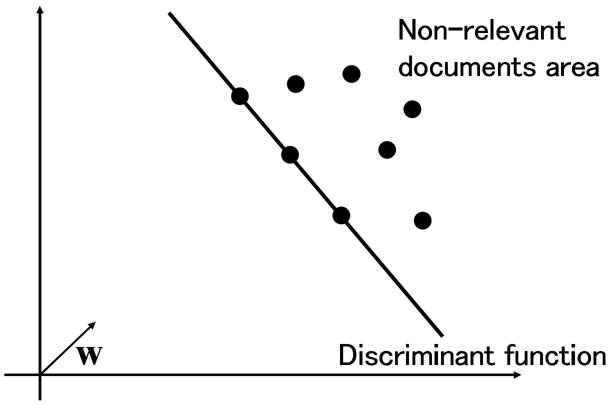


Fig. 3. Determination of non-relevant documents area: Circles denote documents which are checked non-relevant by a user. Solid line denotes the discriminant function.

In relevance feedback, a user has the option of labeling some of the top ranked documents according to whether they are relevant or non-relevant. The labeled documents, along with the original request, are then input to a supervised learning procedure to produce a new classifier. The new classifier is used to produce a new ranking, which retrieves more relevant documents at higher ranks than the original ranking. Non-relevance feedback is used when a user classifies all of the initial top ranked documents as non-relevant.

The relevance feedback based on an SVM assumes both relevant and non-relevant documents which a user has judged. Namely, an SVM, which is a binary classifier, needs both relevant and non-relevant documents. The feedback including only non-relevant documents is not applicable for a two-class SVM classifier.

Non-relevant documents, however, are obtained more easily than relevant documents. In the early stage of relevance feedback, the documents which are retrieved by the system are frequently non-relevant. Using the information on the non-relevant documents may improve the efficiency of document retrieval. In this paper, We propose an efficient retrieval method which uses information on the non-relevant documents only by applying a one-class SVM.

As mentioned in section 2, a one-class SVM clarifies the area of a given class. The area of non-relevant documents in the multidimensional vector space is clarified by a one-class SVM. Therefore, if documents which do not belong to the area of the non-relevant documents are presented, there is a high possibility that a user will judge these documents to be relevant.

The retrieval steps of the proposed method are performed as follows:

Step 1: Preparation of documents for the first feedback

The conventional information retrieval system based on a vector space model displays the top N ranked documents along with a request query to a user. In our method, for the first feedback iteration, the top N ranked documents

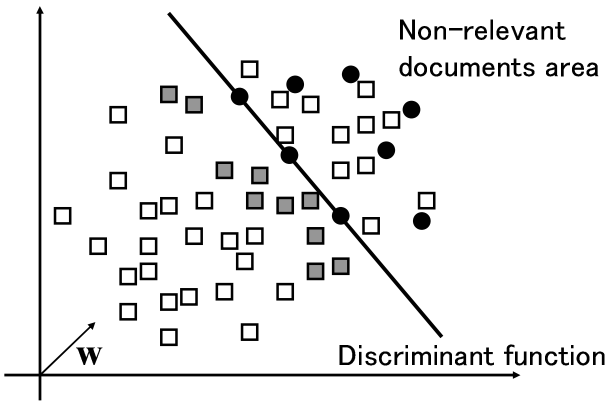


Fig. 4. Mapped non-checked documents in the feature space: Boxes denote non-checked documents which are mapped into the feature space. We show the documents which are represented by gray boxes to a user for the next iteration. These are top N ranked documents classified as "not non-relevant" and close to the boundary defined by the discriminant function.

are selected by using the cosine distance between the request query vector and each document vector.

Step 2: Judgment of documents
The user then classifies these N documents as relevant or non-relevant. In cases that the user labels all documents non-relevant, non-relevance feedback is used. Then go to the next step. If these documents are labeled both relevant and non-relevant, then skip to **Step 5**.

Step 3: Determination of non-relevant documents area
The non-relevant documents area is determined by using a one-class SVM which is learned by non-relevant documents only. (see Figure 3).

Step 4: Discrimination of all documents and information retrieval
The one-class SVM learned in the previous step classifies all documents. The documents which are discriminated as being in the "not non-relevant" are newly selected. From the newly selected documents, the top N ranked documents, which are ranked in order of their distance from the boundary defined by the discriminant function, are presented to the user as the information retrieval results of the system (see Figure 4). Then return to **Step 2**.

Step 5: Shift to Relevance feedback
If both relevant and non-relevant documents are obtained, ordinary relevance feedback is applied.

The non-relevance feedback is intended to present the relevant documents quickly. As mentioned in **Step 4**, the selected documents are discriminated "not non-relevant" and are close to the discriminant function. The reason is that we consider these selected documents are not non-relevant and include given queries because non-relevant documents include given queries in this case. If we select a document far from the discriminant function, the document has no relation to the given queries.

4 Experiments

4.1 Experimental Setting

We conducted experiments for evaluating the utility of our method, as reported in section 3. The document data set we used is a set of articles in the Los Angeles Times which has been widely used in the document retrieval conference, TREC [10]. The data set has approximately 130,000 articles. The average number of words in an article is 526. This data set includes not only queries but also the documents relevant to each query. We used three topics for the experiments as shown in Table 1. These topics have no relevant documents in the top 30 ranked documents for retrieval using an initial query vector.

Table 1. Topics used for experiments

topic	query words	# of relevant doc.
306	Africa, civilian, death	34
343	police, death	88
383	mental, ill, drug	55

We used TFIDF [11], which is one of the most popular methods in information retrieval, to generate the document feature vectors, and the concrete equation [8] of a weight of a term t in a document d w_t^d as shown in the following.

$$w_t^d = L \times t \times u \tag{10}$$

$$L = \frac{1 + \log(tf(t,d))}{1 + \log(\text{average of } tf(t,d) \text{ in } d)} \quad \text{(TF)}$$

$$t = \log(\frac{n+1}{df(t)}) \quad \text{(IDF)}$$

$$u = \frac{1}{0.8 + 0.2\frac{uniq(d)}{\text{average of } uniq(d)}} \quad \text{(normalization)}$$

The notations in these equation are as follows:

- w_t^d is the weight of a term t in a document d,
- $tf(t,d)$ is the frequency of a term t in a document d,
- n is the total number of documents in a data set,

- $df(t)$ is the number of documents including a term t,
- $uniq(d)$ is the number of different terms in a document d.

The sizes N of the groups of retrieved and displayed results developed in **Step 1** in section 3 were set as 10 and 20. In our experiments, we used the linear kernel for a one-class SVM learning, and found a discriminant function for the one-class SVM classifier in the original feature space. The vector space model of the documents is a high-dimensional space. Moreover, the documents which are labeled by a user are small in number. Therefore, the parameter ν (see section 2) is set to have an adequately small value ($\nu = 0.01$). The small ν means a hard margin in the SVM.

For comparison with our approach, two information retrieval methods were used. The first is an information retrieval method that does not use feedback, namely, documents are retrieved using the ranking in vector space model. The second is an information retrieval method using conventional Rocchio-based relevance feedback [6] which is widely used in information retrieval research.

The Rocchio-based relevance feedback modifies a query vector Q_i on the basis of the evaluation of a user, using the following equation.

$$Q_{i+1} = Q_i + \alpha \sum_{x \in R_r} x - \beta \sum_{x \in R_n} x, \tag{11}$$

where R_r is a set of documents which were evaluated as relevant by a user at the ith feedback, and R_n is a set of documents which were evaluated as non-relevant at the ith feedback. α and β are weights for the relevant and non-relevant documents, respectively. In this experiment, we set $\alpha = 1.0$ and $\beta = 0.5$ which are decided experimentally.

4.2 Experimental Results

In this experiment, we evaluate how many relevant documents are presented for each feedback iteration. Here, we describe the relationships of the number of feedback iterations with the number of retrieved relevant documents for the proposed method (One-class), for the retrieval using the initial query vector only (VSM) and for the Rocchio-based feedback (Rocchio). Table 2 shows that the number of presented documents is 20, and Table 3 shows that the number of presented documents is 10.

In Table 2, 1 iteration means that a user has judged the twenty documents and is shown the next twenty documents. Therefore, the user has seen forty documents at this point. In Table 3, 1 iteration means that a user has judged the ten documents and has seen twenty documents altogether.

When the proposed method is used, a user can find a relevant document by seeing forty documents for every topic in Table 2. In other words, if the user judges the twenty documents which are retrieved using the initial query vector, the user can then find a relevant document in the next set of retrieved results. When the retrieval using the initial query vector is applied, the user can find a relevant document by seeing forty documents for topic 306, hundred

Table 2. Number of retrieved relevant documents as a function of the number of iterations (number of presented documents is 20)

topic 306	# of retrieved relevant doc.		
# of iterations*	One-class	VSM	Rocchio
1 (40)	1	1	0
2 (60)	–	–	0
3 (80)	–	–	0
4 (100)	–	–	0
5 (120)	–	–	0

topic 343	# of retrieved relevant doc.		
# of iterations*	One-class	VSM	Rocchio
1 (40)	1	0	0
2 (60)	–	0	0
3 (80)	–	0	0
4 (100)	–	1	0
5 (120)	–	–	0

topic 383	# of retrieved relevant doc.		
# of iterations*	One-class	VSM	Rocchio
1 (40)	1	0	0
2 (60)	–	1	0
3 (80)	–	–	0
4 (100)	–	–	0
5 (120)	–	–	0

*: Number in parentheses is the number of presented documents at this point.

documents for topic 343 and sixty documents for topic 383. However, the user cannot find a relevant document by seeing 120 documents for every topics, when the Rocchio-based method is used. We consider that these results are caused by the feedback method in equation (11). When the relevant documents exist, a useful query vector is created on the basis of the emphasis of the terms in the relevant documents. In the case of the non-relevant documents, however, only the minus term of equation (11) changes.

In Table 3 which shows that the number of presented documents is 10, a user can find a relevant document by judging ten documents for topic 306 when the proposed method is used. This shows that the early feedback on non-relevant documents is effective in the proposed method.

5 Conclusion

In this paper, we proposed the non-relevance feedback method which uses the one-class SVM for enhancement of the information retrieval efficiency. We compared non-relevance feedback with the retrieval using the initial query vector and

Table 3. Number of retrieved relevant documents as a function of the number of iterations (number of presented documents is 10)

topic 306	# of retrieved relevant doc.		
# of iterations*	One-class	VSM	Rocchio
1 (20)	1	0	0
2 (30)	–	0	0
3 (40)	–	1	0
4 (50)	–	–	0
5 (60)	–	–	0

topic 343	# of retrieved relevant doc.		
# of iterations*	One-class	VSM	Rocchio
1 (20)	0	0	0
2 (30)	1	0	0
3 (40)	–	0	0
4 (50)	–	0	0
5 (60)	–	0	0

topic 383	# of retrieved relevant doc.		
# of iterations*	One-class	VSM	Rocchio
1 (20)	0	0	0
2 (30)	1	0	0
3 (40)	–	0	0
4 (50)	–	1	0
5 (60)	–	–	0

*: Number in parentheses is the number of presented documents at this point.

the Rocchio-based feedback. Results of the experiment on a set of articles in the Los Angeles Times showed that the proposed method gave a better performance than the method it was compared with.

In the task of retrieving information at a user's request from large volumes of data, the information which is obtained at an early stage is often what the user does not want. Our future work will focus on the efficient use of such negative information, in various practical problems.

References

1. H. Drucker, B. Shahrary, and D. C. Gibbon, "Relevance Feedback using Support Vector Machines," in *Proceedings of the Eighteenth International Conference on Machine Learning*, pp. 122–129, 2001.
2. IREX Web page http://nlp.cs.nyu.edu/irex/index-e.html.
3. NTCIR Web page http://research.nii.ac.jp/ntcir/index-en.html
4. M. Okabe and S. Yamada, "Interactive Document Retrieval with Relational Learning," in *Proceedings of the 16th ACM Symposium on Applied Computing*, pp. 27–31, 2001.

5. T. Onoda, H. Murata and S.Yamada, "Interactive Document Retrieval with Active Learning," in *Proceedings of International Workshop on Active Mining*, pp. 126–131, 2002.
6. G. Salton, ed., *Relevance Feedback in Information Retrieval*, pp. 313–323. Englewood Cliffs, N.J.: Prentice Hall, 1971.
7. G. Salton and J. McGill, *Introduction to Modern Information Retrieval.* McGraw-Hill, 1983.
8. R. Schapire, Y. Singer, and A. Singhal, "Boosting and Rocchio Applied to Text Filtering," in *Proceedings of the Twenty-First Annual International ACM SIGIR*, pp. 215–223, 1998.
9. B. Schölkopf, J. Platt, J. Shawe-Taylor, A. Smola and R. Williamson, "Estimating the Support of a High-dimensional Distribution," TR 87, Microsoft Research, 1999.
10. TREC Web page http://trec.nist.gov/.
11. R. B. Yates and B. R. Neto, *Modern Information Retrieval.* Addison Wesley, 1999.

Tagging for Intelligent Processing of Design Information

Hideaki Takeda[1], Yutaka Fujimoto[2], Masaharu Yoshioka[3], Yoshiki Shimomura[2], Kengo Morimoto[2], and Wataru Oniki[2]

[1] National Institute of Informatics, 2-1-2 Hitotsubashi, Chiyoda-ku, Tokyo 101-8430, Japan
takeda@nii.ac.jp
[2] Research into Artifacts, Center for Engineering (RACE), The University of Tokyo, Komaba 4-6-1, Meguro-ku, Tokyo 153-8904, Japan
{yutaka_f, simomura}@race.u-tokyo.ac.jp
[3] Graduate School of Information Science and Technology, Hokkaido University
Kita 14 Nishi 9, Kita-ku, Sapporo, Hokkaido, 060-0814, Japan
yoshioka@ist.hokudai.ac.jp

Abstract. This paper describes how to add tags to design documents in order to extract knowledge from information for intelligent design support. Our project called Universal Abduction Studio (UAS) aims to build a new design support system that supports conceptual design by dynamically integrating knowledge in different design domains. This paper focuses on knowledge description form which can be used to capture knowledge from text-based information and then be used for inference for creative design. We propose so-called *design knowledge document* containing both human-readable texts and machine-readable knowledge such as propositions and rules.

1 Introduction

Design support systems have been well developed for geometric and detail design stages. In contrast, those in the conceptual design stage are still far from success. In our opinion, the main difficulty comes from incomplete and insufficient understanding about design knowledge and its operations that play a crucial role in conceptual design. Recently, thanks to development of the Internet technologies, more and more knowledge is accumulated and available electronically. It then becomes an interesting research question how to apply such an enormous amount of diverse knowledge to conceptual design.

Our project called Universal Abduction Studio (UAS)[1] aims to solve this problem in a unique way. The principle of UAS is abduction that leads design processes by integrating knowledge from various domains. In order to apply abduction, we need knowledge in logical or at least semi-logical form. How to describe such knowledge is the main topic of this paper.

In this paper, we first introduce UAS project briefly in Section 2. Then we discuss knowledge representation suitable for our purpose in Section 3. After we investigated knowledge in a book in Section 4, we propose our knowledge representation format in XML in Section 5. We conclude the paper in Section 6.

A. Sakurai et al. (Eds.): JSAI 2003/2004, LNAI 3609, pp. 216 – 225, 2007.

2 Universal Abduction Studio (UAS)

In this section, we explain ideas behind UAS project and basic concepts in it briefly. More detail descriptions are found in [2].

2.1 Abduction in Design

The key issue to build a CAD system capable of supporting the early stages of design, in particular, its creativity aspects, is how to represent design processes. It is no doubt that a creative design process is one of the most intellectual thought processes and is difficult to model. This is not only because generating a creative product or idea itself is hardly inimitable by computers, but also because the knowledge used for creative design is generated, modified, and updated during the process. Once a designer achieves a new creative design after struggles, s/he is able to perform similar designs easily, which implies that her/his knowledge was expanded. We believe that the expansion of knowledge is a mandatory nature of creative design. Creative design therefore has two aspects, i.e., creating a new product and expanding knowledge, and the co-relation between these two is common to various creative activities.

How can we, then, model creative design processes with these two aspects? We have discussed formalization of design processes from the logical viewpoint [3][4][5]. In short, our theory models design as iteration of deduction and abduction (see also Coyne[6], Roozenburg and Eekels[7], while they did not offer any computing mechanisms). In our theory, the core part of design, i.e., creating a new idea or thing can be attributed to "abduction" while ensuring design to deduction. Abduction is thus the crucial part in design.

What is abduction and what can abduction offer as reasoning? Abduction proposed by C.S. Peirce is a logical process to find axiom from theorem [8]. The naïve interpretation is that abduction is an opposite process of deduction. Although this naïve interpretation is somewhat popular within computer science [9], abduction should be interpreted from wider viewpoints and therefore include more various types of reasoning. Schurz [10] collected various types of abductive reasoning and categorized them. In his work, abduction is firstly classified into three, i.e., factual abduction (first-order existential abduction), law abduction, and second-order existential abduction. However, since law abduction seems a sub-category of second-order existential abduction, we regard that distinction of factual and law abduction is the primary classification of abduction. The former concerns discovery of facts and the latter concerns discovery of new laws. "Abduction as inversed deduction" is merely one category of factual abduction.

As we mentioned above, the point of this paper is the dynamics of knowledge when formalizing design processes. Factual abduction infers new facts from given facts (observable facts) with fixed theory (rules). However, as long as we use reasoning with a fixed theory, the ability to create new facts is limited. In addition, although factual abduction can satisfy the primary requirement of abduction ("finding axioms from a theory"), this interpretation does not qualify another important feature of abduction mentioned by Peirce. He explained that abduction can find a "surprising" fact. "Inversed deduction" is insufficient to realize such a process and abduction is

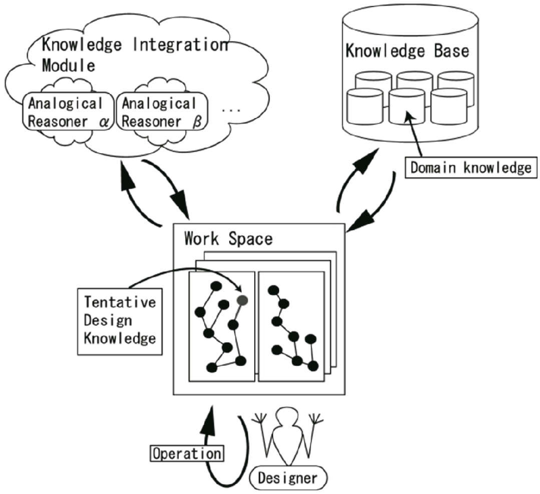


Fig. 1. Fundamental concept of the Universal Abduction Studio

necessarily accompanied by expansion or revision of knowledge [11]. In the domain such as design in which rich knowledge is available, a feasible expansion of knowledge is obtained by integrating existing knowledge [12]. Integration of knowledge here does not mean a simple addition of knowledge, but rather such operations as translation and modification. There seem to exist a number of possible ways to integrate knowledge. Abduction as a method to integrate knowledge can satisfy the two aspects of creative design, i.e., creating a new product and expanding knowledge. Therefore, we believe that abduction can be one model of creative design processes.

2.2 The Architecture of Universal Abduction Studio Systems

A Universal Abduction Studio (UAS) system is a computer environment to support integration of theories (that contain knowledge) from various knowledge domains for creative design. UAS is not a design automation system but a cooperation system that can solve design problems by helping dynamic interaction between a designer and the system. UAS provides a toolbox consisting of a variety of domain knowledge as well as a variety of abductive reasoning mechanisms for knowledge integration. When the designer cannot solve a design problem with knowledge of one domain, the designer chooses a knowledge operation to make correspondences between that domain knowledge and another domain knowledge that the UAS system proposes. Then, the designer estimates and judges whether or not the proposed knowledge should be used. Finally, the designer generates design solutions based on the tentative design knowledge chosen by her / him. The basic feature of the system as an inference system is abduction that can integrate knowledge to proceed design processes.

Figure 1 shows the fundamental concept of UAS. In Figure 1, the designer operates design information and knowledge on the workspace. The knowledge integration module consists of multiple abductive reasoning mechanisms, and the designer chooses one or some of them depending on each design problem. The knowledge base consists of multiple domain knowledge bases and the designer first chooses one to solve a design problem. When the designer cannot solve the design problem, the system reasons about another domain knowledge base that can possibly be integrated with the first domain knowledge. The abductive reasoning system then performs knowledge integration. This fundamental concept requires unified knowledge description among various domain knowledge bases.

3 Knowledge Representation in Design

In this section, we overview knowledge representation in design, and show our basic approach for it.

Knowledge in design is mainly classified into two categories, i.e., knowledge on objects and knowledge on design processes or design procedures [13]. The former is knowledge on how objects are represented and operated, and the latter is knowledge on how designers proceed and complete design.

Many studies focus on object modeling. Typical examples of object modeling are 2D/3D geometric modeling and kinematic modeling. Each object modeling method provides a way of representation based on its aspect. Since any design requires two or more aspects to complete, we should manage multiple object modeling methods so that ontology should be introduced.

Ontology in information systems is introduced in knowledge sharing context. The popular definition of ontology is "an explicit specification of conceptualization"[14]. It provides basic concepts when one wants to represent the target world in some specific context. Each modeling method assumes some basic concepts that are introduced by the theory that the modeling is based on. These concepts can be components of ontology. Some of these concepts are sharable with other modeling methods, and the others are not. Providing an ontology that consists of such sharable concepts helps managing multiple modeling methods. More concrete discussions on ontologies for engineering design are found in [15].

On the other hand, knowledge on design processes has not been investigated well. In object representation, we can assume some background theory that the object representation is based on. Then what kind of theory can we assume as background theory of design process modeling? We have proposed a logical framework for design processes and shown abduction can be the principle for design process [3]. In this framework, abduction corresponds to the process when a new design candidate is created, while deduction corresponds to the process when it is analyzed and validated.

Abduction for design should not closed in a single domain or modeling but should include knowledge from various domains and modeling methods.

Suppose that we are designing "knife which is always sharp", while the following information is provided in the ontology shown in Figure 2. Concepts from different domains and modeling methods are connected in this ontology. Glass and knife may be included in an engineering domain knowledge, while chocolate in a cooking domain knowledge. A possible scenario to design it is as follows. First we make an assumption like "if knife is cut, it is sharp" using the knowledge "if glass is cut, it is sharp" and similarity between glass and knife which is come from relations in the ontology. In like manner, we make an assumption like “if knife is grooved, it is easy to cut” from the knowledge “if chocolate is grooved, it is easy to cut” and similarity between knife and chocolate. The solution is then “grooved knife”. In this example, two fragments of knowledge “if glass is cut, glass is sharp” and “if chocolate is grooved, it is easy to cut” are jointly used in abduction by relating concepts in different domains and modeling methods with the ontology. It should be noted that each modeling method is based on the specific theory so that ontology covered completely can not expected. While ontology provides universally valid relationship among different modeling methods and domains, abduction is expected to support teleological relationship among them as assumptions [12].

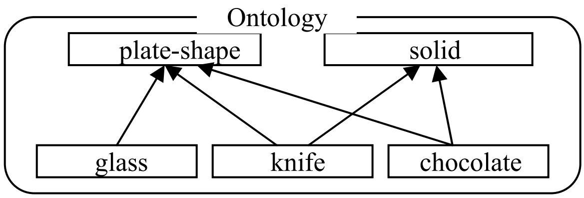


Fig. 2. A sample of ontology for a scenario

4 A Case Study for Knowledge Representation for UAS

As we discussed in the previous section, we expect knowledge for UAS systems as logical or at least semi-logical form because abduction we suppose requires such forms. On the hand, most of knowledge in real design activities is represented in text and figure. In order to know how and what knowledge can be captured from such information sources, we picked up a book on know-how of mechanical design [16] and tried to extract knowledge. Then we analyzed the extracted knowledge.

We extracted pieces of texts that describe information on design processes as candidates of knowledge from the book. The number of pieces extracted is 350. Then we transformed these pieces of texts into if-then rules.

This transformation is not simple. We set a rough criterion to separate if-part and then-part. If-part represents some observation, and then-part represents some action. Even under this criterion, multiple interpretations are observed. For example, “cut glass becomes sharp” can be either interpreted as “if there is glass, cutting it makes it sharp” or “if glass is cut, the glass becomes sharp”. The latter may seem more natural interpretation but it depends on situation. For example, when looking for information of glass as candidate of material, the former rule may be useful.

The other issue is categorization. We investigated the collected rules closely and classified into three. The first category is a collection of rules that have objects as if-part. The rules are furthermore categorized into six sub-categories depending on then-part. Each category includes either “should”, “should not”, “is (are)”, is (are) not”, “there are merits that”, and “there are demerits that”. The second category is those of which if-then have operations to objects. This category is also divided into six sub-categories depending on then-part. Each includes either “it is a good design”, “it is a bad design”, “it needs care”, “it has merits of …”, “it has demerits of …” and “it needs care for …". The third category is those of which include state or situation of object. It is also categorized into two sub-categories. One includes "should" in then-part and the other includes "should not".

We can find some remarks on tagging for design knowledge through this case study. The first one is that texts have naturally multiple interpretations. Objects are easily identifiable but fragments of knowledge like rules are not. We need the different levels of flexibility for annotation. The second is variety of knowledge. Even though we restricted our investigation to rule-style knowledge, meanings of rules are various. The variety probably comes from situations or contexts when we want to use such knowledge. We listed fourteen categories but they are taken from a single book and we should investigate such categories more systematically.

In the next section, we discuss format of tagging for knowledge. We mainly concern the first point in this paper and let the second left for future research.

5 Design Knowledge Document and Its Format

Knowledge for design, in particular, knowledge for design processes is often included in documents written in natural languages. Forming knowledge bases by extracting such knowledge from documents is a possible approach but it requires cost for acquisition and maintenance of knowledge. The latter is especially serious because this approach hardly enables to track changes of documents.

```
<knowledge>
  <head>
    <!-- Properties of this documents are described. -->
  </head>
  <document>
    <!-- Texts with annotations are described. -->
  </document>
  <model>
    <propositions>
      <!-- Facts corresponding annotations are declared. -->
    </propositions>
    <rules>
      <!-- Relations among facts are declared. -->
    </rules>
  </model>
</knowledge>
```

Fig. 3. The abstract structure of design knowledge document

The approach in this paper is knowledge as annotation to texts [17]. We call *design knowledge document* that contains texts and knowledgeable annotations to them. The former is just for human and the latter is mainly for computers but still understandable for human. The benefits of this approach are twofold. One is readability of knowledge. One can easily understand meaning of knowledge since texts can be used as comments for knowledge. This leads to productivity and ease of maintenance of knowledge. We can produce knowledge from existing documents and update knowledge when the corresponding documents are changed. The other is possibility of automatic extraction of knowledge. Because design knowledge documents can be seen as instances of mapping between texts and knowledge, they can be sources to learn this mapping function. We focus on the first point in this paper and let the second point left as a future task.

Figure 3 shows the abstract structure of design knowledge document. The overall structure is formatted as XML. It consists of three major parts, i.e., “head” part, “document” part and “model” part. “Head” part describes general properties of the document. “Document” part describes natural language texts with the specific annotation. Without the annotation, they are just texts in documents. “Model” part describes knowledge related to these texts.

We provide “<word>” tag for text annotation. This tag relates the specific part of texts to concepts in knowledge. The form is

```
<word base="fundamental-form" concept="concept-name"
id="ID"> string</word>
```

String is related to concept *concept-name*, word *fundamental-form*, and *ID*. *Concept-name* is associated with concept of this name. Concept is declared either in model part of the same document or in some ontology. *Fundamental-form* is provided just for natural language processing and search. *ID* is used in model part to refer the specific occurrence of the concept. For example,

```
<word concept=knife id=knife1>this knife</word>
```

declares that there is an occurrence of concept knife named knife1.

Model part consists of two parts, i.e., proposition part and rule part. In proposition part, facts that are believed true in the document are listed. For example,

```
<proposition id="p1">
<predicate concept="cut"/>
<arg idref="knife1"/>
</proposition>
```

<proposition> tag declares a proposition and should include a single <predicate> tag and one or more <arg> tags. Attribute *concept* for <predicate> and <arg> tags is used to specify the corresponding concept, while attribute *idref* is used to occurrence of the concept in the document. The example declares

```
cut(knife1)
```

where knife1 is the occurrence of knife in the document..

```
   <?xml version="1.0" encoding="UTF-8" ?>
   <!DOCTYPE knowledge (View Source for full
doctype...)>
- <knowledge
xmlns:dc="http://purl.org/dc/elements/1.1/"
xmlns:fc="http://www.race.u-tokyo.ac.jp/fc/"
xmlns:rl="http://www.race.u-tokyo.ac.jp/rl/">
- <head>
   <knowledgeid>003</knowledgeid>
   <title>Knowledge on machining setup</title>
   <creator />
   <date>2004-08-13</date>
   <version>1.0</version>
- <origin>
   <dc:title>Hits for mechanical design,
      A second series</dc:title>
   <dc:creator>Hidenori Watanabe</dc:creator>
   <dc:publisher>Nikkan Kogo Shinbunsha
      </dc:publisher>
   <dc:date>1990-08-29</dc:date>
   </origin>
   </head>
- <document>
<word base="fewer"
concept="predicate:fewer">Fewer</word>
<word base="number" concept="upper:number">
numbers</word> of
<word base="setup" concept="upper:setup"
id="setup">
setup</word> and
<word base="change" concept="predicate:change"
id="change">
changes</word> of
<word base="position"
concept="upper:machiing-position"</word>
 id="position">position</word> and
<word base="attitude" concept="upper:attitude"
id="attitude">
attitude</word> in
<word base="machining" concept="upper:machining"
 id="machining">machining</word>
<word base="cause" concept="predicate:cause"
id="cause">
cause</word>
<word base="decrease" concept="predicate:decrease"
 id="decrease1">decrease</word> of
<word base="man-hour"
   concept="upper:time,engineering:man-hour"
id="man-hour">
man-hour</word>
<word base="cost-reduction" concept="upper:cost,
engineering:cost-reduction" id="cost-reduction">
cost reduction</word>
</document>
```

```
- <model>
- <propositions>
- <proposition id="p1">
   <predicate
concept="upper:machining" />
   <arg idref="machining" />
   </proposition>
- <proposition id="p2">
   <predicate
concept="upper:setup" />
   <arg idref="setup" />
   </proposition>
- <proposition id="p3">
   <predicate
concept="predicate:own" />
   <arg idref="machining" />
   <arg idref="setup" />
   </proposition>
...(several lines are omitted)
   </propositions>
- <rl:rules>
- <rl:rule>
- <rl:if>
- <rl:and>
   <rl:atom propositionid="p1" />
   <rl:atom propositionid="p2" />
   <rl:atom propositionid="p3" />
- <rl:not>
   <rl:atom propositionid="p4" />
   </rl:not>
   <rl:atom propositionid="p5" />
- <rl:not>
   <rl:atom propositionid="p6" />
   </rl:not>
   <rl:atom propositionid="p7" />
- <rl:not>
   <rl:atom propositionid="p8" />
   </rl:not>
   </rl:and>
   </rl:if>
- <rl:then>
- <rl:and>
   <rl:atom propositionid="p9" />
   <rl:atom propositionid="p10" />
   <rl:atom propositionid="p11" />
   <rl:atom propositionid="p12" />
   <rl:atom propositionid="p13" />
   <rl:atom propositionid="p14" />
   </rl:and>
   </rl:then>
   </rl:rule>
...(several lines are omitted)
   </rules>
   </model>
   </knowledge>
```

Fig. 4. An example of design knowledge document (translated from descriptions in Japanese)

Rule part is used to declare rule-style knowledge. The example is as follows;

```
<rule>
 <if>
  <atom propositionid="p1"/>
 </if>
 <then>
  <proposition>
     <predicate concept=”sharp”/>
     <arg refid=”knife1”>
   </proposition>
 </then>
</rule>
```

This description is a simple if-then rule. <atom> tag is used to specify a proposition declared in proposition part with propositionid attribute. This example is then declaration of the following rule.

```
If cut(knife1) then sharp(knife1)
```

Figure 4 shows an example of description with this syntax.

By providing design knowledge documents, human can understand and use these documents as usual on one hand, the system can extract their knowledge part and use them in its inference. An example of inference is shown in [1].

6 Summary

In this paper, we discussed knowledge representation in design with a case study and proposed so-called design knowledge document that has an XML-based format including both human-readable texts and computer-understandable knowledge.

There are many issues to be done. The current format for design knowledge document is tentative to need more discussion. Especially Semantic Web approach [18] rather than XML is more suitable for our purpose. We are going to re-write our format to RDF/RDFS/OWL and rule MLs like SWRL[1].

References

[1] Takeda, H., Sakai, H., Nomaguchi, Y., Yoshioka, M., Shimomura, Y., Tomiyama, T.: Universal abduction studio – proposal of a design support envrionment for creative thinking in design –. In Folkman, A., Gralen, K., Norell, M., Sellgren, U., eds.: The Fourteenth International Conference on Engineering Design (ICED 03), Stockholm (2003)

[2] Takeda, H., Yoshioka, M., Tomiyama, T.: A general framework for modeling of synthesis – integration of theories of synthesis –. In: 13th International Conference on Engineering Design – ICED 01, Design Research – Theories, Methodologies, and Product Modelling, Glasgow (2001) 307–314

[1] http://www.daml.org/2003/11/swrl/

[3] Takeda, H., Tomiyama, T., Yoshikawa, H.: A Logical and Computerable framework for reasoning in design, in D. Taylor and L. Stauffer eds., Design Theory and Methodology -- DTM '92 --, pp. 167–174, The American Society of Mechanical Engineers (ASME) (1992).

[4] Takeda, H., Veerkamp, P., Tomiyama, T., Yoshikawa, H.: Modeling design processes, AI Magazine, Vol. 11, No. 4, pp. 37–48 (1990).

[5] Hayashi, K., Takeda, H., Tomiyama, T., Yoshikawa, H.: Analysis and logical formalization of design processes (the third report) -- modeling with circumscription and abduction --, The proceedings of the annual conference of the Japanese Society for Precision Engineering, pp.7-8 (1989) (In Japanese).

[6] Coyne, R.: Logic Models of Design. Pitman Publishing, London (1988)

[7] Roozenburg, N., Eekels, J.: Product Design: Fundamentals and Methods. John Wiley & Sons, Chichester, MA. (1995)

[8] Peirce, C.: Collected Papers of Charles Sanders Peirce. Volume 5. Harvard University Press, Cambridge, MA (1935)

[9] Flach, P., Kakas, A., eds.: Abductive and Inductive Reasoning: Essays on their Relation and Integration. Applied Logic Series. Kluwer Academic Press (2000)

[10] Schurz, G.: Models of abductive reasoning. In Schurz, G., Werning, M., eds.: TPD Preprints Annual. Number 1 in Philosophical Prepublication Series of the Chair of Theoretical Philosophy. The University of Dusseldorf (2002)

[11] Aliseda, A.: Abduction as epistemic change: A peircean model in artificial intelligence. In Flach, P., Kakas, A., eds.: Abductive and Inductive Reasoning: Essays on their Relation and Integration. Applied Logic Series. Kluwer Academic Press (2000)

[12] Takeda, H.: Abduction for design. In Gero, J., Sudweeks, F., eds.: Proceedings of the IFIP WG5.2 International Workshop on Formal Design Method for CAD, Tallinn, Elsevier Science Publishers B.V. (1993)

[13] Tomiyama, T.: From general design theory to knowledge-intensive engineering. Artificial Intelligence for Engineering Design, Analysis and Manufacturing (AIEDAM) 8 (1994) 319–333

[14] Gruber, T.R.: Toward principles for the design of ontologies used for knowledge sharing. Technical Report KSL 93-4, Knowledge Systems Laboratory, Stanford University (1993)

[15] Yoshioka, M., Umeda, Y., Takeda, H., Shimomura, Y., Nomaguchi, Y., Tomiyama, T.: Physical concept ontology for the knowledge intensive engineering framework. (Advanced Engineering Informatics) (Accepted for publication).

[16] Watanabe, H.: Hits for mechanical design, A second series. Nikkan Kogyo Shinbun (1998) (In Japanese).

[17] Yoshioka, M. and Shamoto, Y.: Knowledge Management System for Problem Solving -- Integration of Document Information and Formalized Knowledge --. Proceedings of the 2003 ASME Design Engineering Technical Conference \& Computers and Information in Engineering Conference, The American Society of Mechanical Engineers (ASME), New York, DETC2003/CIE-48217 (CD-ROM), (2003)

[18] Berners-Lee, T., Hendler, J., Lassila, O.: The semantic web. Scientific American (2001)

Application and Analysis of Interpersonal Networks for a Community Support System

Masahiro Hamasaki[1,2], Hideaki Takeda[2,1], Ikki Ohmukai[1,2], and Ryutaro Ichise[2,1]

[1] Soken University, Hayama-cho Miura-gun Kanagawa 240-3630, Japan
hamasaki@grad.nii.ac.jp
http://www-kasm.nii.ac.jp/~hamasaki/
[2] National Institute of Informatics, 2-1-2 Chiyoda-ku Tokyo 101-8430, Japan

Abstract. In this paper, we discuss importance and usefulness of interpersonal network in a community support system. We built a scheduling support system for an academic conference. Our system supports information exchange among participants and information discovery with generating participants' interpersonal network. This system was used in an academic conference called JSAI2003 involving 276 active users. The analysis of the networks reveals that interpersonal networks can promote information exchange among people by indicating existence of people to the others, and that it can also support information discovery by recommendation.

1 Introduction

In this paper, we discuss importance and usefulness of interpersonal network in a community support system based on an analysis of the scheduling support system employed in an academic conference. The system is designed to support not only personal activities in conferences such as scheduling but also communication among participants.

It is important for participants in academic conferences to know which paper or presentation is interesting for them, what kinds of people participates, and which participants share similar concerns. However, it is difficult for them to find such information among a large amount of information on papers and participants provided as conference program because the number of papers is too large to glance at in a short conference period and information on participants is usually very limited or missing. We built a system to support them to find such information in an academic conference as a cooperative scheduling system. We adopt a "person as content" strategy for system architecture. It means that a person is an information source; We treat a person as an information node that is accessible from other users. Such nodes are connected through an interpersonal network. The interpersonal network is built from acquaintanceship of individual participants, and used as a new route to traverse information on participants and papers in this system. Users can reach their interesting papers and participants by tracing links in the interpersonal network. Furthermore the system

A. Sakurai et al. (Eds.): JSAI 2003/2004, LNAI 3609, pp. 226–236, 2007.

helps them by recommending papers and participants as the result of analysis of the interpersonal network. We then analyzed the interpersonal network itself and performance of recommendation based on it.

We organize this paper as follows: In Section 2, we briefly introduce the system that we built for conference support. In Section 3, we explain the result of analysis based on actual data taken in 2003 JSAI conference. In Section 4, we compare our work with other studies and conclude the paper in Section 5.

2 Community-Support System Using Interpersonal Network

We built a collaborative scheduler called *community navigator* and applied it to an academic conference. The primary function of the system is to provide online scheduler during the conference, i.e., users browse the conference program and put papers in which they are interested in their schedule. In order to facilitate this activity, we use the interpersonal network built from

2.1 System Outline

The proposed system is based on an online program of an academic conference. The system provides users with a personal timetable and a portal of this system. Figure 1 shows overview of the system. It generates HTML pages using database and stores data which users input.

The system dynamically generates the following four types of HTML pages: author, paper, session, and personal timetable. Each HTML page is linked mutually based on the relationship stored in the database; a user can browse the generated HTML pages freely and even add a new relationship. The personal timetable is changed based on the relationships added by users.

This system roughly addresses two types of data: resources and links. Resources include three categories: session, presented paper, and person (author, chairperson, and other user). A link shows the relevance from one resource

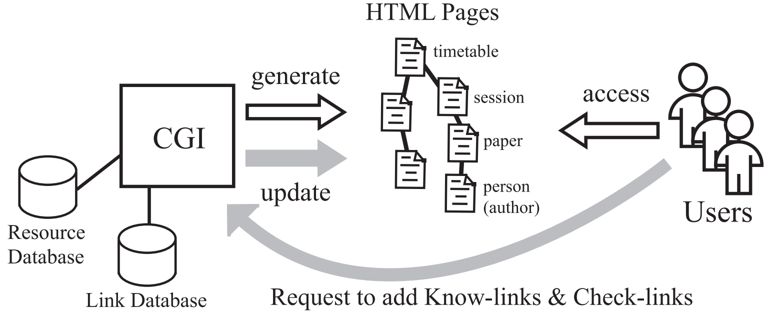


Fig. 1. System Overview

to another. There are five categories on links: *Contain* (session-paper), *Author* (author-paper), *Chair* (chairperson-session), *Check* (user-paper), and *Know* (user-person). *Contains*, *Chairs*, and *Authors* are registered in advance, while *Check* and *Know* can be added by users.

2.2 Adding Links

Creation of personal timetables and the acquaintance lists ("I-know" and "I'm-known-by" lists) are basic functions. When a user finds an interesting paper in the conference schedule or the paper list, she/he can add it to her/his timetable. Then, her/his timetable is updated. Similarly, when a user finds an acquaintance on paper or session pages, she/he can add it to her/his "I-know" list. At the same time, the acquaintance is added to the acquaintance's "I'm-known-by" list. These actions mean additions of a check-link and a know-link, respectively.

Figure 2 is an example of personal page provided for each user. The upper part of the page shows static information like personal information, i.e., her/his presentations and hyperlinks to some services. The rest shows dynamic information i.e., "I-know" and "I'm-known-by" lists.

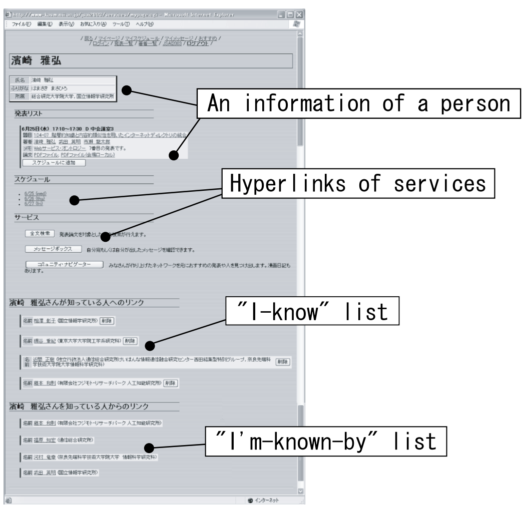


Fig. 2. My Page

An added link is a kind of private information; it is not suitable to the public unconditionally. This system conducts access control using the generated network. The detail information of a person can be accessible for persons who are registered as her/his acquaintance. The similar control is applied for detail information on who "checked" the specific paper.

2.3 Recommendation

We provide information recommendation services using links added by users, i.e., recommendation either for papers or for persons. A user can find interesting papers and persons not only by browsing but also by recommendation.

We adopted two types of recommendation method. One is a method which similar to collaborative filtering like GroupLens [1]. This method finds other persons who have similar concerns of a user and recommends their items, i.e., pap. In our case, we use check-links as information to guess participants' concern. The other is a method using interpersonal network. The method collects acquaintances of the user, and recommends items which are common among them. In this paper, the former is called recommendation using check-link and the latter is called recommendation using know-link.

We explain how to find recommending items. At first, we define three functions i.e., *Check*, *Know*, and *Relate*. If person h_0 adds check-link to paper p_1, $Check(h_0, p_1) = 1$, otherwise $Check(h_0, p_1) = 0$. If person h_0 adds know-link to person h_1, $Know(h_0, h_1) = 1$, otherwise $Know(h_0, h_1) = 0$. $Relate(h_0, h_1) = 1$ if $Know(h_0, h_1) = 1$ or $Know(h_1, h_0) = 1$, otherwise, $Relate(h_0, h_1)$ is 0.

Then we can define $V_{hc}(h_0, h_x)$ and $V_{pc}(h_0, p_x)$ which calculate degrees of recommendation in recommendation using check-link for user h_0 about person h_x and paper p_x, respectively.

$$V_{hc}(h_0, h_x) = \sum_{\{p_x | Check(h_0, p_x) = 1\}} Check(h_x, p_x)$$

$$V_{pc}(h_0, p_x) = \sum_{\{h_x | V_{hc}(h_x, h_0) > 0\}} Check(h_x, p_x)$$

Similarly, $V_{hk}(h_0, h_x)$ and $V_{pk}(h_0, p_x)$ calculate degrees of recommendation in recommendation using know-link for user h_0 about person h_x and paper p_x, respectively.

$$V_{hk}(h_0, h_x) = \sum_{\{h_k | Relate(h_0, h_k) = 1\}} Relate(h_k, h_x)$$

$$V_{pk}(h_0, p_x) = \sum_{\{h_k | Relate(h_0, h_k) = 1\}} Check(h_k, p_x)$$

In the recommendation service, each of four recommendation methods shows at most five candidates of which V values are top five and exceed the predefined threshold.

3 Experimental Results

We applied this system to the academic conference called JSAI2003 (The 17th Annual Conference of The Japanese Society for Artificial Intelligence). In this conference, there are 30 sessions, 259 presentations and 510 authors (including co-authors). About 400 participants joined the conference between Jun 23 and 27, 2003.

3.1 Trace of Users

In this section, we show how users utilize the interpersonal network by tracing of users.

We analyzed pattern of users from access logs of HTTP server [1]. Figure 3 shows major routes which users used.

We can see that routes from schedules to sessions and then those from sessions to papers are the main stream. Each route was used by users about 2000-3000 times. It is a typical access pattern of browsing online schedule of a conference.

The route from a paper page to a person page is used more than 1000 times. Furthermore, the route from a person page to a person page is used more than 1300 times.

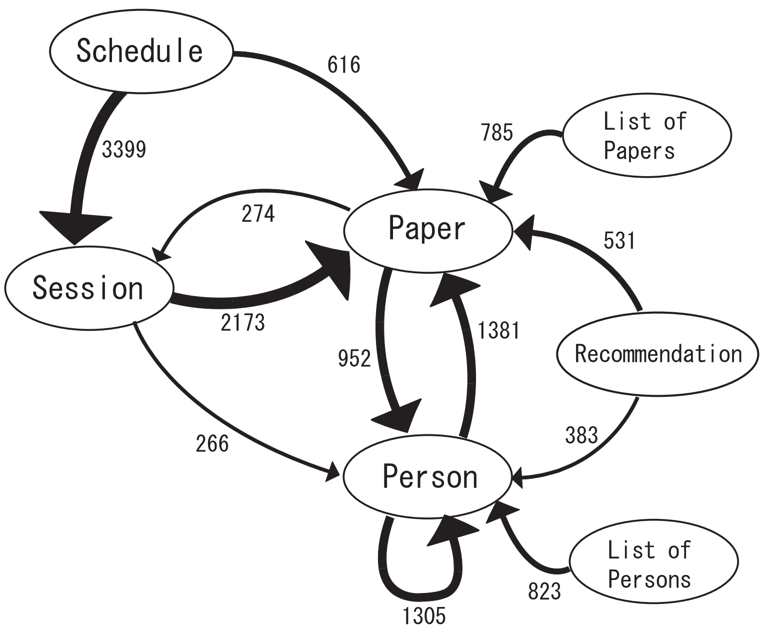


Fig. 3. Access Pattern

Users access to pages through person pages many times. The number is approximately a half of the number of typical access pattern such as routes from a session page to a paper page. And a route from a person page to a person page is also used frequently. The former result indicates that many users access to person as content. The latter result indicates that user utilizes interpersonal

[1] This method cannot trace users when they use back button of web browser and access to pages directly using www bookmark or typing URL. So the number of in-bound links and the number of out-bound links are not the same in each node (Figure 3).

network as an access route among pages. These results indicate a "person as content" strategy supports users to find information in this system.

3.2 Generated Know-Link Networks

In this section, we analyze know-link networks generated by users. In this system, a person can be a tail of a check-link and a know-link, and a head of a know-link. On the other hand, a paper can be head of a check-link. Table 1 shows the number of persons who have tails of know-links (she/he added at least a know-link) and the number of persons who have heads of know-links.

Table 1. Know-Link and Resources

	Person Who Had	Cover Rate
Out-bound Know-link	99 persons	0.18
In-bound Know-link	260 persons	0.47

The number of active users who added links is less than 30 percent of the number of the registered persons. However, 49 percent of persons are included in know-link network. It indicates that interpersonal network can be spread out well even the participants are relatively small.

The system was used by 276 users. 160 users added 1840 check-links add 99 users of them added 840 know-links. Figure 4 shows the co-author network and the know-link network. In these networks, a node is a person and an edge is a relation between persons (co-author or know-link). The two networks shared 135 edges.

The co-author network has 73 clusters, while the network merged with the know-link network has only 5 clusters. It indicates that edges made by know-links connect scattered co-author networks.

Figure 5 shows log-log plot of the cumulative distributions of incoming know-links. It shows a tendency of broad-scale networks that is characterized by a connectivity distribution that has a power law regime followed by a sharp cutoff. This tendency is different from the interpersonal network that Amaral investigated [2]. A possible interpretation of this difference is that cost of adding links. Actions in online systems seem to be less than a real world in cognitive cost.

3.3 Results of Recommendation Using Know-Link

The system provided information recommendation service using two types of recommendation method. One used check-links and the other used know-links. 135 persons used these services.

We investigate whether users accept recommended items or not. In this recommendation service, the system shows all recommended items which are chosen by recommendation methods defined in section 2.3. Users can move to a HTML page of a recommended item and add a check-link or a know-link to their list if they click

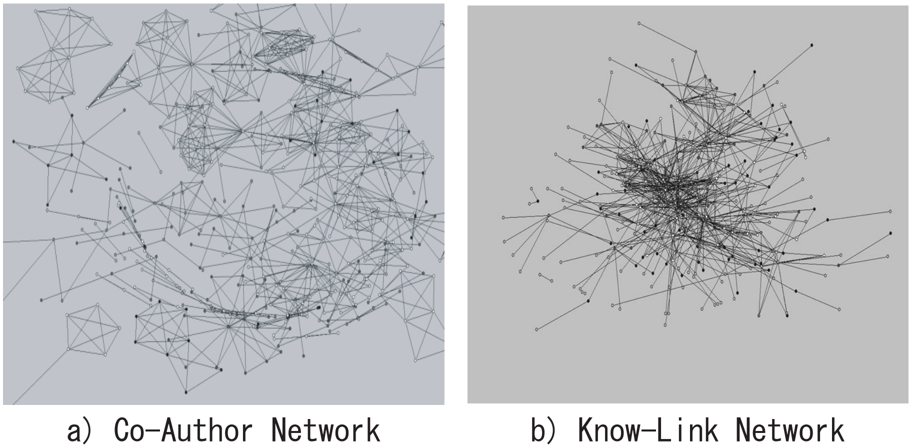


Fig. 4. Co-Author Network and Know-Link Network

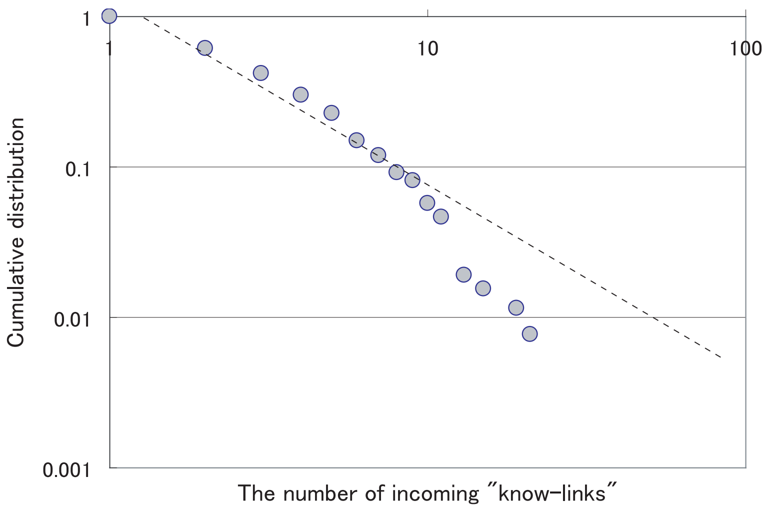


Fig. 5. The Distributions of Incoming Know-Links

"Add a Link" button on the page. To access to a recommended item indicates that user has an interest on it. It means weak accept of a recommendation. Furthermore, user has a strong interest on it if she/he adds a link to a recommended item. It means strong accept of a recommendation. These measurements correspond to "Click" and "Buy" in a paper written by Cosley [3].

Table 2 shows results of recommendation. We show the number of recommended items approximately since we could not count it precisely [2]. Users added 1800 check-links. Seventy-two of them are added by recommendation using check-link and twenty-three of them are added by recommendation using know-link.

Table 2. Weak Accept and Strong Accept in Paper Recommendation

	a) Weak Accept (Rate)	b) Strong Accept (Rate)	a/b
Using Check-Link	347(6.6 – 26.4%)	72(1.4 – 5.6%)	20.7%
Using Know-Link	210(2.1 – 16.4%)	23(0.3 – 1.2%)	10.9%

Table 3 shows averages of in-coming check-links for recommended papers. (a) is the highest and (b) is the lowest. This result indicates that the recommendation method using check-link can find items which many people are interested in, while the method using know-link can find items which less people are interested in. They may be the items which the user and her/his friends are locally interested in.

Table 3. The Difference of Recommendations by Using Check-Links

	Ave. of In-Coming Check-Links
(a) Rec. Using Check-Link	25.8
(b) Rec. Using Know-Link	9.9
(c) All	12.3

"Ave. of Checked-Links" is the average of the number of check-links which a paper has. (a) is about checked papers as a result of recommendation using check-link and (b) is about checked papers as a result of recommendation using know-link.

Cosley suggested that a recommender system is one of decision support tools, i.e., it should help users make a decision whether or not to pursue an item [3]. From this point of view, it is a good way to provide a variety of recommendation method since they can present a variety of selections to users. A recommendation using interpersonal network has a difference from a recommendation using check-link like collaborative filtering. So the recommendation method using interpersonal network (know-link) can be used a one of effective recommendation methods in a community support system.

3.4 Capability of Recommendation Using Know-Link

We analyzed capability of recommendation using interpersonal network based on these results.

[2] It is difficult to count the number of recommended items since they are generated dynamically in this system. In this case, we estimated this value by multiplying the number of access (maximum is the number of access to the recommendation service and minimum is the number of accessed users) by 5 (It is maximum number of recommended items).

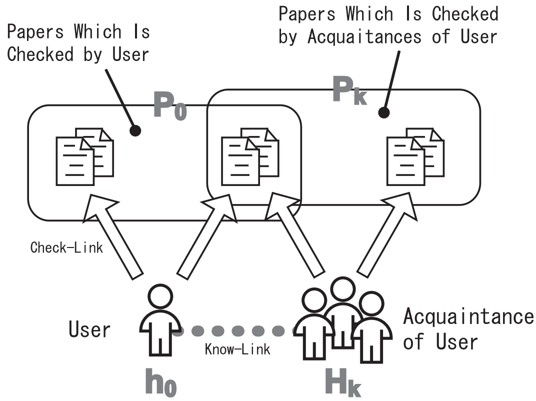


Fig. 6. Relation Among Check-Links of A User and Her/His Acquaintances

We define that persons who have edge of a know-link of user h_0 are $H_k = \{h|Relate(h_0, h) = 1\}$, papers which have a tail of a check-link added by user h_0 are $P_0 = \{p|Check(h_0, p) = 1\}$, and papers which have a tail of a check-link added by H_k are $P_k = \{p|Check(h, p) = 1, h \in H_k\}$. In this case $P_0 \cap P_k$ means papers which user h_0 and her/his aquaitance H_k checked (Figure 6).

According to the data in the experimental use, the rate of papers which user checked are checked by friends of each users is $\frac{|P_0 \bigcap P_k|}{|P_0|} = 0.67$.

The average of added check-links per user is $\overline{|P_0|} = 12.4$, and the average of check-links added by H_k is $\overline{|P_k|} = 52.8$. Incidentally, this value is not the same to $\overline{|H_k|} \times \overline{|P_0|}$ since there is overlap.

The ratio of the number of checked papers of h_0 to the number of papers which H_k also checked is $\frac{|P_0 \bigcap P_k|}{|P_k|} = \frac{|P_0 \bigcap P_k|}{|P_0|} \times \frac{|P_0|}{|P_k|} = 0.67 \times \frac{12.4}{52.8} = 0.16$. From this calculation, we can expect that a probability of strong accept of recommendation using know-link. It is about 16%.

Similarly, we expect the probability in the case of a person recommendation. In this case, P_0 corresponds to H_k and P_k corresponds to $H_{kk} = \{h|Relate(h', h) = 1, h' \in H_k\}$. The ratio of the number of known persons of h_0 to the number of persons whom H_k also know is $\frac{|H_k \bigcap H_{kk}|}{|H_k|} = 0.76$. The average of the number of know-links which user added is $\overline{|H_0|} = 8.5$ and the average of the number of persons whom H_k known is $\overline{|H_{kk}|} = 42.5$. Therefore the ratio of the number of friends of a friend to the number of friends of them is $\frac{|H_k \bigcap H_{kk}|}{|H_k|} = \frac{|H_k \bigcap H_{kk}|}{|H_k|} \times \frac{|H_k|}{|H_{kk}|} = 0.76 \times \frac{8.5}{42.5} = 0.15$. We can expect the probability of strong accept of a person recommendation using know-link is about 15%.

4 Related Work

There are some systems that support to meet and discuss among participants in an academic conference and some of them was managed in academic conferences in real. Ishida et. al. provided *ICMAS Mobile Assistant Project* in *ICMAS'96* [4]. They give a mobile computer to a participant and provide services that support to activate discussion, e.g. e-mail, bulletin board system (BBS), an announcement system about surrounding areas and a supporting system to meet among participants on that mobile computer network. Dey et. al. developed a conference participants supporting system using mobile computers and wireless tags [5]. The system generates personal timetable using positional information of participants. Sumi provided *Digital Assistant Project* in *JSAI2000* [6]. The purpose of this system is support to meet among participants in a conference. A user can get information about the conference in real-time using PDA and information kiosk.

Services using mobile computing like the above systems are one of feasible ways to support users in session. However, it requires a big cost for service providers to prepare original mobile computers and also a lot of efforts for users to use a new device. We aim a community support system which can attract attentions of participants easily using an online program. It is an appropriate way to promote information sharing using asynchronized online community support system like our proposed system because we can start information sharing before a conference, while synchronized systems with mobile computing are suitable to support communication and discussion among participants during a conference.

Jameson provided *UM2001* website which is a conference support system based on an online program [7] [8]. This system recommends users papers using access logs in order to have users make personalized timetable. His system is similar to our system, but our system uses not only papers but also persons for recommendation.

Girgensohn applied the special BBS named *CHI Place* in *CHI2002*[9]. The system is a convenient BBS supports discussion among participants. BBS alone is not sufficient to support discussion in conferences because conferences are too short in time to get to know each other and to make close discussion. Use of Interpersonal network can shorten the process to know each other.

5 Conclusion

This paper reported the community support system based on interpersonal network and analysis. The analysis reveals that characteristics of online interpersonal network are different from real-world interpersonal network. Nevertheless it also showed that interpersonal network is effective in a recommendation. We used interpersonal network just for information access and recommendation, but we believe that interpersonal network has huge potential for information and knowledge sharing. Recently, there are some social networking services, e. g. Friendster [3], Orkut[4] nad Mixi[5]. These services maintain independent

[3] http://www.friendster.com/
[4] http://www.orkut.com/
[5] http://mixi.jp/

interpersonal network. In the future we consider interoperability of interpersonal network on the web using open system technology, e.g., FOAF [10] and XFN[11].

References

1. Konstan, J.A., Miller, B.N., Maltz, D., Herlocker, J.L., Gorden, L.R., Riedl, J.: Grouplens: Applying collaborative filtering to usenet news. Communications of the ACM **40** (1997) 76–87
2. Amaral, L.A.N., Scala, A., Barthelemy, M., Stanley, H.E.: Classes of small-world networks. Applied Physical Science **97** (2000) 11149–11152
3. Cosley, D.: Referee: An open framework for practical testing of recommender system using researchindex. In: Proceedings of the 28th Very Large Databases Conference. (2002)
4. Nishibe, Y., Waki, H., Morihara, I., Hattori, F., Ishida, T., Nishimura, T., Yamaki, H., Komura, T., Itoh, N., Gotoh, T., Nishida, T., Takeda, H., Sawada, A., Maeda, H., Kajihara, M., Adachi, H.: Mobile digital assistants for community support. **19** (1998)
5. Dey, A.K., Futakawa, M., Salber, D., Abowd, G.D.: The conference assistant: Combining context-awareness with wearable computing. In: Proceedings of the 3rd International Symposium on Wearable Computers. (1999) 21–28
6. Sumi, Y., Mase, K.: Conference assistant system for supporting knowledge sharing in academic communities. Interacting with Computers **14** (2002) 713–737
7. Schwarzkopf, E.: An adaptive web site for the um2001 conference. In: Proceedings of the UM2001 Workshop on Machine Learning for User Modeling. (2001) 77–86
8. Jameson, A., Schwarzkopf, E.: Pros and cons of controllability: An empirical study. In: Proceedings of Adaptive Hypermedia and Adaptive Web-Based Systems. (2002) Available from http://dfki.de/∼jameson/abs/JamesonS02.html.
9. Girgensohn, A., Lee, A.: Making web sites be places for social interaction. In: the 2002 ACM Conference on Computer Supported Cooperative Works. (2002) 136–145 Available from http://www.webcollab.com/alee/papers/cscw02.html.
10. Brickley, D., et al.: Friend of a friend (2002) http://www.foaf-project.org/.
11. Celik, T., Meyer, E.A., Mullenweg, M.: Xhtml friends network. In: Demonstration and Poster Proceedings of Hypertext2004. (2004) 48–49 http://gmpg.org/xfn/.

Dynamic Social Simulation with Multi-agents Having Internal Dynamics

Takashi Sato and Takashi Hashimoto

School of Knowledge Science,
Japan Advanced Institute of Science and Technology (JAIST),
1-1, Asahidai, Tatsunokuchi, Nomi, Ishikawa, 923-1292, Japan
{stakashi, hash}@jaist.ac.jp

Abstract. In this paper, we discuss a viewpoint to regard individuals in a society as cognitive agents having internal dynamics, in order to study the dynamic nature of social structures. Internal dynamics is the autonomous changes of an agent's internal states that govern his/her behavior. We first discuss the benefit of introducing internal dynamics into a model of humans and the dynamics of society. Then we propose a simple recurrent network with self-influential connection (SRN-SIC) as a model of an agent with internal dynamics. We report the results of our simulation in which the agents play a minority game. In the simulation, we observe the dynamics of the game as a macro structure itinerating among various dynamical states such as fixed points and periodic motions via aperiodic motions. This itinerant change of the macro structures is shown to be induced by the internal dynamics of the agents.

1 Introduction

The spontaneous social structures in a society, such as institutions, classes, and markets, usually cannot be separated from the individuals in the society, since individuals both shape and are influenced by such structures. The key notion when considering spontaneous structures is *the micro-macro loop*[14][15]. However, we think that this notion alone cannot explain some changes in social structure seen in an actual society. In this paper, we introduce *internal dynamics* in addition to the micro-macro loop and illustrate with a multi-agent simulation where both the social structure at a macro level and the individuals' behavior at a micro level keep changing.

In traditional economics, individuals are often assumed to be isolated from each other, with independent utilities and preferences. On the other hand, Egashira and Hashimoto[3] propose the notion of *socially developmental individuals* whose cognitive frameworks, including utilities and preferences, are shaped through their interaction among themselves. They show the emergence of an institution as a pattern of cognitive frameworks common to the individuals[7].

A. Sakurai et al. (Eds.): JSAI 2003/2004, LNAI 3609, pp. 237–251, 2007.

However, once organized, the institution in their model never changes. In general, if influences from the macro structure to the micro level have a self-enforcement function to regulate the behavior of individuals, it is thought that an institution can emerge and be maintained[1]. But, spontaneous changes of the social structures are not seen in such a case. In reality, social structures change dynamically. Changes in the macro structures are often thought to be caused by changes coming from outside the micro-macro loop, but a mechanism of endogenous change is not explained.

In addition to the idea of socially developmental individuals, we introduce the notion of *internal dynamics*, representing the basic nature of cognitive individuals in a society, in order to understand the endogenous change of social structures. Internal dynamics refers to autonomous changes of the individuals' internal states. Recent cognitive science has developed into clarifying the dynamic nature of cognitive systems. Gelder, for example, advocates that humans are regarded as a kind of dynamical systems, since the complex behavior of dynamical systems can well express cognitive phenomena[17][18]. Varela manifests the importance of structural coupling, which appropriately connects the internal states of a cognitive system to its environment through the interaction between them[19]. These studies place importance on the dynamic change of cognitive systems. In the present study, we also focus on internal dynamics, which has received attention in the field of cognitive science[1]. In the next section, we discuss how internal dynamics is important in considerations of human behavior. Further, we propose a model of an agent with internal dynamics which can be used in multi-agent simulations.

The purpose of this study is to illustrate the importance of viewing individuals in a society as cognitive agents having internal dynamics. In this paper, we perform the following. 1) We propose a simple model of an agent having internal dynamics. This model is expressed by a kind of recurrent neural network. 2) We construct a dynamic social simulation by a multi-agent system which is composed of the agents proposed in 1). Here, a simulation showing the dynamics of social structure is referred to as a dynamic social simulation. We adopt the minority game as a social interaction among the agents. 3) We use this dynamic social simulation to study what causes the macro level dynamics.

The rest of the paper is organized as follows. In section 2, we discuss the internal dynamics and propose a model of agent having internal dynamics. In section 3, a multi-agent system for dynamic social simulation is introduced. Results of the simulations are depicted in section 4. We present a discussion of the results in section 5 and deliver our conclusions in section 6.

[1] An example of a cognitive phenomenon which supports the effect of internal dynamics on cognition is the experiment using a reversible figure. In this experiment, although the figure never changes objectively, the subjective vision of the figure changes with time. This result suggests that cognitive processing is evoked by autonomous changes in the internal states. Moreover, there are some studies about a perception of ambiguous patterns by using chaotic neural network, namely, a network with internal dynamics[11][10].

2 Internal Dynamics

2.1 Importance of Internal Dynamics

From a mechanistic viewpoint, humans can be regarded as a kind of state transition machines. They have internal states that change with external stimuli and return some responses that have one-to-one correspondences with the stimuli.

However, this viewpoint is not always appropriate, since the internal states of humans do not change only in response to external stimuli. It is difficult to explain such characteristics of human behavior as *diversity* and *consistency* by regarding humans as mere state transition machines. The term diversity means here that humans can and often do show various behaviors in the same situation. The sequence of behaviors is usually not random, but has a certain causality. We call this feature of human behavior consistency.

The internal states of humans change even in situations in which the same external stimuli are constantly given or when no external stimuli are given. We refer to this autonomous change of the internal states as *internal dynamics*. By taking the internal dynamics into consideration, we can account for some features of human behavior. Humans can behave variously, even if the same stimuli are given, since their internal states, on which their behavior depends, change autonomously. Accordingly, the human can form a *one-to-many relationship* among a stimulus and his/her responses by means of internal dynamics. The internal states change with actions as well as the external situation. Namely, various influences from the past actions, internal states, and external stimuli are stored in the current internal states. Thus, causality of the human behavior arises, since the actions depend on the internal states and correlate with a history of the past internal states.

2.2 Architecture of Agent Having Internal Dynamics

We conceptualize an architecture of an agent with internal dynamics. As we discussed above, the agent's internal states change autonomously. In addition, the internal states are affected by the agent's past action and the present external stimuli. These assumptions lead to a basic architecture of an agent having internal dynamics, as shown in Fig. 1. In the figure, the agent is regarded as a kind of dynamical system. Thus, we model the agent by means of a dynamical system.

A recurrent network which is regarded as a kind of dynamical system agrees with Gelder's approach of treating humans as dynamical cognitive systems. The recurrent network is known to have various functions such as pattern recognition, motion control, and time series prediction. It is often used in the field of computational cognitive science. Although recurrent networks can produce many behaviors, their computational cost is typically prohibitive for modeling a great number of agents necessary for a large-scale social simulation. On the other hand, the computational cost of a simple recurrent network (SRN) designed by Elman[4] is comparatively less. Furthermore, the SRN is as effective in simulating dynamic phenomena, since it has a powerful ability to learn/predict a time

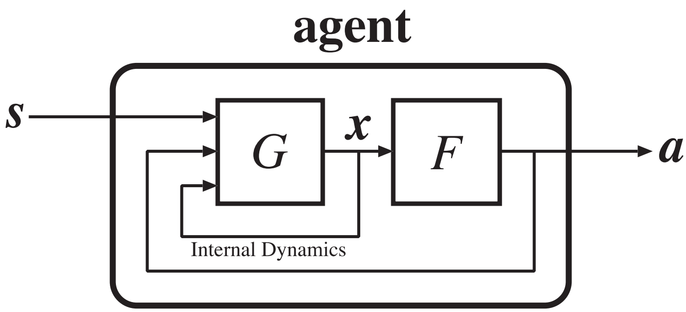


Fig. 1. The basic architecture of agent having internal dynamics. The symbols s, x and a are external stimuli, the agent's internal states and his/her actions, respectively. The boxes labeled by G and F are functions to change the internal states and to decide how the agent behaves when he/she has certain internal states, respectively. The arrows indicate the direction of interactions between the elements.

series[4][5][6]. In spite of these advantages, the SRN is not often used in social simulations.

We propose a concrete agent model corresponding to the basic architecture illustrated in Fig. 1. The model is a modification of the SRN. We call this model a *SRN with self-influential connection* (*SRN-SIC*). Figure 2 shows the proposed architecture of the agent. The SRN has an input layer to accept external stimuli; an output layer to decide the output value based on received signals; and a hidden layer to process input values and to pass them to the output layer. Further, the SRN has a context layer in which each context neuron has one-to-one connections with each neuron of the hidden layer, in order to copy a previous state of the hidden layer. Therefore, the state of the network at a certain time is decided by mixing current stimuli and a history of the past states. Moreover, the SRN-SIC has an additional recurrent connection between the output and the input layers so that the agent decides its own action based on his/her past action.

We show a mathematical form of the SRN-SIC. Each layer has its own index variable: l for recursive output nodes, k for output nodes, j for hidden nodes, i for input nodes, and h for context nodes. The output function of each neuron other than the input and the context neurons is the differentiable nonlinear function L whose range is between -1.0 and 1.0. The function L is defined by

$$L(net) = \tanh(\beta net) \ , \tag{1}$$

where net is the sum of weighted input values, and β decides the nonlinearity of the function L. The output of SRN-SIC is determined by

$$o_k(t) = L(net_k(t)) \ , \tag{2}$$

$$net_k(t) = \sum_{j=0} w_{kj} v_j(t) + \theta_k \ , \tag{3}$$

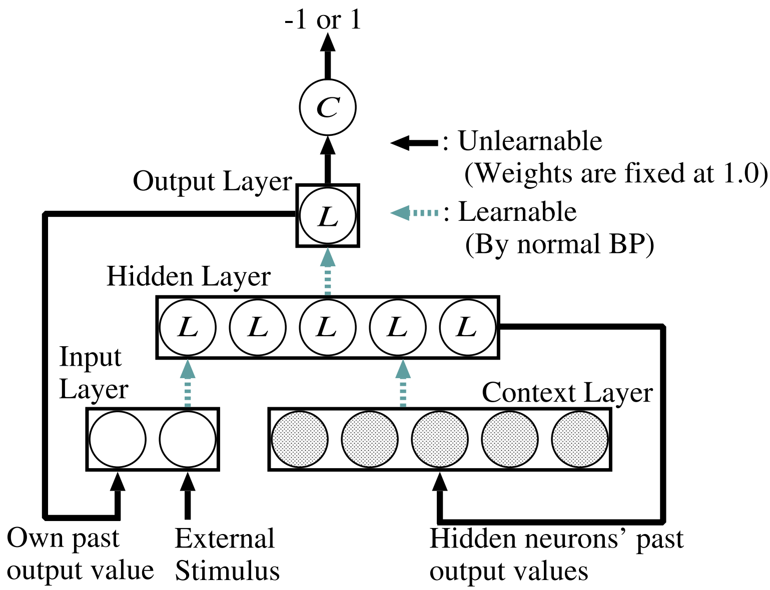


Fig. 2. The SRN-SIC as the proposed architecture of the agent. This is a particular Elman-type network with an additional recurrent connection between the output and the input layers. The symbol L represents a nonlinear function to output a real number between -1.0 and 1.0. The symbol C represents a step function which classifies an output value into two values -1 or 1 in order to be suited for the minority game. Not all connections are shown.

where $o_k(t)$ is the k-th output neuron's value at time t, w_{kj} is the connection weight between the k-th output and the j-th hidden neurons, $v_j(t)$ is the j-th hidden neuron's value at time t, and θ_k is a bias of the k-th output neuron. The hidden neuron's activation is calculated by

$$v_j(t) = L(net_j(t)) \ , \tag{4}$$

$$net_j(t) = \sum_{i=0} w_{ji} x_i(t) + \sum_{h=0} w_{jh} u_h(t) + \sum_{l=0} w_{jl} z_l(t) + \theta_j \ , \tag{5}$$

where w_{ji} is the connection weight between the j-th hidden and the i-th input neurons, $x_i(t)$ is the i-th input neuron's value at time t, w_{jh} is the connection weight between the j-th hidden and the h-th context neurons, $u_h(t)$ is the h-th context neuron's value at time t, w_{jl} is the connection weight between the j-th hidden and the l-th recursive output neurons, $z_l(t)$ is the l-th recursive output neuron's value at time t, and the θ_j is a bias of the j-th hidden neuron. Each value of u and z can be replaced by the past hidden and the past output neuron's activation, respectively. Therefore, the equation (5) is rewritten as

$$net_j(t) = \sum_{i=0} w_{ji} x_i(t) + \sum_{h=0} w_{jh} v_h(t-1) + \sum_{l=0} w_{jl} o_l(t-1) + \theta_j \ . \tag{6}$$

When we consider the context layer as a type of input layer at each time step, the network can be regarded as a kind of feedforward type neural network. Therefore, as a learning method, we adopt the error Backpropagation learning (BP). Each weight of all recurrent connections is fixed at 1.0 and is not adjusted by learning.

3 Multi-agent System for Dynamic Social Simulation

We show a dynamic social simulation by using the multi-agents with internal dynamics proposed in the previous section. In this simulation, we adopt the minority game (MG) proposed by Challet and Zhang[2] as a social interaction among the agents. The game is characterized by the following two basic rules:

1. N (odd) players must choose one out of two alternatives (-1/1 meaning buy/sell, or etc) independently at each step[2].
2. Those who are in the minority side win.

To consider a micro-macro loop in our system, we establish influence from the macro level to the micro level by the following two ways. One is that the previous move of minority side is given to all players as an external stimulus at each step. The other is that all players learn a time series of the past minority move[3].

This simulation is concretely carried out by the following procedure:

1. Each agent independently decides a move (-1 or 1) based on its own past action and the move of minority side at the last play.
2. A current move of minority side is determined from all players' moves.

We call this flow *one step*. All agents learn a time series of the minority moves for the past 100 steps per every 10000 steps. We refer to the 10000 steps between the learning processes as *one turn*.

4 Simulation Results

In this section, we report the results of the multi-agent simulation. In the simulation, the population size of the agents is 101. The SRN-SIC of each agent has one output neuron, five hidden neurons (i.e., there also are context neurons) and two input neurons, as illustrated in Fig. 2. At the beginning of the simulation, all the input values including the feedback input values from the output and the hidden neurons

[2] An output value from the SRN-SIC, which is a real number, is converted to either -1 or 1 in order to correspond to a move in the MG. In this conversion, we regard 0.0 as a border.

[3] Treating the context layer as an input layer, the SRN-SIC can be regarded as a certain type of feed-forward neural network at each time step. Therefore, we can use the error back propagation as the learning method without modification.

are set to be 0.0. The initial connection weights are set to be random real numbers between -0.5 and 0.5, but only recurrent connection weights are fixed at 1.0.

We distinguish several observable dynamics in the micro and the macro levels, as indicated in Table 1.

Table 1. The range of the macro and the micro level

Macro Level		Micro Level	
A time series of minority move	The number of agents who belong to the minority side	Moves of each agent	Internal dynamics of each agent

4.1 Dynamics at Macro Level – Classification of Change Patterns and Itinerant Dynamics –

The time series of the minority move shows various patterns. We classify these into six different patterns, as illustrated in Fig. 3. In order to understand the dynamics at the macro level in detail, we examine in this figure the transition of minority move (-1 or 1) multiplied by the number of agents belonging to the minority side, namely, the winners.

We focus our attention on Fig. 3(b) and (c), in which the minority side never changes. Although all agents continuously receive the same external stimulus, the number of winners changes periodically in Fig. 3(b) and aperiodically in Fig. 3(c). These dynamics imply that the agents can autonomously alter the way they interpret the external information utilizing their internal dynamics, even if the same information is successively given to them. This resembles a human's vision of a reversible figure. Note that the periodic/aperiodic changes happen in one turn in which no learning was executed.

We observe various patterns in the dynamics of the minority move, even in one turn. Figure 4 depicts typical itinerant dynamics at the macro level in one turn[4]. As can be seen, the patterns in the time series of the game itinerate among various dynamical states. The transitions among fixed points and periodic changes are mediated by aperiodic dynamics[5].

The dynamics illustrated in Fig. 4 is observed in one turn. That is to say, it is confirmed that very complex changes at the macro level are induced by the internal dynamics of each agent, though the internal structures of the agents are not modified by learning.

[4] To draw the graphs in Fig. 4, we encode the time series of the minority move. At first, the minority moves, -1 and 1, are coordinated to 0 and 1 as binary digit, respectively. Next, a 20 steps series of the minority move is regarded as a binary fraction. Then, it is converted to a decimal fraction.

[5] These changes resemble *the chaotic itinerancy* proposed by Kaneko and Tsuda[9]. We still could not clarify whether the dynamics observed in our system is precisely chaotic itinerancy.

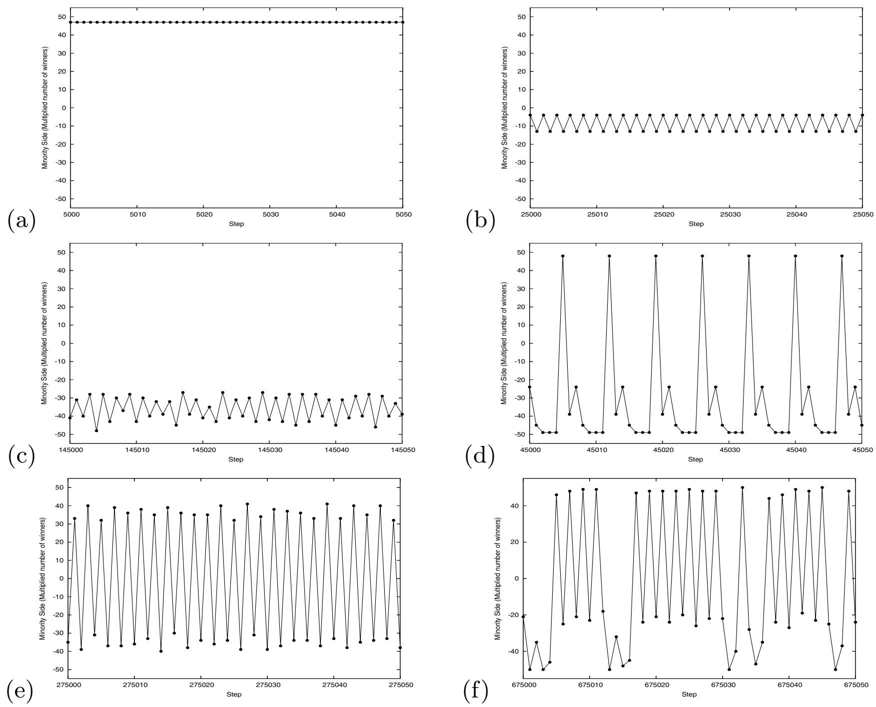


Fig. 3. The patterns of time series of the minority move in different turns. The x-axis is the steps. The y-axis is the minority move (-1 or 1) times the number of winners, i.e., agents in the minority side. The positive (negative) value in the y-axis signifies that the minority move is 1 (-1). (a) Both the minority side and the number of winners are fixed. (b) The minority side is fixed, and the number of winners periodically changes. (c) The minority side is fixed, and the change of the number of winners is aperiodic. (d) Both the minority side and the number of winners show periodic changes. (e) The change of the minority side is periodic, and that of the number of winners is aperiodic. (f) Both the minority side and the number of winners aperiodically change.

4.2 Dynamics at Micro Level – Emergence and Transition of Agent's Strategy –

In this section, we investigate the behavior of the agents. At first, we show that agents obtain particular strategies through learning and interaction in the minority game. A strategy is a way to determine how an agent reacts to external stimuli.

Figure 5 shows examples of two different agents' strategies, expressed as the relationship between the output value and the internal dynamics, namely, the change of two hidden neurons' values. The agent exemplified in Fig. 5(a) acquires a simple strategy which can be denoted by a deterministic finite state transition machine with two states. He/She behaves regularly, depending on the input values. That is to say, there is a one-to-one correspondence between the external

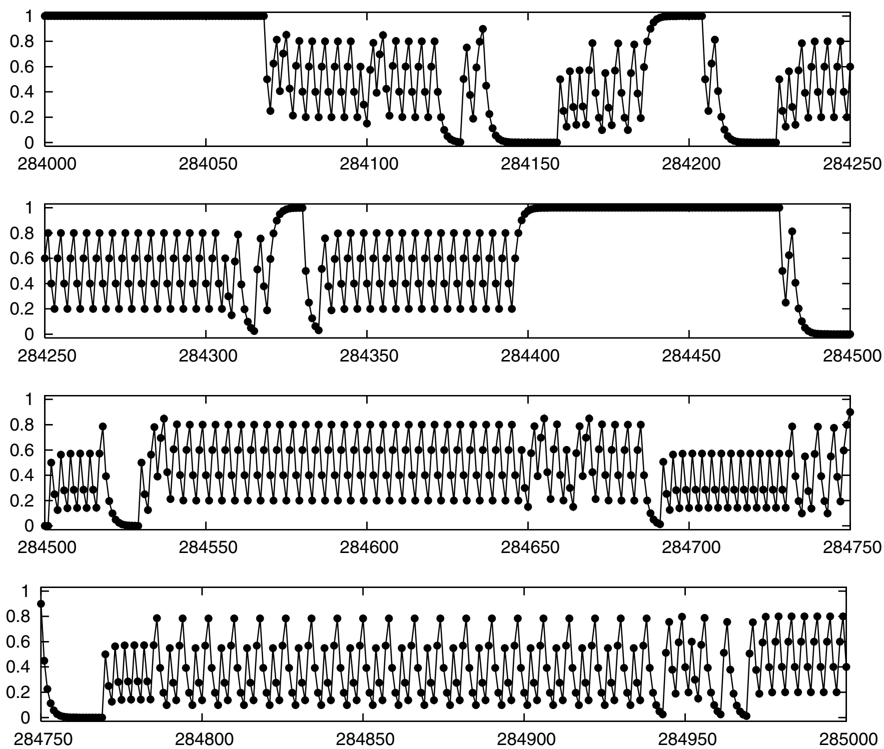


Fig. 4. An example of itinerant dynamics at the macro level in one turn. The x-axis and the y-axis of each figure are the steps and the minority moves converted to real numbers, respectively. The dynamical states of the game change frequently among fixed points and various periodic cycles via aperiodic motions.

stimulus and the action. The other type of strategy, illustrated in Fig. 5(b), accurately uses two rules depending on two kinds of input value, -1 and 1. While the rules described by two closed curves seem simple at a glance, the agent's behavior is complex. The points on the closed curves are so dense that the output sequence of the agent is quasi-periodic. Further, since each closed curve ranges over almost the entire area of the output, the strategy creates a one-to-many relationship from an input to the agent's moves. We also found agents whose strategies are expressed by deterministic finite state transition machines with many states and complex forms like strange attractors.

We illustrate examples of the transitions of two different agents' strategies in several turns in Fig. 6. These strategies vary through the learning process. An interesting agent's behavior is found in the 24th turn. Although the minority move in this turn is fixed on '-1', namely, the same stimulus is continuously given, the trajectory of the agent's internal dynamics drawn in Fig. 6(a) shows a chaotic motion. By using the time series analyses, we confirm that the agent's motion is a low dimensional chaos with weak nonlinearity. In contrast, the other

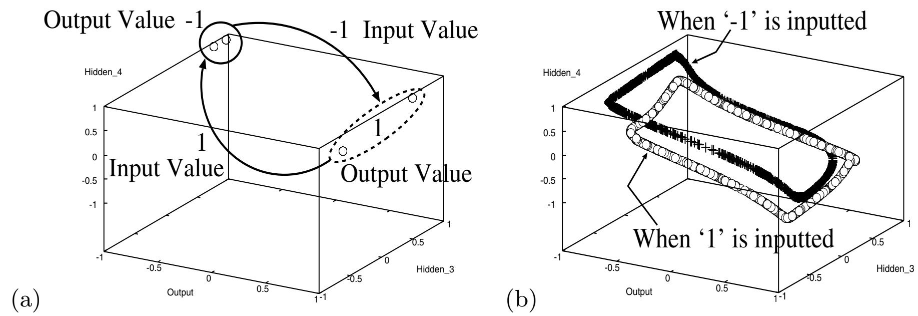


Fig. 5. Examples of two different agents' strategies. The x, y and z-axis of each figure are the values of the output neuron, the third hidden neuron and the fourth hidden neuron, respectively. (a) A strategy described by a simple deterministic finite state transition machine with two states is depicted in the phase space. The small circles show the actual outputs of network. The large circles and the arrows stand for the output values of the agent and the input value that he/she receives, respectively. The dotted circle is the initial state of the agent. The agent behaves periodically. (b) This is a strategy that has two closed curves corresponding to two input values. This means that the agent having the strategy can switch two output sequences according to the external stimuli.

agent in the same turn depicted in Fig. 6(b) acquires a simple deterministic finite state transition machine with three states. In this turn, these agents alter their actions depending on only their past actions. In other words, they attain one-to-many relationships between an input and outputs.

5 Discussion – Causes of Dynamics at Macro Level –

Time series of the minority move show definite features such as fixed points and periodic motions. This suggests that the agents have certain internal structures and form certain relationships with other agents, because the time series of the minority move is decided by the sequence of all agents' moves. Besides, since fixed points and periodic cycles can be described by some rules of dynamical systems, an agents' society showing such dynamics is considered as in some structuralized states with macro level rules. Accordingly, the feature of time series can reflect a macro structure in the agents' society. In our system, the feature of the time series changes with time, as shown in Fig 3 and 4. That is to say, the system realizes the dynamics of the macro structure in the agents' society.

It is thought that the dynamics of the macro structure is brought about by some instability in the system. If so, where is the instability? From the result showing different features of the time series for each turn (Fig. 3), instability must be caused by the learning between turns. Further, there seems to be another instability that is produced by the internal dynamics and interaction of the agents, since the system itinerate among various dynamical states in one turn (Fig. 4). In the following section, we discuss both of the instabilities.

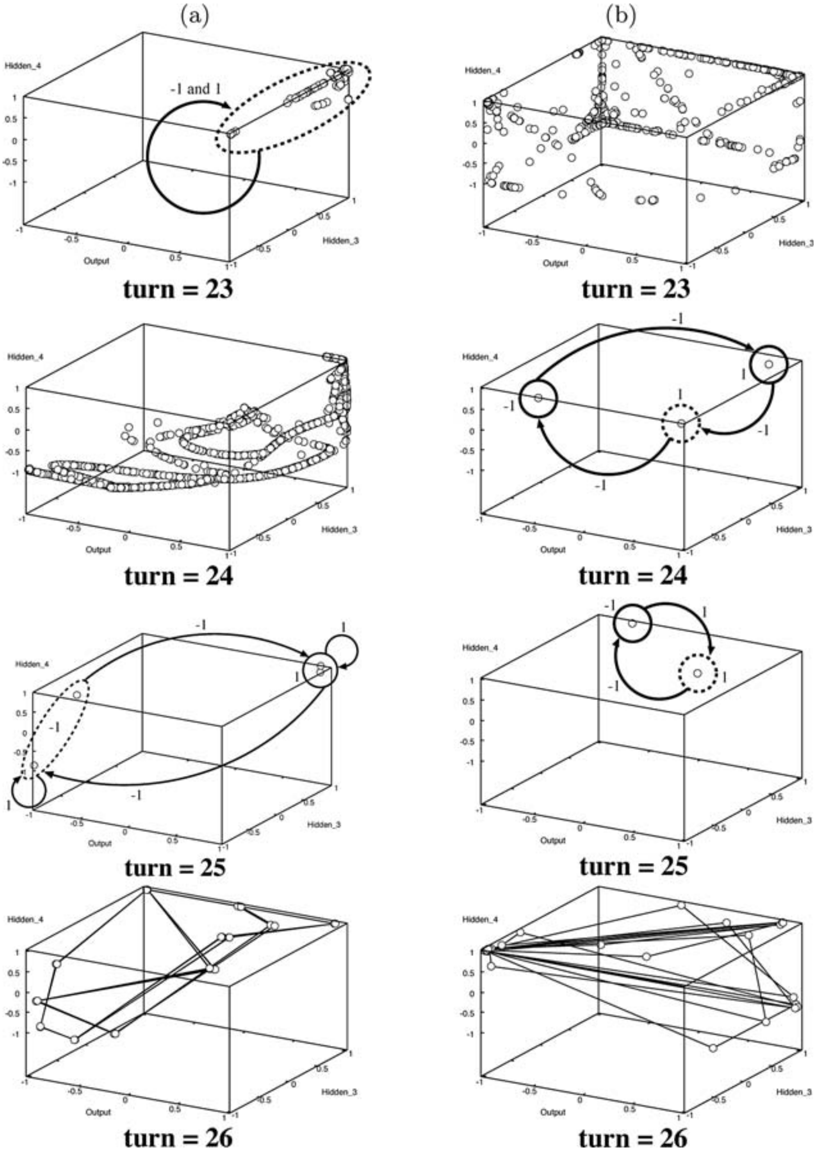


Fig. 6. Transitions of two different agents' strategies in turns 23∼26. All axes are the same as those in Fig. 5. The strategies of each agent vary through the learning process among various deterministic finite state transition machines and complex forms like strange attractors. The agents do not have the same strategy in one turn. For example, in the 24th turn the strategy in (a) forms a strange attractor and in (b) a deterministic finite state transition machine with three states. Although both of the strategies in (a) and (b) in turn 26 are deterministic finite state transition machines with 12 and 30 states, respectively, the number of the states is too many to illustrate. Therefore, we draw the trajectories instead of circles and arrows.

5.1 Instability Between Turns –Effect of Learning–

In the learning process, each agent independently forms a prediction model from the sequence of minority moves of the past 100 steps, to estimate the transition of the game. The model is based on *a static expectation* that the past structure is preserved as is. All agents try to predict the macro structure in the future from a part of the past events. The structure is, however, constructed by all agents whose behavior has been modified by the learning process. Therefore, a static expectation model does not work well to predict the transition of the game.

This is structurally the same destabilization mechanism that is seen in Taiji and Ikegami's studies of the coupled dynamical recognizers[16][8]. There are two agents playing the iterated prisoners' dilemma game in their model. The agents try to make models of their opponents mutually through learning. For each agent, the opponent model used in the previous game is often different from the current opponent. Thus, the dynamics of the game becomes unstable, since the opponent model cannot predict the current opponent's move correctly. Our model can be thought of as an extension from the relationship between two persons in the model of Taiji and Ikegami[16][8] to one among many people. Even though only the moves of the minority side are input to each agent, the minority side is constructed from the moves of all agents. Accordingly, each agent relates to all agents indirectly.

The static expectation is an expression of the agent's bounded rationality. In actual societies, no one can make a complete prediction model that takes the consequences of behavioral changes of all people into consideration. Therefore, the cause of destabilization discussed here is inevitable when social structures are endogenously formed.

5.2 Instability in One Turn –Effect of Chaotic Actions–

To know what feature at the micro level causes the itinerate dynamics at the macro level in one turn, we investigate the configuration of the agents' actions at the micro level. Table 2 shows the configuration corresponding to the classification in section 4.1. When the itinerant dynamics is shown at the macro level, the number of agents with aperiodic actions is much larger than the other cases. The aperiodic action may be chaotic dynamics as indicated in section 4.2. Chaotic dynamics has orbital instability, which expands small differences in the trajectories of agents' actions[12]. Therefore, even a small displacement at the micro level can induce a change in the macro level dynamics.

Let us discuss the aperiodic action from the viewpoint of the relationship between inputs and outputs. A strategy with one-to-many relationships emits aperiodic actions. Periodic actions are, however, also derived from a strategy with one-to-many relationships, as shown in Fig.6(b), the 24th turn. The distinction of these strategies is that the one with aperiodic action, characterized by strange attractors, forms a one-to-infinity relationship from an input to outputs. Accordingly, the condition for the dynamics of macro structure to appear may be that there exists a certain number of agents having a one-to-infinity relationship between an external stimulus and their actions. Internal dynamics is indispensable for obtaining such complex behavior.

5.3 Other Instabilities

In our system, there are also other causes of destabilization. One candidate is the nonlinearity of the SRN-SIC. It is a nonlinear dynamical system and has high dimensional nonlinearity if there are many neurons. Thus, we have to elucidate how this nonlinearity affects an agent's behavior and dynamics of the macro structure by explicating the mathematical structure of the SRN-SIC through experiments such as changing the number of neurons.

A feature of the MG may also be a cause destabilization. In the MG, there is a threshold at half of the players' population. When the number of agents in the minority side is around the threshold, the result of a game changes if only a few players alter their behavior. The similar effect of such a threshold works when output value of the SRN-SIC is divided into '-1' or '1.' Thus, behavior of the

Table 2. The correspondence between changing patterns at the macro level and the configuration of agents at the micro level. The left-hand side of the table indicates the classification at the macro level dynamics (see section 4.1). The first and the second columns are the dynamical states of the minority side and that of the number of winners, respectively. Each number in the right-hand side of the table stands for the number of agents whose actions are in the specific dynamical states. In the case of itinerant macro dynamics (the bottom row), the number of agents showing aperiodic actions is much larger than in the other cases.

Macro Level		**Micro Level**		
Minority Side	The number of Winners	Fixed Point	Periodic Motion	Aperiodic Motion
Fixed Point	Fixed Point	101	0	0
Fixed Point	Periodic Motion	84	17	0
Fixed Point	Aperiodic Motion	50	46	5
Periodic Motion	Periodic Motion	63	38	0
Periodic Motion	Aperiodic Motion	20	77	4
Itinerant Motion	Aperiodic Motion	8	6	87

SRN-SIC is easily changed by a small fluctuation in the input when the output value is around 0.0.

6 Conclusion

In this paper, we have discussed the effectiveness of viewing a human as a cognitive agent having internal dynamics when we account for the emergence and dynamical changes of social structures. We have proposed a model of a social agent having internal dynamics in terms of a simple recurrent network with self-influential connection (SRN-SIC) in order to illustrate the effectiveness concretely. Using a dynamic social simulation considered a micro-macro loop involving such agents, we have shown that complex dynamics emerged at both a micro and a macro levels.

The cause of the macro-level dynamics is conjectured as follows. First, each member of the society does not consider the behavioral change of all the other members to predict the future constructed by them; second, a one-to-infinity relationship between an external stimulus and the actions of each member leads to a chaotic behavior.

Our simulation results substantiate the significance of internal dynamics for forming and maintaining a dynamic social structure. Thus, we conclude that internal dynamics is necessary to form and maintain a dynamic social structure. We also argue that our proposed SRN-SIC is an efficient architecture of a social agent with internal dynamics to construct dynamic social simulations.

We have shown endogenous dynamics of social structures represented by itinerant dynamics, even though the agents' internal structures do not change by learning. It is not yet clear, however, how this dynamics emerged. In further studies, we will clarify the influence of the internal dynamics on agents' behavior and on the macro structure. By solving these problems, we will be able to better perform dynamic social simulations to address the essence of the dynamics in actual societies.

Acknowledgment

The authors wish to thank Susumu Egashira, Takashi Iba, Jun Namikawa and Gen Masumoto for their fruitful discussions. This research was partially supported by a Grant-in-Aid for Scientific Research (No.15700183) from the Ministry of Education, Culture, Sports, Science and Technology of Japan and by the Japan Society for the Promotion of Science.

References

1. Aoki, M., *Towards a Comparative Institutional Analysis*, MIT Press, 2001.
2. Challet, D., Zhang, Y.C., Emergence of cooperation and organization in an evolutionary game, *Physica A*, **246**, 407–418, 1997.

3. Egashira, S., Hashimoto, T., A position of human cognition in social science (in Japanese). In: Nishibe, M. (Ed.), *A Frontier of Evolutionary Economics*, Nihon-Hyouron-Sha, 159–180, 2004.
4. Elman, J.L., Finding structure in time, *Cognitive Science*, **14**(2), 179–211, 1990.
5. Elman, J.L., Distributed representations, simple recurrent networks, and grammatical structure, *Machine Learning*, **7**, 195–225, 1991.
6. Elman, J.L., Learning and development in neural networks: The importance of starting small, *Cognition*, **48**, 71–99, 1993.
7. Hashimoto, T., Egashira, S., Formation of social norms in communicating agents with cognitive frameworks, *Journal of Systems Science and Complexity*, **14**(1), 54–74, 2001.
8. Ikegami, T., Taiji, M., Imitation and cooperation in coupled dynamical recognizers, In: Floreano, D., *et al* (Eds.), *Advanced in Artificial Life*, Springer-Verlag, 545–554, 1999.
9. Kaneko, K., Tsuda, I., Chaotic itinerancy, chaos: Focus issue on chaotic itinerancy, *Chaos*, **13**(3), 926–936, 2003.
10. Nagao, N., Nishimura, H., Matsui, N., A neural chaos model of multistable perception, *Neural Processing Letters*, **12**(3), 267–276, 2000.
11. Nishimura, H., Nagao, N., Matsui, N., A perception model of ambiguous figures based on the neural chaos, In: Kasabov, N., *et al* (Eds.), *Progress in Connectionist-Based Information Systems*, **1**, Springer-Verlag, 89–92, 1997.
12. Ott, E., *Chaos in Dynamical Systems*, Cambridge Univ. Press, 1993.
13. Pollack, J.B., The induction of dynamical recognizers, *Machine Learning*, **7**, 227–252, 1991.
14. Shiozawa, Y., *Consequences of Complexity* (in Japanese), NTT Shuppan, 1997.
15. Shiozawa, Y., *An Introduction to the Economics of Complexity* (in Japanese), Seisansei Shuppan, 1997.
16. Taiji, M., Ikegami, T., Dynamics of internal models in game players, *Physica D*, **134**, 253–266, 1999.
17. van Gelder, T., Port, R., Its about time: An overview of the dynamical approach to cognition, In: Port, R., van Gelder., T. (Eds.), *Mind as Motion: Explorations in the Dynamics of Cognition*, MIT Press, 1995.
18. van Gelder. T., The dynamical hypothesis in cognitive science, *Brain and Behavioral Sciences*, **21**, 615–665, 1998.
19. Varela, F.J., Thompson, E., Rosch, E., *The Embodied Mind –Cognitive Science and Human Experience–*, MIT Press, 1991.

Human-Robot Cooperative Sweeping Using Commands Embedded in Actions

Kazuki Kobayashi[1] and Seiji Yamada[1,2]

[1] The Graduate University for Advanced Studies, 2-1-2 Hitotsubashi, Chiyoda, Tokyo 101-8430, Japan
kazuki@grad.nii.ac.jp
[2] National Institute of Informatics, 2-1-2 Hitotsubashi, Chiyoda, Tokyo 101-8430, Japan
seiji@nii.ac.jp

Abstract. This paper proposes a novel interaction model of a human-robot cooperative task. The model employs CEA (Commands Embedded in Actions), which reduces a human cognitive load because it requires less explicit human-robot communication than direct commanding methods in conventional interaction models. We propose a guideline along which to design robots' actions based on CEA, and apply it to a cooperative sweeping task by a human and a small mobile robot. CEA is experimentally shown to reduce the human cognitive load more than direct commanding methods do in this sweeping task.

1 Introduction

There are robots spreading among people as a progression of its technologies. We can purchase pet robots like AIBO [1] or cleaning robots like Roomba, and interact with them in a home environment. We will see tour-guide robots [2] in a museum in the near future. Robots thus have transferred their scene from industrial environments to home environments. How a home robot interacts with people is one of most important issue to be accepted by people who want to share their time and spaces with robots.

Various researches have been studied in a field of human-robot interaction. Most of the researches have dealt with methods of communication between a human and a robot such as gesture [3,4,5], speech [6,7,8], using control devices such as joysticks or computers [9,10,11], multiple methods [12] and others [13,14,15,16]. These types of interaction focusing into a function of a robot are described in Fig.1 in which the methods of communication correspond to the arc of command, and tasks of a human correspond to the arc of action. This figure assumes that a robot works based on the cycle of sensing, interpretation, and execution. The arc of human's command means that a human gives a robot information which a robot cannot sense or interpret, and the arc of human's action means human's tasks which a robot cannot execute it because of the difficulty of equipping necessary hardware for given tasks. For instance, a commercial cleaning robot such as Roomba can sweep out a room autonomously, however a human

A. Sakurai et al. (Eds.): JSAI 2003/2004, LNAI 3609, pp. 252–265, 2007.

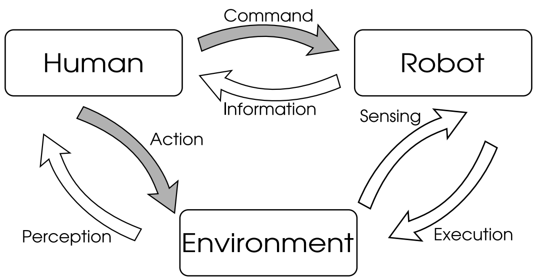


Fig. 1. Conventional Interaction

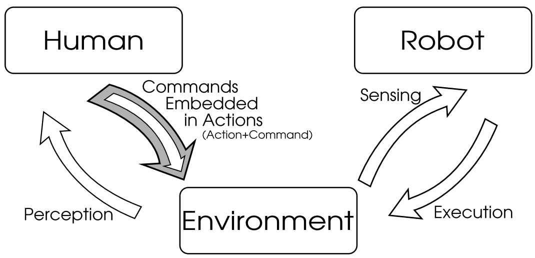


Fig. 2. Suggested Interaction

actually needs to help the robot by removing obstacles because of equipping no hardware for handling the obstacles. Two tasks are assigned to the human: to control the robot by remote controller and to remove the obstacles.

The improvement of the arc of human's command such as gesture recognition, speech recognition and so on are insufficient for an actual task of a human in terms of his/her work-load: the human have to achieve two tasks. We therefore focus on a human's given tasks represented by the arc of human's action and deal with this kind of work as a cooperative task between a human and a robot to reduce the load of a human. We then introduce a new interaction model of Fig.2 in which no direct interaction between a human and a robot are employed. A human can control a robot by executing his/her actions to environment. We call this CEA (Commands Embedded in Actions) by which output and input of a human can be reduced. A human does not need to execute direct commanding to a robot and understand a way for communication with it. The interaction between a human and a robot no longer exists. The existence of a robot is practically transparent for a human, and it leads reduction of a human work-load.

There are studies related our suggested interaction model for a human-robot cooperative task. Their cooperative task is to carry a long or big object by a human and a robot based on manipulator [17,18,19], and outdoor cooperative tasks by a human and a humanoid [20]. These studies are typical instances of CEA. The robots can work well for helping a human by sensing force of a shared object without direct commanding methods. However, interchanging force is only allowed on such cooperation. In contrast, interchanging various information including force is allowed by CEA between a human and a robot. Our aim is to build up a general framework for human-robot cooperation. Therefore, CEA is a novel approach in the research area of human-robot interaction.

After this section, in Section 2, we describe the procedure of interaction design for a human-robot cooperative task. In Section 3, we apply our interaction design to a cooperative sweeping task between a human and a robot. A developing of a behavior-based robot is described in Section 4. Section 5 describes two experiments and shows these results. First experiments evaluate the efficiency of cooperative behavior by using a proposed design procedure. Second one compares CEA using suggested interaction design with direct control using conventional interaction design such as voice or hand manipulation in terms of human's cognitive load. The experimental results are discussed in Section 6. Finally, Section 7 concludes this paper.

2 Method of Interaction Design

We propose the procedure of interaction design for a human-robot cooperative task illustrated in Fig.2 as follow:

1. *Divide a given task into a human's task and a robot's task.*
   (a) Determine a robot's task as maximize robot's autonomy within given cost of hardware design.
   (b) Determine a human's task as assign the task which a robot cannot execute autonomously.
   (c) Determine cooperative behavior and information interchanging between a human and a robot.
2. *Embed commands in actions.*
   Commands including the information of a robot are communicated to a robot as a human executes his/her actions. Actions are determined as minimize the change in space and time from a typical human's action for an assigned task.
3. *Design functions of a robot.*
   Add the functions to a robot to sense or interpret CEA.

CEA can be designed through above procedure. Although details depend on a given task, the procedure can assist interaction design for reducing human work-load. CEA has advantages described as follow:

- *No additional cognitive load in execution:*
  Since CEA has minimum additional actions to typical human's actions to achieve an assigned task, a human does not need additional cognitive load by direct communication with a robot and smoothly does cooperation with a robot by executing only typical actions.
- *No understanding a way for communication:*
  Since direct communication is not necessary in CEA, a human does not need to understand communication protocol like gesture commands, special

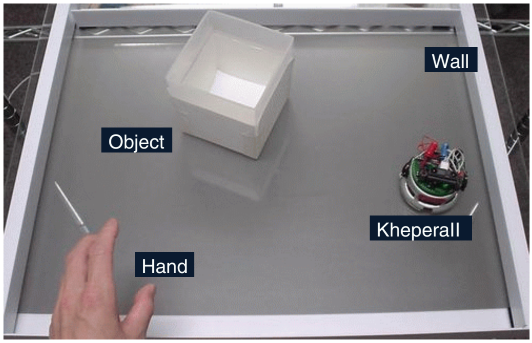


Fig. 3. Experimental environment

speech commands. CEA thus releases a human from learning protocol and training to communicate with a robot.

3 Interaction Design on Cooperative Sweeping

We deal with a human-robot cooperative sweeping by a human and a small mobile robot as a cooperative work. A goal of the cooperative task is to sweep out a desk including the region of under an object. In this section, we first describe an experimental environment and a specification of small robot. Then we apply the procedure of our interaction design to the cooperative sweeping.

3.1 Environment and Robot

Fig.3 shows an experimental environment where a human and a robot work cooperatively. This environment simulates a place used by a human routinely such as a desktop. A desk swept out by a robot has a flat surface and a wall which encloses the region for keeping a robot not to fall.

We use a small mobile robot KheperaII. The robot has eight infrared proximity and ambient light sensors with up to 100mm range, a processor Motorola 68331 (25MHz), 512 Kbytes RAM, 512 Kbytes Flash ROM, and two DC brushed servo motors with incremental encoders. The program written by C-language runs on the RAM.

3.2 Interaction Design

The procedure of our interaction design is applied for cooperative sweeping as follow:

1. *Divide a sweeping task into a human's task and a robot's task.*
   (a) A robot's task is to sweep out a desk autonomously with strategy of a random turn. The hardware resource of a robot is equal to that of a commercial sweeping robot.

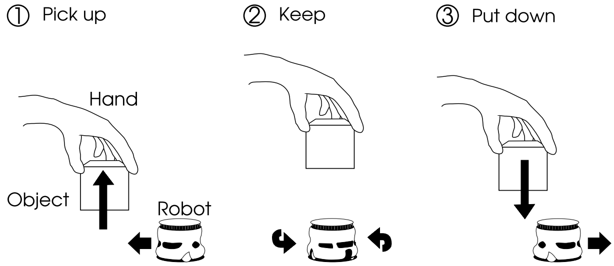


Fig. 4. Cooperative sweeping by CEA

(b) A human's task is to move an object because a robot cannot move an object by itself.
(c) Cooperative behavior is that a robot sweeps out the region under an object when a human moves the object.

2. *Embed commands in actions*
   A command sent by a human makes a robot sweep out the region under an object. This command is made by human actions to achieve his/her task. Fig.4 shows CEA in which the human's action has no changes in the trajectory from the typical one, and it has little additional time to keep picking up. A human does not need to move an object to another place.
3. *Design functions of a robot.*
   An extra infrared sensor which measures the distance in vertical direction is added on a robot's general I/O turret to sense CEA.

The detail of designed CEA in Fig.4 is described as follow:

1. A human picks up an object when a robot approaches an object. Then a robot comes into the region under the object picked up.
2. While a robot is in under an object, it keeps turning at the object's edge.
3. A robot goes out of the region when a human approaches an object to it.

4 Behavior Design of Mobile Robot

Our robot performs obstacle avoidance, going forward (when no obstacles are on its front) and turning for random direction. We use sweeping with selecting directions at random because it has lower cost than a calculated sweeping using a map. Many methods for region covering have been developed [21], and most of them need precise position of a robot. However, our robot cannot use reliable methods like dead reckoning. We hence consider that the random sweeping is adequate to our study because the most of consumer sweeping robots adopt this kind of method.

A robot is implemented by behavior-based approach, and we adopt subsumption architecture [22] to manage following behaviors.

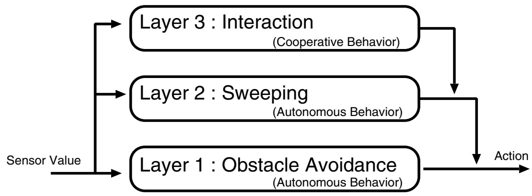


Fig. 5. Subsumption architecture

- Obstacle avoidance.
- Autonomous sweeping while no object is sensed.
- Interactive sweeping while an object is above a robot.

Fig.5 shows the robot's behaviors into the three layers in subsumption architecture. Each layer asynchronously checks the applicability of behaviors and executes applicable ones. Higher layers suppress lower layer's behaviors, and lower layers have more reactive behaviors. The behaviors of each layer consist of multiple actions. When the system obtains multiple outputs, it generally selects the highest layer's action. Each layer has output frequency of action to control the robot smoothly. We set the frequency as obstacle avoidance: an action by 5msec, sweeping: 10msec, interaction: 5msec, obstacle avoidance and interaction occur most frequently.

5 Experiments

Two experiments are conducted. First experiments are measurement of sweeping time to evaluate the cooperative sweeping behavior by comparing with normal autonomous sweeping behavior. Second experiments are measurement of human cognitive loads to compare the load between CEA and conventional direct control of a robot.

5.1 Sweeping Time Measurement in Simulation

The purpose of a simulation is to confirm that the time of cooperative sweeping is shorter than that of autonomous sweeping. Cooperative sweeping is the cooperative behavior in the procedure 1-(c) of interaction design. Autonomous sweeping is a normal sweeping strategy of a robot. It is important to clarify that the cooperative behavior employs an efficient strategy, not an arbitrary inefficient strategy, to apply CEA.

Simulator
We develop an original khepera simulator illustrated in Fig6. It can represent a sweeping area as arbitrary size cells in a two-dimensional plane. A cell changes its color when a robot crosses it. A cell changing its color means that the cell was swept by a robot. Therefore, this simulator is appropriate for a sweeping task. It can simulate functions and behavior of a robot almost completely including

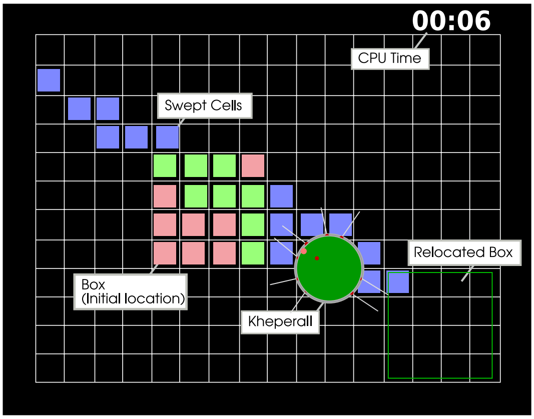


Fig. 6. Developed Khepera Simulator

vertical infrared sensors and subsumption architecture. It also simulates behavior of manipulating obstacles such as moving box. The simulation program is written in Ruby language with GTK-2.0 library.

Methods

In cooperative sweeping, a box is picked up and kept it until all the cells of under the box are swept by a robot and then it is put down. After that, the box is not moved any longer. In autonomous sweeping, on the other hand, a box is relocated to a given location when the experiments begin, and then the box is put down to the initial location when the cells of the initial location of the box are swept out. After that, the box is neither moved any longer. Two policies of the relocation for autonomous sweeping are prepared. The first policy is to relocate a box to a corner of a sweeping area. It makes a robot to sweep easily because of a simple structured area. The second one is to relocate a box to the place near to the initial location of the box. It makes a human to move a box easily because of the minimum quantity of the movement. Autonomous sweeping is estimated by averaging these policies.

The sweeping area has 16×12 cells, and 3×3 cells approximately correspond to the area of a robot. Sweeping time until the robot covers the 98% of all cells is measured in three cell sizes of boxes: 3×3, 4×4 and 5×5. These are measured 10 times in a given location of a box. The initial location of a robot is determined at random in each measuring. Larger sizes of boxes (such as 6×6) are not employed because these conditions make a free space too small and narrow for a mobile robot to fully sweep there. The conditions of a simulation are described in Table1. Cooperative sweeping and autonomous sweeping have the same number of box's locations. The number of measuring in autonomous sweeping is duple of cooperative sweeping because two policies of the relocation are employed.

Table 1. Conditions of the simulation

box size	sweeping method	number of box's locations	number of measuring
3×3	cooperative	15	150
	autonomous	15	300
4×4	cooperative	15	150
	autonomous	15	300
5×5	cooperative	8	80
	autonomous	8	160

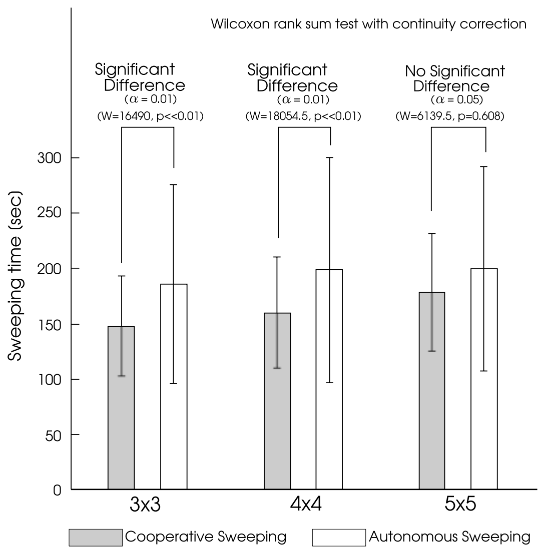


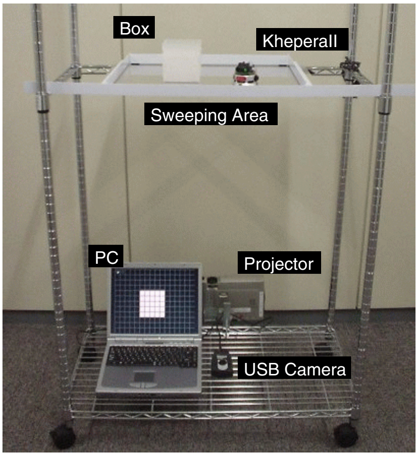
Fig. 7. Results of averaged sweeping time in each size of a box

Results

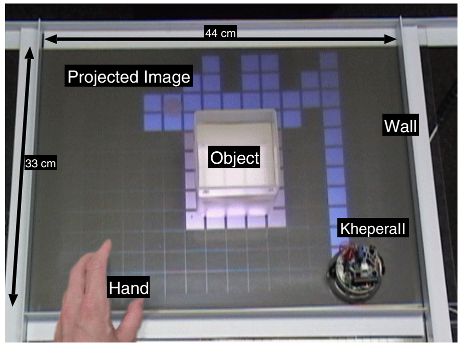
Fig.7 shows average times of sweeping in each size of boxes, error bars indicating their standard deviation for cooperative sweeping and autonomous one. The sweeping time in this figure is CPU time which corresponds to the frequency of the used CPU resources by the simulation program. Each pare of average times are tested by Wilcoxon's rank sum test with continuity correction. As a result, the 3×3 ($p << 0.01, \alpha = 0.01, W = 16490$) and the 4×4 ($p << 0.01, \alpha = 0.01, W = 18054.5$) boxes have significant differences of sweeping time. The 5×5 ($p = 0.608, \alpha = 0.05, W = 6139.5$) has no significant difference.

5.2 Human Cognitive Load Measurement

The purpose is to confirm that our interaction design reduces work-load of human. We measure human cognitive loads to evaluate human work-loads, and compare the load between CEA and conventional direct control of a robot. The experiments are performed on a developed experimental system.



(a) Experimental system



(b) Sweeping area

Fig. 8. Experimental system and Sweeping area

Experimental System

Fig.8 (a) shows the experimental system which can indicate a robot's trajectory. This system consists of a sweeping area and a projection system. In experiments, a human interacts with a robot on the sweeping area in Fig.8 (b) indicating a swept location by the projection system including a personal computer, a projector, and a USB camera. The projection system detects a robot's location using a picture of two beams of infrared LEDs equipping on the robot. The robot's location is calculated by image processing in the picture from the camera, and then an image indicating the trajectory is created with the location. This image is ultimately projected on the sweeping area. The projected image also includes small square cells. A cell is lit when a robot enters the cell in real time. These small cells therefore express the trajectory of a robot. The sweeping area having a width of 44 cm and a height of 33 cm is divided into 16 × 12 cells. Cells of 3 × 3 approximately correspond to the area of a robot.

Methods

Two direct control methods are chosen as typical one without extra devices such as remote controller and a human can create a command by moving his/her body. These control methods are shown in Fig.9, and the detail is described as follow:

EC : A robot performs sweeping by CEA.
SC : A robot sweeps with receiving a sound command by hand clapping.
HC : A robot sweeps with receiving a command by hand as blocking its line.

The robots receiving such commands are prepared by adding new sensors such as microphone or modifying the program of a robot. When a robot controlled by direct command receives a command from a human, it turns from 90° to 180°

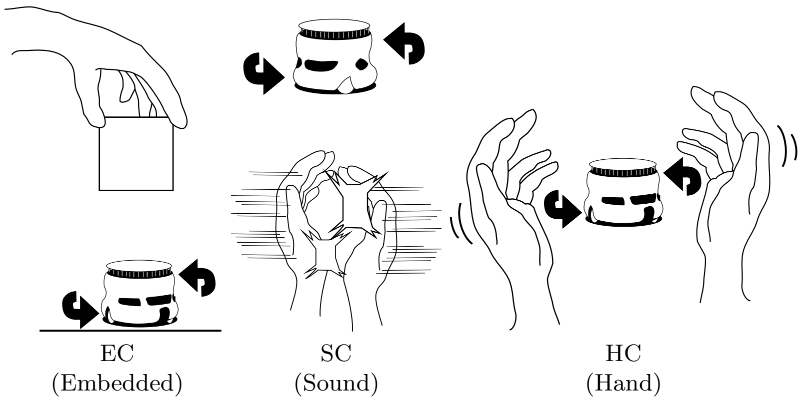


Fig. 9. Three types of interactions

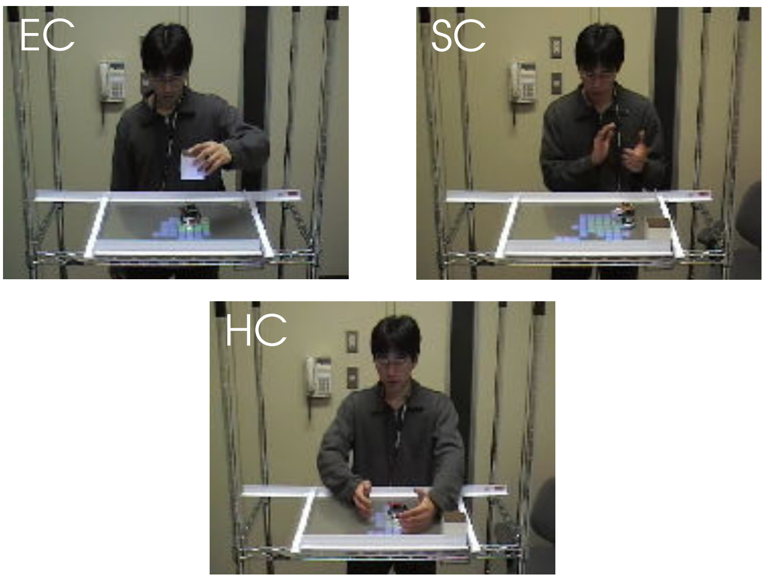


Fig. 10. Experimental Appearance

randomly. A robot controlled by CEA also turns from 90° to 180° randomly when it runs out from under an object.

Measurement starts when a robot enters the region of under an object, and it continues until all cells of the region are swept. A box whose size is 4×4 cells is employed because sweeping time of the region is appropriate for subjects and the measurement. In the EC, subjects keep to pick up an object until all the cells of under the box are swept. In the SC and HC, first, a human relocates a box to a corner of the sweeping area, and then send a command for a robot to be turn by making sounds or approaching their hand to it when it is likely to run out from the region of a box.

We use a dual task method to measure human cognitive load. Subjects have to do mental arithmetic as a secondary task while controlling the robot as a primary task [9,23]. They count backwards by three from a randomly selected three-digit number vocally. We obtain the average number of correct answers per second, and evaluate the human's cognitive load for controlling each robot. Subjects are required to calculate as quickly as possible, and to prioritize the controlling a robot over the counting. They practice controlling robots and the counting well before experiments begin. The order of EC, SC and HC for each subject is determined at random, and these experiments are recorded three times respectively for a subject. Subjects are also measured counting ability without operations of a robot before a measuring of EC, SC, or HC respectively. The counting ability is the number of correct answers of the counting for 30 seconds. Fig.10 shows the experimental appearance.

Results

Subjects are eight men and four women between the age of 22 and 32. Each subject has three scores: EC, SC, and HC. A score is the average of normalized number of correct counting answers per second for a subject. The normalization is to divide the correct answers per second by correct answers per second without operations of a robot. Fig.11 shows averaged scores, standard deviations, and differences tested by Dunnett's test. Each EC, SC and HC is the average of all subject's scores. The number 1.0 means counting ability of each subject without operations of a robot. EC has the highest average. The difference between EC and HC has a significant

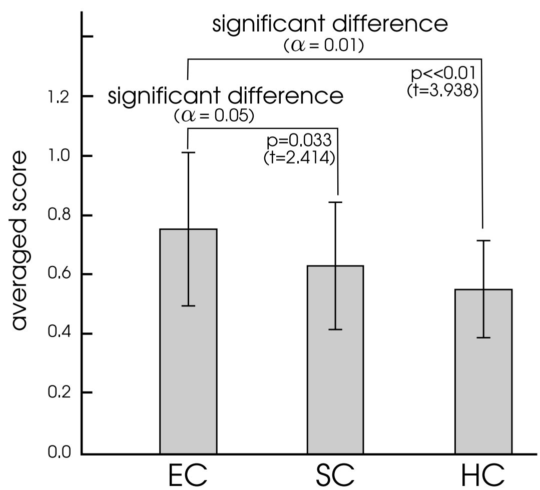


Fig. 11. Result of scores and differences

difference ($p << 0.01, \alpha = 0.01, t = 3.938$). The difference between EC and SC also has a significant difference ($p = 0.033, \alpha = 0.05, t = 2.414$).

6 Discussion

The results of the first experiments show that the cooperative sweeping behavior makes the whole area sweeping efficient, and it depends on proportion of the size of box and the size of the whole sweeping area. This effect however is available only for the robot's random sweeping strategy, and can be applicable not only in the case of using CEA, but in the case of using the direct control methods. We conform that such cooperation shortens the sweeping time significantly. It is clear that the cooperative behavior does not employ an arbitrary weak strategy to apply CEA. These results suggest that the interaction design procedure can apply CEA to a given task without decreasing its efficiency.

The results of the second experiments show that the CEA reduces a human cognitive load in comparison with other direct commanding methods. CEA has a low cognitive load because of minimizing cost of sending commands and also shortening the trajectory of moving a box. Therefore, CEA plays the significant role in a human-robot cooperative task.

Physical loads concern the cognitive loads. Each commanding method accompanies motions of human arms. Hence, the measured cognitive loads include a load of the motions and a load of human attention. However, we consider that the effects of human motions on the measured cognitive loads keep a minimum because the motions of human arms are intuitive and natural. Actually, the subjects have no choice except to clap his/her hands or to move his/her arms in the experiments. Therefore, the difference between CEA and direct control methods is attention for a robot. A subject has to repeat the cycle of observation of a robot and execution of moving his/her arms in the experiments of SC and HC. Contrastively, in the experiments of EC, a subject does not need to concentrate his/her attention on a robot, and he/she can interact with environment as Fig.2. In addition, the experiments are set to be fair deal between CEA and direct control methods in terms of controlling a robot without specific devices.

7 Conclusion

We proposed a novel interaction model of a human-robot cooperative task. The model employed CEA (Commands Embedded in Actions), which reduced a human cognitive load because it required less explicit human-robot communication than direct commanding methods in conventional interaction models. We developed a guideline along which to design robots' actions based on CEA, and applied it to a cooperative sweeping task by a human and a small mobile robot. The first experiments for measuring of sweeping time were conducted by a developed simulator to evaluate the determined cooperative sweeping behavior by comparing normal autonomous sweeping behavior. The second experiments for conforming reduction of a human work-load on our interaction design were conduced on the sweeping task. Human cognitive loads were examined to evaluate human work-load while a human interacted with a robot, and loads between CEA and direct control of a robot were compared. The results of the experiments showed that the interaction design procedure was able to apply CEA to a given task without decreasing its efficiency, and the CEA minimized a human cognitive load in comparison with other direct commanding methods. Therefore, CEA led by our interaction design method plays the significant role in a human-robot cooperative task.

References

1. Hornby, G.S., Takamura, S., Yokono, J., Hanagata, O., Yamamoto, T., Fujita, M.: Evolving robust gaits with aibo. In: Proc. of IEEE International Conference on Robotics and Automation (ICRA'00). (2000) 3040–3045
2. Burgard, W., Cremers, A.B., Fox, D., Hahnel, D., Lakemeyer, G., Schulz, D., Steiner, W., Thrun, S.: The interactive museum tour-guide robot. In: Proc. of the Fifteenth National Conference on Artificial Intelligence (AAAI/IAAI). (1998) 11–18
3. Triesch, J., von der Malsburg, C.: Robotic gesture recognition. In: Proc. of the Bielefeld Gesture Workshop. (1997) 233–244
4. Waldherr, S., Romero, R., Thrun, S.: A gesture based interface for human-robot interaction. Autonomous Robots **9** (2000) 151–173
5. Marrone, F., Strobel, M.: Cleaningassistant - a service robot designed for cleaning tasks. In: Proc. of Advanced Mechatronic Systems (AIM'01). (2001)
6. Breazeal, C., Aryananda, L.: Recognition of affective communicative intent in robot-directed speech. Autonomous Robots **12** (2002) 83–104
7. Lauria, S., Bugmann, G., Kyriacou, T., Klein, E.: Mobile robot programming using natural language. Robotics and Autonomous Systems **38** (2002) 171–181
8. Kajikawa, S., Hiratsuka, S., Ishihara, T., Inooka, H.: Robot position control via voice instruction including ambiguous expressions of degree. In: Proc. of IEEE Int. Workshop on Robot and Human Interactive Communication (ROMAN'03). (2003)
9. Crandall, J.W., Goodrich, M.A.: Characterizing efficiency of human-robot interaction: A case study of shared-control teleoperation. In: Proc. of IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS'02). (2002)
10. Katagami, D., Yamada, S.: Active teaching for an interactive learning robot. In: Proc. IEEE Workshop Robot and Human Interactive Communication (RO-MAN'03). (2003)
11. Khamis, A.M., Rodríguez, F.J., Salichs, M.A.: Remote interaction with mobile robots. Autonomous Robot **15** (2003)
12. Severinson-Eklundh, K., Green, A., Hüttenrauch, H.: Social and collaborative aspects of interaction with a service robot. Robotics and Autonomous Systems **42** (2003) 223–234
13. Ono, T., Imai, M.: Reading a robot's mind: A model of utterance understanding based on the theory of mind mechanism. In: Proc. of Seventeenth National Conference on Artificial Intelligence. (2000) 142–148
14. Butler, J.T., Agah, A.: Psychological effects of behavior patterns of a mobile personal robot. Autonomous Robots **10** (2001) 185–202
15. Kaplan, F., Oudeyer, P., Kubinyi, E., Miklosi, A.: Robotic clicker training. Robotics and Autonomous Systems **38** (2002) 197–206

16. Fong, T.W., Nourbakhsh, I., Dautenhahn, K.: A survey of socially interactive robots. Robotics and Autonomous Systems **42** (2003) 143–166
17. Hayashibara, Y., Sonoda, Y., Takubo, T., Arai, H., Tanie, K.: Assist system for carrying a long object with a human – analysis of a human cooperative behavior in the vertical direction –. In: Proc. of 1999 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS‘99). (1999)
18. Arai, H., Takubo, T., Hayashibara, Y., Tanie, K.: Human-robot cooperative manipulation using a virtual nonholonomic constraint. In: Proc. of 2000 IEEE International Conference on Robotics and Automation (ICRA‘00). (2000)
19. Nakai, K., Kosuge, K., Hirata, Y.: Control of robot in singular configurations for human-robot coordination. In: Proc. of IEEE Int. Workshop on Robot and Human Interactive Communication (ROMAN‘02). (2002) 356–361
20. Yokoyama, K., Maeda, J., Isozumi, T., Kaneko, K.: Application of humanoid robots for cooperative tasks in the outdoors,. In: Proc. of IEEE/RSJ IROS Workshop on Explorations towards Humanoid Robot Applications. (2001)
21. Choset, H.: Coverage for robotics - a survey of recent results. Annals of Mathematics and Artificial Intelligence **31** (2001) 113–126
22. Brooks, R.A.: A robust layered control system for a mobile robot. IEEE Journal of Robotics and Automation **2** (1986) 14–23
23. Condron, J.E., Hill, K.D.: Reliability and validity of a dual-task force platform assessment of balance performance : Effect of age, balance impairment, and cognitive task. Journal of American Geriatrics Society **50** (2002) 157–162

Discovery of Skills from Motion Data

Kosuke Makio, Yoshiki Tanaka, and Kuniaki Uehara

Graduate School of Science and Technology, Kobe University, Japan

Abstract. In this paper, we discuss how to discover "*skills*" from motion data. Being able to understand how a skilled person moves enables beginners to make better use of their bodies and to become experts easier. However, only few attempts have so far been made for discovering skills from human motion data. To extract skills from motion data, we employ three approaches. As a first approach, we present association rule approach which extracts the dependency among the body parts to find the movement of the body parts performed by the experts. The second is an approach that extracts frequent patterns (motifs) from motion data. Recently, many researchers propose algorithms for discovering motifs. However, these algorithms require that users define the length of the motifs in advance. Our algorithm uses the MDL principle to overcome this problem so as to discover motifs with optimal length. Finally, we compare the motions of skilled tennis players and beginners, and discuss why skilled players can better serve.

1 Introduction

Recently, motion capture technology has been widely used to create CG images in movies, to analyze movements in the sports science field, to recognize a sign language, and so on. Among these applications, human motion analysis especially receives many researchers' attention. Now, there are two approaches for human motion analysis: the first approach is image data processing, which obtains motion data as an image time series. The second approach is using motion capture systems that track markers attached to body parts, and record human motion as multi-dimensional time series.

In the approach of image data processing, there is no need to attach markers to the body. However, we must extract feature points from the image time series, and relate each image to a human model created in advance. There are two types of human models. One is the 3D Model, in which human body parts are represented in 3D shapes[1]. Another one is the view-based model, in which feature points are extracted from image time series data of human motion[2,3]. Both models have the problem that it is difficult to extract the feature points with accuracy and a correct correlation between the image data and the model cannot be created.

On the other hand, in the approach of using motion capture systems, markers are attached to the human body parts, and the system tracks them and records their 3D coordinates. This approach has a number of advantages. It eliminates

A. Sakurai et al. (Eds.): JSAI 2003/2004, LNAI 3609, pp. 266–282, 2007.

the need of extracting the feature points and, in general, has a lower computational cost than the approach of image data processing. In addition to these advantages, we can analyze motion data more precisely by this system. For these reasons, we adopt the approach of motion capture system.

We use an optical motion capture system that consists of infrared cameras and PC. By using this system, motion data capturing process is performed as follows: an actor with 18 markers (Fig. 1(a)) performs some actions in the space surrounded by six infrared cameras. Then, 3D coordinates of markers are computed from the images recorded by the cameras. Finally, the 3D time series of human motion are obtained (Fig. 1(b)). This figure represents the movement of right hand while pitching.

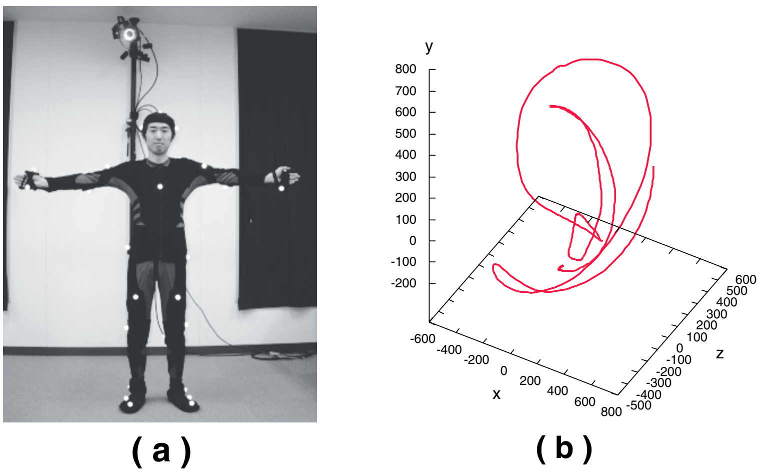


Fig. 1. (a) Our motion capture system and an actor who puts on markers. (b) An example of 3D time series of human motion.

In this paper, we aim to discover *skills* from motion data captured by our system. Here, skill is defined as the technique to do some exercise well. For example, a good tennis player has a special skill to shot strong services. By discovering skills from motion data, more effective practices can be possible and one can become an expert more easily. To achieve this goal, we employ three approaches: as a first approach, we present association rule approach. Human motions have cooccurrence relations between body parts, and we can find these relations by extracting association rules from motion data. The association rules, which are extracted from motion data of experts, can explain how experts perform their movement. The second is an approach that extracts *motifs* from motion data. Motifs are defined as frequent patterns that are previously unknown. To find motifs, it is difficult to decide the length of motif appropriately. Recently, many researchers propose algorithms for discovering motifs. However, these algorithms require that users should define the length of the motifs in advance. Our algorithm uses the MDL principle to overcome this problem, so as to get motifs with optimal length.

Finally, we compare the motions of different persons. In the music domain, Widmer et al.[4] try to discover skills from music performances. They discover the differences of performance style between concert pianists by analyzing relations between their tempo and loudness. Inspired by their research, we experiment with extraction of the characteristics of skilled tennis players and beginners, and the reasons why skilled players can make better services are revealed.

2 Association Rule Discovery Algorithm

Association rules are used to find dependency relations between elements. For example, an association rule "$A \Rightarrow B$" means "if A occurs, then B will likely also occurs". In order to extract association rules from motion data, we have to segment motion data into primitive motions, which are minimum elements of the motion, and assign a unique symbol to each of them.

To segment motion data, we define the points, where the direction of movement changes, as the breakpoints. These breakpoints can be detected where the speed of the body part on one of the X, Y, Z axes becomes 0. However, this definition of the breakpoint is too sensitive to slight motions, such as hand shaking. These slight motions lead to many "meaningless" primitive motions that are too short or too small for extraction of association rules. For this reason, we use Minimal Distance/Percentage Principle (MDPP)[6] to remove "meaningless" breakpoints. In MDPP, the breakpoints are compared with temporal and spatial thresholds, and then only meaningful breakpoints are extracted.

After motion data is segmented into primitive motions, the distances between primitive motions are measured. However, motion data, as well as speech data, has an inconvenient problem: even the same motions of the same person have some different moves and different finishing times. In the case of speech recognition, the same phrase has some different wave patterns and different finishing time even if the phrase is spoken by the same person in a similar way. In order to measure the distance between primitive motions, we must address this problem. Thus, we use Dynamic Time Warping (DTW) as a distance function[5], which is developed in the field of speech recognition. For computing the distance with aligning time lags, DTW uses dynamic programming techniques.

Next, we classify the similar primitive motions into a cluster, and consider that all the motions in a specific cluster represent the same primitive motion. The primitive motions are clustered by the nearest neighbor based algorithm as follows:

1. Primitive motions are given incrementally.
2. Distances between the new primitive motion and the already existing clusters are calculated.

3. The nearest distance is extracted and compared to a given threshold.
4. If the distance is smaller than the threshold, the new primitive motion is grouped into the cluster. Otherwise, a new cluster consisting only of the new primitive motion is created.

For extracting association rules from motion data, each primitive motion is labeled with one unique symbol corresponding to its cluster. Then, motion data can be converted to symbols sequences, and association rules are extracted from them. For example, in Fig. 2, the motion time series data of right wrist is converted to the symbols subsequence "$ABBBBDDDCCB\cdots$".

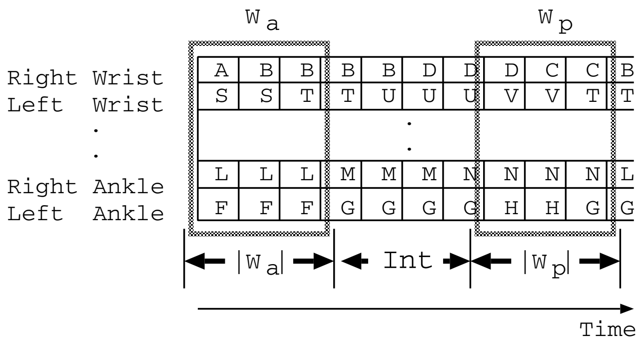


Fig. 2. Symbols sequences transformed from time series

To extract dependency relations, we will introduce two types of motions: motions which have an influence on other motions are called "*active motions*", and ones which are influenced by the active motions are called "*passive motions*". In order to find the dependency relation between active motions and passive motions, we use two windows W_a (active) and W_p (passive) with fixed interval length of Int. Fig. 2 presents an example. By sliding two windows on the symbol streams, combinations of two symbol patterns are extracted, and then the combinations that appear frequently are considered as association rules, which represent strong dependency relations among body parts.

We tested our algorithm on the motions of "rumba", a type of social dance. One of the association rules that has been found is shown below:

$$if \begin{matrix} {}'\,left\ ankle\ 166\,' \\ {}'\,right\ wrist\ 48\,' \end{matrix}\ then\ {}'\,left\ ankle\ 167\,' \quad (1)$$

This rule represents a dependency relation: "if one puts the center of gravity on the left foot while moving the right hand down, then one will move the left foot up". The motion, "moving the left foot up" after "putting the center of gravity on the left foot" precisely captures a characteristic of rumba. However, "moving the right hand down" is not a rumba's characteristic but a habit of the person being tested. We must take notice that the extracted association rules represent not only the dependency relations but also the habits of the person.

3 Motif Discovery Algorithm

There are two issues relative to the association rule method introduced in the previous section, that need to be addressed. One is that dividing a motion data by the breakpoints makes too many small primitive motions. Thus, a significant amount of computations is needed to find association rules, and only short and meaningless rules are found from small primitive motions. For example, we cannot find the association rule, such as "if one moves forward two steps, then one will move backward one step". Because a "step" movement is a frequent pattern composed of many primitive motions, and repetitions of this pattern cannot be detected if a single primitive motion differs. The other one is that a wrong division leads to wrong classifications, and eventually, "false" association rules are extracted from a motion data. This problem makes extracting the skill more difficult. In this section, we introduce an approach that extracts frequent patterns from time series[7], called motifs[9]. We extract motifs with an optimal length from motion data, and consider motifs as the characteristic motion.

3.1 Dimensionality Reduction of Motion Data

To discover motifs from motion data, some problems on multi-dimensional time series data must be addressed. One is that a huge amount of computation is required. Secondly, a more complex algorithm must be developed to deal with multi-dimensional time series data. In order to solve these problems, we reduce multi-dimensional time series data into one-dimensional time series data. This makes the motif discovery algorithm simpler, and decreases the computational cost.

However, in the dimensionality reduction process, the loss of the information must be as small as possible. At least, the information representing frequent characteristic patterns should be preserved. Therefore, we adopt Principal Component Analysis (PCA)[11]. PCA is a method to find the characteristic variables (Principle Component) in the multi-dimensional data. PCA is widely used in statistical field. If PCA is applied to m-dimensional time series data C of length n, m principle components are derived. The first principle component $\dot{C}$ contains most of the information of C. We use this first principle component $\dot{C}$ to find motifs instead of using the original time series C.

3.2 Symbolization of Motion Data

To make our approach effective, it is important to dynamically determine the length of motifs. Thus we use MDL principle[12] as a criterion in determinating this length. The MDL principle is used to estimate the optimality of a stochastic model. The MDL principle states that the best model to describe a set of data is the model that minimizes the description length of the entire data set. Here, for a time series data, we consider the best model as a motif.

In order to use the MDL principle to determine the length of the motif, our method transforms one dimensional time series data into a sequence of symbols that represent the dynamics of the data. In the transformation process, we introduce an algorithm based on SAX (Symbolic Aggregate approXimation) representation[10] (see Fig. 3).

SAX representation is a vector expression obtained by dividing a time series data into some segments and calculating the average value of each segment. By applying SAX representation, one-dimensional time series data $\dot{C} = \dot{c_1}, \cdots, \dot{c_n}$ can be represented as $\bar{C} = \bar{c}_1, \cdots, \bar{c}_w$. Note that $\bar{c}_1, \cdots, \bar{c}_w$ is the mean value of each segment, and the number of segments w is set by the user.

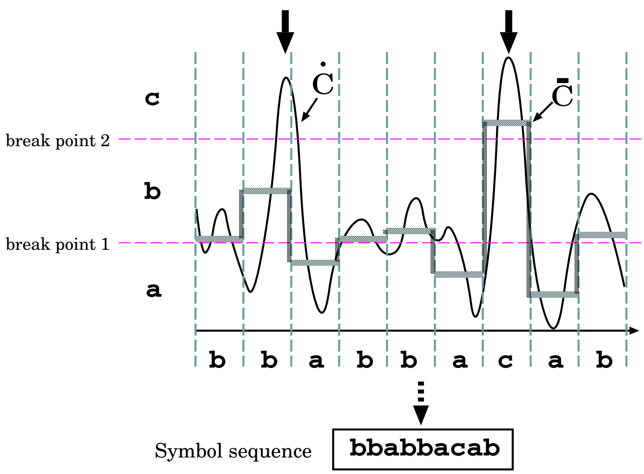


Fig. 3. Transformation time-series into a symbol sequence based on SAX representation

Next, mean values series $\bar{C}$ is transformed into a symbols sequence $\tilde{C}$ as follows: the mean values series are segmented into regions by "break points", and then an unique symbol is assigned to each region. Here, break points are determined such that each symbol has an equal probability of frequency in $\tilde{C}$[10], and the number of break points is set by the user. For example, in Fig. 3, two breakpoints are determined and three regions are provided. Then, every element of $\bar{C}$ is transformed into the symbol assigned to the region where the value of element belongs to. For example, the first element of $\bar{C}$ is transformed into "*b*", since it belongs to the region "*b*" between the break point 1 and 2. Eventually, $\bar{C}$ is transformed into the symbols sequence "*bbabbacab* $\cdots$".

3.3 Estimating Extracted Patterns Based on MDL Principle

For evaluating the optimality of the pattern extracted from the symbols sequence by the MDL principle, the description length of the pattern must be defined. We define the description length of a pattern denoted by SC as follows:

$$DL(SC) = log_2 n_p + n_p log_2 s_p \tag{2}$$

Note that n_p is the length of SC, and s_p is the number of unique symbols in SC.

In addition, the description length $DL(\tilde{C}|SC)$ needs to be defined. This is the description length of $\tilde{C}$ where SC is replaced by a single symbol. The description length $DL(\tilde{C}|SC)$ is calculated as follows:

$$DL(\tilde{C}|SC) = log_2 \acute{n}_a + \acute{n}_a log_2(s_a + q) \quad (3)$$

Here, $\acute{n}_a$ is the length of $\tilde{C}$, s_a is the number of unique symbols in SC, and q is the frequency of appearance of SC in $\tilde{C}$. In this equation, $log_2 \acute{n}_a$ represents the number of bits required to encode the number of symbols in $\tilde{C}$, and $\acute{n}_a log_2(s_a + q)$ represents the number of bits required to encode the number of unique symbols in $\tilde{C}$. Finally, MDL estimation function $MDL(\tilde{C}|SC)$ is defined as follows:

$$\begin{aligned} MDL(\tilde{C}|SC) &= DL(\tilde{C}|SC) + DL(SC) \\ &= log_2 \acute{n}_a + \acute{n}_a log_2(s_a + q) \\ &\quad + log_2 n_p + n_p log_2 s_p \end{aligned} \quad (4)$$

3.4 Discovering Motifs from Motion Data

In this section, we present the algorithm of discovering motifs from symbols sequences. By using sliding window technique, the most frequent pattern represented by a symbols subsequence can be easily extracted as follows: first, we prepare an "analysis window" with pre-defined length. By shifting the window on the symbols sequence, all "symbols subsequences" are extracted. Second, the frequency of appearance is counted for each symbols subsequence, and then we can find the most frequent pattern represented by symbols subsequences.

Then, MDL estimation function of the most frequent pattern is computed, and the frequent pattern is replaced by a single symbol. For example, in Fig. 3, the most frequent pattern "*bba*" is found from the symbols sequence "*bbabbacab* $\cdots$". Then MDL estimation function of the pattern "*bba*" is computed, and each pattern "*bba*" appearing in "*bbabbacab* $\cdots$" is replaced by a new symbol "*á*". Eventually, symbols sequence "*bbabbacab* $\cdots$" is transformed into "*áácab* $\cdots$". To find motifs even if the length of motifs is longer than the length of the analysis window, these processes are repeated until all patterns that are extracted from the symbols sequence are different from each other.

We consider the pattern that minimizes MDL estimation function as the motif candidate. The motif candidate may represent the characteristics of $\tilde{C}$ most accurately. But, we must take into account that the information of original the time series data is partly lost because of symbolization, and motif candidates must be reexamined with respect to the original time series data. In order to obtain the motif with accuracy, the ADM algorithm[13], a clustering algorithm for time series data, is applied to the subsequences of the original time series data corresponding to motif candidates. The distances between all the subsequences are

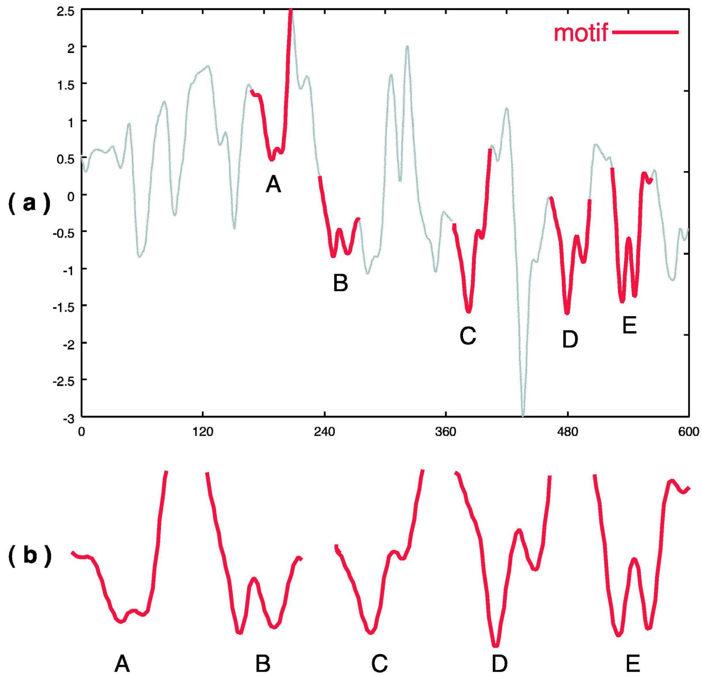


Fig. 4. (a)The motif discovered in an 1-dimension time series that represents "neck movement while running".(b)Normalized motif.

calculated using the Euclidean distance. Then, the subsequences are grouped in clusters such that the distance between any two elements in the same cluster is smaller than a pre-defined threshold. Finally, the subsequences belonging to the cluster that has the largest number of elements are extracted as motifs.

3.5 Experimental Results

We applied our algorithm to non-periodic time series data (see Fig. 4(a)). The motifs extracted from this data are depicted in thick line. At a first glance, it seems that the subsequence A and C have different shapes (Fig. 4(a)). The subsequence A and C have a rapid drop followed by a sharp increase, which generates a "V" shape. On the other hand, the other subsequences have the different shapes: they decrease rapidly, describe a parabola, and then finally increase sharply. This dynamics generate a "W" shape. However, all the normalized subsequences show same shapes which generate a "W" shape (Fig. 4(b)). This result proves our algorithm can extract the motifs effectively.

Next, we tested the validity of our algorithm for multi-dimensional time series data. Fig. 5 shows the motifs extracted from the "walking" motion data of right foot. In the Y-coordinate time series, each subsequence corresponds to the motif representing the characteristics of the motion appropriately, because they have

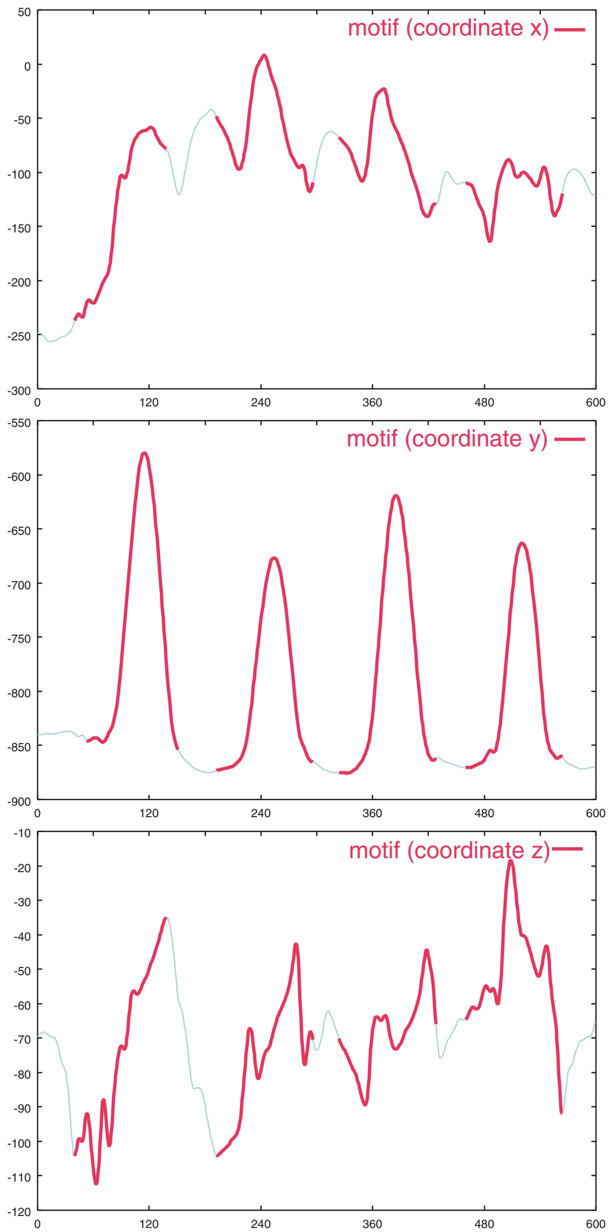


Fig. 5. An example of discovered motif from the original 3-dimensional time series, "feet movement while walking." From above, the figures represent coordinate "X", "Y","Z", respectively.

very similar shapes. On the other hand, in the X-coordinate and Z-coordinate time series, the shapes of first and fourth subsequences is different from second and third ones. With the data represented in Fig. 5, the PCA estimates that the Y-coordinate time series is the most characteristic element, whereas X-coordinate and Z-coordinate time series are not so. That is why the Y-coordinate time series is mainly focused on. This PCA's estimation can be verified from the standpoint of human motion analysis. The strongest characteristic of the "walking" motion appears in the Y-coordinate time series that represents vertical motion, next the X-coordinate time series data that represents left-and-right motion, finally the Z-coordinate time series data that represents backward-and-forward motion. This result shows that our algorithm can appropriately extract motifs from multi-dimensional time series data.

4 Comparison of the Skills

In the previous section, we have discussed about finding the motion characteristics of a person. In this section we discuss the differences between the motions of different persons. That is, we want to extract the motion characteristics of skilled sport players and beginners. For improving the beginner's skill, there are two problems that must be addressed. First, it is indispensable for him/her to learn the rhythm of the motion. To take a simple example, a tennis novice cannot make his/her best swing even if he/she traces skilled person's motions perfectly because he/she cannot adjust the timing of his/her maximum power. Second, an accurate synchronization of all body parts is indispensable to perform a good motion. To solve these problems, we had two experiments: first, we analyzed the relationship between trajectory and velocity. Then, we analyzed the posture of the motion to reveal the relationship between body parts. Finally, at the end of this section, we introduce our skill extraction algorithm that helps us to find the differences between the motions of different persons.

4.1 Analysis of Motion Velocity

In this section, we analyze the relationship between the trajectory and the velocity of the right hand in a tennis service action. The motion data is captured from 4 persons including 2 beginners and 2 skilled players by our motion capture system. Fig. 6 shows that there is a clear difference of timing between beginners and skilled players: the beginner's tennis racket reaches the peak of velocity after it is swung down, while the skilled player's racket reaches the peak when it is swung up. That is, the beginners cannot make their most powerful shot because of bad timing, while the skilled players can do it on exact timing. However, the differences between beginners and skilled players are not only timing. In the next section, we analyze the motion data from a different point of view.

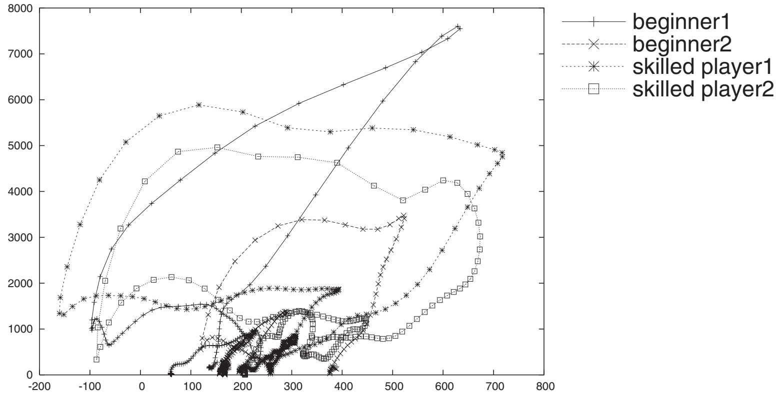


Fig. 6. The relation between velocity and coordinate of right hand movement while servicing

4.2 Analysis of Posture

A motion is a combination of the movements of all body parts. Therefore, the relationship between the body parts, or the posture is one of the key factors for explaining the skill. In this section, we analyze the posture of the motions of tennis service used in the previous section. Fig. 7 tells us that the skilled players swing their arms in different ways than the beginners do in two points: first, the beginner's graphs have long-wise ellipse form, while the skilled player's graphs have a sector form. This sector form means that the skilled players swing down the racket with their arms stretched, and their services are shot from a higher position, having more chances to make a better serve. Second, the skilled player's graph has a small round shape on the top right. The skilled players bring up their rackets with their arm bended in a slower motion preparing the hitting action carefully and then, suddenly accelerate in a powerful hitting movement. On the other hand, the beginner does not perform such a movement. This difference is made by the fact that by bending their arm, one can accelerate the racket for a longer time while moving the hand forward.

4.3 Extracting Characteristic Motions

As we explained in the previous section, the differences between the skilled players and the beginners are the key factors for explaining skills. In this section, we present an algorithm that help us to find such differences by visualization techniques. By dividing motion data into subsequences, and then mapping them to 3D space by MDS (Multi Dimensional Scaling)[15] techniques, the similarity relationship among them can be visualized. MDS is a technique that can reveal the structure of multi-dimensional data by mapping data to lower dimension space. Then we can easily find the differences among the motion data. The algorithm is described follows.

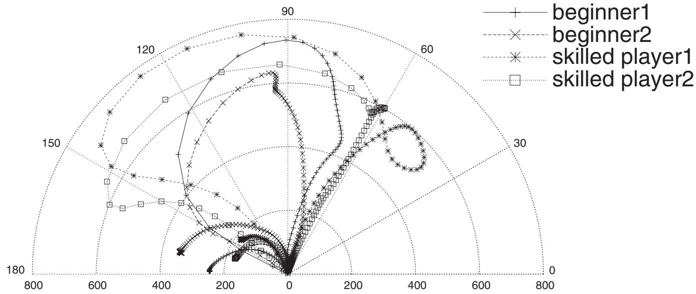


Fig. 7. The posture differences between skilled players and beginners

Let C be a time series data that represents the movement of a person's body part:

$$C = c_1, c_2, \ldots, c_N \tag{5}$$

$$c_k = (x_k, y_k, z_k) \tag{6}$$

Note that x_k, y_k, z_k are X-coordinate, Y-coordinate and Z-coordinate of the body part respectively. By using a sliding window that has a pre-defined length of W, C is divided into m subsequences $C^1, C^2, \ldots, C^n$:

$$\begin{aligned} C^1 &= c_1, c_2, \ldots, c_n \\ C^2 &= c_{1+W_{int}}, c_{2+W_{int}}, \ldots, c_{n+W_{int}} \\ &\vdots \\ C^n &= c_{1+(n-1)W_{int}}, c_{2+(n-1)W_{int}}, \ldots, c_N \end{aligned}$$

Note that W_{int} is the interval between sliding windows.

Then, the similarities between all pairs of subsequences are calculated. To calculate the similarities properly, we must decide what similarity (or dissimilarity) measure we use. We want the similarity measure that can calculate how similar the body part's movements are. For our purpose, the Euclidean distance that is a standard distance (= dissimilarity) measure is not appropriate because it uses only the coordinates of body part's movement, and it is strongly affected by "where players are standing" than "how players move their body part". Thus, instead of using the Euclidean distance, the similarity is calculated by the following method: assume that C^a and C^b are subsequences of motion time series data.

$$\begin{aligned} C^a &= c_1^a, c_2^a, \ldots, c_W^a \\ C^b &= c_1^b, c_2^b, \ldots, c_W^b \end{aligned}$$

To remove effects of the coordinates, C^a and C^b are transformed to the vector sequences $\hat{C^a}$ and $\hat{C^b}$:

$$\begin{aligned} \hat{C^a} &= \overrightarrow{c_1^a c_2^a}, \overrightarrow{c_2^a c_3^a}, \ldots, \overrightarrow{c_{W-1}^a c_W^a} \\ \hat{C^b} &= \overrightarrow{c_1^b c_2^b}, \overrightarrow{c_2^b c_3^b}, \ldots, \overrightarrow{c_{W-1}^b c_W^b} \end{aligned}$$

Then, the similarity between vector sequences is measured by summing up similarities between vectors of each sequence. By using dynamic programming techniques, the pairs of corresponding vectors are determined so as to maximize the total similarity. The similarity between vectors is calculated as follows:

$$Sim(\vec{v_1}, \vec{v_2}) = Dist(\vec{v_1}, \vec{v_2}) \times Weight(\vec{v_1}, \vec{v_2}) \times (-1) \quad (7)$$

$$Dist(\vec{v_1}, \vec{v_2}) = \frac{Dist_{angle}(\vec{v_1}, \vec{v_2}) + Dist_{length}(\vec{v_1}, \vec{v_2})}{2} \quad (8)$$

$$Dist_{angle}(\vec{v_1}, \vec{v_2}) = \frac{\theta}{2\pi} \quad (9)$$

$$Dist_{length}(\vec{v_1}, \vec{v_2}) = \frac{||\vec{v_1}| - |\vec{v_2}||}{Max(|\vec{v_1}|, |\vec{v_2}|) \times LegthThreshold} \quad (10)$$

$$Weight(\vec{v_1}, \vec{v_2}) = \left(\frac{|\vec{v_1}|}{\sum_k^{V^1} |\vec{v_k}|} + \frac{|\vec{v_2}|}{\sum_k^{V^2} |\vec{v_k}|}\right) \times \frac{1}{2} \quad (11)$$

Note that $\vec{v_1}$ and $\vec{v_2}$ are the vectors which are contained in the vector sequence V^1 and V^2 respectively, and θ is the angle between v_1 and v_2. $Sim(\vec{v_1}, \vec{v_2})$ and and $Dist(\vec{v_1}, \vec{v_2})$ are the similarity and the distance between $\vec{v_1}$ and $\vec{v_2}$ respectively. $Dist_{angle}(\vec{v_1}, \vec{v_2})$ means the difference of the angles, and $Dist_{length}(\vec{v_1}, \vec{v_2})$ means the difference of the length of vectors. *LengthThreshold* is the threshold of difference of the lengths. $Weight(\vec{v_1}, \vec{v_2})$ represents how $Sim(\vec{v_1}, \vec{v_2})$ accounts for total similarity.

After the similarities between all pairs of vector sequences are calculated, we visualize the similarity relationship among all vector sequences by using Principal Coordinates Analysis[14]. Principal Coordinates Analysis is a method of MDS, which can visualize the similarity relationship between multi-dimensional data. The coordinates used in visualization are computed as follows: first, the similarity matrix S is prepared.

$$S = \begin{bmatrix} s_{11} & s_{12} & \cdots & s_{1j} & \cdots & s_{1n} \\ s_{21} & s_{22} & \cdots & s_{2j} & \cdots & s_{2n} \\ \vdots & \vdots & \ddots & \vdots & & \vdots \\ s_{i1} & s_{i2} & \cdots & s_{ij} & \cdots & s_{in} \\ \vdots & \vdots & & \vdots & \ddots & \vdots \\ s_{n1} & s_{n2} & \cdots & s_{nj} & \cdots & s_{nn} \end{bmatrix} \quad (12)$$

Note that s_{ij} is the similarity between the ith vector sequence and the jth vector sequence. Secondly, the similarity matrix S is transformed by an operation called "double centering" described in equation (13) and (14). The resulted matrix, called $\bar{S}$, has an eigenvalue 0 and the corresponding eigenvector is $[1, 1, \ldots, 1]^t$.

$$\bar{S} = (\bar{s_{ij}}) \quad (13)$$

$$\bar{s_{ij}} = s_{ij} - \frac{1}{n}\sum_k s_{ik} - \frac{1}{n}\sum_k s_{kj} + \frac{1}{n^2}\sum_k \sum_l s_{kl} \quad (14)$$

Finally, eigenvalues and eigenvectors of $\bar{S}$ are calculated. The elements of eigenvector that corresponds to the largest eigenvalue represent X-coordinates used in the visualization. That is, an eigenvector $E = [e_1, e_2, \ldots, e_n]^t$ gives the coordinates of the kth vector sequence as e_k. In the same way, Y-coordinates and Z-coordinates are given by the eigenvectors corresponding to the second largest eigenvalue and the third largest eigenvalue respectively.

We applied our algorithm to the motion data of tennis service that is used in the previous section. Fig. 8 shows the similarity relationship between the

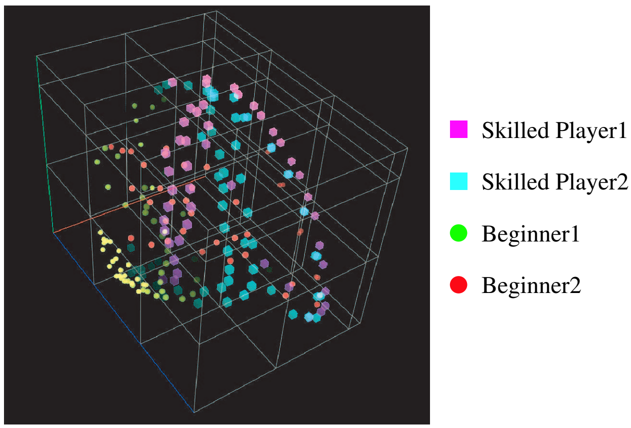


Fig. 8. 3D plotted points of subsequence

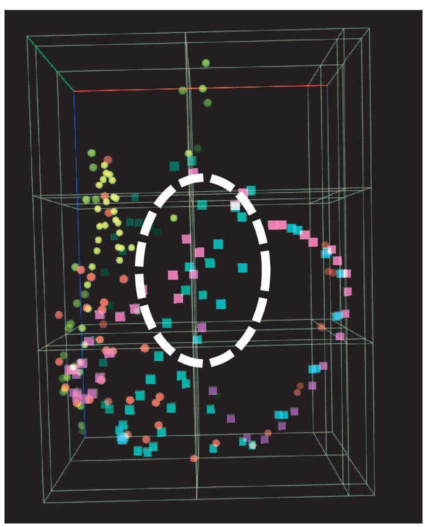


Fig. 9. The group consists mainly of skilled player's points

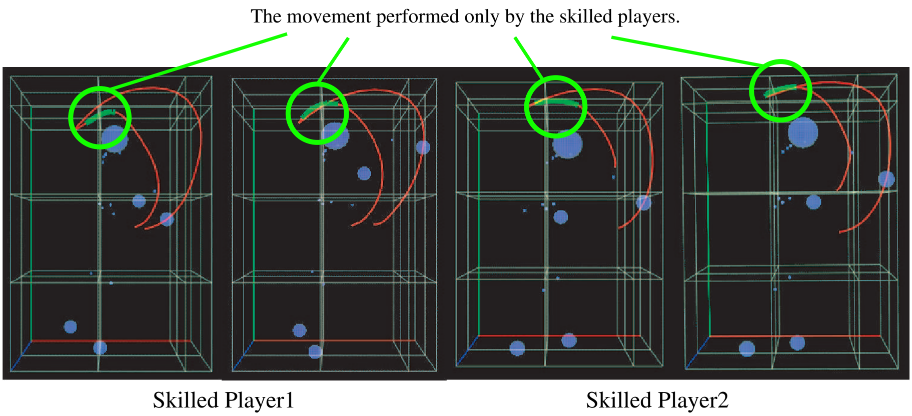


Fig. 10. The movement performed by the skilled players

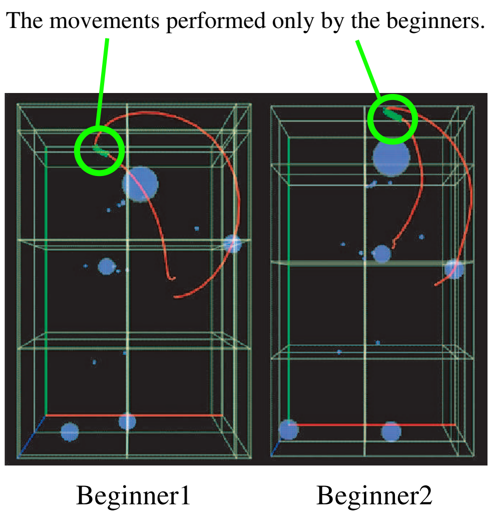


Fig. 11. The movement performed by the beginners

subsequences of motion time series data that represents the movement of right hand. Square point and round point represent a subsequence of skilled player's motion and a subsequence of beginner's motion respectively. The shorter distance between points means that the subsequences corresponding to them have more similar movements. To find the differences between skilled players and beginners, the distributions of the skilled player's points and of the beginner's points are compared. We can find three types of groups from Fig. 8: a group that consist mainly of skilled player's points, a group that consist mainly of beginner's points, and a group that skilled player's and beginner's points are equally contained. The first group represents movements performed mainly by skilled players. Such a group is circled in Fig. 9, and the group corresponds to the movement represented by the thick line in Fig. 10. This means that only skilled tennis players move their racket back strongly before they hit the ball.

On the other hand, the movement that only beginners do is represented by thick line in Fig. 11. It is obvious that beginners do not move their right hand like the skilled players do. This result demonstrates that our algorithm can help us to easily find the differences between skilled players and beginners easily.

5 Conclusion

In this work we have presented two algorithms that can analyze the structure of human motion. First, we introduced an algorithm for extracting association rules between the movements of human body parts. Then, we introduced an algorithm for extracting motifs from multi-dimensional time series data. In addition, we tried to find the differences of the skill between the skilled players and the beginners with sports science viewpoint. In future works, we intend to develop a technique for finding the technical differences from beginner's motion data automatically, and to create a system that can advise them to improve their perfomances.

References

1. D. M. Gavrila and L. S. Davis, "3D Model-based Tracking of Humans in Action: A Multi-view Approach", Proc. of Computer Vision and Pattern Recognition, pp.73-80 (1996).
2. G. R. Bradski and J. Davis, "Motion Segmentation and Pose Recognition with Motion History Gradients", Proc. of IEEE Workshop on Detection and Recognition of Events in Video (2001).
3. A. F. Bobick and J. Davis, "The Recognition of Human Movement Using Temporal Templates", IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol.23, No.3, pp.257-267 (2001).
4. G. Widmer, S. Dixon, W. Goebl and E. Pampalk, "In Search of the Horowitz Factor", AI Magazine, Vol.24, Issue 3, pp.111-130 (2003).
5. R. Osaki, M. Shimada and K. Uehara, "A Motion Recognition Method by Using Primitive Motions", Proc. of 5th IFIP 2.6 Working Conference on Visual Database Systems, pp.117-128 (2000).
6. T. Mori and K. Uehara, "Extraction of Primitive Motion and Discovery of Association Rules from Human Motion", Proc. of 10th IEEE International Workshop on Robot and Human Communication, pp.200–206 (2001).
7. Y. Tanaka and K. Uehara, "Discover Motifs in Multi Dimensional Time-Series Using the Principal Component Analysis and the MDL Principle", Proc. of 3rd International Conference on Machine Learning and Data Mining in Pattern Recognition, pp.252-265 (2003).
8. Y. Tanaka, K. Iwamoto and K. Uehara, "Discovery of Time-Series Motif from Multi-Dimensional Data Based on MDL Principle", Machine Learning (to appear).
9. J. Lin, E. Keogh, S. Lonardi and P. Patel, "Finding Motifs in Time Series", Proc. of the 2nd Workshop on Temporal Data Mining, pp.53-68 (2002).

10. J. Lin, E. Keogh, S. Lonardi and B. Chiu, "A Symbolic Representation of Time Series with Implications for Streaming Algorithms", Proc. of the 8th ACM SIGMOD Workshop on Research Issues in Data Mining and Knowledge Discovery, (2003).
11. D. B. Heras, J. C. Cabaleiro, V. B. Perez, P. Costas and F. F. Rivera, "Principal Component Analysis on Vector Computers", Proc. of VECPAR, pp.416–428 (1996).
12. J. Rissanen, "Stochastic Complexity in Statistical Inquiry", World Scientific Publishing (1989).
13. D. Shasha and T. Wang, "New Techniques for Best-Match Retrieval", ACM Trans. Information Systems, Vol.8, No.2, pp.140–158 (1990).
14. J. C. Gower, "Some Distance Properties of Latent Root and Vector Methods Used in Multivariate Analysis", Biometrica, Vol.53, pp.325–328 (1966).
15. J. B. Kruskal and M. Wish, "Multidimensional Scaling", Sage Publication (1978).

Emergence and Evolution of Linguistic Communication

EELC 2004, the First International Workshop on Emergence and Evolution of Linguistic Communication, was held in association with JSAI annual conference 2004, in Kanazawa, Japan.

Because of this colocation the first workshop was planned only by Japanese though of course the call for submissions was announced worldwide. The program committee were as follows; Chair: Satoshi Tojo (JAIST), Co-chair: Koiti Hasida (AIST), Takaya Arita (Univ. Nagoya), Takashi Hashimoto (JAIST), Takashi Ikegami (Univ. Tokyo), Tetsuo Ono (Future Univ. Hakodate), and Akito Sakurai (Keio Univ.).

The followings are our motivation. Rules of natural languages such as usage, grammar, and vocabulary change diachronically dependent upon the social situations of the language community. This workshop focused on those language phenomena concerning language changes and evolution, that is, emergence, pidginization, and creolization, from the viewpoints of social, evolutionary, computational linguistics. Thus, we expected that the workshop would contribute to the joint discussion among those who share this common interest.

Relevant themes were stated as Language change, Language emergence, Language acquisition, Second language acquisition, Multi-agent model of communication, Lingua Franca, Pidgin and creole, and other computer simulation concerning language dynamics.

Thus far, the similar topics have been included in EVOLANG (International Conference on the Evolution of Language), though we intended to concentrate more on linguistic aspects of human and/or artificial agents' communication.

At the workshop, J. R. Hurford of University of Edinburgh, UK, was invited to give a talk. Besides, fourteen technical papers were accepted to present, the topics of which were diverse including many branches of linguistic communication.

In this volume, we will show the selected papers among them.

July 2005 Satoshi Tojo

A. Sakurai et al. (Eds.): JSAI 2003/2004, LNAI 3609, p. 285, 2007.

The Emergence of Language in Grounded Adaptive Agents and Robots

Angelo Cangelosi[1], Thomas Riga[1], Barbara Giolito[1,2], and Davide Marocco[1,3]

[1] Adaptive Behaviour & Cognition Research Group
School of Computing, Communications and Electronics
University of Plymouth (UK)
acangelosi@plymouth.ac.uk
http://www.tech.plym.ac.uk/soc/research/ABC/
[2] Department of Human Studies
Universita' del Piemonte Orientale - Vercelli (Italy)
barbara.giolito@infinito.it
[3] Artificial Life and Robotics Lab
Institute of Cognitive Sciences and Technology - CNR Rome (Italy)
davide.marocco@istc.cnr.it

Abstract. We present a computational modeling approach to language based on an integrative view of the agent's cognitive system. The emergence of linguistic abilities (both evolutionarily and developmentally) is strictly dependent on, and grounded in, other sensorimotor behaviors and cognitive abilities. Linguistic simulations imply the use of groups of autonomous agents that interact via language games to exchange information about the environment. The agents' coordinated communication system is not externally imposed by the researcher, but emerges from the interaction between agents. We present a series studies on grounded simulation adaptive agent and on evolutionary and epigenetic robots.

1 Modeling Language with Grounded Agents and Robots

The grounding of linguistic symbols in the organism' cognitive system, and indirectly in the physical and social environment in which individuals live, is one of the most important issues in recent experimental and computational approaches to language. This is normally referred as the Symbol Grounding Problem [14]. In cognitive science, psychological experiments have focused on the relationship between language and perception [2,11] and language and action [13]. These empirical investigations show a strong interdependence between language development and perceptual and embodiment factors. In robotics and artificial intelligence, various models have been proposed to ground language in neural networks [5,12,16] and in interactive robots [23,24,25]. Moreover, computational approaches to the evolution of language [7] commonly deal with the issue of symbol grounding.

In this paper we present a computational modeling approach to language based on an integrative view of the agent's cognitive system. This work has mainly been

A. Sakurai et al. (Eds.): JSAI 2003/2004, LNAI 3609, pp. 286 – 294, 2007.

developed at the "Adaptive Behaviour & Cognition" research group of the University of Plymouth. All models reviewed here are characterized by the fact that the emergence of linguistic abilities (both evolutionarily and developmentally) is strictly dependent on, and grounded in, other behaviors and abilities. These include sensorimotor skills (e.g. action categories), cognitive abilities (e.g. categorical perception), neural mechanisms, and social and evolutionary factors. Agents are able to build an intrinsic link between the linguistic symbols (words) they use to communicate and their own sensorimotor and cognitive representations (meanings) of the perceptual and sensorimotor interaction with the external world (referents). This approach is called "grounded adaptive agent modeling" for the emergence of language [4]. Linguistic simulations with such agents imply the use of groups of autonomous agents that interact via language games to exchange information about the environment. It also entails the fact that their coordinated communication system is not externally imposed by the researcher, but emerges from the interaction between agents. The paper will present a series of past and current studies on grounded adaptive agent models. Some will be based on simulated sensorimotor agents, others on evolutionary and epigenetic robots.

2 Emergence of Nouns and Verbs in Adaptive Agents

Models based on adaptive agents simulate a multi-agent scenario in which individuals use a lexicon to communicate about their environment. Various adaptive agent models have been used to model the evolutionary emergence of language (e.g. [3,15]). Here we focus on grounded agent models, where the individual's neural network is used to control all sensorimotor, cognitive and linguistic abilities. In previous studies by Cangelosi and Parisi [6] we showed that agents were able to evolve the ability to use a syntactic lexicon consisting of proto-verbs (names of actions) and proto-nouns (names of objects). The lexicon was externally imposed by the researcher. In these models, the analysis of the agents' neural representations can highlight the neural mechanisms responsible for the integration of language, cognition and action. For example, categorical perception analyses [4,5,6] have shown that when agents use verbs and nouns, the similarity space of the representations of verbs is enhanced and optimized with respect to that of nouns. In addition, synthetic brain imaging techniques [1,8] have shown that the neural representations of syntactic word classes are sensitive to the level of integration of linguistic information and sensorimotor knowledge. In particular, verbs produce more neural activity in the regions of the network that specialize for the integration of sensorimotor information. Nouns are more active in the modules dedicated to the processing of sensory (e.g. visual) information only.

Recent simulations have extended this model to include (i) the cultural acquisition of language through communication with other agents, (ii) the linguistic production of nouns and verbs. The main aim of such a simulation consists in testing some hypothesis on the emergence of nouns and verbs. We hypothesize that there is a semantic contribution to the grammatical distinction between nouns and verbs. In particular, we suggest that some semantic aspects of the grammatical category of nouns are related to the objects to which nouns refer, and that some semantic aspects of the grammatical category of verbs are related to the actions that subjects are performing. In other words, our hypothesis is that the distinction among nouns depends - at least in part - on the variations of the objects with which organisms are interacting, while the distinction among verbs depends - at least in part - on the variations in the actions that organisms are performing on these objects. A term is a noun because its semantic aspect depends on the objects it refers to, whereas another term is a verb because its semantic aspect depends on the actions it refers to. Our intention is not to deny the existence of an additional 'grammatical' basis for the distinction between nouns and verbs, but to suggest that a part of this distinction is based on a sort of semantic differentiation. Moreover, we suppose that such a differentiation originates from the evolutionary interaction among organisms and between the agents and their environment.

To test such a hypothesis, we used agents whose neural controller consists of: 4 proprioceptive units, 25 visual units and 4 language units in the input level; 2 layers of hidden units; and 8 output units (4 motor and 4 linguistic units). Agents are subject to an evolutionary algorithm for 5000 generations. Each generation consists of a no-language object manipulation task (push/pull 2 different objects) and 10 other linguistic epochs (various combinations of visual images and noun/verb linguistic inputs). During the first 1000 generations, agents are evolved solely for their ability to perform the object manipulation task. The linguistic tasks are introduced at generation 1001 and last until generation 5000.

Preliminary results on the emergence of nouns and verbs indicate a tendency to evolve a lexicon based on both verbs and nouns. Although only 2 populations out of 15 evolve a full verb-noun lexicon (see Figure 1 for a sample population), in all other populations a great deal of correct verbs or nouns also emerges. These results show that agents evolve an ability to differentiate between the two word classes. This confirms that at least a partial grounding on the noun-verb distinction can be originated from the evolutionary use of sensorimotor information and interaction with the environment. New simulations are currently underway to improve the robustness of the results and produce more verb-noun languages. These will mainly focus on the modification of the neural network architecture, as suggested in a related model on verb-noun control in modular neural networks [4].

3 Emergence of Communication in Evolutionary Robots

Evolutionary robotics has been successfully applied to the synthesis of robots able to exploit sensorimotor coordination [18] and signaling capabilities [20]. Most of the properties of evolutionary robotics (e.g. sensorimotor coordination, social interaction, evolutionary dynamics, use of neural controllers) also are beneficial for modeling the emergence of grounded communication. For example, in recent simulations, robotic agents first evolve an ability to create action categories by interacting with physical objects.

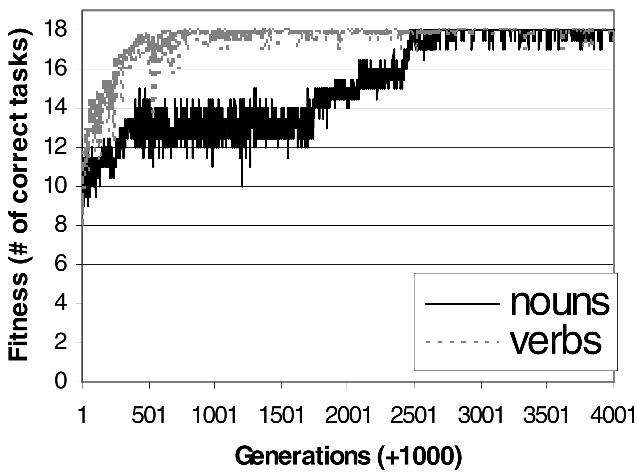


Fig. 1. Emergence of a verb-noun lexicon in a population of adaptive agents

Recent linguistic experiments by Marocco, Cangelosi and Nolfi [17] have shown that the ability to form categories from direct interaction with the environment constitutes the ground for subsequent evolution of names of objects (e.g. nouns). In this model, agents are implemented with simulated robots provided with a 3-segments arm with 6 degrees of freedom (DOF) and extremely coarse touch sensors. Agents use proprioceptive information behavior to actively explore the environment and build categories, on the basis of tactile information, and the communication, about the type of objects that are in it. The environment consists of an open three-dimensional space in which one of two different objects is present in each epoch. The two objects used in this simulation are a sphere and a cube.

The sensory system consists of a simple contact sensor placed on the body that detects when this body collides with another and proprioceptive sensors that provide the current position of each joint of the arm. The controller of each individual consists of an artificial neural network in which, in addition to proprioceptive sensors, two sensory neurons receive their input from the other agents. The output layer has motor neurons, which control the actuators of the corresponding joints, and two additional output neurons, which encode the signal to be communicated to the other agents. The two linguistic units work as a small winner-takes-all cluster, where the neuron with the highest activation is set to 1 and the other to 0. This means that, in addition to the proprioceptive information, agents also receive in input a 2-bit signal produced by some other agent in the population, such as the parent or any agent from the population (linguistic comprehension task). The protocol of interaction and communication between agents was systematically changed. Before they act as speaker, agents undergo a linguistic production task. That is, each agent is put in the environment and asked to interact with the object. The value of the two output neurons in the last cycle of the epoch is saved and used as the signal produced to "name" the object. A genetic algorithm is used to evolve the behavior of agents.

The evolutionary robotics model was used to run a series of experiments on the role of various social and evolutionary variables in the emergence of shared communication. The first independent variable of the experimental design is the selection of speakers: each agent receives communication signals solely from its own parent or from any individual of the population. This looks at the role of different social groups of speakers in facilitating shared communication. The second independent variable is the time period in which communication is allowed: agents can communicate right from the initial random generation or only after the pre-evolution of the ability to touch/avoid the two objects. Through this variable it will be possible to investigate the initial behavioral and cognitive abilities necessary to evolve communication. The simulation results show that populations evolve stable shared communication (i.e. using two different signals are produced for the two different objects) mostly when the parents act as speakers and when signaling is introduced in the second stage [17]. Additional analyses of results support the following findings: (a) the emergence of signaling brings direct benefits to the agents and the population, in terms of increased behavioral skill and comprehension ability (but the agents' fitness does not assess the ability to communicate well); (b) there is a benefit in direct communication between parents and children, not only because of kinship mechanisms, but also because parents produce more stable and reliable input signals; (c) the pre-evolution of good sensorimotor and cognitive abilities permits the establishment of a link between production and comprehension abilities, especially in the early generations when signaling is introduced.

A second model has been tested, in order to simulate the emergence of different types of syntactic categories. In the previous model, the signal associated to each object can be simultaneously interpreted as the name of the object ("noun") or as the name of the action ("verb"). For example, when a parent produces the signal X at the end of its interaction with the sphere, the child can interpret it as "sphere" (which indicates the object) or as "touch" (which indicates the action to be performed). In the new model, the touch/avoid actions are not rigidly linked to any specific object. The agents can touch and avoid both the sphere and the cube, depending on the task context. This is defined by the parent's language and the fitness formula. When a parent produces a signal X after having touched the sphere, the child must interpret it as the "touch" verb and touch any object regardless of its shape. Thus this signal can be clearly considered as the name of the action, which is a typical case of verbs. Our aim was the identification of significant differences between the verb lexicon of the second simulation and the signaling system in the first simulation. Such comparisons provided useful insights on the evolutionary transition from signaling to syntactic languages reinforcing the fact that the pattern of results in the first simulation correspond to that of the evolution of verbs, as observed in the adaptive agent model of verbs and nouns [6].

4 Imitation and Language in Epigenetic Robots

More recently, new experiments by Riga and Cangelosi [9] have focused on the developmental emergence of communication in epigenetic robots. In this study simulated robots observe and execute actions via imitation learning, while using an artificial language to communicate about the names of actions and objects. Analyses of the model, through synthetic brain imaging techniques, highlight the specific role of sensorimotor knowledge in the processing of the words for actions.

We developed a computer simulation of two robots embedded in a virtual reality world (Figure 2), that accurately models physical constraints present in real-world situations, using the physics engine Open Dynamics Engine. Each robot has 12 degrees of freedom and consists of two 3-segment arms attached to a torso and a base with 4 wheels. The teacher robot has preprogrammed behavior to manipulate objects; the imitator agent learns to perform the same actions by observing the teacher executing them. The imitator approximates the teacher's movements using an on-line mimicking algorithm, resulting in movement dynamics that are processed by a neural network, which memorizes action patterns and enables for their autonomous execution. Visual preprocessing is simulated, thus the agent receives joint angles directly in input instead of analyzing visual input to compute joint angles. This integrated system models the mechanism responsible for matching perceived actions with equivalent self-generated actions and the way in which actions are memorized and successively reproduced autonomously. It addresses the question of how agents learn to perform actions on objects using simple imitative mechanisms like mimicking.

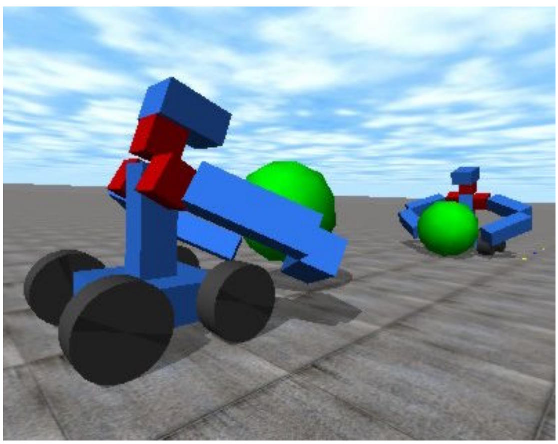


Fig. 2. Imitator and learning robots in epigenetic robot simulation

In a first simulation the imitator agent learned to perform actions when receiving a linguistic description of them. Furthermore it learned to give a linguistic description of actions performed by the teacher agent. The neural network controlled both motor and linguistic behavior. Descriptions consisted of a verb, indicating the action, and a noun, referring to the object involved. Brain imaging studies (Pulvermueller, 2003; Cappa & Perani, 2003) reveal that language comprehension and production activates different brain areas for verbs and nouns. In particular they indicate that the left temporal neocortex exhibits strong activation in tasks involving lexical-semantic processing, while additional regions of the left dorsolateral prefrontal cortex are recruited during the processing of words related to actions. We applied the synthetic brain imaging techniques initially develop by Arbib and colleagues [1] to analyze the internal neural network structure that emerged during training.

The results showed that noun processing involved the area responsible for object recognition while verb processing recruited neurons in both the object recognition area and the area responsible for motor program execution. The neurons in the object recognition area differentiated, by specializing in either noun processing or motor program execution. The fact that the agents, in absence of a linguistic input, performed a different default action for every object caused the object recognition area to have a double function: categorizing the objects and bootstrapping a default action in absence of linguistic input. This simulation supports the view that language builds on existing neural structures responsible for action execution and observation [13,22]. Linguistic representations are embodied: they build on sensorimotor representations. In fact, embodiment effects can be detected when the imitator agent observes its teacher: it performs very small movements in synchrony with the observed action.

During the second simulation agents learned to perform basic actions by mimicking them, while simultaneously learning words corresponding to these actions. Furthermore they learned higher-level composite behaviors by receiving linguistic descriptions containing these previously acquired words. The agents merged basic actions into a composite action by transferring the neural grounding of the words referring to basic actions to the word indicating the higher-level behavior. This process of grounding transfer [21] grounds words, known exclusively from linguistic descriptions, on the neural level by adapting neural activations of the words contained in the description.

The imitator robot, during training, learned the basic actions of opening and closing their left and right arms (upper arms & elbows), lifting them (shoulders), and moving forward and backward (wheels), together with the corresponding words. At the 50th epoch it received 1st level linguistic descriptions of combined actions, consisting in a new word and two known words referring to basic actions. A combined action consisted for example in grabbing the object in front of them and was described like: “close_left + close_right = grab”. Grounding was transferred from “close_left” and “close_right” to “grab”. Consequently, when the agent was given the command “grab” it successfully executed the combined action of pushing its arms towards the object and grabbing it. At the 100th epoch it received second level descriptions, in which a defining word was itself learned exclusively from a linguistic description. Following the example of before, we combined grabbing and moving forward into carrying: “move_forward + grab = carry”. Also at this level grounding was successfully transferred to the new word, enabling the agent to correctly perform the action of carrying on hearing the word “carry”: it pushed both arms against the object and moved forward, effectively exhibiting the behavior of carrying the object. The system learned several of these combined actions simultaneously, and also four-word definitions and grounding transfers of up to three levels have been realized.

The second simulation sheds light on language as a cognitive enhancer, as a means through which new behaviors can be acquired quickly and effortlessly, building on experience accumulated by previous generations of agents. The importance of cultural transmission in cognitive development is highlighted. Our long-term goal is to develop a framework for training robots by demonstration, using both imitation and a natural language inter-face, enabling for a neuro-robotic approach to investigating imitation as a precursor of communication.

5 Conclusion

Adaptive agent models and evolutionary and epigenetic robots can significantly contribute to a better understanding of the strict interdependence between language and perceptual, motor and cognitive capabilities. Such models of language emergence have important scientific and technological implications for research in language and communication. In robotics and artificial intelligence, they provide new approaches and algorithms for the development of autonomous interactive systems. In cognitive science, these models permit a deeper understanding of the psychological and cognitive bases of language and its grounding in perceptual and sensorimotor abilities. Finally, in linguistics and other disciplines interested in language origins, agent and robotics models allow the simulation of the evolutionary emergence of language and the test of language origin hypothesis [7].

References

1. Arbib M.A., Billard A., Iacoboni M., Oztop E. (2000). Synthetic brain imaging: grasping, mirror neurons and imitation. *Neural Networks*, 13: 975-997.
2. Barsalou L. (1999). Perceptual symbol systems. *Behavioral and Brain Sciences*, 22: 577-609.
3. Cangelosi A. (2001). Evolution of communication and language using signals, symbols, and words. *IEEE Transactions in Evolutionary Computation*. 5(2): 93-101
4. Cangelosi A. (2004). The sensorimotor bases of linguistic structure: Experiments with grounded adaptive agents. In S. Schaal et al. (Eds.), *Proceedings of the Eighth International Conference on the Simulation of Adaptive Behaviour: From Animals to Animats 8*, Cambridge MA, MIT Press, pp. 487-496
5. Cangelosi A., Greco A., Harnad S. (2000). From robotic toil to symbolic theft: Grounding transfer from entry-level to higher-level categories. *Connection Science*, 12(2): 143-162
6. Cangelosi A., Parisi D. (2001). How nouns and verbs differentially affect the behavior of artificial organisms. In J.D. Moore, K. Stenning (Eds.), *Proceedings of the 23rd Annual Conference of the Cognitive Science Society*, London: LEA, 170-175.
7. Cangelosi A., Parisi D. (2002). *Simulating the Evolution of Language*. London: Springer.
8. Cangelosi A., Parisi D. (2004). The processing of verbs and nouns in neural networks: Insights from synthetic brain imaging. *Brain and Language*, 89(2): 401-408
9. Cangelosi A. & Riga T. (submitted). The grounding of language in sensorimotor agents and robots. *Artificial Intelligence Journal*, Special issue on grounding language in the real world,
10. Cappa S.F., Perani, D. (2003). The neural correlates of noun and verb processing. *Journal of Neurolinguistics*, 16 (2-3): 183-189.
11. Coventry K.R., Garrod S.C. (2004). *Saying, Seeing and Acting: The Psychological Semantics of Spatial Prepositions*. Psychology Press. Hove and New York
12. Dyer M.G. (1994). Grounding language in perception. In V. Honavar, L. Uhr (Eds.), *Artificial Intelligence and neural networks: Steps toward principled integration*. Boston: Academic Press.

13. Glenberg A., Kaschak M. (2002). Grounding language in action. *Psychonomic Bulletin & Review*, 9(3), 558-565.
14. Harnad S. (1990). The symbol grounding problem. *Physica D*, 42: 335-346
15. Hashimoto T., Ikegami T. (1996). Emergence of net-grammar in communicating agents, *BioSystems*, 38: 1-14
16. Joyce D., Richards L., Cangelosi A., Coventry K.R. (2003). On the foundations of perceptual symbol systems: Specifying embodied representations via connectionism. *Proceedings of the 5th Intl. Conference on Cognitive Modeling (ICCM 2003)*. Bamberg
17. Marocco D., Cangelosi A., Nolfi S. (2003), The emergence of communication is evolutionary robots. *Philosophical Transactions of the Royal Society London – A*, 361: 2397-2421
18. Nolfi S., Floreano D. (2000) *Evolutionary Robotics*. Cambridge, MA: MIT Press.
19. Pulvermueller F. (2003). *The Neuroscience of Language. On Brain Circuits of Words and Serial Order*. Cambridge University Press.
20. Quinn M. (2001). Evolving communication without dedicated communication channels. In J. Kelemen & P. Sosík (Eds.), *Proceedings of ECAL01*, pp. 357-366, London: Springer.
21. Riga T., Cangelosi A., Greco A. (2004). Symbol grounding transfer with hybrid self-organizing/supervised neural networks. *IJCNN04 International Joint Conference on Neural Networks*. Budapest, July 2004
22. Rizzolatti G., & Arbib M. (1998). Language within our grasp. *Trends in Neuroscience*, 21: 188-194.
23. Roy D., Pentland A. (2002), Learning words from sights and sounds: A computational model, *Cognitive Science*, 26: 113-146.
24. Vogt P. (2002). The physical symbol grounding problem. *Cognitive Systems Research*, 3(3): 429-457.
25. Steels L. (2002). Grounding symbols through evolutionary language games. In A. Cangelosi & D. Parisi (Eds.), *Simulating the Evolution of Language*. London: Springer Verlag.

Exposure Dependent Creolization in Language Dynamics Equation

Makoto Nakamura[1], Takashi Hashimoto[2], and Satoshi Tojo[1]

Graduate School of {[1]Information, [2]Knowledge} Science,
Japan Advanced Institute of Science and Technology,
Nomi, Ishikawa, 923-1292, Japan
{mnakamur,hash,tojo}@jaist.ac.jp

Abstract. The purpose of this paper is to develop a new formalism of language dynamics so that creole may emerge. Thus far, we modified the *transition probability* of the dynamics so as to change in accordance with the distribution of population of each language at each generation, and in addition, we introduced a new parameter called *exposure rate* with which infants were exposed to other languages than mother tongues. Thus, we could observe creolization under limited conditions. In this paper, we revise the learning algorithm in our model, considering the amount of language input during the language acquisition period. Thus, the transition probability depends not only on the exposure rate but also on the amount of language input. With this model, we show that creolization occurs only when the influence of mother tongues and the socially dominant languages balance.

1 Introduction

In general, children correctly inherit language from their parents and/or neighbors during their acquisition period, even though it has not yet been clarified how children correctly deduce the underlying grammatical rules and acquire the same language as their mothers'. In other cases, some children whose parents speak a pidgin may acquire another new language called a creole. Pidgin and creole are defined as two different stages of language change [1,2]. Pidgin is a simplified tentative language spoken in multilingual communities. Creole is a full-fledged new language which children of the pidgin speakers obtain as their native language. Some properties of creoles imply the existence of an innate universal grammar.

Linguistic studies are going to have clarified why and how creoles emerged. Observing actual pidgins and creoles, linguists have argued that creoles would appear under a specific environment like a pidgin community [1,2,3]. From the linguistic efforts, it is clear that the emergence of creole is affected by contact with other languages, the distribution of population of each language and similarities among the languages. In population dynamics [4], by parameterizing these elements, we could derive conditions for the emergence of creole from the

A. Sakurai et al. (Eds.): JSAI 2003/2004, LNAI 3609, pp. 295–304, 2007.

theoritical and numerical analyses [5], and then could contribute to specify the function of the universal grammar.

Thus far, we revised the language dynamics by Nowak et al. [6] in such a way that the transition rates changed according to the distribution of population of each grammar at each generation. In addition, we introduced an *exposure rate* which assesses an extent that a child is exposed to other languages than that of his/her parents. Using this approach, we have shown the emergence of a creole when multiple parental languages are similar in some way [7,8]. We improved our model to exclude *fitness* that dominated the ratio of offsprings with regard to communicability [9]. We observed such unnatural phenomena that the creole emerged even when children learned language only from their parents. In this paper, we will present a new formalism to remedy this problem.

In Section 2, we describe our previous model. In Section 3, we present the new formalism and in Section 4 we define a creole in population dynamics. Section 5 reports our experiments. We conclude in Section 6.

2 Language Dynamics Equation Without Fitness

In this section, we breafly explain our previous model and consider the emergence of creole in population dynamics.

In response to the language dynamics equation by Nowak et al. [6], we assume that any language is classified into one of a certain number of grammars. Thus, the population of language speakers are distributed to a finite number (n) of grammars $\{G_1 \dots G_n\}$. Let x be a ratio of speakers of each language. Then, the language dynamics is modeled by an equation governing the transition of language speakers among languages.

In the language dynamics equations, the similarity matrix S and the transition matrix $\overline{Q}(t)$ play important roles. The similarity matrix $S = \{s_{ij}\}$ denotes the probability that a sentence of G_i is accepted by G_j. The transition matrix $\overline{Q}(t) = \{\overline{q}_{ij}(t)\}$ is defined as the probability that a child of G_i speaker obtains G_j by the exposure to his/her parental language and to other ones. Being different from the definition by Nowak et al., our definition of $\overline{Q}(t)$ depends on the generation parameter t, as well as the S matrix and a learning algorithm.

Because Nowak et al. assume that language speakers bear offsprings in proportion to their successful communication, they embed a fitness term in their model which determines the birth rate of each language group. Our model excludes the fitness on the assumption that in the real world creoles do not emerge because creole speakers have more offsprings than that of other pre-existing languages. We have already shown the difference between the models with and without fitness [9], in which the latter becomes:

$$\frac{dx_j(t)}{dt} = \sum_{i=1}^{n} \overline{q}_{ij}(t)x_i(t) - x_j(t) \ . \tag{1}$$

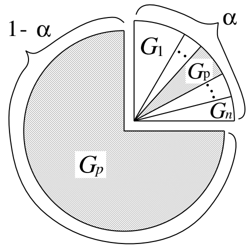


Fig. 1. The exposure rate α

3 Language Acquisition and Transition of Population

In this section, we propose a new transition matrix $\overline{\overline{Q}} = \{\overline{\overline{q}}_{ij}(t)\}$. Our approach takes account of a probability distribution of the number of acceptable sentences for each grammar against the number of input sentences during acquisition term. Firstly, we explain the learning algorithm. Secondly, we represent the transition matrix $\overline{\overline{Q}}$.

3.1 Learning Algorithm

In some community, a child learns language not only from his/her parents but also from other adults, whose language may be different from the parental one. In such a situation, the child is assumed to be exposed to other languages, and thus may acquire the most efficient grammar in accepting those language input. In order to assess how often the child is exposed to other languages, we divide the language input into two categories: one is from his/her parents and the other is from other language speakers. We name the ratio of the latter *exposure rate* α. This α is subdivided into the smaller ratios corresponding to those other languages, where each ratio is in proportion to the population of the language speakers. An example distribution of languages is shown in Fig. 1. The child of G_p speaker is exposed to G_p at the rate of the shaded part, that is $\alpha x_p + (1 - \alpha)$, and the ratio of a non-parental language G_i comes to be αx_i.

Our learning algorithm resolves Niyogi [10]'s problem that there is an unrealistic Markov structure which implies that some children cannot learn certain kinds of language. From the viewpoint of a universal grammar that all conceivable grammars of human beings are restricted to a finite set, language learning is considered as a choice of a plausible grammar from them. The following algorithm realizes such learning as: 1) In a child's memory, there supposed to be a score table of grammars. 2) The child receives a sentence uttered by an adult. 3) The acceptability of the sentence is tested for each grammar. The grammar which accepts the sentence scores 1 point. 4) 2) and 3) are repeated until the child receives a fixed number (w) of sentences, that is regarded as enough for the estimation of the grammar. 5) The child adopts the grammar with the highest score.

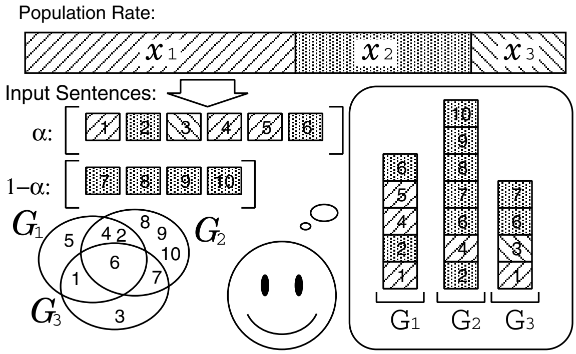


Fig. 2. The learning algorithm

The distribution of population and the exposure rate α determine the rate of adult speakers of each language to which the child is exposed, while the S matrix determines the acceptability of a sentence. In Fig. 2, we show an example where a child of G_2 speaker obtains G_2 after the exposure to a variety of languages. The child receives sentences, that are numbered boxes from 1 to 10. The input sentences are divided into two sets according to the exposure rate α. One of the sets consists of sentences of all grammars. The number of the sentences of each grammar is proportional to the rate of population of the grammar. For example, the child hears sentences 1, 4 and 5 uttered by G_1 speakers. The other consists of sentences of his/her parents. Therefore, these sentences are acceptable by a particular grammar. Because his/her parental grammar is G_2, for example, the sentences 7 to 10 are randomly chosen from the language of G_2. The child counts acceptable sentences for each grammar. The sentence 1 can be accepted by G_3 other than G_1, while it is uttered by a G_1 speaker. The Venn diagram in Fig. 2 represents that each language shares sentences with others. In this case, because the sentence 1 is acceptable both by G_1 and by G_3, the child adds 1 to both of the counters in his/her mind.

3.2 Revised Transition Probability

Suppose that children hear sentences from adult speakers depending on the exposure rate and on the distribution of population. A probability that a child whose parents speak G_i accepts a sentence by G_j is expressed by:

$$U_{ij} = \alpha \sum_{k=1}^{n} s_{kj} x_k + (1-\alpha) s_{ij} \ . \tag{2}$$

After receiving a sufficient number of sentences for language acquisition, the child will adopt the most plausible grammar that is estimated by counting a number of accepted sentences by each grammar. This learning algorithm is simply represented in the following equation. Exposed to a variety of languages in proportion to the ratio of adult speakers, children whose parents speak G_i will adopt G_{j^*} in the following manner:

$$j^* = \underset{j}{\operatorname{argmax}}\{U_{ij}\} \ . \tag{3}$$

When the children hear w sentences, a probability that a child of G_i speaker accepts r sentences with G_j is given by a binomial distribution,

$$g_{ij}(r) = \binom{w}{r}(U_{ij})^r(1-U_{ij})^{w-r} \ . \tag{4}$$

On the other hand, a probability that the child accepts less than r sentences with G_j is

$$h_{ij}(r) = \sum_{k=0}^{r-1}\binom{w}{k}(U_{ij})^k(1-U_{ij})^{w-k} \ . \tag{5}$$

From these two probability distributions, the probability that a child of G_i speaker accepts k sentences with G_j, while less than $k-1$ sentences with the other grammars, comes to $g_{ij}(k)\prod_{l=1,l\neq j}^{n} h_{il}(k)$. For a child of G_i speaker to acquire G_j after hearing w sentences, G_j must be the most efficient grammar among n grammars; viz., G_j must accept at least $\lceil \frac{w}{n} \rceil$ sentences. Thus, the probability $\overline{\overline{q}}_{ij}$ becomes the sum of the probabilities that G_j accepts $w, w-1, \cdots, \lceil \frac{w}{n} \rceil$ sentences. Because each of the sentences is uttered by a speaker and is accepted by at least one grammar, there must be a grammar which accept $\lceil \frac{w}{n} \rceil$ or more out of w sentences. Thus, if G_j accepts less than $\lceil \frac{w}{n} \rceil$ sentences, the child does not acquire G_j. Therefore, $\overline{\overline{q}}_{ij}$ becomes:

$$\overline{\overline{q}}_{ij}(t) = \frac{\displaystyle\sum_{k=\lceil \frac{w}{n} \rceil}^{w}\left\{g_{ij}(k)\prod_{\substack{l=1\\ l\neq j}}^{n} h_{il}(k) + R(k)\right\}}{\displaystyle\sum_{m=1}^{n}\left[\sum_{k=\lceil \frac{w}{n} \rceil}^{w}\left\{g_{im}(k)\prod_{\substack{l=1\\ l\neq m}}^{n} h_{il}(k) + R(k)\right\}\right]} \ , \tag{6}$$

where $R(k)$ is the sum total of the probabilities that the child choose G_j when one or more other grammars accept the same number of sentences as G_j. When there are m candidate grammars including G_j, the probability becomes the one divided by m.

4 Creole in Population Dynamics

Creole is considered as a new language. From the viewpoint of population dynamics, we define a creole as a transition of population of language speakers. Creole is a language which no one spoke in the initial state but most of people come to speak at a stable generation. Therefore, a creole is expressed to such a grammar G_c that: $x_c(0) = 0, x_c(t) > \theta_c$, where $x_c(t)$ denotes the rate of the population of G_c at a convergent time t, and θ_c is a certain threshold to be regarded as a dominant language. In this paper, we set $\theta_c = 0.9$ through the experiments.

We have mainly observed the behavior of the model of three grammars. Suppose the size of language is the same and each sentence of the language is chosen with a uniform probability, the similarity matrix can be expressed as such a symmetric matrix that:

$$S = \begin{pmatrix} 1 & a & b \\ a & 1 & c \\ b & c & 1 \end{pmatrix} . \tag{7}$$

Here, we regard G_3 as a creole grammar, giving the initial condition as $(x_1(0), x_2(0), x_3(0)) = (0.5, 0.5, 0)$. Therefore, the element a denotes the similarity between two pre-existing languages, and b and c are the similarities between G_1 and the creole, and between G_2 and the creole, respectively.

5 Experiments

In this section, we show the experimental result of our model. We examine the conditions that creole appears and comes to be dominant in combinations of the S matrix and α.

5.1 Emergence of Creole

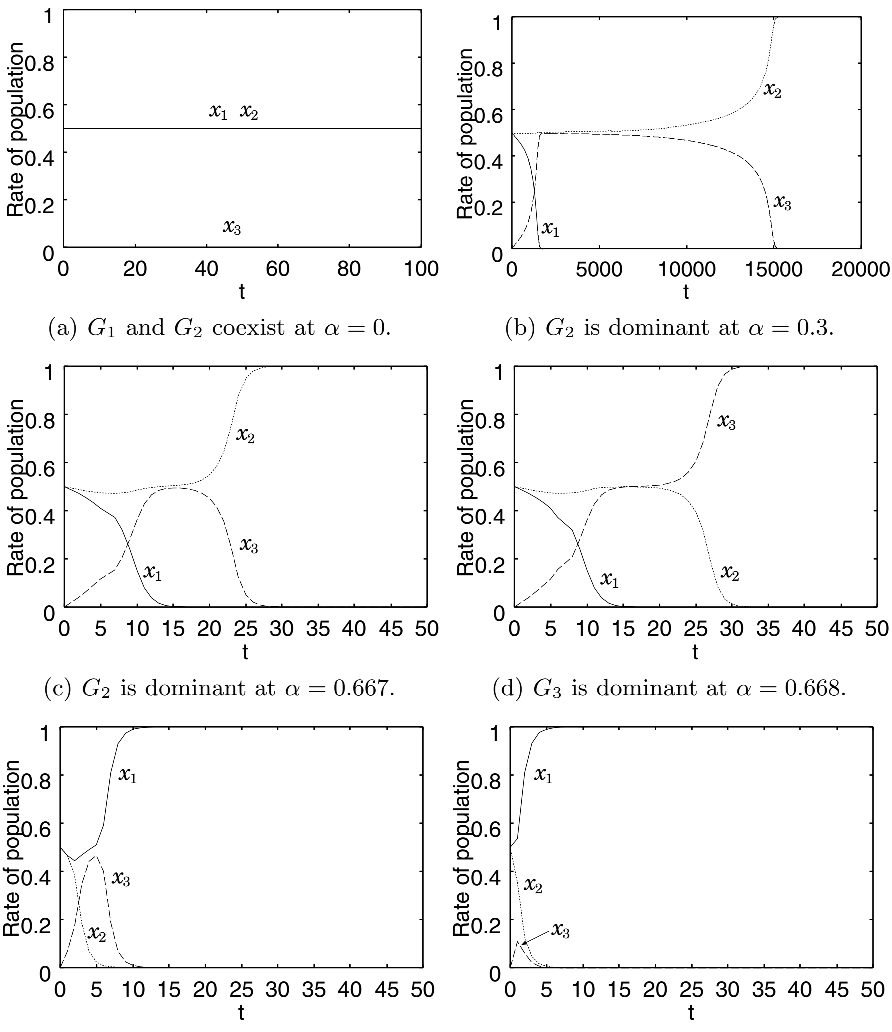
In Fig. 3, we show the result of our model. We arbitrarily set the S matrix to $(a, b, c) = (0, 0.45, 0.35)$, in which the pre-existing grammars G_1 and G_2 do not share any sentence. We gave the number of input sentences $w = 30$ that was found to be large enough for language acquisition in three grammars. The exposure rate α is examined at the range from 0 to 1.

In case $\alpha = 0$, children learn a language only from their parents. Accordingly, Fig. 3(a) shows that both populations of G_1 and G_2 hardly transmit. In the previous model [9], we found a problem that a creole coexists with other languages at $\alpha = 0$. However, we come to resolve the problem.

According to the increase of α, x_3 rises gradually though $x_3(0) = 0$, while x_1 falls down to 0 in Fig. 3(b). However, x_3 declined in further generations and eventually disappeared. Because the transition of population depends on Eqn 2, we can approximately compare the directions of transition of population among grammars with Eqn 2. Eqn 8 expresses an expansion of Eqn 2 at $a = 0$.

$$U = \begin{pmatrix} (1-\alpha) + \alpha(x_1 + bx_3) & \alpha(x_2 + cx_3) & (1-\alpha)b + \alpha(bx_1 + cx_2 + x_3) \\ \alpha(x_1 + bx_3) & (1-\alpha) + \alpha(x_2 + cx_3) & (1-\alpha)c + \alpha(bx_1 + cx_2 + x_3) \\ (1-\alpha)b + \alpha(x_1 + bx_3) & (1-\alpha)c + \alpha(x_2 + cx_3) & (1-\alpha) + \alpha(bx_1 + cx_2 + x_3) \end{pmatrix} \tag{8}$$

Although the population is shared among only G_1 and G_2 at the initial generation, the increase of α makes the transition from G_1 and G_2 to G_3 active.



(a) G_1 and G_2 coexist at $\alpha = 0$.

(b) G_2 is dominant at $\alpha = 0.3$.

(c) G_2 is dominant at $\alpha = 0.667$.

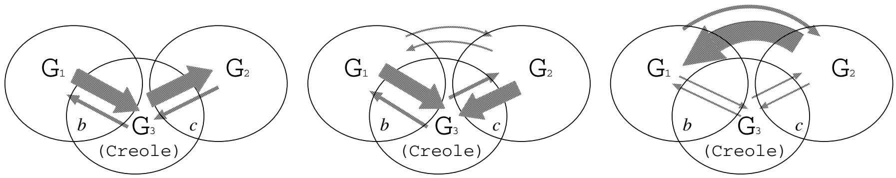
(d) G_3 is dominant at $\alpha = 0.668$.

(e) G_1 is dominant at $\alpha = 0.835$.

(f) G_3 is dominant at $\alpha = 1$.

Fig. 3. The transition of dominant language by changing α ($(a, b, c) = (0, 0.45, 0.35)$)

Because $x_3 \simeq 0$ and $U_{13}, U_{23} > 0$ at early generations, x_1 and x_2 start flowing into x_3. Moreover, because $U_{13} > U_{23}$, x_1 is easier to flow into x_3 than x_2. Once x_3 has earned a certain rate of population, U_{13} becomes greater than U_{31} and the outflow of x_1 to x_3 accelerates, while $U_{23} \simeq U_{32}$. After x_1 mostly diminished, the difference between U_{23} and U_{32} is expanded as the difference between $(cx_2 + x_3)$ and $(x_2 + cx_3)$, that is x_3 and x_2. Therefore, the difference between the two population rates determines which of the corresponding languages becomes dominant. In case $\alpha = 0.667$ (See Fig. 3(c)), because x_2 is barely more than x_3 at a point of generation at which x_1 mostly disappeared, G_2 finally dominates



(a) G_2 becomes dominant ($0 < \alpha \leq 0.667$). (b) G_3 becomes creole ($0.668 \leq \alpha \leq 0.834$). (c) G_1 becomes dominant ($0.835 \leq \alpha \leq 1$).

Fig. 4. Flow of the population by changing α value ($(a, b, c) = (0, 0.45, 0.35)$)

the community. These flows of the population is shown in Fig. 4(a). We can see that the larger α, the solution converges at the earlier generations in Fig. 3(b) and Fig. 3(c). On the contrary, we have encountered that the solution did not converge in realistic time at a small α.

In case of $(cx_2 + x_3) > (x_2 + cx_3)$ at the point at which x_1 mostly disappeared, x_3 rises to 1, that is, G_3 becomes dominant at $\alpha = 0.668$. This is the emergence of creole, as shown in Fig. 3(d). Similarly, Fig. 4(b) depicts the process of creolization that in addition to the inflow of population of G_1 the transition from G_2 to G_3 outstrips the outflow of population of G_3.

Further increasing α, we can observe G_1 becomes dominant although it loses the population at the very first in small α. Also, let us pay attention to U_{12} and U_{21} in Eqn 8. At the early generation, x_3 has not earned enough population yet. When α is large enough like Fig. 3(e), U_{12} is larger than U_{13}. Large α represents that children of G_1 speakers grow up, hearing sentences of G_2 in the almost same rate as those of G_1. Therefore, the direct transition between G_1 and G_2 occurs at large α. We show the flows of the population between G_1 and G_2 in Fig. 4(c). Thus, we can regard our experimental result is that creoles are not the easiest to emerge at $\alpha = 1$. This result adequately remedied our fallacious expectation [7].

5.2 Dominant Language and Creole

In the previous section, we showed the emergence of a creole, and quantitatively considered the process of creolization. We observed that a creole appears within a certain area of α which must be large enough but less than the value at which the direct transition between pre-existing languages becomes mainstream. The value of a in Eqn 7 concerns the direct transition, while it is fixed to 0 in the previous experiment. The next experiment aims at drawing a diagram as to which language would be dominant in various values of the similarity between the pre-existing languages.

By parameterizing a in Eqn 7 and α, we examined the dominant grammar at the convergent generation. The parameter region in which each grammar dominates is shown in Fig. 5. In the figure, the region of asterisk ($*$) denotes none of the languages becomes dominant. Namely, either the solution converged to the coexistence of a few languages, or the solution could not converge at over a million generations. As we mentioned in the previous experiment, with small values of α the solution hardly converges. On the contrary, in the upper side of

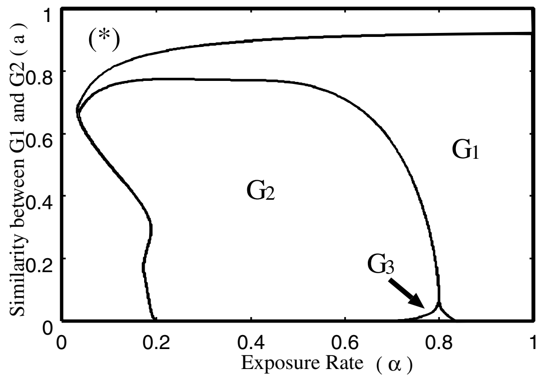


Fig. 5. Distribution of dominant grammars $((a, b, c) = (a, 0.45, 0.35))$

the asterisk region, the pre-existing languages coexist because those languages are regarded as an almost identical language at very high value of a.

The previous experiment was examined along with the horizontal axis at $a = 0$. At the bottom of the figure around $\alpha = 0.8$, G_3 (creole) becomes dominant. Thus, the lower value of a, the easier the creole emerges. In other words, a similarity between pre-existing languages implies the ease of surviving of the languages. This result is consistent with that of our previous model [7] that a creole may emerge if the pre-existing languages are not similar to each other, but to the newly appeared language.

6 Conclusion

In this study, we proposed the modified Q matrix of the language dynamics equation [6], where children may migrate to non-parental languages, estimating the number of sentences of probable grammars. In our previous work, we had a problem that children happened to acquire a new language (creole) even when language was given only by their parents [9]. It seems that this unsatisfactory situation was caused by the fact that the learning algorithm could not adjust the amount of language input. We introduced a new parameter concerning this adjustment and examined the behavior on how children guess plausible grammars. As a result, we could show that creolization rarely occurred in high values of the exposure rate α, no less in low values. In a high value of α, children tended to select a pre-existing dominant language, and in a low value they certainly learned parental language; thus, we could contend that creole might emerge between the influence of mother tongues and that of the socially dominant language.

In our future study, we will examine the relation between the amount of language input and the creolization. The preliminary examination showed that the fewer the language input, the easier the creolization occurred. In addition, we need to consider refining the learning algorithm and need to establish a more reliable theory on language similarity.

References

1. Arends, J., Muysken, P., Smith, N., eds.: Pidgins and Creoles. John Benjamins Publishing Co., Amsterdam (1994)
2. DeGraff, M., ed.: Language Creation and Language Change. The MIT Press, Cambridge, MA (1999)
3. Bickerton, D.: Language and Species. University of Chicago Press (1990)
4. Weibull, J.: Evolutionary Game Theory. The MIT Press, Cambridge, MA (1995)
5. Cangelosi, A., Parisi, D., eds.: Simulating the Evolution of Language. Springer, London (2002)
6. Nowak, M.A., Komarova, N.L., Niyogi, P.: Evolution of universal grammar. Science **291** (2001) 114–118
7. Nakamura, M., Hashimoto, T., Tojo, S.: Creole viewed from population dynamics. In: Proceedings of the Workshop/Course on Language Evolution and Computation in ESSLLI. (2003) 95–104
8. Nakamura, M., Hashimoto, T., Tojo, S.: The language dynamics equations of population-based transition – a scenario for creolization. In Arabnia, H.R., ed.: Proceedings of IC-AI'03, CSREA Press (2003) 689–695
9. Nakamura, M., Hashimoto, T., Tojo, S., Smith, K.: The effect of fitness in the emergence of creole. In: Proceedings of EVOLANG5. (2004) 85
10. Niyogi, P.: The Informational Complexity of Learning from Examples. PhD thesis, Massachusetts Institute of Technology (1996)

Evolution of Birdsong Grammars

Kazutoshi Sasahara and Takashi Ikegami

Department of General Systems Studies, Graduate School of Arts and Sciences,
University of Tokyo
3-8-1, Komaba, Meguro-ku, Tokyo 153-8902, Japan
{sasahara,ikeg}@sacral.c.u-tokyo.ac.jp

Abstract. Recent studies on song birds reveal that their vocal communication has some common features with human language. In particular, a songbird called Bengalese finch has interesting vocal communication in which courtship song is arranged by finite-state grammar [6]. It has been hypothesized that the song grammar may have evolved as a result of sexual selection [10]. In order to explore the evolution of the song grammars, we model the co-evolution of male and female finches by asymmetric finite-state automata. In this paper, we demonstrate that song grammars could evolve by simple communication. We observe that a transition from lower complexity to higher complexity song grammars is driven by the changing of male birds' courting strategy.

1 Introduction

Language is the most complex communication system, in which vocal phonemes are arranged recursively in one-dimensional manner with precise articulation. Hauser, Chomsky and Fitch insist that a core property of the faculty of language in the narrow sense (FLN) is '*recursion*', which is a recursive computational property to generate discrete infinity from a finite set of elements [4]. And they hypothesize that FLN is unique to human species. However, as they mention, a homologous computational ability with FLN can be found in other species. For example, navigation capacity of insects; after searching, insect can find out the shortest path between their nest and destination [2]. In addition, as we see in this paper, the recursive rule set for courtship song can be found in a song bird called Bengalese finch (*Lonchura striata var. domestica*) [6]. If a recursive rule set enables optimal strategies for many biological situations, we count insects' optimal search behavior and birds' courtship song as the evidences for recursion.

Furthermore, studies on song birds have been revealed that their vocal communication has some common features with human language [1][9]. According to the analysis by C.Hocket, a characteristic of human language can be divided into 10 categories [5]: '*use of the vocal-auditory channel*', '*arbitrariness*', '*spontaneous usage*' and '*duality*' can be found in the vocal communication of songbirds. In addition to these features, the male Bengalese finches have a recursive rule set for singing a courtship song (in this paper, we call it song grammar), which is homologous with FLN. In terms of the recursion for vocal communication, song

A. Sakurai et al. (Eds.): JSAI 2003/2004, LNAI 3609, pp. 305–314, 2007.

grammar is much similar to human language than the navigation capacity of insects. Therefore, a study on the evolution of the vocal communication of the Bengalese finch is a good starting point to understand the origin and evolution of language.

In this paper, we focus on the communication of the Bengalese finch based on Okanoya's experiments. The courtship songs of the male Bengalese finch consist of a combination of chunks, each of which is a sequence of sound element, can be described by finite-state automata (FAs) [6]. In addition, it has been observed that more complex songs are preferred by females and that their songs are still more complex than those of their ancestors, White-Backed Munia (*Lonchura striata*) [8]. On the grounds of these facts, Okanoya hypothesizes that complex song grammar may have evolved as a result of '*sexual selection*' [17][10][11]. In order to explore this hypothesis, we study the co-evolution of males' song grammars and females' preferences by a synthetic approach which represents birds as asymmetric finite-state automata (FAs).

2 Model

In this section, we describe the main concepts of our modeling. We model the co-evolution of males and females as a communication game [3][14]. A key of the modeling is the way in which female birds gauge the complexity of the songs they hear. There is a suggestive experimental result: while complex songs generated by a FA enhance the reproductive behavior of females, the frequency of this reproductive behavior tends to be relatively low when the females are listening to monotonous or random songs [10]. This indicates that the female birds can discern arrangement of song elements, and distinguish interesting songs from monotonous or random ones.

Consequently, we assume that female birds may be sensitive for arrangement of chunks and may have innate preferences for phrasing and rhythm in a song, so that they can gauge songs according to her preferences [12]. As a results, the song grammars of males may have become complex due to the diversity of females' preferences. We introduce the following interaction to model this process. The female interjects in synchrony with the male song (by wagging her tail or chirping softly, for example), measuring how many interjections succeed according to her preferences before she evaluates her satisfaction with the song. [1] In this model, we call such interaction "*song-interjection*" communication. The concrete specification of our model is briefly described bellow.

2.1 Song Grammars and Female Preferences

The song grammar of a male bird is expressed as a finite-state automaton: $G = (Q, \Sigma, \Delta, \delta, \lambda, q_0)$, where Q is a finite set of states, q_0 is an initial state,

[1] The female Zebra finch (*Taeniopygia guttata castanotis*) has been observed to chirp in synchrony with the male song. However, no such observation has yet been made for the female Bengalese finch.

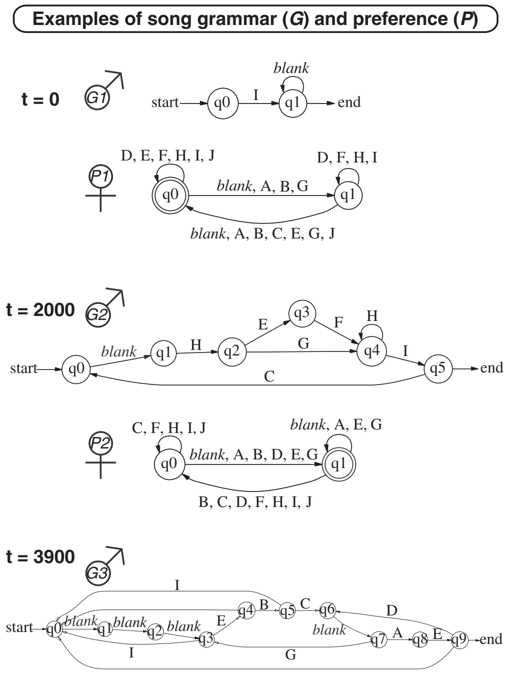


Fig. 1. Examples of song grammars (G) and preferences (P): $G1$, $G2$ and $G3$ shows male song grammars and $P1$, $P2$ show female preferences. $G1/P1$ and $G2/P2$ are well-suited pairs that yielded good communication at t = 0 and t = 2000, respectively. $G3$ shows a grammar more complex than $G1$ or $G2$. The double circle represents a accepting state.

Σ is a finite set of input symbols, Δ is a finite set of output symbols, δ is a state transition function, $Q \times \Sigma \rightarrow Q$, λ is an output function and $Q \times \Sigma \rightarrow \Delta$ [7]. In this model, $\Delta = \{blank, A, B, ..., J\}$, where each letter represents a song chunk and "*blank*" represents a silent interval between chunks. The bunch of combinatorial chunks between blanks expresses a phrase and the whole output sequence expresses a courtship song. A male bird arranges the chunks in accordance with his song grammar.

On the other hand, the preference of the female bird (i.e. interjection action) is expressed by a FA: $P = (Q, \Sigma, \delta, q_0, F)$, where Q, Σ, δ, and q_0 are the same as above, and F is a set of accepting states, which is a subset of Q. This expresses a female's preference for the phrasing, rhythm, and arrangement in a courtship song. She changes her internal state by listening to the song and interjects if she is in an accepting state. Examples of male and female FAs are shown in Fig.1. As we see in Fig.1, the diversity of males and females are expressed as the diversity of the shapes of the FAs.

Examples of communication

```
t = 0
G1(N^node=2,L_song=10,LI=0.67) vs. P1(N^node=2)
I__I_I_I_I
_#__##__##
score = 0.38

t = 2000
G2(N^node=6,L_song=13,LI=0.67) vs. P2(N^node=2)
_HEFIC_HGI_HE
#_#___#_#_#_#
score = 0.43

t = 3900
G3(N^node=10,L_song=148,LI=0.63) vs. P3(N^node= 8)
___I___EC___EBI___EBI___I___EBC_GEC___I___I___I___EBC_GEC___EC___EBI...
###_###__###___###___###_###___##__###_###_###_###___##__###__###___...
score = 0.74

t = 2000
G4(N^node=9,L_song=28,LI=0.64) vs. P4(N^node=2)
GEGEJC_IGEJC_EGEGIC_IGIC_IJC
##___#_#___#__##_##_#_##_#_#
score = 0.0
```

Fig. 2. Examples of communication: $G1$ vs. $P1$, $G2$ vs. $P2$ and $G3$ vs. $P3$ lead to suitable communications. $G4$ vs. $P4$ shows an example of poor communication.

2.2 Communication

Each male bird attracts a female at random and sings a song for length L_{song} according to his grammar G. Since a male's song must be a signal for tempting a female, '*novelty*' is an important factor in a courtship song [15][16][13]. So, we assume that females may pay attention to the novelty of males' song as follows:

(i) There must be at least one chunk in the song which is not *blank*.
(ii) The female must make at least one mistake in interjecting. This is because a song to which a female can interject perfectly (i.e. which is perfectly predictable by her FA) is boring for her.

Unless these conditions are satisfied, the birds involved are not eligible candidates for mating. Listening to a novel song that fulfills the above conditions, a female bird interjects to the song in accordance with her preference P, and evaluates her contentment. Each male bird can sing for a length up to L_{song}^{max} and to several females within this length, where L_{song}^{max} denotes the maximum length he can sing. For example, if a male bird has $L_{song} = 10$ and $L_{song}^{max} = 50$, he can attract five females. Each time step, every male bird behaves in the above way. On the other hand, the female interjects to all songs that she hears and that she is satisfied with according to her preference and the above novelty criteria. Some examples of communication are illustrated in Fig.2.

After a pair of birds communicate, their communication is assigned a score calculated as follows:

$$S = \frac{1}{3}\{\frac{1}{N_{interj}^{th}} min(N_{interj}^{succ}, N_{interj}^{th}) + \frac{N_{interj}^{succ}}{N_{interj}^{all}} + \frac{N_{chunk}}{L_{song}}\}, \quad (1)$$

where, $0 \leq S \leq 1$. In total, the communication score (1) considers the evaluation of both quantity and quality of interjection, and the richness of song elements. The first term denotes the evaluation of the number of successful interjections; it is proportional to the number of successful interjections N_{interj}^{succ} below the threshold N_{interj}^{th}. If $N_{interj}^{succ} \geq N_{interj}^{th}$, the female bird's evaluation is saturated and the first term becomes 1. The second term denotes the success rate of interjection; the ratio between the total number(N_{interj}^{all}) and successful number(N_{interj}^{succ}) of interjections. The third term denotes the fraction of non-empty chunks in a song.

2.3 Evolutionary Dynamics

According to (1), females select the males with the highest score as their mating partners. Assuming that they produce offspring in proportion to their communication score (i.e. affinity between the male and female), the number of offspring is calculated as $C_{offs} \cdot S$. Then, their offspring's genders are randomly assigned and they are added into the system as new child birds.

Since child birds study songs from their fathers or may have similar song preferences to their mothers as a result of their upbringing, their characters become similar to those of their parents. Therefore, in our model child birds inherit the FAs which are similar to their parents, changed according to the following genetic mutation operations:

(a) **Arrow Mutation:** Change the transitions of the FA with the number of nodes remaining fixed.
(b) **Node Mutation:** Change the number of nodes (± 1) and then add or remove arrows as required.
(c) **Random Mutation:** A new FA is made at random.

These (a)-(c) express the possible inaccuracy in child birds inheriting their parents' characteristics, song grammars G and preferences P. In addition, the following mutation is performed in the male bird population only:

(d) **Song Mutation:** Change L_{song} (± 5), and change L_{song}^{max} (± 2)

In this artificial ecosystem, each bird has a life time T_{life}, after which they are removed from the system. In order to limit the maximum number of birds in the system, some birds are removed due to a fixed ecological capacity of $C_{echo} \cdot (N^{male} + N^{female})$.

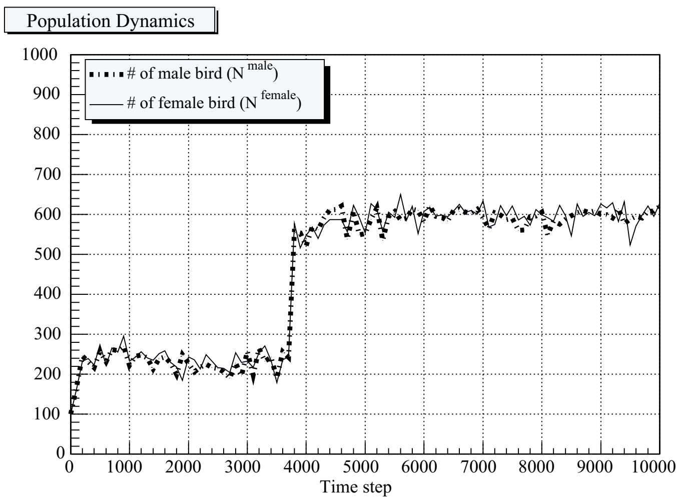


Fig. 3. Population dynamics: Step-like evolution is observed. After the critical period around t = 3700, the number of birds rapidly increases.

3 Simulation Results

We describe the typical results of this artificial evolution. The parameters of our simulations were set as follows. The initial populations of males and females were 100, respectively. Every male bird had a FA constructed randomly with $N_{node} = 2$, $L_{song} = 10$ and $L_{song}^{max} = 50$. The maximum length of song was 500. Meanwhile, every female bird also had a randomly constructed FA with $N_{node} = 2$. Two examples of initial FAs are shown in the top of Fig.1. Other significant parameters were $N_{interj}^{th} = 100$, $C_{offs} = 3.5$, $T_{life} = 5$ and $C_{echo} = 0.3$.

3.1 The Changing of Male Birds' Courting Strategy

We see in Fig.3 a step-like evolution where the population increased rapidly at around $t = 3,700$ and then remained almost constant.[2] After this period, we can find a change of strategy in the courtship behavior of both males and females.

Shortly after that period, a trend in preferring longer songs emerged. Once such a trend appeared in the system, the character of males and females was drastically changed as may be seen in Fig.4 and Fig.5. In Fig.4, we see a rapid increase in the length of males' songs, L_{song}. Before this period, the male birds sang relatively short songs, even if the maximum song length was much higher.

[2] Whether or not we see such a stepwise change depends on both the parameters N_{interj}^{max} and C_{offs} that affect the number of offspring. The step-like evolution is observed in wide region of these parameters, provided they are not too big or too small (i.e. the production rate is not too high or too low).

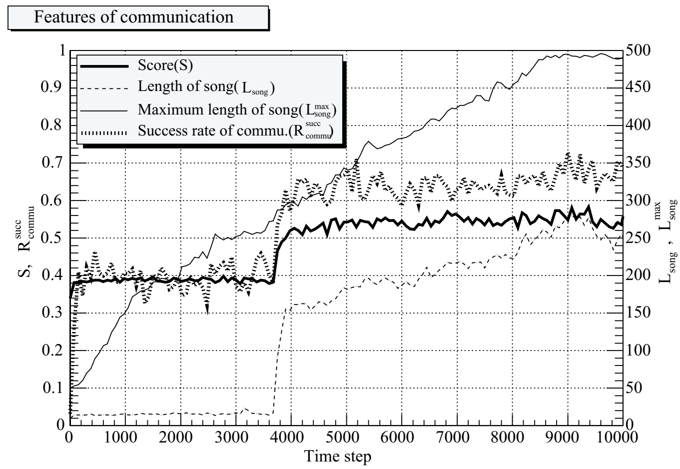


Fig. 4. Features of communication: Shortly after the critical period t = 3700, a trend in preferring longer songs emerged. Subsequently, communication scores and successful communication rates increased.

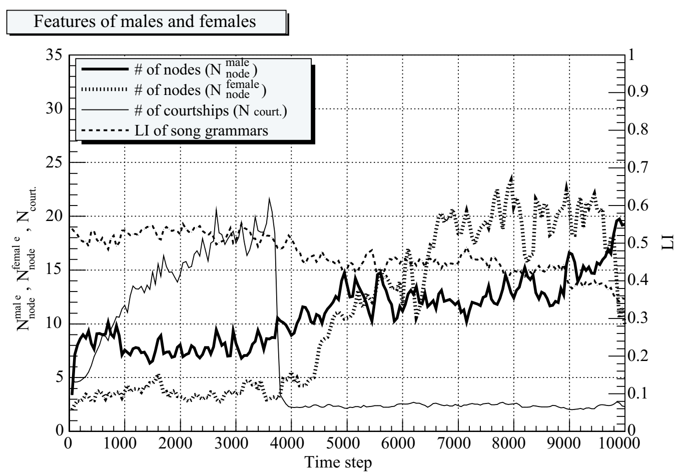


Fig. 5. Features of males and females: The number of times the male birds sang to females N_{court} and the number of nodes of females $N_{node}female$ drastically changed at the critical time t = 3700. On the other hand, the number of nodes of males and the linearity of song grammars LI changed gradually over time.

They could find partners successfully by singing to many females. This is clear from the N_{court} in Fig.5, which represents the average number of courtship per a female bird. Before this period, the communication scores were low in Fig.4, but N_{court} was very high in Fig.5.

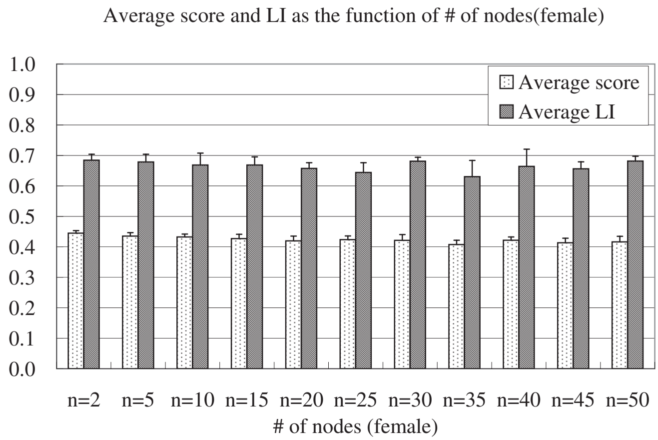


Fig. 6. Average score and LI with changing the number of nodes (female): There is no correlation between the complexity of song grammars and the degree of female preferences

In particular, Fig.5 shows that the male birds began to sing longer and more complex songs as the number of FA nodes increased. Subsequently, the female birds had a tendency to become sensitive to the arrangement of song chunks and also had FAs with increasing numbers of nodes.

The male birds never sang for the maximum length L_{song}^{max} shown in Fig.4. This results from a kind of dilemma. The male birds wanted the females to listen to their songs to get high scores, but if their affinity was bad, the cost of failure would be more serious. Therefore, we conclude that the male birds evolved a survival strategy as a result of co-evolution, avoiding the risky behavior of singing to only one female a song of length close to L_{song}^{max}.

On the whole, the communication evolved to become quite successful judging from the change in average communication score and the success rate of communication which represents the frequency of non-zero scores in Fig.4.

3.2 The Complexity of Song Grammars and Female Preferences

Here we define LI, the linearity of a song grammar for the following discussion:

$$LI \equiv N_{node}/N_{arrow}, \tag{2}$$

where N_{arrow} is the number of arrows leaving a node. If $N_{node} = N$, this value ranges between $1/N \leq LI \leq 1$ as N_{arrow} varies from N^2 to N. More complex FAs have lower values of LI.

In the initial state, only simple song grammars like $G1$ exist in the system. As the evolution proceeds, the song grammars became more complex. In Fig.1, more complex song grammars $G2$ and $G3$ are shown. Such song grammars that include some branches can arrange non-deterministic chunks, thereby avoiding perfect interjection, which is a prerequisite for novelty criteria (ii). However, if LI is less than 0.5, song grammars have more than 2 branches per node and it becomes difficult for female birds to successfully interject. In Fig.5, LI decreases gradually from 0.6 to 0.4 over time and LI never reaches 0 or 1. Such song grammars are not too simple and not too stochastic. This indicates that the song grammars have to be understandable by female birds.

In addition, we investigated the relationship between the complexity of song grammar (LI) and the degree of female preference (the number of nodes). In Fig.6, each 2 bins denote the average score and LI from different simulation runs, as the function of the fixed node numbers of female's P. We can see that both the complexity of song grammars and the communication score aren't affected by the degree of female preferences. This shows that the song grammars may evolve to become complex even with simple female preferences.

4 Conclusions

In our simulations, we observed: that song grammars could evolve to become complex via relatively brief, '*novel song-interjection*' communications; that the transition from lower complexity to higher complexity song grammars could be associated with the change of male birds' courting strategy.

By introducing a complexity measure for song grammar LI, we studied the effect of females' preference P. We found that the change of LI reflected the cognitive ability of the female birds; the complexity evolved by sexual selection must be understandable by females.

Lastly, we confirmed that the female birds with the relatively small number of node, could induce the complexity of song grammars as much as one with the larger number of node. This indicates that the complication of song grammars doesn't necessarily call on the complexity of female birds.

The birdsongs don't have rich meaning; however the forms (i.e. song grammars and female preferences) have great diversity and which enable rich communication without meaning. Our simulation results indicate that simple interjection by female induces diversity and complexity in song grammar.

References

1. J. Aitchison. *The Seeds of Speech: Language Origin and Evolution.* Cambridge University Press, 2000.
2. N. Chomsky, R. Huybregts, H.C.V. Riemsdijk, N. Fukui, and M Zushi. *The Generative Enterprise Revisited: Discussions With Riny Huybregts, Henk Van Riemsdijk, Naoki Fukui and Mihoko Zushi.* Walter De Gruyter Inc., 2004.
3. T. Hashimoto and T. Ikegami. Emergence of net-grammar in communicationg agents. *Biosystems*, 38:1–14, 1996.
4. M.D. Hauser, N. Chomsky, and W.T. Fitch. The faculty of language: What is it, who has it, and how did it evolve? *Science*, 298:1569–1589, 2002.
5. C.F. Hockett. The origin of speech. *Scientific American*, 203:88–96, 1969.

6. E. Honda and K. Okanoya. Acoustical and syntactical comparisons between songs of the white-backed munia(lonchura striata) and its domesticated strain, the bengalese finch(lonchura striata var. domestica). *Zoological Sciecne*, 16:319–326, 1999.
7. J.E. Hopcroft and J.D. Ullman. *Introduction to Automata Theory, Languages and Computation.* Addison Wesley, 1979.
8. T. Hosino and K. Okanoya. Lesion of a higher-order song control nucleus disrupts phrase-level complexity in bengalese finches'. *NeuroReport*, 11:2091–2095, 2000.
9. S.B. Michael and J.D. Allison. What songbirds teach us about learnig? *Nature*, 417:351–358, 2000.
10. K. Okanoya. Sexual display as a syntactical vehicle. In Alison W., editor, *The Transition to Language*, pages 46–63. Oxford University Press, 2002.
11. K. Okanoya. *Kotorino Uta kara Hito no Kotoba he (in Japnaese).* Iwanami Syoten, 2003.
12. K. Sasahara and T. Ikegami. Coevolution of birdsong grammar without imitation. In B. Wolfgang, C. Thomas, D. Peter, T.K. Jan, and Z. Jens, editors, *Advances in Artificial Life, 7th European Conference on Artificial Life*, pages 482–490. Springer, 2003.
13. K. Sasahara and T Ikegami. Song grammars as complex sexual displays. In *Artificial Life IX: Proceedings of the 9th International Conference on the Simulation and Synthesis of Living Systems.* The MIT Press, 2004.
14. L. Steels. *The Talking Heads Experiment. Volume 1. Words and Meanings.* Antwerpen, 1999.
15. J. Suzuki and K. Kaneko. Imitation game. *Physica D*, 75:328–342, 1994.
16. G.M. Werner and P.M. Todd. Too many love songs: Sexual selection and the evolution of communication. In Alison W., editor, *Fourth European Conference on Artificial Life*, pages 434–443. MIT Press, 1997.
17. A. Zahavi and A. Zahavi. *The Handicap Principle.* University Press, 1997.

Homophony and Disambiguation Through Sequential Processes in the Evolution of Language

Caroline Lyon, Chrystopher L. Nehaniv, Sandra Warren,
Bob Dickerson, and Jean Baillie

School of Computer Science, University of Hertfordshire, College Lane, Hatfield,
Hertfordshire AL10 9AB, United Kingdom
{C.M.Lyon,C.L.Nehaniv}@herts.ac.uk

Abstract. Human language may have evolved through a stage when words were combined into structured linear segments, before these segments were used as building blocks for a hierarchical grammar. This hypothesis is approached by examining the apparently ubiquitous prevalence of homophones. It suggests how, perhaps contrary to expectation, communicative capacity does not seem to be adversely affected by homophones, and how it is that they can be routinely used without confusion. These facts are principally explained by disambiguation through syntactic processing of short word sequences. Local sequential processing plays an underlying role in language production and perception, a hypothesis that is supported by evidence that small children engage in this process as soon as they acquire words. Experiments on a corpus of spoken English calculated the entropy for sequences of syntactically labelled words. They show there is a measurable advantage in decoding word strings when they are taken in short sequences, rather than as individual items. This suggests that grammatical fragments of speech could have been a stepping stone to a full grammar.

1 Introduction

"Don't think, but look!"

-L. Wittgenstein [24, sec. 66]

The usage of speech sounds lies at the core of the human ability to communicate, and to manipulate the world around through language. (See [24,15].)

The limited range of sounds that other creatures can make contrasts markedly with the much wider range and combinatorial use of phonetic elements in human speech. The physiological changes to the vocal tract that were necessary to enable the production of speech sounds has concomitant disadvantages, but the value of the mechanisms exapted or adapted to support language appears to have outweighed these problems [8].

If human language had been designed to a teleological programme, we might have expected that there would be an optimum number of phonemes that provided the basis for speech. However, we find that the number of phonemes in

A. Sakurai et al. (Eds.): JSAI 2003/2004, LNAI 3609, pp. 315–324, 2007.

human languages is diverse, varying from about 12 to well over 100, e.g. Hawaiian with 13, Khoisan !Xu with 141 [10]. There is massive redundancy. Some phonetic elements that can serve as particularly salient distinguishing features, such as clicks or ejectives, only occur in a subset of human languages. We see here not survival of the fittest, but survival of the many, varied, fit.

We might also have expected a one-to-one mapping between sounds and meanings. Indeed, recent mathematical models addressing how language might have evolved take this approach and show how a limited number of phonemes can be combined to produce an indefinitely large number of unambiguous words [16,17]. Nowak and collaborators assert that "ambiguity ... is the loss of communicative capacity that arises if individual sounds are linked to more than one meaning" [16, p. 613], that absence of word ambiguity is a mark of evolutionary fitness, and that "for a given phonemic matrix, the maximum fitness of a language increases exponentially with word length" [17, p. 158].

However, these models do not reflect language in the real world. Seemingly ubiquitous homophony is common in English as in other languages, though it is certainly not the case that a shortage of phonetic elements leads to a need for the same sounds to have multiple meanings. Many of the most frequently used words are ambiguous homophones (for example: *to, too, two; there, their; I, eye*). In spite of the theoretical possibilities of exploiting combinatorial properties of a set of phonemes, this does not in practice necessarily occur, yet communicative capacity does not seem to be adversely affected. We find homophones in the speech of small children [22] and observe the slippage of language into forms with more homophones ([21],[20, p. 5]).

Another way the models differ from linguistic reality is that they do not address costs associated with increases in word length. Indeed, standard information-theoretic methods (e.g. [3]) can be applied to optimize word length with respect to production costs. This generally results in frequent words having a shorter length (less costly to produce) and less frequent words having longer length, in order to minimize expected signalling costs (cf. [3,14]). Evolutionary considerations suggest that this should be the case for language and communicative signalling in general [14], even without any homophony.

2 Analysis of Homophones

We can analyse homophones in two groups: those in which the homophonous forms are the same grammatical parts-of-speech (type 1), and those in which they are different parts-of-speech (type 2). In English, and other languages the second class is much the larger [7]. Taking the smaller class of type 1 homophones first, semantic information and contextual may be necessary to distinguish these words. They may be distinct concepts spelt differently, such as *hair* and *hare*; or distinct concepts spelt the same such as *(river) bank* and *(money) bank*. They may have common ancestry, and been subject to a gradual semantic shift. For instance *to stamp* can have the distinct meanings to stamp a foot, or to stamp a letter. Linking these two meanings was a stage when letters were sealed with a heavy stamp. Homophonous forms may also be variations on a theme, as in the example from Wittgenstein of the word *game* [24, sections 66-76]. He points out that there is nothing common to all meanings of the word, but rather a complicated network of similarities, overlapping and criss-crossing. This class of homophones with the same parts-of-speech has been the subject of mathematical modelling, for example by Wang et al. [20], where there seems to be an implicit assumption that they are *content words.*

However, the much larger class of homophones that are different parts of speech raise significant issues and deserve further scrutiny. Homophonous forms are frequently *function words* – often having little or no direct semantic and referential content, such as pronouns and prepositions – and the fact that we can disambiguate them with such facility provides clues to our underlying syntactic abilities. Typically, one or both lexical items in a type 2 homophone pair are function words. For example, the words *to / too / two* are used and understood correctly by children very early on. It is clear that disambiguation must be through contextual processing, and this contextual processing seems to be mainly based on relations with adjacent words (for example: *me too, two sweets, to the swing*). The subconscious use of grammatical categories can explain how the appropriate lexical item is selected. Without invoking a full grammar, short word sequences, grammatical fragments, can be acceptable or not.

3 Perception and Production of Syntactically Correct Phrases

There is an ongoing debate as to how children acquire syntactic knowledge (see, e.g. [19,5,23]), but there is a general consensus that children from a very young age are aware of syntactic categories. Infants are aware of prosodic clues to syntactic elements, and can exploit them in the processing of speech [13,4]. For instance in English, children use correct word order as soon as two words are produced [12]. This helps to explain how young children can understand phrases and sentences with homophonous terms: local syntactic constraints are employed as soon as words are acquired. Older speakers as much as infants are implicitly aware of syntactic categories. The fact that many could not explicitly define these categories does not detract from the proposition. In the same way, we can estimate the distance to a remote object implicitly using optical rules that we cannot explicitly formulate.

If we accept this proposition, then we can see that the disambiguation of homophones will often be based on the admissibility or otherwise of neighbouring parts-of-speech. For instance, consider *their / there. "their"* is a possessive pronoun typically followed by a noun or noun phrase. *"there"* is not usually followed by a noun or noun phrase, but typically by a verb, adverb or preposition:

Their adventures made a good story.
Their thrilling exploits amazed us.

There are many more to come.
They went there quickly.

The alternative forms *their / there* cannot be confused, because of local syntactic disambiguation. For homophonic *function words* like these, there is little or no content in them to aid disambiguation, nor is it necessary.

4 Experiments on the Efficient Decoding of Word Strings

The observations made so far suggest that processing short, syntactically labelled word sequences could play an underlying role in speech production and perception. To test this hypothesis, we carried out experiments to see if there was an advantage in processing words as short strings rather than as individual items.

Using Information Theoretic tools we have investigated the efficiency of decoding word sequences segmented in different ways. The concept on which these experiments are based is that we can measure the entropy of a sequence, and a decline in entropy is associated with an increase in predictability, an improvement in the efficiency of decoding and comprehensibility [18]. For a simple introduction to this concept see [9, p. 170]. A standard reference is [3].

Taking the proposition that we are implicitly aware of syntactic categories or part-of-speech tags, we investigate whether tag strings are more easily decoded if they are taken in short sequences rather than as single items. In the rest of this paper we take the term "tag" to mean "part-of-speech tag". If we find that entropy declines as we take tags in pairs and triples, this would indicate that processing of short sequences is likely to have developed with improved understanding of speech. In turn, this would help explain how homophonous words are routinely used without confusion: they are disambiguated by being taken in conjunction with neighbouring words.

For our experiments, we take the Machine Readable Spoken English Corpus MARSEC, organized by Arnfield [11,1]. About 26,000 words are used. MARSEC includes prosodic annotation, which we are not using in the current experiments. The corpus includes unscripted news commentary, scripted news and lectures. This can be considered well formed language, not like informal conversation. Experiments are planned on other types of spoken language, and on larger corpora.

The first step in the experiment is to map words onto part-of-speech tags. This was done using a version of the CLAWS tagger (supplied by the University of Lancaster) described by Garside [6]. The CLAWS tagset was mapped onto a smaller customized tagset consisting of 26 part-of-speech tags (Appendix B).

The next stage is to measure the entropy in four cases: with no statistical information, then with information on single tags, tag pairs and tag triples.

Taking the symbol H as entropy, H is the average number of bits needed to determine a symbol (tag). We need to find:

- H_0 : entropy with no statistical information, all symbols equi-probable.
- H_1 : entropy from information on the probability of single symbols occurring.
- H_2 : entropy from information on the probability of 2 symbols occurring consecutively.
- H_3 : entropy from information on the probability of 3 symbols occurring consecutively.

In general, H_n measures the uncertainty about the nth symbol that remains if the preceding $n-1$ symbols are known.

5 Description of Entropy

Intuitively, we are looking at how much extra information about the part-of-speech tag we have when we take contextual information into account. Take for example the partial sentence:

We	see	a	complicated	network
$\langle pronoun \rangle$	$\langle verb \rangle$	$\langle article \rangle$	$\langle adjective \rangle$	$\langle noun \rangle$

Consider the prediction of the tag of the word "complicated". With no statistical information, we can only say that all tags are equally probable. With information on the probability of single tags occurring we can make a better estimate. If we have information on tag pairs this is improved further, and we can better predict the tag of the word "complicated" if we know it is preceded by the tag $\langle article \rangle$. With information on tag triples we can again make a further improvement in prediction, also knowing that the tag of "complicated" is followed by a noun. By calculating the entropy, we have a metric to quantify our intuitive understanding.

5.1 Formula for Entropy

In mathematical terms let $\mathcal{T}$ be a tagset, and X be a discrete random variable taking values x in $\mathcal{T}$. The probability that X takes symbol x as its value, is $p(x)$

$$p(x) = probability(X = x)$$

$$H(x) = -\sum_{x \in \mathcal{A}} p(x) * \log_2 p(x)$$

Since $p(x)$ is a probability, $p(x) \leq 1$, so $\log_2 p(x)$ is negative, (or zero if $p(x) = 1$). The minus sign at the start of the formula cancels this out. The derivations for the formulae for H_0, H_1, H_2, H_3 are given in Appendix A.

Table 1. Entropy measures for 26,001 tagged words of the MARSEC corpus, with 26 tags, compared to entropy of sequences of independently generated random numbers with 26 equi-probable values

Speech representation	H_0	H_1	H_2	H_3
Entropy of 26001 tags	4.70	4.10	3.31	2.99
Entropy of 26001 random numbers	4.70	4.70	4.68	4.16
Entropy of 500K random numbers	4.70	4.70	4.70	4.68

5.2 Results

Shannon showed that the entropy of a sequence will decline as more contextual information is taken into account, if there are dependencies between neighbouring items. The n-gram entropy H_n measures average uncertainty of the next symbol extending over n adjacent symbols when the preceding $n-1$ symbols are known. It will always be the case that $H_n \leq H_{n-1}$: there cannot be less information when an extra term is taken into account. However, if there are no dependencies between adjacent terms (for instance, if the sequence is random) then the entropy will stay at a similar level. If entropy declines as more context is taken into account, this indicates that there are dependencies between neighbouring terms.

Applying this analysis to the tagged MARSEC corpus, and to sequences of random numbers for comparison, we get the results shown in Table 1.

The figure for H_3 for the tagged corpus in Table 1 means that having identified the part-of-speech for two consecutive words in an utterance, there are on average about 3 bits of uncertainty in the next tag. Additional information, e.g. from the phonetic stream, would further reduce this uncertainty in the disambiguation of homophones.

We find a decline in entropy between H_0 and H_3. As information is taken over more adjacent tags the uncertainty decreases, comprehensibility increases. This suggests that processing of short sequences was likely to emerge in evolutionary changes, as it would be correlated with an improvement in communicative success.

As we have a relatively small corpus, we cannot find H_4 and beyond, since errors from undersampling will give misleading results. The entropy of the sequences of independently generated random numbers for H_n as n increases should not decline, and it can be seen from Table 1 that there is some slight distortion with only 26,000 items for an alphabet size of 26.

6 Conclusion

We have analysed the level of uncertainty in processing short word sequences from a corpus of transcribed English speech. The results lend credence to the theory that sequential processing plays a role in the perception of language. This can help to account for the fact that we do not seem to have any difficulty in disambiguating homophones from different syntactic categories: the possible interpretations of the homophonic form are limited, usually to one, by local sequential and contextual constraints.

One of the purposes of this paper is to open up a discussion on the distribution of homophones in other languages, and to investigate whether syntactic processing of short word sequences is universally advantageous, or just a phenomenon found in a limited group of languages.

At a low level, sequential processing by primitive neural elements plays a key role in the production of human speech [8], and similar sequential processes also operate at higher levels of speech production and comprehension. We see that observed combinations of phonemes are controlled by sequential regulators, phonotactic rules. Then groups of phonemes are combined into syllables, and syllables into words subject to morphophonemic constraints. The focus of this paper has been the next level, where short word sequences are processed. We find that syntactic categories are needed for disambiguation before we move up to a full, sentence based hierarchical grammar. This is consistent with the hypothesis that language may have evolved through an intermediate stage of structured linear segments, before these segments were themselves used as building blocks for a hierarchical grammar [9].

It is hard to resist the temptation to conjecture about how initial moves towards structured language may have begun. Consider, for instance, a hunting scenario in the distant past. Suppose the leader of one tribe can say things like: “Hide in the tree where I waited with your father while I chase the deer past you down the gully”. Compare this to the utterance of the leader of another tribe that can only string words together in an unstructured way: “Hide tree wait father chase deer gully”. If one of the aims of speech is to transfer information, then an advantage is conferred by the ability to use structured language rather than unstructured strings of words. To use short sequences of words that constitute meaningful phrases and clauses may be an initial step in this process. Sequential processes can then contribute to disambiguation between homophones and between syntactic categories in early stages in the emergence of linguistic communication.

Experiments are planned to extend the study of these phenomena in larger corpora and also to measure the decrease in entropy due to the context before and after a item in a sequence.

References

1. S Arnfield. *Prosody and Syntax in Corpus Based Analysis of Spoken English.* PhD thesis, University of Leeds, 1994.
2. T C Bell, J G Cleary, and I H Witten. *Text Compression.* Prentice Hall, 1990.

3. T M Cover and J A Thomas. *Elements of Information Theory.* John Wiley & Sons, Inc., 1991.
4. P Eimas. The perception and representation of speech by infants. In J Morgan and K Demuth, editors, *Signal to Syntax.* Lawrence Erlbaum, 1996.
5. C Fisher. The role of abstract syntactic knowledge in language acquisition: A reply to Tomasello. *Cognition*, 82(3):259–278, 2002.
6. R Garside The CLAWS word tagging system. In R Garside, G Leech and G Sampson, editors, *The Computational Analysis of English: A Corpus Based Approach.* Longman, 1987.
7. Jinyun Ke, Feng Wang, and Christophe Coupe. The rise and fall of homophones: A window to language evolution. In *Proceedings of 4th International Conference on the Evolution of Language*, 2002.
8. P Lieberman. On the nature and evolution of the neural bases of human language. *Yearbook of Physical Anthropology*, 2002.
9. C Lyon, B Dickerson, and C L Nehaniv. The segmentation of speech and its implications for the emergence of language structure. *Evolution of Communication*, 4(2):161–182, 2003.
10. I Maddieson. *Patterns of sounds.* Cambridge University Press, 1984.
11. MARSEC. *Machine Readable Spoken English Corpus.* `http://www.rdg.ac.uk/AcaDepts/ll/speechlab/marsec/` .
12. R Mazuka. Can a grammatical parameter be set before the first word? Prosodic contributions to early setting of a grammatical parameter. In J Morgan and K Demuth, editors, *Signal to Syntax.* Lawrence Erlbaum, pp. 313-330, 1996.
13. J Morgan and K Demuth. Signal to syntax: An overview. In J Morgan and K Demuth, editors, *Signal to Syntax.* Lawrence Erlbaum, pp. 263-283, 1996.
14. C L Nehaniv, Meaning for observers and agents, *IEEE International Symposium on Intelligent Control / Intelligent Systems & Semiotics (ISIC/ISAS'99).* IEEE Press, pp. 435–440, 1999.
15. C L Nehaniv, The making of meaning in societies: Semiotic and information-theoretic background to the evolution of communication. In B Edmonds and K Dautenhahn, editors, *Proc. AISB Symposium: Starting from Society - the application of social analogies to computational systems.* Society for the Study of Artificial Intelligence and Adaptive Behaviour, pp. 73–84, 2000.
16. M A Nowak, N L Komarova, and P Niyogi. Computational and evolutionary aspects of language. *Nature*, 417:611–617, 2002.
17. J B Plotkin and M A Nowak. Language evolution and information theory. *Journal of Theoretical Biology*, 205:147–159, 2000.
18. C E Shannon. Prediction and entropy of printed English (1951). In N J A Sloane and Aaron D Wyner, editors, *Shannon: Collected Papers.* IEEE Press, 1993.
19. M Tomasello and P Brooks. Early syntactic development: a construction grammar account. In M Barrett, editor, *The Development of Language.* Psychology Press, Hove, UK, 1999.
20. William Wang, Jinyun Ke, and James Minett. Computational studies of language evolution. In *Proceedings of COLING*, 2002.
21. P Warren, M Rae, and J Hay. Goldilocks and the three beers. In *9th Australian International Conference on Speech Science and Technology*, 2002.
22. S Warren. *Phonological Acquisition and Ambient Language: A Corpus Based, Cross-Linguistic Exploration.* PhD thesis, University of Hertfordshire, UK, 2001.

23. S Waxman and D Markow. Words as invitations to form categories: evidence from 12- to 13-month-old-infants. *Cognitive Psychology*, 29(3), 1995.
24. L Wittgenstein. *Philosophical Investigations*. Blackwell, 1953. Translated by G. Anscombe.

Appendix A

Derivation of the Formula for Calculating Entropy

This is derived from Shannon's work [18] on the entropy of symbol sequences. He produced a series of approximations to the entropy H of written English, taking letters as symbols, which successively take more account of the statistics of the language.

H_0 represents the average number of bits required to determine a symbol with no statistical information. H_1 is calculated with information on single symbol frequencies; H_2 uses information on the probability of 2 symbols occurring together; H_n, called the n-gram entropy, measures the amount of entropy with information extending over n adjacent symbols[1]. As n increases from 0 to 3, the n-gram entropy declines: the degree of predictability is increased as information from more adjacent symbols is taken into account. If $n-1$ symbols are known, H_n is the conditional entropy of the next symbol, and is defined as follows.

b_i is a block of $n-1$ symbols, j is an arbitrary symbol following b_i
$p(b_i, j)$ is the probability of the n-gram consisting of b_i followed by j
$p_{b_i}(j)$ is the conditional probability of symbol j after block b_i, that is $p(b_i, j) \div p(b_i)$

$$\begin{aligned} H_n &= -\sum_{i,j} p(b_i, j) * \log_2 p_{b_i}(j) \\ &= -\sum_{i,j} p(b_i, j) * \log_2 p(b_i, j) + \sum_{i,j} p(b_i, j) * \log_2 p(b_i) \\ &= -\sum_{i,j} p(b_i, j) * \log_2 p(b_i, j) + \sum_{i} p(b_i) * \log_2 p(b_i) \end{aligned}$$

since $\sum_{i,j} p(b_i, j) = \sum_i p(b_i)$.

Appendix B

Description of the Tag Set

The tagset used in these experiments is derived from CLAWS4, mapped onto a smaller set of 26 classes. They are as follows:

[1] This notation is derived from that used by Shannon. It differs from that used, for example, by Bell, Cleary and Witten [2].

- article - singular e.g. “a”
- determiner - singular or plural “the”
- predeterminer e.g. “all”
- pronominal determiner e.g. “some”
- pronominal determiner - singular e.g. “this”
- proper noun
- noun - singular
- noun - plural
- pronoun - singular
- pronoun - plural
- relative pronoun
- possessive pronoun
- verb - singular
- verb - plural
- auxiliary verb - singular
- auxiliary verb - plural
- existential “here” or “there”
- present participle
- past participle
- infinitive “to”
- preposition
- conjunction
- adjective
- singular number “one”
- adverb
- exceptions

The tagging process includes the identification of common phrases or idioms, which are then treated as single lexical items. For instance, “of course” is tagged as an adverb.

Mirroring, Deixis, and Interaction Topology in the Emergence of Shared Vocabularies

Jean Baillie, Chrystopher L. Nehaniv, Patrick Quick,
Attila Egri-Nagy, and Sandra Warren

School of Computer Science, University of Hertfordshire,
College Lane, Hatfield, Herts AL10 9AB, United Kingdom
{J.Baillie, C.L.Nehaniv}@herts.ac.uk

Abstract. Neuroscientists have suggested that the mirror-neurons in our primate ancestors may have provided a substrate for the emergence of language in humans. Simulation studies of the emergence of language, using minimal implementations of proposed mechanisms, are a way to assess their explanatory power for the emergence and evolution of communication. In this work, we study the emergence and stability of linguistic labelling in a communities of agents with mirror-neuron mechanisms for associating deictic reference with speech utterances. These minimal agents possessing a built-in mirror-neuron style temporal recurrent neural network architecture are capable of perceiving and carrying out deixis ('pointing') to refer to others in their group, as well as producing and perceiving utterances of another agent in their group. They are able to generate and learn temporally extended phonetic utterances ('names') and associate these to deictic referents. Thus, the agents utter what they hear, and tend refer to the same entities as another agent that they watch when it points. Previous work has shown the emergence and stability of arbitrary names generated by the agents in certain fixed topologies of interaction. In this work, we systematically study the effects of different interaction topologies on the dynamics of convergence to a common vocabulary in the population, and its stability over time. Results show that certain topologies of interaction to be more conducive than others to the emergence of a stable vocabulary. Moreover, some topologies of interaction (such as cycles) are seen to yield instability and to amplify feedback given the mirror-neuron system. Linguistic convergence and change bear similarity to those of natural language. Homophony and multiple referents of particular proto-words may also emerge. In light of results, we suggest that mechanisms for confirming joint-attention and for suppression of mirroring could play an essential role in maintaining stability in the emergence of linguistic reference.

1 Preliminaries and Motivation

The debate over the existence and role of a possible innate language acquisition device (e.g. [1,2,3,4]) for human language as opposed to language-readiness based on other, more general cognitive capabilities (e.g. [5,6,7]) motivates the study of the origin and maintenance of language and language-like phenomena in animals (e.g. [8,9,3,10]) and artificial agents, including very simple ones, (e.g. [5,11,12,6,13,14,15,16]).

A. Sakurai et al. (Eds.): JSAI 2003/2004, LNAI 3609, pp. 325–334, 2007.

We study the development and dynamics of the use of naming in the course of interaction of communities of agents endowed with simple deixis and 'speech' production/perception capabilities and possessing only extremely simple connectionist cognitive capabilities. Deixis—or pointing—via gestures or gaze direction, or also by the use of other signals, including words like *this* or *those*, can serve as a substrate for joint attention that grounds the development of linguistic and social abilities in humans as they grow from pre-linguistic infants into adults (cf. [17,18,19]). Roboticists have begun to implement the mechanisms of joint attention (such as gaze detection and deictic gaze) as steps toward achieving social competencies in robots [20,21,22]. Deixis will also serve as a useful substrate component for the implementation and development in a social context of linguistic capabilities in robots and agents. Moreover, deixis-based learning can later allow reference to absent entities [16]. Major open questions (cf. [15]) in the emergence and evolution of linguistic communication include understanding how the following can emerge in the course of evolution of language and its precursors: (1) negation, (2) predication (perhaps approached via a topic-comment stage), (3) compositional syntax, (4) shared referencing and usage conventions, and (5) reference to spatially and temporally removed entities. In the study of issues (4) and (5), deixis plays particularly important roles.

Connectionist Architecture to Mirror Speech and Deictic Reference. Our agents mirror speech and deixis of agents they attend to; they start with no pre-existing vocabulary, and may generate new speech output when there is no perceived response to their deictic acts. Even in this case, and without a reinforcement feedback mechanism, time-delay recurrent connectionist architecture, deixis, and mirroring may support the emergence of shared vocabularies of names for the referents of the deictic acts.

DRAMA [13] is a time-delay recurrent neural network which uses Hebbian update rules[1] and which was designed for dynamic control and learning of autonomous robots. It is used here to study the development over time of a simple proto-language consisting of shared – or somewhat shared – names. Connections in DRAMA are associated with two weight parameters: a confidence factor, ω, modeling the frequency of correlated activation of any two units; and a time parameter τ which makes correlations between delayed and simultaneous occurrences of different input patterns. The network is fully recurrent, with asymmetric directed links, and the weight parameters record separately the spatial and temporal features of the input patterns. For a full account of the architecture, see [13]. In the case studies described in [13,14] which involved learner and teacher robots, not only was it necessary to make sensor/actuator associations but it was also important that agents 'remember' the delay experienced between the events thus associated. This is exploited here to associate deixis with temporally proximal utterances, and each phoneme in an uttered 'proto-word' with the following one. In our agents the sensory array consists of an event-detector which reads data from a designated agent's deixis and speech (via an actuator output buffer). An event is deemed to have occurred if the value of an input data signal exceeds the previous value of that signal (stored in the sensor) by a predefined threshold. Although the incoming data is real-valued, as

[1] The basic Hebb rule is that "neurons that fire together wire together". In DRAMA this 'wiring' encodes not only their tendency to fire together, but also a particular time-delay interval between the firing of two neurons.

indeed is the threshold value, the actual values transmitted to the DRAMA network are binary—a DRAMA input node takes the value 1 if an event has been detected and 0 otherwise. Perception of each action (deixis or utterance of a phoneme) and its actuation occur using the same nodes. This gives one simple realization of a mirror-neuron model.

2 Model and Methods

The Paradigm. We wish to construct a community of interacting DRAMA agents in which individual agents may focus their attention on various others (including themselves) and in which agents may join or leave the community at will. As in human communities, agents may have little control over which other agents are focusing on them—indeed there may be no-one listening at all—and may change their own focus over time. They may also miss a communication, even from the agent on which they are focusing, if their attention was not engaged at the critical moment. Against such a background, we wish to observe the development of naming in the agents through *interaction games* [5,23] involving deixis and utterances, and in which the DRAMA agents are themselves referents.

In the particular experiments reported on here, there are five DRAMA agents involved in the game. Each agent has a sensor array, and an output buffer to which the values of its network's nodes at the end of each processing cycle are written, effecting deictic reference and speech. In these simple experiments an agent focuses on exactly one other agent in the community throughout an interaction game. The number of agents taking part is fixed; agents joining and leaving the community is not implemented here but is left for future studies. Agents may or may not perceive deixis and speech from the agent they are focused on, depending on the details of the update processing.

Interaction. The set-up generally follows [16], in which, in contrast with [5] where processing is sequential, agents here are operating concurrently, with random scheduling of time updates for each agent forming an integral part of the communication. It is a feature of the architecture that the values of DRAMA nodes decay by a fixed factor over time (that is, over a number of processing cycles, eventually dying altogether) in the absence of any event that 'refreshes' them. If a *read* operation of an agent's output buffer by a particular sensor is not scheduled for a number of processing cycles, a data value may well have decayed to below the point where an event is deemed to have occurred and so may not be detected when its reading is eventually scheduled.[2]

At the outset, each agent may generate a referent to which it points, and then 'listens' for a response through its sensor array. In practice for any given agent there is a good chance that no other agent is in fact listening, so any response received will not, from an external perspective, be a response in the strict sense at all, though the agent itself has no knowledge of that. Agents know which other agent they are listening to but not who is

[2] Previously we used *asynchronous* scheduling depending on emergent time-slicing in the operating system [16,24], but the current system supports concurrency in a more controlled manner and yields similar results. Moreover, it allows results to be reproduced by fixing the random seed and time-slicing mechanisms.

listening to them. If no response is received within a (predetermined) number of cycles, the agent might itself point again and generate utterance (creating a ‘name’), though with a low probability. Intuitively, we can think of agents as ‘asking’ one another for the name of the referent. They point at a referent and listen for a response. If a name is heard then it may be learned; if not, then one might be generated and uttered, and (possibly) learned by any listening agents. The communicative interaction is permitted to continue for a fixed number of cycles (per agent: running processes are non-deterministically interleaved) after which all nodes are reset to zero and the process starts again with a newly generated referent.

Tracking the Development of Vocabularies. Interaction continues over thousands of cycles and the values of the two sets of connections for each agent are sampled and stored every cycle. At the end of each cycle, retrieval tests are run on stand-alone agents, created by loading the recorded state of the agents (that is, the connections) from the interaction sessions and proto-words or invariants based on them are logged. Tracking the time evolution of each agent’s vocabulary is effected in a non-concurrent environment at each time-step; a clone (from a particular moment) of each agent is treated as a ‘linguistic informant’ and ‘interviewed’ separately from the others, with no further learning permitted during this phase. It is prompted with a referent and its responses recorded over time. The interrogation process is repeated for each referent. The trajectory of its each agent’s vocabulary development and change over time is thus recorded.

Neural Organization. In the application described in this paper each sensor array and its associated DRAMA agent consists of 14 network nodes (though of course it is not necessary for the sensor to have the same number of nodes as the network, and also there may be many sensors in an application). Notionally, the nodes are classified in two ways: nodes 0-4 are deictic, and nodes 5-14 are phonetic output. Events on the output nodes are mapped onto the elements of a phonetic array; the phonetic ‘utterances’ are output when a corresponding node ‘lights up’, that is, takes the value 1. The resulting successive sounds are referred to as ‘proto-words’, and as ‘names’ if they are associated with deictic acts. Here there is no positive or negative reinforcement: agents have no feedback as to whether linguistic behavior is “successful” (in contrast to [5]). The temporal recurrent neural networks merely associate perceived deixis to utterances that it senses as occurring proximately in time.

Deixis and Speech Output. Each agent has a built-in ability to point at any other agent (or itself). The interaction for the DRAMA agents works as follows: a referent is randomly generated within the agent, that is to say, one, and only one, of the deictic nodes is set to 1. For example, agent 3 is pointed to when the third node is set. Any agent attending to that agent perceives the pointing as deixis to the appropriate referent agent. As stated above, each agent listens to one other agent and the speech output from that agent is taken in via the listening DRAMA agent’s sensor, where events may be detected. This may or may not happen immediately, depending on when the process is next scheduled. If in response to deixis (that is, while the referent node still has a positive value) events are detected on the output nodes, then the association between the referent and the output is learned. If there is no response from others, a response from the agent itself may be generated, since each agent perceives its own deixis. Utterances are not atomic but have duration and may be uttered over several time steps. A newly generated name takes the following form: two of the agent's own output nodes are successively set to 1, that is, one of the *consonant* nodes 5-9 and then one of the *vowel* nodes 10-14, ensuring (one might assume) that at retrieval time one consonant and one vowel are generated (such is the arrangement of the phonetic array onto which these nodes are mapped, though this is arbitrary). The consonants are b, p, f, k, t and the vowels a, e, i, o, u. After a fixed number of processing cycles each agent's nodes are reset to 0 and for several more cycles the network operates on inputs of patterns of 0's, flushing the system, before the process is repeated with a newly generated referent.

Measuring Convergence: A Proto-Word Invariant. To assess the similarity of words we use the following *invariant of proto-words*. If $w = a_1 \ldots a_n$ is word we calculate an integer value the following way: $v(w) = \sum_{i=1}^{n-1} |o(a_{i+1}) - o(a_i)|$, where $o(a)$ gives the ASCII value of letter a. In all cases, w includes a non-empty prefix and suffix consisting of silence (denoted by the symbol $*$). The invariant takes similar values for proto-words that have similar transitions between phonemes (and between silence and a phoneme). Generally the invariant will take higher values in proto-words including many transitions (e.g. for most long proto-words). The invariant is coarse, but fine enough to indicate convergence; moreover, it is unaffected by temporal shift in the onset of utterances. Similar proto-words used by agents for a referent will have similar invariant values; therefore clustering of invariant values is suggestive of convergence to a shared proto-word for a given referent.

Topologies. We track the invariant on the utterances elicited from agents the community. In this study all possible fixed topologies of intercommunication for 5 agents in a single community have been considered where each agent attended to itself and exactly one other agent (see Figure 1, in which arrows from an agent A to another agent B point from a listener to a speaker, whose utterances and deixis may be perceived). In mirror-neuron fashion, each listener's own deictic reference and speech neural nodes can be activated by pointing and utterances of the speaker.

3 Experimental Results

For each topology, five runs of the system each consisting of 5000 concurrent time-steps for each agent were run. Time evolution of proto-word invariants of the proto-word associated to each possible referent agent by each member of the community were recorded. Due to the nature of the invariant, having no utterance for a particular referent yields a value close to zero, while longer proto-words tend to include many transitions and thus yield high values of the invariant. Clustering of the graphs of the invariants from each speaker is indicative of lexical convergence in the population, while spread would be indicative of many different proto-words being used for the same referent. A characteristic sample for topology **2cycleA** is shown in the figure. Each graph shows data for a fixed referent; different lines indicate the changing values of the invariant (vertical axis) of name for the referent used by the different agents over time (horizontal

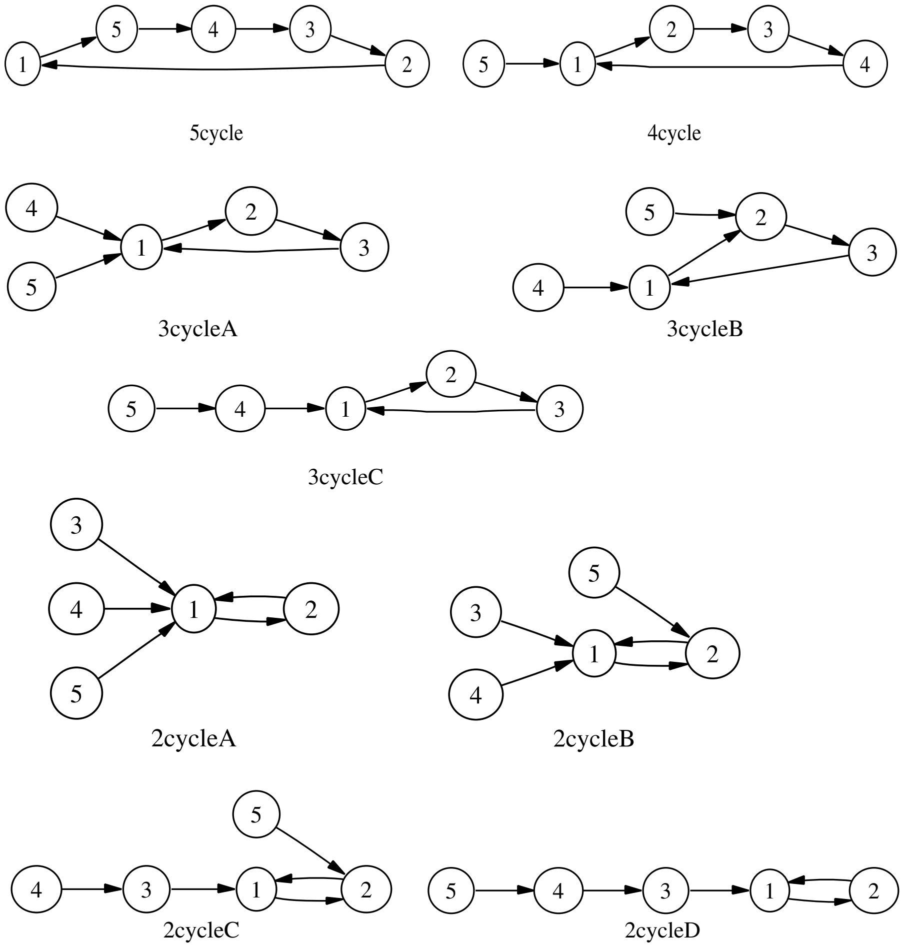


Fig. 1. All 9 possible intercommunication topologies for communities of 5 agents where each agent attends to exactly one other agent (and itself)

axis). The proto-words at the end of 5000 time-steps used by agents 1 to 5 for a given referent are listed to the right of the graphs (where $*$ indicates silence).

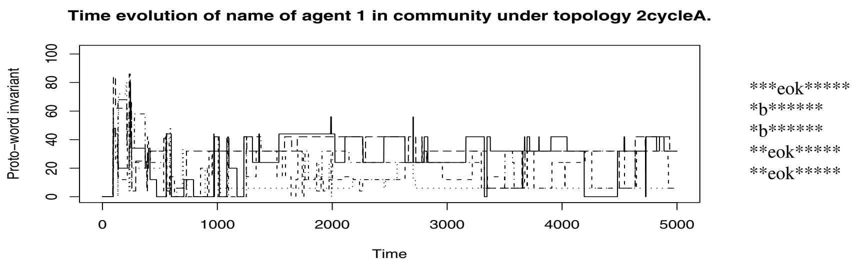
Several topologies – especially those with long cycles (**5cycle**, **4cycle**) – showed initially self-reinforcing unstable fluctuations in which proto-words rapidly increased in length during the first 1000 time-steps or so, but then settled into generally quite short, and more-or-less generally shared referential proto-words. Other topologies exhibited more initial stability but not necessarily better long term convergence.

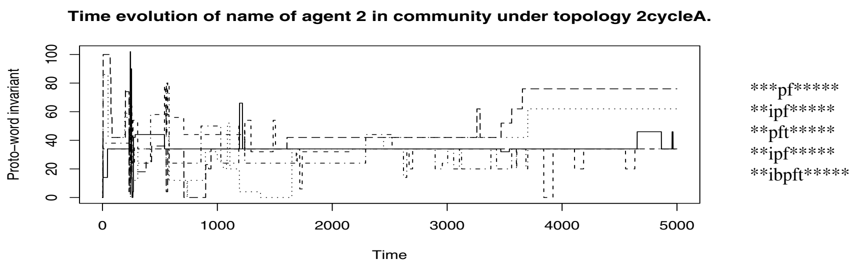
Due to space restrictions results from more runs and these other topologies are not shown here. Similar to our previous work [16,24] we observed some degree of convergence (in the form of shared vocabulary) to common proto-words for all possible referents, loss and replacement of words, as well as phonetic drift, continual change, and also sometimes homonymy, but also names with multiple referents.

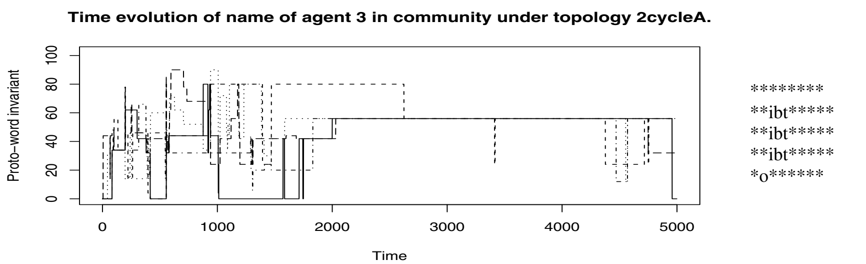
4 Discussion of Results and Conclusions

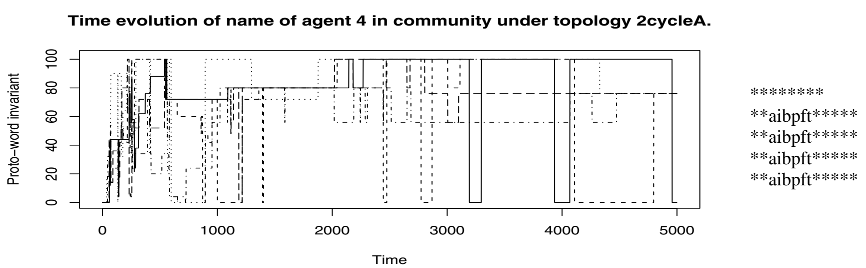
In the development of language and early word acquisition in children, "pointing connects a visual referent to the concurrent sound stream so that a relation of identity exists between these two aspects of the infant's perceptual experience" [25]. We have implemented this kind of association using simple agents organized in various fixed topologies. From result graphs some convergence of naming in the resulting proto-language can be seen in each experiment and topology. The connectionist architecture used has properties very close to those exhibited by mirror neurons, which may have played a role in the evolution of human language readiness [26,7]. Mirror neurons in the premotor cortex of monkeys fire when a certain affordant action is perceived as well as when it is used. This happens whether the actor is the animal itself or another who is being observed by the animal (see references cited for more details). Similarly in this implementation of the DRAMA architecture, agents make no distinction between a name that has been heard and one that was generated by the agent itself; in both cases the same neurons fire. Also, in observing or generating deictic acts the same neurons fire. Through interaction associations of temporal patterns of deixis and utterance are learned. Although syllables are initially generated in consonant-vowel pairs, they have temporal extent and are later not necessarily spoken in that order. In the data, we see the emergence of more complex articulations than just simple consonant-vowel pairs, although this had not been intended. Moreover, we also observe slow or abrupt phonetic drift in the proto-words, and the replacement of one word by another in all or part of the community. Similar phenomena of convergence and lexical drift and phonetic change as seen in the data are also well-known from historical linguistics of natural language (see, for example, [27] for a good introduction). Even without an explicit feedback mechanism reinforcing any 'successful' communicative acts, many phenomena of community linguistic ontogeny are apparent. It is an interesting research question to characterize exactly how much reinforcement is really necessary for achieving various aspects of linguistic phenomena. Notions of connecting signalling to costs and benefit for individual agents acting on the world might be necessary to achieve highly sophisticated communication systems (cf. [23]), but minimal mirroring, deixis and associative temporal learning already are seen here to yield some characteristic phenomena of language.

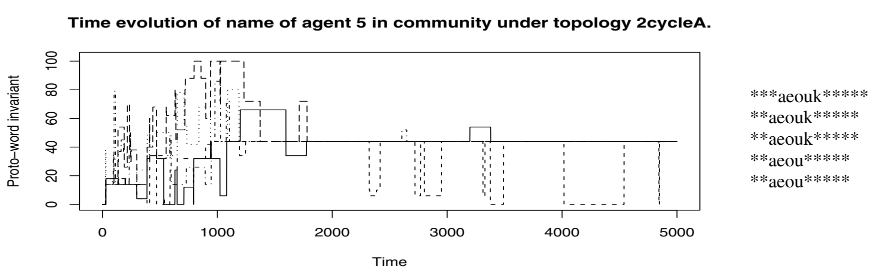
However, the results tend to confirm that the topology of interactions may have strong effects of the dynamics of proto-word stability, on length (apparently due to uncontrolled feedback effects arising from mirroring with no inhibition), and on convergence patterns over time. Semantic drift generally cannot be observed in these experiments since the deictic mirroring tends not to permit much ambiguity of referent, but we would expect it with less tight grounding (cf. [28] where ranges of sensor values are used instead of deixis). An exception occurs when a proto-word is uttered by a speaker agent although the listener agent has pointed to a different referent: since the proto-word is associated by the listener to a different referent than by the speaker, this led sometimes to homonymy or near homonymy. Several unforeseen results occurred with these minimal agents: (1) *Homophonous names* can result when an agent pointing to another heard an utterance referring to a third agent, due to associative learning. Also, (2) use of names with *multiple referents* (or group referents) arose due to occasional activation of multiple deictic nodes through mirroring, whose joint activation becomes associated to


Time evolution of name of agent 1 in community under topology 2cycleA.
Proto−word invariant
0
20
40
60
80
100
0
1000
2000
3000
4000
5000
Time
eok**
*b******
*b******
eok***
eok***


Time evolution of name of agent 2 in community under topology 2cycleA.
Proto−word invariant
0
20
40
60
80
100
0
1000
2000
3000
4000
5000
Time
pf**
ipf***
pft***
ipf***
ibpft***


Time evolution of name of agent 3 in community under topology 2cycleA.
Proto−word invariant
0
20
40
60
80
100
0
1000
2000
3000
4000
5000
Time
****
ibt***
ibt***
ibt***
*o******


Time evolution of name of agent 4 in community under topology 2cycleA.
Proto−word invariant
0
20
40
60
80
100
0
1000
2000
3000
4000
5000
Time
****
aibpft***
aibpft***
aibpft***
aibpft***


Time evolution of name of agent 5 in community under topology 2cycleA.
Proto−word invariant
0
20
40
60
80
100
0
1000
2000
3000
4000
5000
Time
aeouk**
aeouk***
aeouk***
aeou***
aeou***

any concurrently perceived utterance. *Attentional mechanisms* and *inhibition of lateral activation* of mirror-neurons (e.g. among deictic nodes for possible referents) was not modelled, but it seems likely that this would modulate the influence of processes resulting in (1) and (2). The dynamic instability and lack of complete convergence for some speakers to a common proto-word for a referent in some cases studied here suggests that more than just mirroring of deixis and utterance is necessary for very stable referential vocabularies to emerge and be maintained. Other difficulties observed with some interaction topologies suggest that *selective attention* might be beneficial. We hypothesize that selective *suppression* of mirroring, e.g. inhibition of any utterance sufficiently different from the one the agent already associates to a given referent, should enhance the stability of naming systems. As mentioned, in some cases, homonymous utterances – identical names for the different agents – emerged, as did utterances having multiple referents (names for subsets of the group), which, since there was no reinforcement regarding successful singular unambiguous reference could persist and proliferate. Reference negotiated with *confirmation* of joint-attention (cf. 'checking' in [25]) between speakers and listeners should also enhance the distinctness of utterances with different reference (reducing homonymy and 'spread' of reference to multiple referents). Thus, in addition to dynamic topologies with communities of minimal linguistic agents that may come and go, it is appropriate to study how restrictions on attention could prevent instability caused by uncontrolled mirroring of all observed utterances and deictic actions.

References

1. N. Chomsky, *Language and Mind.* Harcourt, Brace and World, 1968.
2. N. Chomsky, *Reflections on Language.* Pantheon Books, 1975.
3. D. Bickerton, *Language and Species.* Chicago, 1990.
4. S. Pinker and P. Bloom, "Natural language and natural selection," *Brain and Behavioural Sciences*, vol. 13, pp. 707–784, 1990.
5. L. Steels, "A self-organizing spatial vocabulary," *Artificial Life*, vol. 2, no. 4, pp. 315–332, 1996.
6. J. Hurford, C. Knight, and M. Studdert, eds., *Approaches to the Evolution of Language: Social and Cognitive Bases.* Cambridge University Press, 1998.
7. M. A. Arbib, "The mirror system, imiation, and the evolution of language," in *Imitation in Animals and Artifacts* (K. Dautenhahn and C. Nehaniv, eds.), MIT Press, 2002.
8. S. Savage-Rumbaugh and K. E. Brakke, "Animal language: Methodological and interpretative issues," in *Readings in Animal Cognition* (M. Bekoff and D. Jamieson, eds.), pp. 269–288, MIT Press, 1996.
9. L. M. Herman and S. N. Austad, "Knowledge acquisition and asymmetry between language comprehension and production: Dolphins and apes as general models for animals," in *Readings in Animal Cognition* (M. Bekoff and D. Jamieson, eds.), pp. 289–306, MIT Press, 1996.
10. I. M. Pepperberg, "Allospecific referential speech acquisition in grey parrots (P*sittacus erithacus*): Evidence for multiple levels of avian vocal imitation," in *Imitation in Animals and Artifacts* (K. Dautenhahn and C. Nehaniv, eds.), MIT Press, 2002.
11. L. Steels, "The synthetic modelling of language origins," *Evolution of Communication*, vol. 1, no. 1, 1997.

12. L. Steels, "Synthesising the origins of language and meaning using co-evolution, self-organization and level formation," in *Approaches to the Evolution of Language* (J. Hurford and M. Studdert-Kennedy, eds.), Cambridge University Press, 1998.
13. A. Billard and G. Hayes, "DRAMA, a connectionist architecture for control and learning in autonomous robots," *Adaptive Behavior*, vol. 7, no. 1, pp. 35–63, 1999.
14. A. Billard and K. Dautenhahn, "Experiments in learning by imitation - grounding and use of communication in robotic agents," *Adaptive Behavior*, vol. 7, no. 3/4, 1999.
15. C. L. Nehaniv, "The making of meaning in societies: Semiotic and information-theoretic background to the evolution of communication," in *AISB Symposium: Starting from Society: the application of social analogies to computational systems*, Society for the Study of Artificial Intelligence & the Simulation of Behaviour, 2000.
16. J. Baillie and C. Nehaniv, "Deixis and the development of naming in asynchronously interacting connectionist agents," in *Proc. First International Workshop on Epigenetic Robotics*, vol. 85 of *Lund University Cognitive Studies*, (Lund, Sweden), pp. 123–129, 17-18 September 2001.
17. M. Scaife and J. Bruner, "The capacity for joint visual attention in the infant," *Nature*, vol. 253, pp. 265–266, 1975.
18. G. Butterworth, "The ontogeny and phylogeny of joint visual attention," in *Natural Theories of Mind* (A. Whiten, ed.), Blackwell, 1991.
19. C. Moore and P. J. Dunham, eds., *Joint Attention: Its Origins and Role in Development*. Erlbaum, 1995.
20. B. Scassellati, "Imitation and mechanisms of joint attention: A developmental structure for building social skills on a humanoid robot," in *Computation for Metaphors, Analogy, and Agents* (C. L. Nehaniv, ed.), vol. 1562, pp. 176–195, Springer Verlag, 1999.
21. K. Dautenhahn, "Trying to imitate – a step towards releasing robots from social isolation," in *Proc. From Perception to Action Conference (Lausanne, Switzerland, Sept 7-9, 1994)*, pp. 290–301, IEEE Computer Society Press, 1994.
22. K. Dautenhahn, "I could be you — the phenomenological dimension of social understanding," *Cybernetics and Systems*, vol. 25, no. 8, pp. 418–453, 1997.
23. C. L. Nehaniv, "Meaning for observers and agents," in *IEEE International Symposium on Intelligent Control/Intelligent Systems & Semiotics*, pp. 435–440, IEEE, 1999.
24. J. Baillie and C. Nehaniv, "Deixis and the development of a proto-language in asynchronously communicating connectionist agents," 27-30 March 2002. Paper presented at the *Fourth International Conference on the Evolution of Language*, Harvard.
25. G. Butterworth, "Pointing is the royal road to language," in *Pointing: Where Language, Culture, and Cognition Meet* (S. Kita, ed.), pp. 9–33, Lawrence Erlbaum, 2003.
26. G. Rizzolatti and M. A. Arbib, "Language within our grasp," *Trends in Neurosciences*, vol. 21, no. 5, pp. 188–194, 1998.
27. R. J. Jeffers and I. Lehiste, *Principles and Methods for Historical Linguistics*. MIT Press, 1979.
28. L. Steels, "The origins of syntax in visually grounded robotic agents," *Artificial Intelligence*, vol. 103, pp. 1–24, 1998.

A Role Sharing Model of Language Areas

Yoshihisa Shinozawa[1] and Akito Sakurai[1,2]

[1] Keio University, Hiyoshi, Kohoku-ku, Yokohama 223-8522, Japan
[2] CREST, Japan Science and Technology Agency
shino@ae.keio.ac.jp

Abstract. We propose a role sharing model of language areas in which Broca's area is for categorizing symbols used to represent rules stored and retrieved in other language areas. For example, at the syntactical level, the other language areas store rules represented with terminal symbols and also rules represented with non-terminal symbols, whereas Broca area invents non-terminals and forms abstract rules at language acquisition phase and converts a terminal symbol to a non-terminal corresponding to it to get an appropriate rule at performance phase. The model role of Broca's area is supposed to be essential but minimal to support human language faculty. Under this assumption, the emergence of Broca's area is hypothesized to have triggered the evolution of language supposing that the other mechanisms are fully evolved. The argument is based on the recent fMRI study of KE family members and related findings on a mutation of FOXP2 gene.

1 Introduction

Broca's area has been thought to be and is proven to be the center of syntactic information processing ([6]). But some other areas might be involved in it since there are many cases that aphasic patients recover to some degree with Broca's area still damaged([12]); many experimental studies show that areas other than Broca's are evoked during syntactic tasks. Questions arise: how is syntax processing shared among brain areas and is it related to human evolution?

Half of the four-generational KE family members suffer from specific language impairments ([11]) and are revealed by fMRI study to have different brain activation patterns than unaffected KE family members and normal subjects ([10]). The language impairment is characterized by "inability to generate syntactic rules such as those for tense, number, and gender," specifically by difficulty of converting present verbs to past forms pronounced more for regular verbs than irregulars ([3],[4]). fMRI study shows "the affected members showed significant underactivation relative to the unaffected members in Broca's area and its right homolog" ([10]).

Based on these facts, we may conclude that large part of concrete rules are stored in other areas than Broca's, relatively abstract rules may be stored in Broca's area, or they are in the other area and Broca's area engages at least in managing the abstract rules. A concrete rule, here, means a rule represented

A. Sakurai et al. (Eds.): JSAI 2003/2004, LNAI 3609, pp. 335–344, 2007.

with phonemes for word construction, or a rule represented with words for sentence construction, whereas an abstract rule is a rule represented with phonetic categories or a rule represented with syntactic categories.

The inactivation or non-existence of Broca's area in the affected KE family members is caused by a mutation in FOXP2 gene ([8]), that is, the point mutation co-segregates with the disorder. Moreover, it is strongly suggested that the gene has been the target of selection in recent human evolution ([2]). These derive no logical consequence but still tempt us to assume that Broca's area might be developed after other language areas were, that the abstract rules became usable after concrete rules, or that human language capability was increased at the time when Broca's area was formed or became active.

There is one defect in the above argument. Affected KE family members spend their daily lives with sufficient capability of communication. It is natural to believe that they produce sentences which were not heard beforehand, that is, they use rules with variables where arbitrary words can be put into. The difference between the general inflection rules that are absent in the affected KE family members and the sentencial rules that may be present in them is that the latter rules are dependent on semantics which might influence the internal representations of words and rules. It is commonly observed that aphasic patients understand, for example, "ready for dinner" but cannot understand "a pencil is on the book." Therefor the defect can be removed naturally if we can assume that words in the brain associate with their meaning and simple sentences are formed based on semantic similarity to the exemplar sentences we know.

In the following, we consider syntactic rules rather than phonetic rules. The difference is only what abstraction level we are addressing.

2 Abstract and Concrete Rules

Concrete rules in this paper are rules written in terminal symbols, or object-level symbols, and abstract ones are in non-terminal symbols, or meta-level symbols.

Concrete rules are rules formed based on representations that assign different symbols to different words, such as "*cat* may follow *a dog chases*" or "*a dog barks* is grammatical but *a dog bark* is not." Usually these are not considered "rules" since they will not be applied to other cases. In natural languages, though, even among terminal symbols there exists some similarity measure, so that we can apply a rule to similar situations, for example we would erroneously infer that "cat" can replace "rat" since they are similar in their forms.

Abstract rules are rules written in non-terminal symbols or syntactic category names, such as "N precedes V," which are rules in common sense.

By abstract rules, we mean in this paper rules with non-terminal symbols or category names rather than rules with variables. Variables are versatile tool to represent abstract rules. But since we do not yet have common understandings on how to represent variables in artificial neural networks and natural neural networks, and grammars are written in non-terminal symbols in formal language theory, we decided to use category names in abstract rules. Note that artificial neural networks have been under intense research to make it eligible for symbols naturally ([7]).

3 Hypothesis

We propose a hypothesis asserting that Broca's area categorizes symbols used to represent rules in language areas, or more precisely, groups together syntactically related symbols used to represent rules stored or retrieved in language areas, gives them new representations, stores them, and allow language areas to use them. The hypothesis is, intuitively, minimal in the sense that we cannot describe syntactic rules without the function and we cannot divide it further.

In Section 1, we left room of having two alternative hypothesis for what and how Broca's area manage rules. One is that Broca's area has the abstract rules in it whereas the other is that Broca's area only concerns with managing abstract rules. For the former one, we had better hypothesize that the management mechanism, if exists, lies in Broca's area too since it seems to be natural, although without any other possible rationalization.

We in this paper take a position in between so that we will assume that some management mechanism exists in Broca's area without which abstract rules are not activated in a proper way. This implies that when Broca's area is damaged after abstract rules are well formed, abstract rules could work with diminished functionality, or some other area than Broca's can take its place, and that when Broca's are does not exit in learning period abstract rules might not be well developed. These inferred results well conform with the above observations of affected KE family members and aphasic patients. Note that we will not make a claim on where abstract rules exist, which may be in Broca's area or in some other cortical areas.

We assume that language areas other than Broca's are responsible for concrete rules with generalization limited within proximity introduced by defining distance measure between representations of the symbols.

4 Conceptual Model

We propose a conceptual model consisting of two sets of networks. One of the sets R_1 is responsible for learning and applying rules, i.e., a model of brain circuitry of acquiring and applying grammatical rules, regardless of the rules being concrete or abstract, phonological or syntactic, while the other set R_2 is for mappings from concrete (word or sound) representations to categorical representations obtained by observing behaviors of R_1 or activities of elements in R_1, which we propose as a model of Broca's area.

R_1 works without help of R_2. It acquires grammatical rules by listening to (watching in case of signs) flow of natural language inputs. It basically acquires concrete rules which are formed word by word basis or phoneme by phoneme basis with naturally introduced similarity.

When R_1 works with help of R_2, it can learn and process abstract rules represented with category names. When learning proceeds, the flow of word names are converted to flow of category names, for example, noun or verb, and are learnt by R_1 in as much like as word-name sequence learning. For example, by using R_1 with R_2 it could learn the grammaticality of "N V N" not only the grammaticality of "a dog chases a cat."

An important point is that the rule learning mechanism in R_1 is applicable to both of word name sequences and category name sequences. The difference between R_1 with R_2 and R_1 without R_2 at learning is that the former is supplied with category name but the latter is not; the difference at performance is that the former can use the abstract rules to judge grammaticality or to construct new sentences based on abstract grammar rules but the latter cannot use (or only poorly use) them.

Although R_1 without R_2 realizes concrete rules only, it still has generalization capability by utilizing similarity of representation of constants (each word), so that for example it could infer "picked" from "pick" with knowing that "kicked" is the past form of "kick" and "pick" is similar to "kick." R_1 with R_2 can generalize independently of representation of constants, for example, it infers that the past form of certain set of verbs irrespective of their representations is obtained by adding "ed" to them.

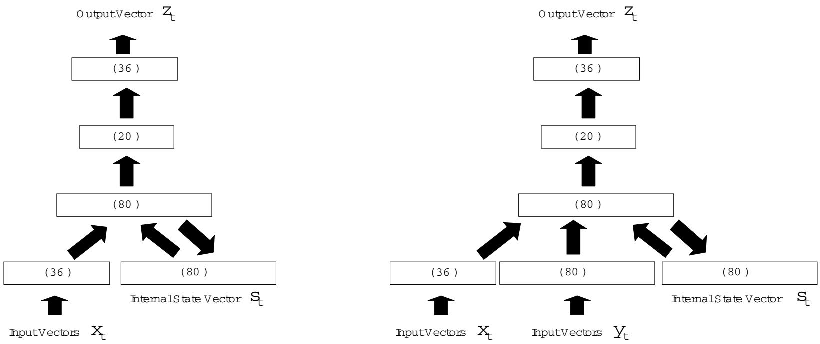
5 Simulation Model

We conducted numerical experiments for the conceptual architecture. We adopted an Elman (Figure 1(a)) network for R_1, which is a simple recurrent neural network and has been used to explain a network's ability to acquire grammars, but with a slight modification (Figure 1(b)):

$$\begin{cases} s_{t+1} = f\left(w_s \cdot s_t + w_x \cdot x_t + w_y \cdot y_t\right) \\ z_t \quad = \quad g\left(s_t\right) \end{cases}$$

where x_t and y_t are the input vectors $\in \mathfrak{R}^{n_x}$ and $\mathfrak{R}^{n_y}$, respectively, s_t the internal state vector $\in \mathfrak{R}^{n_s}$, z the output vector $\in \mathfrak{R}^{n_z}$, f is the standard sigmoid function (applied componentwise to a vector) $x \mapsto 1/\left(1+\exp\left(-x\right)\right)$, and g is the standard one-hidden layer network with the sigmoid function as the output function. For the moment, inputs where y_t appear is set constantly 0-vector. The Elman network is used to predict a sequence, starting by letting s_0 being 0, and then by calculating z_{t+1}on observing $x_1, \ldots, x_t$. The Elman's idea is to train the network to output x_{t+1}on observing $x_1, \ldots, x_t$ by using the standard error backpropagation algorithm. x_t's represent words and are locally coded so that only one component is 1 and the others are 0, which can be any linearly independent vectors.

R_2 is just a table in this simulation to convert locally coded word representation to observed representations. We use the word "observed representation" in this section instead of "category" since in the experiment, although we intended



(a) An Elman Network (b) A Proposed Network

Fig. 1. Recurrent Neural Network

so, we cannot say in advance that the representation we use really represent syntactic categories.

R_2 observers R_1 when R_1 is learning to predict the next word x, and calculates the average of s_t of R_1 over the repetitions to get $y(x)$ which is an entry to the table, that is the average of the hidden layer representations when the specified word is to be output. The way the values defined is slightly different from that of Elman's ([1]) but in essence it is equivalent and produces simliar clusters as shown in the following experiments.

Our experiments went as follows.

1. Generate 20,000 grammatical sentences randomly from the grammar. The grammer is shown in Figure 2(a), which is almost the same as [1]. The examples of the sentences are shown in Figure 2(b). One half of the sentences are used as learning exemplar sentences and the other half are used for evaluation of the networks.
2. (First learning) Train R_1, with y_t being 0-vector, to predict the next word to come for the prefix of the exemplar sentences as in [1].
3. Calculate the internal representation, that is, a long time average of s_t for each word to come next, and stores it in a table (which is assumed to be R_2).
4. (Second learning) Train R_1 again but this time with y_t being $y(x_t)$ retrieved from R_2 (Figure 1(b)).
5. (Third learning) Modify R_1 so that it outputs a vector representing y_{t+1}. Train this R_1' with inputs, where x_t is assumed, being set a 0-vector, with y_t being $y(x_t)$ retrieved from R_2 (that is categorical information), and with a target being $y_{t+1} = y(x_t)$.

6 Simulation Results

For each learning experiment, R_1 is trained to predict the next word or category to come for the prefixes of the exemplar sentences. We evaluated the networks by

S → NP VP .
NP1 → PronN1 | N1 | N1 RC
NP2 → PronN2 | N2 | N2 RC
VP → V | V NP2
RC → whom NP1 VP | who VP
N1 → boy | girl | cat | dog | boys | girls | cats | dogs
N2 → boy_ | girl_ | cat_ | dog_ | boys_ | girls_ | cats_ | dogs_
PronN1 → John | Mary
PronN2 → John_ | Mary_
V1 → chase | feed | see | hear | chases | feeds | sees | hears
V2 → walk | live | walks | lives

(a) Grammer

John lives.
Mary walks.
Cats walk.
A boy sees Mary_.
A girl feeds a dog_.
A girl sees John_ who walks.
John who lives feeds cats_.
John whom Mary sees feeds a dog_.
Boys who live see Mary who walks.
Dogs whom John who walks sees walk.

(b) Examples

Fig. 2. Grammar

10,000 sentences other than the ones used for learning and calculated prediction accuracy. Since the word representation is locally coded, we can interpret the prediction output value of each unit as the degree of preference to that word; i.e. the highest output designates the most preferred word, the second highest designates the second most preference, and so on. We repeated each experiment ten times and calculated the averages and the standard deviations of prediction accuracy. The results are summarized in Table 1.

Table 1. Prediction accuracy after each learning

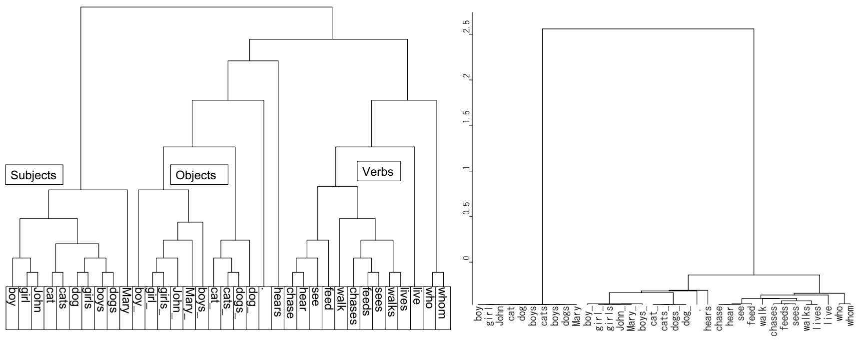
	Average	S.D.
First Learning	4.31	0.45
Second Learning	3.80	0.022
Third Learning R_1'	5.24	0.35

6.1 The First Learing Experiments

The first learning is designed to confirm the R_1 without R_2 can learn word sequences. Prediction accuracy of the network was 4.31 (S.D. 0.45). R_1 without R_2 can predict the next word representation x_{t+1} based on $x_1, \cdots, x_t$ the prefix of sentences. As is the same in [1], the network acquired concrete rules, such as "*sees* follows *a dog* or *a cat*, but does not follow *dogs* or *cats*."

A clustering analysis, similar to the one mentioned in [1], was conducted. The result is shown in Figure 3(a) which shows only how clusters are formed and in Figure 3(b) with proper scaling. The words are roughly classified into three categories, subjects, objects, and verbs where each category is formed of the words in the same syntactic category.

The result shows that although the network acquires concrete rules without being given categorical information, the network activity reflects the syntactic category of the word to come next.



(a) Relative Relations. (b) In Proper Scale.

Fig. 3. Dendrogram

6.2 The Second Learing Experiments

The second learning experiment is to know if R_1 with R_2 can learn word sequences better than R_1 without R_2. When the second learning was completed, prediction accuracy of the network was 3.8 (S.D. 0.022), averaged over ten experiments, which is slightly better than the first learning result (4.31). The network, when categorical information is given, acquires concrete rules more accurately. The reason might be that the learning utilized additional information (may not be correct) included, as shown in Figure 3, in y_t retrieved from R_2.

In addition, we conducted the following experiments to know how severe damage a lesion of R_2 causes to prediction accuracy.

(1) This is the case for the damaged R_2. We assume that R_2 is damaged and the categorical information y_t can not be retrieved from R_2 correctly, when R_1, which is trained with R_2, is to predict next word representation x_{t+1}. We simply set y_t to be a constant vector of the weighted average of $y(x)$ (weights are proportional to occurrence frequencies) over all x's.
(2) This is the case for the damaged word information storrage. We assume that R_2 works normally and so the categorical information y_t is retrieved from R_2 as is usual, but word representation x_t is not retrieved correctly, that is, only the categorical inforamation is used for inputs when the network is to predict the next word. We simply set the word representation to be a 0-vector for all x_t's.

In the first case, prediction accuracy (calculated as the rank) was 11.0 (S.D. 1.09) with setting aside predictions of very low confidence, that is, the outputs of very small magnitude, which were considered as "no answer." This result is worse than the second learning result (3.80) and shows that once R_1 learns with the help of R_2, then its performance degrades very much when R_2 suffers heavy damage and the categorical information is not supplied from R_2.

In the second case, the prediction accuracy of the network was 9.0 (S.D. 0.18). The result is worse than the second learning result but slightly better than the first case above. The result shows that once R_1 is trained with R_2, even though the word information is absent, it can still predicts the next word with the help of R_2 by receiving category information from it. It is interesting to note that a lesion of R_2 causes more serious damage to prediction accuracy than word information itself.

The results might might be interpreted as to explain that rules can be used (or the sentences are yet understood) in recovering stages with Broca's area still damaged.

6.3 The Third Learning Experiments

The third learning experiment is to check if the network can learn abstract rules, which are represented with category names. Prediction accuracy of $R_1^{'}$ was 5.27 (S.D. 0.35), where this time the network output is considered in rank k if the correct word comes in the k-th in the array of words arranged increassing order of the Euclidean distance between the network output and the internal representation of words. The result is worse than the first learning because representations of words in the same category tend to be similar. The result may be interpreted that the distinction of words is still reflected in the internal representations and $R_1^{'}$ utilized it, so that the predictions made by $R_1^{'}$ are in essence based on the words, not the categories.

We designed another experiment for the same purpose. Each word is assigned the average of the internal representation of the ones in the same category as the word, where the categories are: subjects (N1 ∪ PronN1), objects (N2 ∪ PronN2), V1, V2, who, whom, .(period) . The test exemplar sentences are transfered to the list of the above defined value and are fed to $R_1^{'}$ and the prediction is judged correct when the output value is closest to the above defined value of the category to which the correct word to come belongs. The results are shown in Table 2.

Prediction accuracy of $R_1^{'}$ was 84.5% on average. The network can learn the rules which predict subjects, objects and verbs. The prediction for period and relative pronouns are relatively low, simply because just after nouns period and relative pronouns may come with similar frequency in exemplar sentences.

Table 2. Prediction accuracy of $R_1^{'}$

	Prediction accuracy
N1∪PronN1	100.0%
N2∪PronN2	100.0%
V1,V2	93.9%
who,whom	52.7%
.	56.1%
Average	84.5%

The result clearly shows that abstract rules in category names, which are in this experiment the average internal representation of the words in the category, are acquired in the network $R_1^{'}$.

7 Discussions

We proposed hypothesis that Broca's area categorizes symbols used to represent rules and language areas other than Broca's are responsible for concrete rules. We constructed the conceptual model based on our hypothesis and conducted experiments.

As is anticipated, the network learns word sequences, that is, concrete rules. After learning word sequences, the network activity reflects syntactic categories. Figure 3 shows that syntactic clusters are formed properly. The network can acquire concrete rules more accurately when categorical information is supplied from R_2. A lesion in R_2 causes sever degeneration of predictive accuracy. The network predicts the next word with very low accuracy when R_2 is damaged and category information can not be supplied from R_2. An interesting thing is that when the word information is not supplied to R_1 but information from R_2 is still given, the network performance is similar. Using categorical information retrieved from R_2, the network learns category name sequences, that is, abstract rules.

These results cannot be a proof but indicates that our hypothesis is a candidate to explain the role of Broca's and other language areas.

8 Conclusions

We proposed a model of syntax processing shared by Broca's area and other language areas, which is based on the recent findings on SLI of KE family and its fMRI results, and speculated that the assumed responsibility of Broca's area is the key to the evolution of language faculty deeply related to a mutation of FOXP2 gene.

In the model, Broca's area is modeled to categorize symbols used to represent rules stored and retrieved in other language areas. The model conforms to the findings on SLI of KE families and aphasic patients and the result of simple simulations of the model was consistent with them, too.

References

1. Elman, J. L., Finding Structure in Time, *Cognitive Science*, 14, 179–211 (1990).
2. Enard, W., Przeworski, M., Fisher, S.E., Lai, C.S.L., Wiebe, V., Kitano, T., Monaco, A.P. and Paabo, S., Molecular evolution of FOXP2, a gene involved in speech and language, *Nature*, 418, 869–872 (2002).
3. Gopnik, M., Feature-blind grammar and dysphasia, *Nature*, 344, 715(1990).
4. Gopnik, M. and Crago, M.B., Familial aggregation of a developmental language disorder, *Cognition*, 39, 1–50 (1991).

5. Hauser, M.D., Chomsky, N., and Fitch, W.T., The Faculty of Language: What Is It, Who Has It, and How Did It Evolve?, *Science*, 298, 1569–1579 (2002),
6. Hashimoto, R. and Sakai, K.L., Specialization in the left prefrontal cortex for sentence comprehension. *Neuron*, 35, 589–597 (2002).
7. Jagota, A., Plate, T., Shastri, L. , and R. Sun (eds), Connectionist symbol processing: Dead or alive?, *Neural Computing Surveys*, 2, 1-40 (1999).
8. Lai, C.S., Fisher, S.E., Hurst, J.A., Vargha-Khadem, F., and Monaco, A.P.A., forkhead-domain gene is mutated in a severe speech and language disorder, *Nature*, 413, 519–523 (2001).
9. Lai, C.S.L., Gerrelli, D., Monaco, A.P., Fisher, S.E., and Copp, A.J., FOXP2 expression during brain development coincides with adult sites of pathology in a severe speech and language disorder, Brain, 126(11), 2455–2462 (2003).
10. Liegeois, F., Baldeweg, T., Connelly, A., Gadian, D.G., Mishkin, M., and Vargha-Khadem. F., Language fMRI abnormalities associated with FOXP2 gene mutation, *Nature Neuroscience*, 6 (11), 1230–1237 (2003).
11. Vargha-Khadem, F., Watkins, K., Alcock, K., Fletcher, P., and Passingham, R., Praxic and nonverbal cognitive deficits in a large family with genetically transmitted speech and language disorder, *Proceedings of the National Academy of Sciences USA*, 92, 930–933 (1995).
12. Watkins, K. E., Dronkers, N. F., and Vargha-Khadem, F., Behavioural analysis of an inherited speech and language disorder: comparison with acquired aphasia, *Brain*, 125(3), 452–464 (2002).

The Evolution of Writing Systems: Against the Gelbian Hypothesis

Tadao Miyamoto

Graduate School of International Cultural Studies
Tohoku University, Sendai, Miyagi, Japan
tmiyamoto@mail.tains.tohoku.ac.jp

Abstract. This paper discusses three issues concerning the evolution of writing systems [8]. First, the paper critically examines Gelb's [10] hypothesis which characterizes the development of writing systems as an evolution from primitive to advanced, with the most advanced system being alphabetic. Second, this paper discusses some possible reasons why the evolution of writing systems often deviates from the simplistic path proposed by Gelb. Finally, this paper compares three major writing systems, Maya, Egyptian, and Chinese, with a view to examining why, unlike Maya, both the Egyptian hieroglyphic and Chinese logographic writing systems developed extremely large inventories of symbols.

1 Gelb's Claim

Linguists have generally ignored writing systems under the assumption that they are little more than crude reflections of spoken language. Recently, however, this topic has gained ground as a legitimate area of scientific inquiry [7, 24].

Gelb's [10] study is one of the earliest investigations into writing systems and remains the most oft-cited work on this subject. Gelb's main claim is that writing systems develop along a path from primitive to sophisticated, with degree of abstractness serving as the main index of sophistication. More specifically, Gelb divides written communication into two broad stage-based categories: the forerunners to writing and full writing systems. Within this framework, writing conventions that depend on descriptive-representational devices would be considered forerunners to writing. For instance, the ancient Middle Eastern token system, where stone tokens were used as mnemonic devices for the purpose of commercial transaction [see 21, 22], would be considered a forerunner to writing.

Also among the forerunners to writing are systems that use semasiographic or pictographic word signs devoid of any phonetic value. To capture the significance of phoneticization, Gelb cites the example of the Sumerians, who were able to break entirely away from the conventions of the descriptive-representational device by assigning phonetic value to each of their pictographic signs, which were, in turn, the forerunners to cuneiform scripts.

A. Sakurai et al. (Eds.): JSAI 2003/2004, LNAI 3609, pp. 345 – 356, 2007.

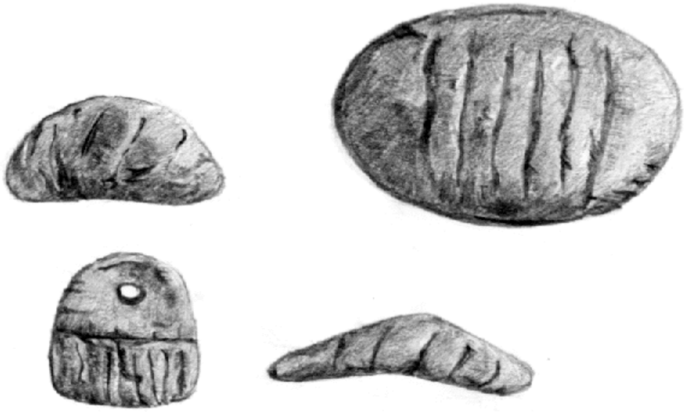


Fig. 1. Stone tokens used for commercial transaction in the ancient Middle East. Each token represented a specific commercial value.

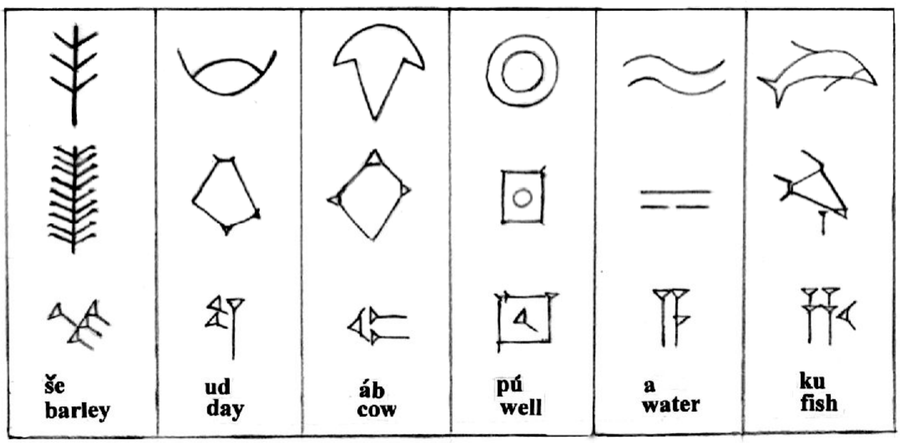


Fig. 2. Cuneiform letters (shown at the bottom) developed from pictograms (shown at the top)

Gelb subdivides full writing systems into three stage-like categories: logo-syllabic, syllabic, and alphabetic. In discussing the stage of full writing, Gelb makes the highly controversial claim that the so-called consonantal writing systems used by many Afro-Asiatic languages are not in fact consonantal, but syllabic. Guided by this assumption, Gelb groups not only pure syllabic systems, such as Linear B, but also Egyptian hieroglyphic writing into the syllabic sub-category. However, as noted by many scholars, including Coulmas [6] and Sproat [24], this claim is motivated by Gelb's attempt to create a neat and tidy picture of the evolution of writing in which alphabetic writing emerges as the most advanced of full writing systems. Gelb explicitly states that "in reaching its ultimate development writing, whatever its forerunners may be, must pass through the stages of logography, syllabography, and alphabetography in this, and no other order" [10: 201].

A further problem with Gelb's hypothesis is the assumption of uni-directionality in the evolution of writing systems. Gelb claims that "[t]here is no reverse development: an alphabet cannot develop into a syllabary, just as a syllabary cannot lead to the creation of logography" [10: 201]. However, there is clear historical counterevidence to this claim. For example, the Ethiopians, by significantly enlarging the inventory of signs, transformed a borrowed alphabetic system into a syllabic one.

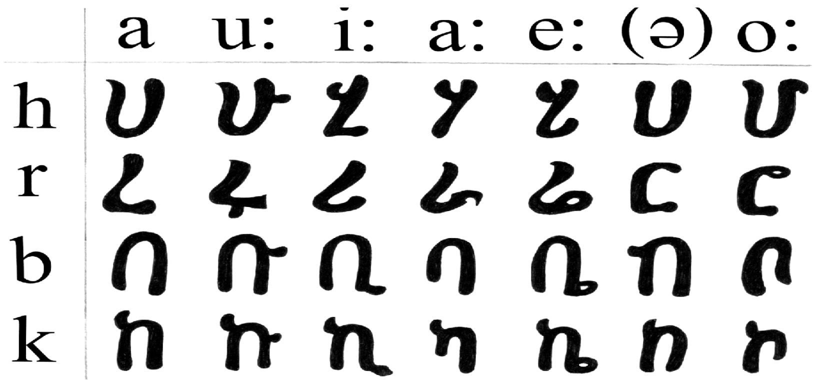


Fig. 3. Part of the Ethiopic syllabary, created by subscripting vowel symbols onto the original alphabetic symbols, shown in the leftmost column, all of which assumed the vowel /a/ with no subscript

A third major problem with Gelb's claim is that the history of writing attests several obvious exceptions to his evolutionary assumption — perhaps most notably, the Chinese writing system, which has remained basically unchanged for nearly four thousand years [9]. Importantly, Gelb's alphabetico-centrism is unable to account for why certain writing systems, such as cuneiform, have followed fairly closely the evolutionary path he prescribes, while other writing systems, such as Chinese, exhibit hardly any evolutionary development.

2 Evolutionary Motivations

It is interesting to consider why the evolution of writing systems seldom occurs in the manner proposed by Gelb. When the history of writing systems is carefully examined, the development of even the Latin alphabetic system fails to confirm Gelb's evolutionary hypothesis. Its origin can be traced back only to the small set of mono-consonantal Egyptian hieroglyphic symbols adopted by the Proto-Sinaitic writing system [see 19], but never to Egyptian hieroglyphic writing *per se*, which Gelb falsely assumes to be logo-syllabic.

Also, as noted above, the basic principles underlying the Chinese writing system, which is truly logo-syllabic [9], have remained virtually unchanged throughout its entire history.

Fig. 4. Proto-Sinaitic script found at Serabit el-Khadim. The script is historically based on the set of 24 Egyptian mono-consonantal hieroglyphic letters.

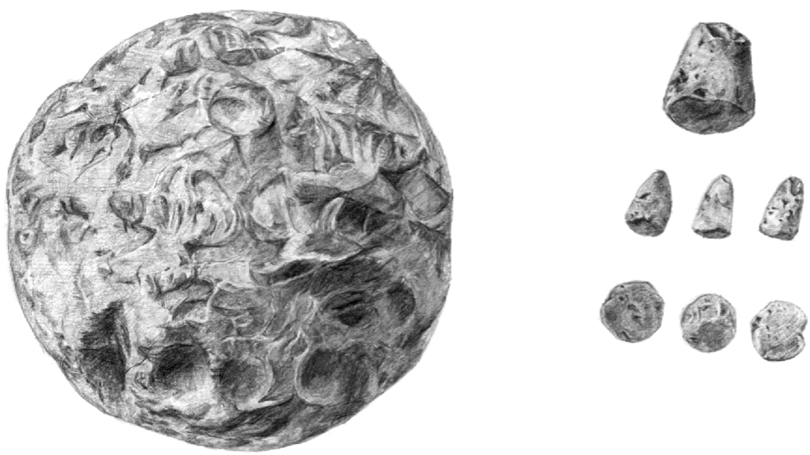


Fig. 5. Bulla and tokens. The impressions on the surface of the bulla were eventually replaced with pictograms depicting specific transactions.

In contrast, even though cuneiform never reached Gelb's proposed evolutionary goal of alphabetism, its development largely corresponds to his account. Cuneiform originated as a token system. Matters relating to commercial transaction were depicted on the surface of bullae, as seen in Fig. 5; these depictions were, in turn, stylized as pictograms (see Fig. 2).

Using these pictograms as its major components, a logographic cuneiform writing system emerged. Following a variety of modification processes adopted by lingua franci, the original logographic system eventually morphed into a syllabic system around the time it was adopted by the Persians [26]. Although cuneiform never evolved into alphabetic writing, Persian syllabograms tended to function doubly as alphabetic symbols. In a similar vein, even though the Chinese writing itself stubbornly defied significant evolutionary changes, the Chinese writing adopted by neighboring peoples has shown extensive changes, as witnessed in the development of *ido* in Korea [25] and *kana* in Japan [see 15].

As the above examples clearly attest, there is considerable diversity in the evolution of writing systems, contradicting Gelb's neat and tidy picture and begging the question of what motivates such discrepancies. The basic conceptual flaw in Gelb's theory stems from the false assumption that the more advanced the writing system, the smaller the inventory of characters (see also McLuhan [18] for a similar idea). We argue that the tendency to adopt progressively smaller linguistic units as the basis for a writing system, hence, reducing the size of symbol inventories, is just one of a range of possible forces that may influence the development of writing systems.

Basically, all communication systems strive for optimal efficiency in the encoding and decoding of messages. This goal may be accomplished in a variety of ways, with gains and losses for production and comprehension depending on the paths taken. Reducing the size of a character set is rational only from the viewpoint of production or, more specifically, from that of acquisition. From the viewpoint of comprehension or reading, however, this strategy may be problematic, generating a constraint that works against the development of a minimal inventory of symbols.

Psycholinguistic studies may suffice to illustrate this point. For example, Kaiho [13] shows that Chinese characters rich in visual complexity, that is, those with a relatively large number of strokes, are processed more quickly and accurately than less complex characters with a smaller number of strokes. Another experimental study, Rozin, Poritsky and Sotsky [20], suggested that American urban black children with dyslexic tendencies would, with the support of intensive instruction, perform far better in reading comprehension with Chinese characters than with the English alphabet [see also 14, 15].

Hence, visual complexity, which is characteristic of logographic and logo-syllabic systems, enhances reading comprehension, while visual simplicity, manifested to a greater extent in alphabetic systems, does not. By implication, the visual complexity associated with having a large inventory of symbols should also enhance reading comprehension. Consequently, there is no compelling psycholinguistic reason why a writing system should develop into an alphabetic system.

Even at the socio-cultural level, certain factors work against the development of simple, alphabetic systems. According to Coulmas [6], the main functions of writing are related to memory, interaction, distancing, reification, and social control. Of these functions, both reification (i.e., endowing a message with authoritative power) and social control (i.e., regulating the social conduct of subordinates) ought to be better served by a writing system which is visually rich and structurally complex—that is, a system that cannot be acquired in a matter of mere hours.

Communication can be either horizontal (i.e., reciprocal and bi-directional) or vertical (i.e., non-reciprocal and uni-directional). Writing systems, in turn, may develop to facilitate one or the other type of communication. To better accommodate horizontal communication, a writing system would tend to assume a simple form, as historically witnessed in Greek society [5]. By contrast, to promote vertical communication, a writing system would tend to favour complexity, as seen in Egyptian, Maya, and Chinese societies.

In the latter group of societies, the most efficient way to conduct vertical communication must have been to endow messages with authoritative power. The best way to fulfill this reifying function was to possess a system whose properties and functions were beyond the full grasp of the general public. At the height of Egyptian civilization, for instance, its population reached approximately four million, only 1% of which is estimated to have been literate [1]. The privileged 1% were those with sufficient leisure time to acquire the complex writing system. (The situation was basically the same in Maya society, although possibly less extreme [2]).

Similarly, a highly effective means of regulating the social conduct of the general public is via a communicative medium. If the function of social control has already been built into the medium itself due to its incomprehensibility to the vast majority of the public, the authority would have no motivation to alter the principles and functions of the medium. Hence, in Egyptian, Maya, and Chinese society, there were no serious attempts to simplify the writing system.

Finally, it is useful to consider the force that promotes the development of writing, which is perhaps best illustrated by the example of cuneiform. Throughout its three-thousand-year history, cuneiform passed through the various stages outlined by Gelb. That is, it evolved from pictographs to logographs, and in turn from a logo-syllabary to a syllabary, and so on to what might be considered a quasi-alphabetic form of writing [6].

Fig. 6. Fragment of an Old Persian text, in which some syllabic cuneiform letters are used as alphabetic consonants, losing the vowel traditionally associated with these symbols

While Gelb would consider this development to represent a natural evolutionary process, we attribute it to the process of linguistic *borrowing*—a force that may be likened to the gene mutations driving the emergence of favorable traits in the evolution of a species.

When the writing system of one language is borrowed by another, significant structural changes are entailed in the system, particularly if the two languages differ typologically. Such was the case in the development of cuneiform. Over the course of three millennia, cuneiform was used to write approximately 15 languages, most of which were lingua franci. In brief, cuneiform was originally used to write Sumerian, an agglutinative language without inflection. At this stage, cuneiform was purely logographic, that is, symbols were used to represent words [26]. Around 1900 BCE, Sumerian disappeared, the Sumerian-speaking public having been absorbed by Akkadian-speaking invaders. As a means of writing Akkadian, an inflecting language, logograms are cumbersome. Hence, the Akkadians transformed the logographic system into a logo-syllabic one, making extensive use of syllabograms to mark inflection. This trend intensified when cuneiform was adopted to write Old Persian, a true Indo-European inflecting language. Consisting of only 41 signs, the Old Persian system was very much at the threshold of alphabetic writing.

The basic strategy taken by the borrowers was "simplification based on phoneticization," that is, the phonetic representation of words wherever possible. This strategy is akin to the one employed in the development of *ido* in Korean and *kana* in Japanese. In the case of cuneiform, however, the simplification processes spanned three millennia and involved a range of typologically different languages.

In summary, while the development of cuneiform largely corresponds to Gelb's hypothesis, his alphabeto-centrism is overly simplistic and fails to account for the many different forces that promote or constrain the development of writing systems.

3 A Comparison of Egyptian, Maya, and Chinese Writing

In this section, we briefly compare Egyptian, Maya and Chinese writing, in an effort to determine why, in contrast to Maya, the number of symbols in both the Chinese and Egyptian writing systems increased beyond the point of practicality. From a linguistic viewpoint, this superfluous growth can be attributed to what we call an *unwholesome use of conflation and punning*. However, such an insane enlargement of character sets can also be ascribed to a socio-cultural factor. The following elaboration will shed some more light on the nature of writing and the evolutionary forces that govern it, which are by no means linear.

The Egyptian hieroglyphic system includes two types of signs: phonograms and semagrams. Phonograms may be mono-, bi-, or tri-consonantal signs or so-called phonetic complements, which are used to reinforce the pronunciation of consonantal signs. Semagrams are of two types: determinatives, which lack any phonetic value, and logograms, which express lexical meanings. The Egyptian writing system is an intricate mixture of these phonetic and semantic symbols [4].

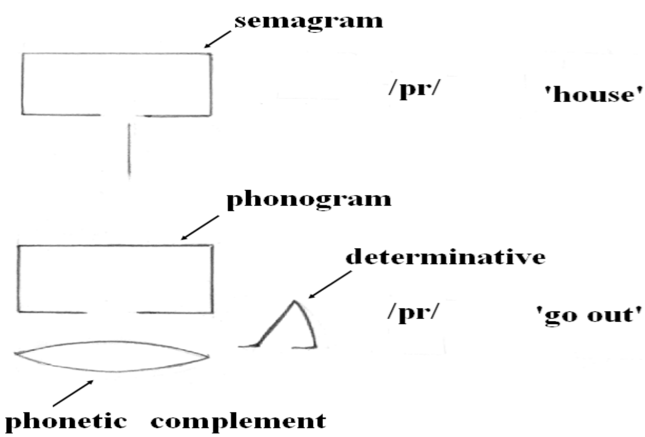


Fig. 7. Example of the intricate combination of phonograms and semagrams in Egyptian hieroglyphic writing

The Maya writing system is similar to the Egyptian in that it employs both phonograms, all of which are basically syllabograms, and semagrams [11]. Interestingly, it also possesses phonetic complements even though determinatives are rare in Maya writing.

Fig. 8. Maya combination of the semagram (BALAM, jaguar) and the syllabic phonograms (/ba/ and /ma/), which are used here as phonetic complements

No explanation may be required for Chinese writing, the basic characteristics of which can be accounted for by *Rokusho* or the Six Principles of Character Formation [see, for instance, 9].

Historically, the Egyptian hieroglyphic system originated in 3000 BCE with about 1,000 symbols. In the Middle Egyptian period, the number of signs was reduced to about 700 due to codification imposed by scribes. During the period of the Greco-Roman invasions, however, the number of signs increased to approximately 5,000 [16].

Chinese writing appeared no later than the Shang dynasty, which began in 1766 BCE, with about 2,500 characters. By the time of the Han dynasty, the number of characters increased to nearly 10,000. By the twelfth century CE, the number of characters had soared to 23,000; by the eighteenth century, there were as many as 49,000 characters [see 7, 23].

Maya society reached its cultural height during the Classic Period from 250 to 900 CE [27]. The writing system had eight or so hundred symbols in its inventory; however, at any given time, the size of the active inventory consisted of a much smaller number of glyphs [12, 17].

While the historical background of Chinese and Egyptian writing differs, both writing systems developed unnecessarily large inventories of symbols. In both cases, the means of creating such a large number of characters was *conflation*. Conflation is a character formation process, by which two simple characters are *glued together* to form a complex character, as seen in Fig. 9.

We hypothesize that there are two types of conflation: *wholesome* and *unwholesome*. Wholesome conflation is based on the semantic-phonetic compound principle of *Rokusho*.

Chinese characters may be simplex or complex [see 15]. A complex character consists of two simple characters—or, more precisely, two radicals—one of which represents the semantic content of the whole character and the other its phonetic value, as seen in Fig. 10.

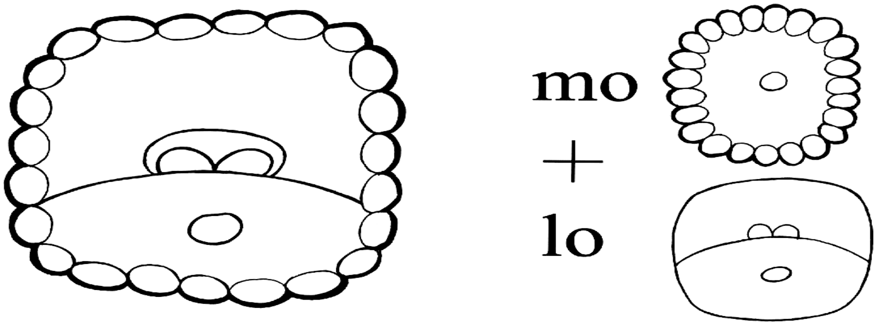


Fig. 9. A rare use of *conflation* in Maya word formation for the month name, *mol(o)*

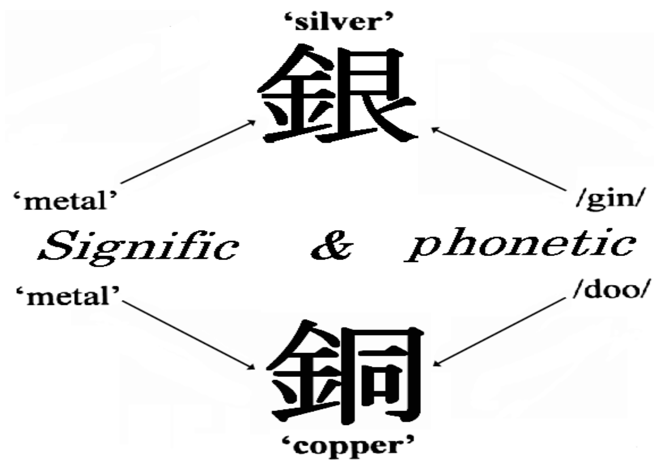


Fig. 10. An instance of wholesome conflation, based on the combination of a *signific* and a *phenetic*, which specifies the phonetic values associated with the whole character

This type of conflation is wholesome in the sense that phonetic values are *visible*, hence each character is pronounceable. Over 90% of common Chinese characters are formed on the basis of this principle. (Wholesome) conflation was rarely employed as a character formation method in the Maya writing system [see 3], and never used in Egyptian writing prior to the Greco-Roman period [see 16].

Unwholesome conflation is basically an orthographic or *visual pun*. In Chinese, it is manifested in an overuse of the so-called compound ideographic principle of *Rokusho*: a combination of two semantic radicals, rather than the more common combination of semantic and phonetic radicals. For example, the combination of the 'knife' and 'nose' characters yields a character which symbolizes the punishment of chopping off a criminal's nose.

Fig. 11. An instance of unwholesome conflation in Chinese word formation

Fig. 12. An instance of unwholesome conflation in Japanese word formation

The lexical meaning of the character is accessible only to those who know the etymological background of the whole character. This method is similar to that used in the formation of Japanese *kokuji*. Recall how the *kokuji* for "elevator girl" ("female" staff who used to be seen in large department stores, assisting customers in going "up" and "down" the elevator) is formed.

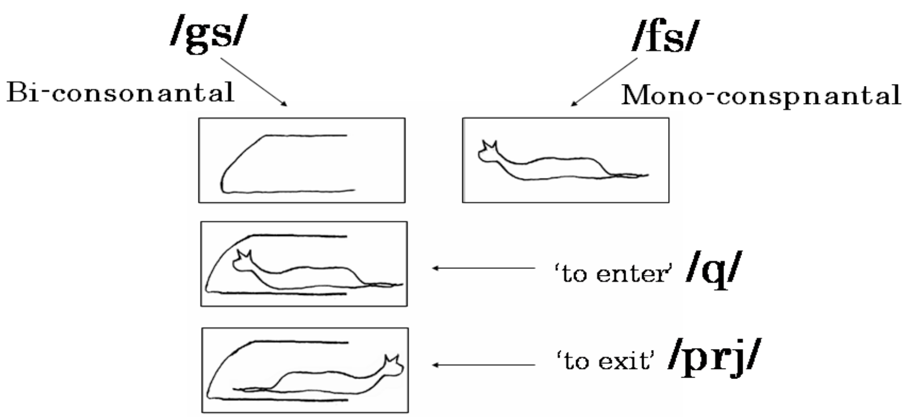


Fig. 13. An instance of unwholesome conflation in Egyptian hieroglyphic writing. The mono- and bi-consonantal signs, /f/ and /gs/, are used purely for a visual punning effect.

Unwholesome conflation was a major force in the enlargement of the character inventory in Chinese writing. A similar kind of orthographic punning was rampant in Egypt during the period of the Greco-Roman occupations [see 16].

The innovators were Egyptian religionists deliberately attempting to make their religious texts incomprehensible to invaders. In the case of Chinese, the inflation of characters may have arisen as a result of societal elites trying to outdo one another in elegance. In essence, the true function of writing as a communicative system was forgotten in both societies. These extravagances highlight the dangers of leaving a communicative system in the hands of a small, isolated, socially elite group.

4 Conclusion

In conclusion, this paper suggests that in developing a unified account of the evolution of writing systems, it is important to consider a variety of linguistic, psycholinguistic, and socio-cultural factors.

References

1. Allen, J.P.: Middle Egyptian: An Introduction to the Language and Culture of Hieroglyphs. Cambridge University Press, Cambridge (2000)
2. Coe, M.D., Kerr, J.: The Art of The Maya Scribes. Harry N. Abrams Inc., New York (1998)
3. Coe, M.D., van Stone, M.: Reading the Maya Glyphs. Thames & Hudson, New York (2001)
4. Collier, M., Manley, B.: How to Read Egyptian Hieroglyphs. University of California Press, Los Angeles (1998)
5. Cook, B.F.: Reading the Past: Greek Inscriptions. University of California Press, Los Angeles (1987)
6. Coulmas, F.: The Writing Systems of the World. Blackwell Publishers, Cambridge (1989)
7. Coulmas, F.: Writing Systems: An Introduction to Their Linguistic Analysis. Cambridge University Press, Cambridge (2003)
8. Daniels, P., Bright, W.: The World's Writing Systems. Oxford University Press, New York (1996)
9. DeFrancis, J.: The Chinese Language: Fact and Fantasy. University of Hawaii Press, Honolulu (1984)
10. Gelb, I.J.: A Study of Writing. 2nd edn. University of Chicago Press, Chicago (1963)
11. Houston, S.P.: Reading the Past: Maya Glyphs. University of California Press, Los Angeles (1989)
12. Houston, S., Mazariegos, W.C., Stuart, D. (eds.): The Decipherment of Ancient Maya Writing. University of Oklahoma Press, Norman (2001)
13. Kaiho, H.: Kanji Joho Shori Kisei o Megutte [Information Processing of Kanji]. Keiryo Kokugogaku 11 (1979) 331-340
14. Kess, J.F., Miyamoto, T.: Japanese Psycholinguistics: A Classified and Annotated Research Bibliography. John Benjamins Publishers, Amsterdam (1994)
15. Kess, J.F., Miyamoto, T.: The Japanese Mental Lexicon: Psycholinguistic Studies of Kana and Kanji Processing. John Benjamins Publishers, Amsterdam (1999)

16. Loprieno, A.: Ancient Egyptian: A Linguistic Introduction. Cambridge University Press, Cambridge (1995)
17. Macri, M.J., Looper, M.G.: The New Catalog of Maya Hieroglyphs. Volume One: The Classic Period Inscriptions. University of Oklahoma Press, Norman (2003)
18. McLuhan, M.: The Gutenberg Galaxy. University of Toronto Press, Toronto (1962)
19. Robinson, A.: The Story of Writing: Alphabets, Hieroglyphs & Pictograms. Thames & Hudson, London (1995)
20. Rozin, P., Poritsky, S., Sotsky, R.: American Children with Reading Problems Can Easily Learn to Read English Represented by Chinese Characters. *Science* 171 (1971) 1264-1267
21. Schmandt-Besserat, D.: The Envelopes that Bear the First Writing. Technology and Culture 21 (1980) 371-374
22. Schmandt-Besserat, D.: From Tokens to Tablets: A Re-evaluation of the So-called "Numerical Tables." Visual Language 15 (1981) 321-344.
23. Senner, W.M. (ed.): The Origins of Writing. University of Nebraska Press, Lincoln (1989)
24. Sproat, R.: A Computational Theory of Writing Systems. Cambridge University Press, Cambridge (2000)
25. Taylor, I., Taylor, M.M.: Writing and Literacy in Chinese, Korean, and Japanese. John Benjamins Publishers, Amsterdam (1995)
26. Walker, C.B.F.: Reading the Past: Cuneiform. University of California Press, Los Angeles (1987)
27. Webster, D.: The Fall of the Ancient Maya: Solving the Mystery of the Maya Collapse. Thames & Hudson, New York (2002)

The Emergence and Evolution of Graphical Productions

Nicolas Fay

ATR Media Information Science Laboratories, 2-2-2 Hikaridai,
Keihanna Science City, Kyoto 619-0228, Japan
nfay@atr.jp

Abstract. To study the development of graphical conventions, members of a simulated community were asked to play a series of graphical interaction games with partners drawn from the same pool. Once established, the community arrived at a set of conventional graphical referring expressions. The present paper offers a qualitative analysis of this interactive process, documenting the convergence and symbolization of participants' initially iconic graphical productions. This global process is contrasted with the local process evident among pairs who interact in isolation. Consistent with an evolutionary perspective, I argue that the graphical conventions that evolved within the simulated community are optimized representations, developed via a dynamic, interactive process.

1 Introduction

Whether reading a book or a road sign, graphical communication takes place everyday, and more often than not, several times a day. But where do these graphical productions come from, and how did their meaning become fixed within the wider community? The research reported addresses these fundamental questions by studying the evolution of graphical conventions within a simulated community of drawers.

It is widely believed that writing systems evolved from iconic representations to their current symbolic state (see Tversky, 1995). However, there is disagreement about the process of symbolization. Tversky (1995) argues that changes in representational form occur as a consequence of changes in writing medium (e.g. shells to bronze to paper). I believe this viewpoint underestimates the social nature of humans, where learning is often an interpersonal, as opposed to intrapersonal, process.

Prior work supports this position, demonstrating that interaction, rather than repetition, facilitates the transformation of icons to symbols (Fay, Garrod, Lee & Oberlander, 2003). In the Fay et. al. (2003) study pairs of participants communicated a series of concepts using only graphical means. The changing representation of the concept 'computer monitor' is illustrated in Figure 1.

What is initially an iconic representation of a computer monitor develops, through an interactive process of adaptation and entrainment, into a simplified, symbolic form (a schematic drawing of a computer case). The observed refinement and convergence of the graphical referring expression is consistent with that found in linguistic communication research (Clark & Wilkes-Gibbs, 1986; Garrod & Anderson, 1987). Figure 1 illustrates the emergence of a *local* convention; a graphical referring expression that meets interlocutors' local needs, analogous to the personal, often idiosyncratic

A. Sakurai et al. (Eds.): JSAI 2003/2004, LNAI 3609, pp. 357 – 361, 2007.

language used by close friends. In the present paper I extend this work to account for the evolution of *global* conventions; graphical productions that meet the communicative needs of the wider community.

A simulated community was created through a series of one-to-one interactions among partners drawn from the same pool. Out of these local interactions a set of global graphical referring conventions were established. These global conventions emerged on account of the polarization of a dominant graphical referring scheme within the community. In the following sections I document this process.

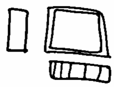
Person 1	Person 2	Person 1	Person 2	Person 1	Person 2
					

Fig. 1. A pair's changing representation of the concept 'computer monitor' over 6 games (from Fay, Garrod, Lee & Oberlander, 2003)

2 Method

Fay et. al.'s (2003) graphical referential communication task was employed. In this task pairs of participants are asked to communicate a series of concepts (12 from 16) using only graphical means. To ensure the task is not overly simple, a graphically confusable set of concepts was developed (i.e. theatre, art gallery, museum, parliament, Robert De Niro, Arnold Schwarzenegger, Clint Eastwood, drama, soap opera, cartoon, television, computer monitor, microwave, loud, homesick, poverty). Like the game 'Pictionary', participants are prohibited from speaking, or using text or numbers in their drawings.

Partners played six consecutive games using the same item set, with their role as drawer or identifier alternating from game to game[1]. Having concepts recur, and partners exchange drawing and identifying roles, facilitated an investigation of the evolution and co-ordination of graphical representations. Drawing was done on a standard whiteboard, the images recorded by digital camera.

A simulated community was created through a series of one-to-one interactions among partners drawn from the same pool. Over 7 rounds, each participant interacted with the other members of the community. The structure of the community is illustrated in Figure 2.

It was so designed that the community could first establish itself at Round 4. This was the earliest point the community could converge upon a conventional referring scheme. For example, if person 1 influences person 2 (Round 1), person 2 influences person 3 (Round 2) and person 3 influences person 8 (Round 3), person 1 and 8 will share some interactive history upon meeting in Round 4 (i.e. person 8 stands to have

[1] Although drawing and identifying roles alternated from game to game, participants were permitted to draw in either role, e.g. identifiers could graphically question or request clarification from the drawer.

ROUND	PAIR	PAIR	PAIR	PAIR
1	1&2	3&4	5&6	7&8
2	1&4	3&2	5&8	7&6
3	1&6	3&8	5&2	7&4
4	1&8	3&6	5&4	7&2
5	1&3	2&4	5&7	6&8
6	1&5	2&6	3&7	4&8
7	1&7	2&8	3&5	4&6

Fig. 2. Structure of pair-wise interactions among members of the simulated community

been indirectly exposed to the referring scheme of person 1 via person 3). Thus, Rounds 1-3 represent pre-convergence games whereas Rounds 4-7 represent post-convergence games.

3 Results

Several independent quantitative analyses were carried out to ascertain the emergence of a conventional graphical referring scheme (see Fay, Garrod, MacLeod, Lee & Oberlander, 2004). Rather than document these here, I shall instead present a set of example drawings that act as illustrations. Figures 3, 4 and 5 detail the 8 community members' changing representation of the concept 'cartoon' at Rounds 1, 4 and 7 (i.e. pre-convergence, convergence and post-convergence rounds respectively).

First, observe the reduction in graphical complexity across Rounds of the task. This is especially clear when pre-convergence (Round 1) and post-convergence drawings (Rounds 4 and 7) are contrasted. Not only do pre-convergence drawings contain more ink (one measure of visual complexity), they are composed of a greater number of distinct semantic elements (e.g. number of characters, objects etc.) when compared with post-convergence drawings. Thus, just as local refinement and symbolization of graphical expressions was observed among interacting pairs (see Figure 1), Figures 3 to 5 illustrate a comparable global process among members of the simulated community.

Notice also that graphical complexity is not reduced arbitrarily across rounds. Through interaction community members converge upon a specific representation, selected from the set of competing alternatives produced in Round 1. In the example provided, this is a simplified Mickey Mouse-like depiction of cartoon, characterized by two large circular ears above the head. Unlike the initial depictions of cartoon (Round 1, persons 5 and 6), the derived representations (Rounds 4 and 7) have been stripped of the unnecessary framing of the television and antenna, leaving only the salient properties of the image.

As community members encounter new partners, the Mickey Mouse-like depiction of cartoon becomes increasingly prevalent (2 instances in Round 1, as opposed to 4 in round 4 and 5 in Round 7). The proliferation of this scheme ensures that interlocutors are increasingly exposed to this referring expression, which in turn places them under mounting pressure to adopt this dominant scheme. It is as a consequence of this pressure, brought about by partners' interaction within a closed community, that graphical conventions are established. This finding is consistent with those reported by Garrod and Doherty (1994) in their study of the development of linguistic conventions.

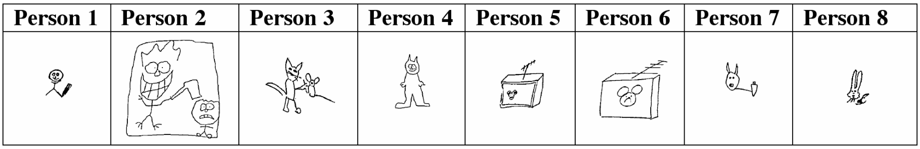


Fig. 3. Community members' drawings of the concept 'cartoon' in Game 1 of Round 1

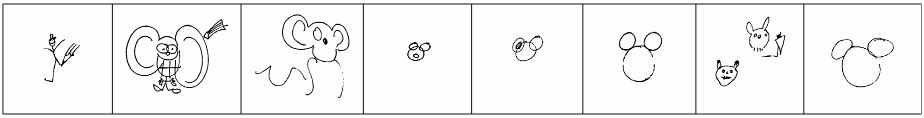


Fig. 4. Community members' drawings of the concept 'cartoon' in Game 1 of Round 4

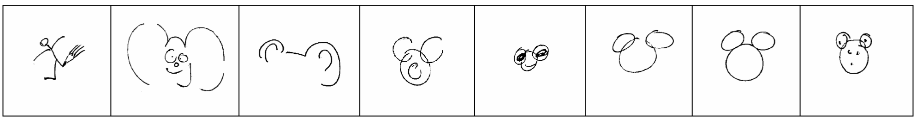


Fig. 5. Community members' drawings of the concept 'cartoon' in Game 1 of Round 7

4 Discussion

In the present paper I have documented the emergence and evolution of graphical productions. This was accomplished by creating a simulated community, composed of the pair-wise interactions of members drawn from within the same pool. Through interaction, as opposed to repetition, the community members negotiated a conventional graphical referring scheme. As the referring scheme was internalized, the need for elaborate external representation was reduced. Accordingly, community members' initially iconic representations were refined, resulting in a set of simplified, symbolic graphical productions.

Convergence upon a set of uniform graphical conventions was brought about by pressure to adopt the dominant referring scheme. Once the community was established, pressure to conform to the dominant scheme was exerted though participants' pair-wise interactions. This pressure to mimic our interlocutor's behaviour is well established, both linguistically and non-linguistically (Pickering & Garrod, in press; Chartrand & Bargh, 1999). Furthermore, as more community members adopted the dominant scheme, others were placed under increasing pressure to do likewise. Thus, the dominant scheme flourished within the community.

In several respects this process is comparable to that exhibited by interacting pairs (see Fay et. al., 2003), where interaction causes refinement, symbolization and convergence of graphical referring expressions. However, within a closed community this process takes place at a global level, as opposed to the local process evident among pairs who interact in isolation. As discussed, the polarization of a referring scheme within the community exerts global pressures that do not exist among interacting pairs, who need only satisfy their partner's personal communicative needs. Garrod and Doherty's (1994) finding that community members exhibit more closely coupled linguistic entrainment when compared with interacting pairs illustrates this global/local distinction.

Up until now I have discussed how graphical conventions evolve via an interactive process among the members of a closed community. This, however, leaves open the question of why a community converges upon one representation as opposed to another. Consistent with an evolutionary perspective, I argue that representations are selected on the basis of *fitness*. Recent work supports this viewpoint, showing that representations developed in communities offer distinct processing advantages (faster response times) when compared with those developed among unchanging pairs (Fay, Garrod, MacLeod, Lee & Oberlander, 2004). Discovering exactly what properties contribute to the *fitness* of a representation (e.g. semantic/syntactic complexity, iconicity, or systematicity) is an area of ongoing research.

Finally, while the research reported focuses upon the development of novel conventional representations, it can also account for the changing form of established representations. I predict that external factors (e.g. environmental change, or interaction with members of other communities) will introduce noise, provoking a reorganization of the existing referring system. Specifically, the accommodation of new concepts may require that older concepts be altered or discarded. This would explain the evolving nature of both writing systems and language.

Acknowledgements

This research was supported in part by the National Institute of Information and Communications Technology, Japan, and the ESRC and EPSRC (grant L323253003).

References

Chartrand, T. L., & Bargh, J. A. (1999). The chameleon effect: The perception-behavior link and social interaction. *Journal of Personality and Social Psychology,* 76, 893-910.

Clark, H. H., & Wilkes-Gibbs, D. (1986). Referring as a collaborative process. *Cognition,* 22, 1-39.

Fay, N., Garrod, S., Lee, J., & Oberlander., J. (2003). Understanding interactive graphical communication. *Proceedings of the 25th Annual Conference of the Cognitive Science Society, 2003,* pp. 384-389.

Fay, N., Garrod, S., MacLeod, T., Lee, J., & Oberlander., J. (2004). Design, adaptation and convention: The emergence of higher order graphical representations. *Proceedings of the 26th Annual Conference of the Cognitive Science Society, 2004,* pp. 411-416.

Garrod, S., & Anderson, A. (1987). Saying what you mean in dialogue: a study in conceptual and semantic co-ordination. *Cognition*, 27,181-218.

Garrod, S., & Doherty, G. (1994). Conversation, co-ordination and convention: an empirical investigation of how groups establish linguistic conventions. *Cognition*, 53, 181-215.

Pickering, M., & Garrod, S. (in press). Towards a mechanistic theory of dialogue. *Behavioral and Brain Sciences*.

Tversky, B. (1995). Cognitive origins of graphic conventions. In F. T. Marchese (Editor). Understanding images. pp. 29-53. New York: Springer-Verlag.

Complex Vocal Behavior and Cortical-Medullar Projection

Kazuo Okanoya[1,2,4], Sayaka Hihara[3,4], Naoko Tokimoto[1],
Yasuko Tobari[1], and Atsushi Iriki[3,4]

[1] Okanoya Lab., Faculty of Letters, Chiba University,
1-33 Yayoi-cho, Inage-ku,
Chiba 263-8522, Japan
{Okanoya,Ntokimot,Ytobari}@cogsci.L.chiba-u.ac.jp
[2] Japan Science and Technology Agency,
Kawaguchi Center Building, 4-1-8, Honcho, Kawaguchi-shi,
Saitama 332-0012, Japan
[3] Cognitive Neurobiology Lab., Tokyo Medical and Dental University,
1-5-45 Yushima, Bunkyo-ku,
Tokyo 113-8549, Japan
{Iriki,Hihara}@cnb.tmd.ac.jp
[4] Brain Science Institute, RIKEN,
2-1, Hirosawa, Wako,
Saitama 351-0198, Japan
{Okanoya,Hihara,Iriki}@brain.riken.jp

Abstract. We argue that the intentional control of vocal organ is the most basic predisposition for vocal learning and thus for language acquisition. Anatomical substrates for intentional vocal control are the direct cortical-medullar projections that connect face motor cortecies and the nucleus retro-ambiguus. We ask how such projections may be reinforced in non-vocal learners including macaque monkeys and rodents by behavioral manipulations. We hypothesize how such connections may be prepared in humans.

1 Introduction

Vocal learning independently evolved several times in vertebrates [1-3]. Vocal learning refers to motor learning of spectro-temporal features of sound signals that are used in intra- and inter species communication by conspecific members. Most animal vocalizations are innately determined. Vocal learning, experience dependent, categorical, and long-lasting modification of vocal output, do not occur in most animals and only a few species of animals including whales, bats, birds, and humans that.

1.1 Anatomical Pre-adaptations

Are there any specific anatomical substrates that correlate with the faculty of vocal learning? One of the candidates for this question is the direct cortical-medullar pathway for articulation and breathing. In humans, a part of motor cortex directly projects

A. Sakurai et al. (Eds.): JSAI 2003/2004, LNAI 3609, pp. 362 – 367, 2007.

to the medullary nuclei, the nucleus ambiguus and the nucleus retro-ambiguus [4]. This projection is absent in the squirrel monkey and chimpanzee. Jurgens [5] thus assumes that this projection exists only in humans among primates.

Similarly, there is a direct cortical-medullar pathway for articulation and breathing in the zebra finch, a species of songbirds, but a similar projection in pigeons do not exist. Most of pigeon vocalizations are considered to be innate [6,7]. Considering these evidence, we can hypothesize that this projection exists in the species that show vocal learning while it is absent in the species without vocal learning.

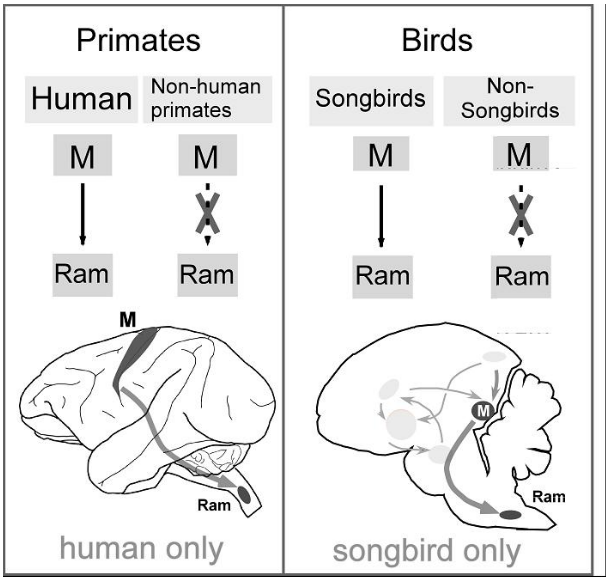


Fig. 1. In primates, humans are the only species that was shown to have direct pathway connecting face motor regions of the cortex and medullary respiratory structures. In birds, zebra finches and Bengalese finches are shown to have such pathway, but pigeons do not.

1.2 Plasticity in the Anatomical Predisposition

While this projection exist only in a limited number of species, it may be possible, that this projection is simply very faint in most of species. Deacon [8] introduced anecdotal story of a zoo seal that learned to mimic speeches of drunken persons. This seal had a brain inflammation as young and Deacon suspect that during the process of recovery the cortical medullar projection might be reinforced in this particular animal. If such cases could occur, by training animals to perform spontaneous vocalization while they are young, we may be able to reinforce this pathway and induce vocal learning in a species that was said to be non vocal learners. A report on the spontaneous vocal differentiation in Japanese macaques [9] and that on the spontaneous construction of nested hierarchical structure in a species of rodent, Degus, are suggestive of such possibility although no anatomical data were available at present [10].

2 Vocal "Naming" of Objects by Japanese Macaques

We observed that when trained to use a rake to retrieve a distant food, monkeys began to vocalize "coo" calls spontaneously. They did so especially when the preparation of the rake tool by the experimenter delayed [9]. To further investigate this phenomenon, we systematically manipulated behavioral contexts by presenting the tool or food whenever the monkey made a vocalization regardless of the types of the calls. In one experimental situation, the experimenter gave a food at a distance when the monkey produced a coo call (Call A). By the second coo call (Call B) the experimenter placed a rake tool to the monkey. The monkey could retrieve the food by the rake. In another, the experimenter gave the rake to the monkey beforehand. A food was placed at a distance when the monkey vocalized a coo call (Call C). Likewise, we never tried to differentiate the calls (Fig. 2).

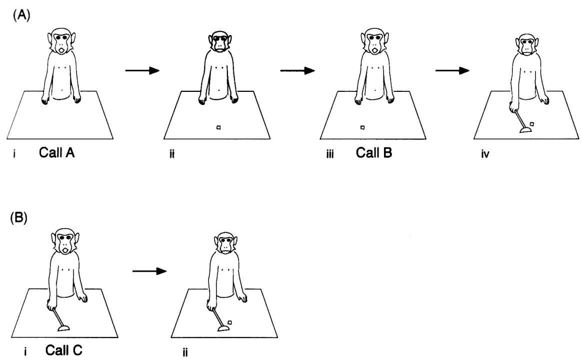


Fig. 2. Experimental settings. (A) Condition 1: When the monkey produced a coo-call (Call A), the experimenter put a food reward on the table, but out of his reach. When the monkey again vocalized a coo-call (Call B) the experimenter presented the tool within his reach. The monkey was then able to retrieve the food suing the tool. (B) Condition 2: Initially, the tool was presented within the monkey's reach on the table. When the monkey vocalized a coo-call (Call C) the experimenter set a food reward within reach of the tool. From Hihara et al. (2003) [9].

2.1 Reward Context and Vocal Plasticity

After 5 sessions of trainings, the monkeys eventually used acoustically distinct types of calls when they asked for the tool (Call 2) or food (Call 1 and 3). The calls used to ask for the tool was longer and higher pitched than the ones used to ask for the food.

We argue that the different reward conditions (food or tool) set up different emotional contexts for the monkeys. Different emotional contexts, in turn, affected the production of coo calls differently for the tool or food situations. Since the tool training activated the neo-cortex very highly, the calls were associated with different behavioral contexts.

Thus, the calls became categorized and emotionally differentiated calls gradually became categorical vocalizations. Through this process, we suspect the emotional coo calls changed into categorical labels denoting the behavioral situation.

3 Spontaneous Construction of Hierarchical Structures in Vocal Trained Rodents

In another study, when Degus (a species of rodent) were trained to vocalize in order to obtain food reinforcement, we observed they spontaneously constructed hierarchical self-embedded structure [10]. We trained Degues in a vocal operant conditioning task and the subjects learned to vocalize in order to get food within 2 months of time. During this period, Degues spontenously began to construct curious objects in their home cage. With a large dust-bath dish, a medium sized food cup, and a small toy ball, they spontaneously constructed a triplet "Chinese box" like structure of "the ball into the cup into the dish." This particular behavior occurred only during the period of operant training (Fig. 3).

(a) (b)

Fig. 3. (a) A "Chinese box" constructed by the degu. (b) From left to right: Sand bath, food cup, ball, and the degu. From Tokimoto & Okanoya (2004) [10].

3.1 Hierarchical Manipulation in Animals

It has been assumed that the ability to combine multiple objects hierarchically to construct self-embedded structures is restricted to primates including humans. We interpreted this as an opposite case for the monkey example. Vocal operant training probably put a heavy load to the brain of the Degus and required co-activation of medullary, limbic, and cortical vocal related areas. This prepared hierarchically organized behavior in general and Degus could utilize this for their "play" like, or "tool-use" like behavior.

Greenfield et al. [11] found parallels between developmental patterns of phonological and grammatical constructions and combinatorial activity with nesting cups in children. When 3 cups were given to children they performed one of the following 3 activities; 2 of the 3 cups were put together to form duplex (pair), 3 cups were put together one by one to form a triplet (pot), and the smallest one and the medium one were put together to form a subassembly and this subassembly was put into the largest one (subassembly). Children at 11 months mostly performed only the pair strategy but gradually the proportion of the pot strategy increased until 24 months and after that, the subassembly strategies gradually increased [11]. This time course corresponded with emergence of two-word sentences and more elaborated grammatical sentences. She extended this finding to the possibility of neurally based developmental homology between language and action.

A similar behavior was observed in tool-trained Japanese macaques [12]. These monkeys sit on a monkey chair and were trained to use a large rake to retract a food that was out of its reach without the rake. The monkeys were then presented with a situation in which the large rake was also out of reach but they can use a small rake to get the large rake. Without training, the monkeys succeeded in using the small rake to get the large rake, and then using the large rake to get the food. This behavior also requires the understandings of "meta tool" and the relationship between the tools.

3.2 Vocal Plasticity and Hierarchical Manipulation

These observations suggested that when a proper context exists, these animals exhibit a faculty of hierarchical operation. Whether or not this faculty is actually related with phonological and grammatical constructions require another perspective besides behavioral observation.

4 What Are "Pre-adaptation" for Cortical-Medullar Projection in Humans?

Among vocal learners, birds, bats, and whales seem to have reasonable pre-adaptation Birds and bats have to control their breathing while they are flying and this requirement was probably the pre-adaptation to the intentional vocal control. Similarly, whales need to control breathing while they submerge in the water and this led to the intentional vocal control. But why humans gained intentional control of vocalizations while no other primate species evolved such functions? One of the explanations had been that of Homo-Aquarius, but here we suggest an alternative.

4.1 Baby Cry and Origin of Cortical-Medullar Pathway

Human babies are conspicuous among primates in that they emit high- intensity, long-lasting cries right after the birth. Such cries are obviously maladaptive in wild animals in that it could easily attract predators. We suggest therefore the infant cry must be a behavior obtained after ancestral humans acquired social and cultural skills to protect themselves from predation risk [13]. The first cry after the birth has function to eliminate amniotic fluid, but cry continues long after that.

Likewise, nidicolous species of birds emit very loud begging calls and isolation calls that emit parental behavior including feeding and protection of the hatchlings. Begging calls are so loud that it easily can attract predators. Therefore, parents have to emit strong parental behavior to avoid predation of hatchlings and themselves. Birds could do this because they acquired cortical control of breathing to adapt for flying.

We analyzed syntactical and phonological developmental changes in baby cry to show how the pattern becomes complex as the baby grow. At least four stages of cry development could be identified; each may be associated with respective anatomical changes. Baby cries begin as a regular repetition of a stereotyped vocal unit. At this stage, the cry probably is controlled by midbrain vocal center only. Babies continue to cry such vocalizations until two weeks after birth, after which cries begin to be more irregular, showing variable patterns of phonologies and rhythmic patterns. At this stage, limbic influence will gradually becomes stronger. Mothers often can identify what babies want based on the pattern of cries.

During the interaction with mothers, the cry probably became more adaptive by allowing cortical control so that mothers can be more preciously governed by baby cry. Infant cries thus may function to train cortical-midbrain connections necessary for more intentional vocal output that eventually resulted in the acquisition of language.

5 Conclusions

Taken together, we suspect heavy cognitive load associated with vocal behavior may prepare for the direct anatomical connection between cortical and medullary vocal centers necessary for vocal learning. To further investigate this process, we need to show how these cognitive loads could actually affect anatomical structures.

References

1. Boughman, J. W.: Vocal learning by greater spear-nosed bats. Proc. R. Soc. Lond. B, 265 (1998) 227-233
2. Catchpole, C. K., Slater, P. J. B.: Bird Song. Cambridge University Press, Cambridge (1995)
3. Payne, R.: Among Whales. Simon & Shuster, Inc., New York (1995)
4. Kuypers, H. G. J. M.: Corticobulbr connection to the pons and lower brain-stem in man. Brain, 81 (1958) 364-88
5. Jurgens U.: A study of the central control of vocalization using the squirrel monkey. Med. Eng. Phys., 7-8 (2002) 473-477
6. Wild, J. M.: Descending projections of the songbird nucleus rubustus archistriatalis. J. Comp. Neurol. 338 (1993) 225-241
7. Wild, J. M., Li, D., Eagleton, C.: Projections of the dorsomedial nucleus of the intercollicular complex (DM) in relation to respiratory – vocal nuclei in the brainstem of pigeon and zebra finch. J. Comp Neurol. 377 (1997) 392-413
8. Deacon, T. W. The symbolic species. Norton & Company, New York (1997)
9. Hihara S, Yamada H, Iriki A, Okanoya K. Spontaneous vocal differentiation of coo-calls for tools and food in Japanese monkeys. Neurosci Res. 45 (2003) 383-9
10. Tokimoto, N. Okanoya, K.: Spontaneous construction of “Chinese Boxes” by Degus (Octodon degu): a rudiment of recursive intelligence? Japanese Psychological Research, In press (2004)
11. Greenfield, P. M.: Language, tools, and the brain: The ontogeny and phylogency of hierarchically organized sequential behavior. Behavioral and Brain Science, 14 (1991) 531-595
12. Hihara, S., Obayashi, S., Tanaka, M., Iriki, A. Rapid learning of sequential tool use by macaque monkeys. Physiology & Behavior, 78 (2003) 427-434
13. Okanoya, K., Ichii, H., Ushijima, H.: Baby cry as a pre-adaptation to language. Abstract for the 4th congress of the evolution of language. (2002)
14. Okanoya, K.: Functional and structural pre-adaptations to language: Insight from comparative cognitive science into the study of language origin. Japanese Psychological Research, In press (2004).

Logic and Engineering of Natural Language Semantics 2004

This part of this volume is made up of revised and refined versions of the papers submitted to the First International Conference of Logic and Engineering of Natural Language Semantics, LELNS2004, held in Kanazawa, Japan, on May 31, 2004. LELNS2004 featured theories of *Dynamic Semantics* and their applications to natural language analysis and natural language engineering. Dynamic Semantics is one of the most vivid and promising areas of current research in natural language semantics. Some representative approaches within this general area are:

– *Discourse Representation Theory (DRT)*,
– *Segmented Discourse Representation Theory (SDRT)*,
– *Dynamic Predicated Logic (DPL)*,
– *Context Change Semantics (CCS)*, and
– *Update Semantics (US)*.

This section is made up of eight papers presented at LENLS. They were revised several times by the authors, with input from the Organizing Committee: Norihiro Ogata (Osaka University), Yasuo Nakayama (Osaka University), Katsuhiko Yabushita (Naruto University of Education), and Eric McCready (Univeristy of Texas).

I wish to express my gratitude to all the members of the Organizing Committee, all those who submitted papers, and all the members of the Orginizing Committee of the International Workshops of the Annual Conference of JSAI.

I also thank the JSAI for providing the opportunity to hold an internatoinal workshop on Dynamic Semantics in Japan.

Toyonaka, October 2004 Norihiro Ogata

A. Sakurai et al. (Eds.): JSAI 2003/2004, LNAI 3609, p. 371, 2007.

Dynamic Predicate Logic of Dependent Questions and Answers

Norihiro Ogata

Faculty of Language and Culture, Osaka University
1-8 Machikaneyama, Toyonaka, Osaka, Japan
ogata@lang.osaka-u.ac.jp

1 Introduction

In the progress of formal semantics of natural language, one of the main themes is formal semantics of questions and answers, such as [18], [22], [5], [23], [20], [4], [9], [7], [15], [12], [19], [26], [17], [3], [14], [13], [1], [2], and [29][1]. [18] proposed that the semantics of a question is a set of propositions that can serve as possible

[1] According to [28], in the main question theories, the denotations of the n-ary question formula $?\bar{x}\varphi$ are defined as follows:

- (Hamblin-Karttunen) $[\![?\bar{x}\varphi]\!]^{w,g}_{HK} = \{\{u \in W|\, [\![\varphi]\!]^{w,g[\bar{d}/\bar{x}]} = [\![\varphi]\!]^{u,g[\bar{d}/\bar{x}]} = 1\}|\bar{d} \in D^n\}$
- (Higginbotham 1996) $[\![?\bar{x}\varphi]\!]^{w,g}_{H} = \{\{u \in W|\, [\![\varphi]\!]^{w,g[\bar{d}/\bar{x}]} = [\![\varphi]\!]^{u,g[\bar{d}/\bar{x}]}\}|\bar{d} \in D^n\}$
- (Groenendijk-Stokhof) $[\![?\bar{x}\varphi]\!]^{w,g}_{GS} = \{u \in W|\forall\bar{d} \in D^n.\, [\![\varphi]\!]^{w,g[\bar{d}/\bar{x}]} = [\![\varphi]\!]^{u,g[\bar{d}/\bar{x}]}\}$
- (Aloni) $[\![?\bar{x}\varphi]\!]^{w,g}_{A} = \{\{u \in W|\, [\![\varphi]\!]^{w,g[\bar{c}(w)/\bar{x}]} = [\![\varphi]\!]^{u,g[\bar{c}(u)/\bar{x}]}\}|\bar{c} \in \prod_{i\in n}(\wp(x_i))\}$, where $\wp : \mathbb{N} \to CC$ (conceptual perspective) and CC is a set of individual concepts such that $\forall w \in W.\forall a \in D.\exists! c \in CC.c(w) = a$.

In [1], the definition of the above equation is confused with respect to type. [17] define an answer ψ to $?\bar{x}.\varphi$ iff $[\![\psi]\!]^g \subseteq [\![?\bar{x}\varphi]\!]^g$, while [1] defines it as $[\![\psi]\!]^g \in [\![?\bar{x}\varphi]\!]^g$. Therefore, I have redefined the above semantics of Aloni.

[3] defines the semantics of questions as follows:

- $[\![?\lambda x\varphi]\!] = \{p|p = [\![\alpha(\lambda x.\varphi)]\!]\} \wedge {}^{\vee}p \wedge (\exists x\Box({}^{\vee}p \to \varphi) \vee \Box({}^{\vee}p \to \neg\exists x\varphi)$

[11] defines constituent questions in the framework of dynamic epistemic logic:

- $(w, \sigma) \models^g_R ?x\varphi \Leftrightarrow (w, \sigma) \models_R \forall x(\Box\varphi \vee \Box\neg\varphi)$
- $(w, \sigma) \models^g_R \Box\varphi \Leftrightarrow \forall v \in \sigma.\forall h.\forall x \in Var.(w, g(x))R(v, h(x)).(v, \sigma) \models^h_R \varphi$

R is an equivalence relation, $w = \langle D_w, I_w\rangle$ an epistemic possibility and σ an information state, i.e., a set of epistemic possibilities.

[14] proposes a dynamic semantics of questions and answers, as follows: let $c \subseteq W \times W$ be a *context*, and $w_1, w_2 \in W$ possible worlds.

- (answer) $c[\varphi!] = \{\langle w_1, w_2\rangle \in c|\, [\![\varphi!]\!]_{w_1} = [\![\varphi_2]\!]_{w_2} = 1\}$
- (question) $c[\varphi?] = \{\langle w_1, w_2\rangle \in c|\, [\![\varphi!]\!]_{w_1} = [\![\varphi_2]\!]_{w_2}\}$

[2] also propose a dynamic semantics of questions and focus, But, unfortunately, it is not accurate, although its ideas are interesting.

A. Sakurai et al. (Eds.): JSAI 2003/2004, LNAI 3609, pp. 372–382, 2007.

answers. Basically, [22] follows this stance. I call these *answer-set theories* of questions and answers. [20], [15], and [17] are also basically *answer-set theories*, although they introduce the partitions of possible worlds instead of simple sets of propositions. However, from the point of view of dynamic semantics, questions and answers clearly change their contexts. A question changes a context to the *questioned context*. It verifies only its answers, or another relevant question, as in (1). This means questioned contexts can be nested. On the other hand, an answer changes the questioned context to a recovered context, as in (2), where the recovered context is the first questioned context.

(1) A: Do you like coffee?; B: What do you mean?

(2) A: Do you like coffee?; B: What do you mean?; A: You seem to be sleepy.; B: Okay, I'll have a coffee.

Therefore, questions make the contexts stack-like structures. Firstly, to handle these stack-like structures, I will propose a dynamic semantics of questions and answers by expanding *Dynamic Predicate Logic* (*DPL*) by [16], called $DPL_{?¿}$, by introducing *parameterized Herbrand base stacks* as contexts. In particular, $DPL_{?¿}$ will handle questions which I call *dependent questions*[2] such as *Wh-GQ questions* ([13]) and *multiple Wh-questions* ([24]). To handle dependent questions and their answers, the parameterized structure of parameterized Herbrand base stacks and their operations are exploited.

Secondly, as [29], [12] and others argue, the *resolvedness* is sensitive to the intentions of the questioner. For example, in Tokyo, the question 'Where am I?' and the answer 'Tokyo' is non-sense in a normal situation, whereas the same question and the answer 'Roppongi' (a name of a district of Tokyo) is informative. [29] treats this *decision problem* by exploiting the decision theory. But I think it is relevant to the presupposition of the *situated* question, and this presupposition is formalized by the common knowledge between the questioner and the answerer, as in [25]. Therefore, it will be treated as a parameter of *questioning context*, i.e., the context before questioning.

Thus, in this paper, questions and answers are regarded as update functions of contexts. I call this view the *update theory* of questions and answers. This is a quite different view of questions and answers, compared with the previous researches enumerated above.

Section 2 will classify *dependent questions* and their answers.[3] Section 3 will define $DPL_{?¿}$. Section 4 will apply $DPL_{?¿}$ to some significant cases of dependent questions and their answers.

2 Dependent Questions and Their Answers

Many papers study dependent questions and their answers. First, I classify dependent questions as follows:

[2] This term is used in a sense similar to [23]'s notion of *multiple questions*.

[3] In this paper, I will not treat wh-GQ questions such as 'Which student did few/many/at least two/... professors see? On this issue, see [13], [30], or [27].

(3) a. ∀-who type: e.g., Who does every boy admire?
b. ¬∃-who type: e.g., Who does no boy admire?
c. who-∀ type: Who admires every boy?
d. who-who type: e.g., Who admires who?[4]
e. who∼who type: e.g., ∗ What did who read?

Secondly, I classify possible answers to dependent questions into the following types:

(4) a. Single answers: e.g., Agent Smith
b. Exhaustive pair-list answers: e.g., Bill admires John, and Ken Sue.
c. Single pair-list answers: e.g., Bill admires John.
d. Functional answers: e.g., (Every boy admires) his mother.
e. Partial pair-list answers.

Although this is not an exhaustive list, I sum up and classify *binary dependent questions* as in the following table (OK means that it is semantically acceptable; on the other hand, NG means that it is not semantically acceptable):

Question Type	Single	Exhaustive Pair-list	Single Pair-list	Functional	Partial Pair-List
∀-who	$OK_{2,7}$	$OK_{2,7}$	NG	$OK_{2,7}$	NG_6
¬∃-who	[OK]	$NG_{2,7}$	NG	$OK_{2,7}$	–
who-∀	$OK_{2,7}$	$NG_{2,7}$	NG	OK_2	–
who-who	NG_8	$OK^5_{5,8}$	NG_8	OK_8	OK_6

where the items with subscript 2 are from [7], the items with subscript 5 [13], the items with subscript 6 [19], the items with subscript 7 [6], the items with subscript 8 [8][5], and the items enclosed by [], my informants' judgments.

[4] According to [21], the following focus patterns of the answers of a dependent question are possible: for question *Who kissed who?*,

(i) $[\text{Larry}]_{TF}$ kissed $[\text{Nina}]_{FF}$.

(ii) $[\text{Larry}]_{FF}$ kissed $[\text{Nina}]_{FF}$.

TF is *Topic-Focus* (or contrastive topic), and FF is *Focus-Focus* (the narrow sense of focus) in terms of [21]. Naturally, this paper does not consider answers of type (i) focus pattern, since (i) already presuppose the possible answers of 'Who did Larry kiss?' which means the dependency between the arguments of the answer has already specified in the discourse. I am interested in the dependency between non-specific arguments. Therefore, many linguists report that usually type (ii) answers are exhaustive pair list, but not single answers.

[5] In some contexts, single pair-list is acceptable. For example, uttered in a context where there are only two individuals, the following is semantically acceptable: (i) Q: Who hit who first? A: Tom hit Ben.

3 $DPL_{?¿}$

3.1 The Problem of the Logical Form of Dependent Questions

First, consider the logical forms of ∀-who type questions. They are of the form:

(5) $\forall x.\varphi_1(x) \rightarrow ?y.(\varphi_2(y) \wedge \varphi_3(x, y))$

In dynamic semantics, $\forall x.\varphi$-type formulas are tests, that is, they cannot update the input states. This fact is problematic, since I adopt the view that questions changes the input states to their *questioned states*. To avoid this problem, we have two choices: (i) the revision of the dynamic semantics of $\forall x.\varphi$, or (ii) the introduction of *Skolem function variables* $(\xi, \zeta, ...)$. I adopt (ii), since I do not want to make revisions to the base form of DPL. The resulting formula is as follows:

(6) $?\xi.\forall x.\varphi_1(x) \rightarrow \varphi_2(\xi(x)) \wedge \varphi_3(x, \xi(x))$

This formula can not only update the input states in the dynamic semantics, but also the semantic value is the same with (5) in the static semantics.

Second, consider the logical forms of who-who type questions. They are of form:

(7) $?x.\varphi_1(x) \begin{Bmatrix} \rightarrow \\ \wedge \end{Bmatrix} ?y.(\varphi_2(y) \wedge \varphi_3(x, y))$

One problem is the choice of → or ∧ of the right of the first question formulas. The first 'who' work as "sorting keys" of answers, as [24] points out, as follows. For question (8a), answer (8b) is appropriate, but (8c) is not.

(8) a. Who kicked who?
b. John kicked Mary, Bill Sue, Mary John, and Sue Bill.
c. # Mary was kicked by John, Sue Bill, John Mary, and Bill Sue.

Therefore, for each answer of the first 'who', its total answer replies. This fact urges us to choice → instead of ∧, since → corresponds to the function of sorting keys, and furthermore, as the first 'who' works as a sorting key, the ?-operator of (8) has a kind of exhaustive listing sense, which I distinguish from the normal ?-operator, and which I write as $?^{\forall}$, as in (9).

(9) $?^{\forall}x.\varphi_1(x) \rightarrow ?y.(\varphi_2(y) \wedge \varphi_3(x, y))$

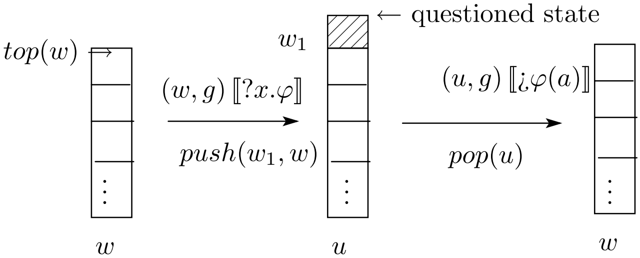
Therefore, for each possible value of the first 'who', its questioned state is decided. To formalize this property, I introduce *parametrized possible worlds*, which functions a function from individuals to possible worlds. I define a parametrized possible world as a pair $\langle w, f \rangle$ of possible world w and function $f : D_w \rightarrow W$. Instead of $appl(\langle w, f \rangle, a)$, I write it by $w_{f(a)}$, which is a possible world.

Secondly, the second 'who' must be *Skolemized* similarly with the case of (5), since in (9), $?y$ is in the scope of → and → is a test in DPL. This means that $?y$ cannot update the input states. Therefore, the resulting formula is as follows:

(10) $?\xi.?^{\forall}x.\varphi_1(x) \rightarrow \varphi_2(\xi(x)) \wedge \varphi_3(x, \xi(x))$

3.2 The Dynamic Interpretation of Dependent Questions and Answers

I propose that the interpretation of a question is a function from input states to their questioned states. More strictly, the interpretation of a question is a function which pushes the questioned state of the top of the input state to the top of the input state. The questioned states verify only the correct answers. Its answer checks or tests the questioned state. See the following figure:



If the answer is verified by the questioned state and the verifiable condition is only the answer, the questioned state will be popped from the top of the input state. Otherwise, it fails. Therefore, input states are stacks.

3.3 The Basic Definitions on $DPL_{?¿}$

I define the language of $DPL_{?¿}$ and its dynamic semantics, as follows:

- Let $c \in Con$ (constants), $x \in Var$ (variables), $f \in Fun(n)$ (n-ary function symbols), and $\xi \in SVar$ (Skolem function variables). Then a term $\tau \in TERM$ is defined by the following BNF grammar:
  $\tau ::= x|c|f(\tau_1, \ldots, \tau_n)|\xi(\tau_1, \ldots, \tau_n)$
- Let $R \in Rel(n)$ (n-ary relation symbols) and Φ be the set of atomic formulas, i.e., the set of formulas of form $\tau_1 = \tau_2$, $\xi = S$ ($S \in Sk(n)$ is an n-ary Skolem function symbol) or $R(\tau_1, \ldots, \tau_n)$, and $p \in \Phi$. Then a formula $\varphi \in L$ is defined by the following BNF grammar:
  $\varphi ::= p|\varphi_1 \wedge \varphi_2|\neg\varphi|\varphi_1 \vee \varphi_1|\varphi_1 \to \varphi_2|\varphi_1 \equiv \varphi_2|\neg\varphi|\exists x.\varphi|\forall x.\varphi|?\bar{\xi}.\varphi|?\bar{x}.\varphi|?^{\forall}\bar{x}.\varphi|¿\varphi$,
  where $¿\varphi$ is an answer formula.
- $\mathsf{A} \subseteq TERM$ is a set of agents.
- A model is a tuple $\mathfrak{A} = \langle W, W_0, D, (R_G)_{G \subseteq \mathsf{A}}), I\rangle$, where
  1. W is a set of possible world stacks generated by set of possible world W_0 and functions $F : D \to W_0$, defined as the least class satisfying the following conditions:
     (a) $\Lambda \in W$ (the empty stack),
     (b) $w \in W_0, u \in W_0 \cup W \Rightarrow push(w, u) \in W$,
     (c) $w \in W_0, f \in F, u \in W_0 \cup W \Rightarrow push(\langle w, f\rangle, u) \in W$,
     and $top(push(w, u)) = w$ and $pop(push(w, u)) = u$. $\langle v, f\rangle$ is a function from $dom(f) \to W_0$, called a *parametrized possible world*.
  2. $I : (\forall w \in W_0.\forall n \in \mathbb{N}.(Rel(n) \to (D_w)^n) \cup (Con \to D_w) \cup (Fun(n) \cup Sk(n) \to {}^{D_w}D)$.

3. $R_G \subseteq W_0 \times W_0$ (the group epistemic accessibility relation between possible worlds)

– A context $c \in C$ is a function such that:
  1. if $G \subset \mathsf{A}$, then $c(G) \subseteq W_0$ and it is the largest fixed point of equation:[6]

$$X = \bigcap_{a \in G} \{w \in W_0 | \forall u.wR_a u \Rightarrow u \in c(G) \cap X\}$$

  2. $c(\mathit{speaker}), c(\mathit{hearer}), c(\mathit{adressee}) \in \mathsf{A}$ and $c(\mathit{speaker}) \neq c(\mathit{addressee})$

– Let $g, h, i, k \in \mathcal{A}$ and $w, u, v \in W$. The semantics $[\![\varphi]\!] \subseteq (C \times W \times \mathcal{A}) \times (C \times W \times \mathcal{A})$ and $[\![\tau]\!]_g$ of L is defined by recursion on τ and induction on φ:
  1. $[\![x]\!]_{w,g} = g(x)$;
  2. $[\![c]\!]_{w,g} = I(w)(c)$;
  3. $[\![f(\tau_1, \ldots, \tau_n)]\!]_{w,g} = I(w)(f)([\![\tau_1]\!]_{w,g}, \ldots, [\![\tau_n]\!]_{w,g})$
  4. $[\![S]\!]_{w,g} = I(w)(S)$;
  5. $[\![\xi]\!]_{w,g} = g(\xi)$;
  6. $(c, w, g)\, [\![\tau_1 = \tau_2]\!](c, w, g)$ iff $[\![\tau_1]\!]_{w,g} = [\![\tau_2]\!]_{w,g}$;
  7. $(c, w, g)\, [\![R(\tau_1, \ldots, \tau_n)]\!](c', u, h)$ iff $c' = c$, $u = w$, $h = g$, and $([\![\tau_1]\!]_{w,g}, \ldots, [\![\tau_n]\!]_{w,g}) \in I(w)(R)$;
  8. $(c, w, g)\, [\![\varphi_1 \wedge \varphi_2]\!](c', u, h)$ iff there is some (c'', v, i) such that $(c, w, g)\, [\![\varphi_1]\!]\, (c'', v, i)$ and $(c'', v, i)\, [\![\varphi_2]\!](c', u, h)$;
  9. $(c, w, g)\, [\![\varphi_1 \rightarrow \varphi_2]\!](c', u, h)$ iff $u = w$ and $h = g$ and for all (c'', v, i) such that $(c, w, g)\, [\![\varphi_1]\!](c'', v, i)$, there is (c''', t, j) such that $(c'', v, i)\, [\![\varphi_2]\!](c''', t, j)$;
  10. $(c, w, g)\, [\![\varphi_1 \equiv \varphi_2]\!](c', u, h)$ iff $u = w$ and $h = g$ and $(c, w, g)\, [\![\varphi_1 \rightarrow \varphi_2]\!]\, (c', u, h)$ and $(c, w, g)\, [\![\varphi_2 \rightarrow \varphi_1]\!](c', u, h)$;
  11. $(c, w, g)\, [\![\neg\varphi]\!](c', u, h)$ iff $c' = c$, $u = w$, $h = g$ and there is no (c'', v, i) such that $(c, w, g)\, [\![\varphi]\!](c'', v, i)$;
  12. $(c, w, g)\, [\![\exists x.\varphi]\!](c', u, h)$ iff for some $d \in D_{top(w)}$, $(c, w, g[a/x])\, [\![\varphi]\!](c', u, h)$;
  13. $(c, w, g)\, [\![\forall x.\varphi]\!](c', u, h)$ iff $c' = c$, $u = w$, $h = g$ and for all $d \in D$, there is (c'', v, i) such that $(c, w, g[d/x])\, [\![\varphi]\!](c'', v, i)$;
  14. $(c, w, g)[\![?\bar{x}.\varphi]\!](c', u, h)$ iff $u = \mathit{push}(v, w)$ and v is the possible world such that $\exists!\bar{a} \in D^n_w.(w, g[\bar{a}/\bar{x}]) \models \varphi \Rightarrow (v, g[\bar{a}/\bar{x}]) \models .\varphi$ & $\forall\psi \not\equiv \varphi.(v, g[\bar{a}/\bar{x}]) \not\models \psi$, and $h = g[\bar{a}/\bar{x}]$, where $(v, g[\bar{a}/\bar{x}]) \models \varphi$;
  15. $(c, w, g)[\![?^{\forall}\bar{x}.\varphi]\!](c', u, h)$ iff $u = \mathit{push}(v_f, w)$ where v_f is the parametrized possible world such that for each $\bar{a} \in D^n_w.(w, g[\bar{a}/\bar{x}]) \models \varphi \Rightarrow (v_{f(\bar{a})}, g[\bar{a}/\bar{x}]) \models .\varphi$ & $\forall\psi \not\equiv \varphi.(v_{f(\bar{a})}, g[\bar{a}/\bar{x}]) \not\models \psi$, and $h = g$; $\forall(c, v, g) \models \forall\bar{x}.\varphi$ and for other formula $\psi \neq \varphi$, $c, v, g \not\models \forall\bar{x}.\psi$;
  16. $(c, w, g)[\![?\bar{\xi}.\varphi]\!](c', u, h)$ iff there are Skolem functions $\bar{f}$ such that $(c, top(w), g[\bar{f}/\bar{\xi}] \models \varphi$ and $(c, w, g[\bar{f}/\bar{\xi}])\, [\![\varphi]\!](c', u, h)$;
  17. $(c, w, g)\, [\![¿\varphi]\!](c', u, h)$ iff for all v such that $(c, w, g)\, [\![\varphi]\!](c', v, h)$:
     (a) $u = \mathit{pop}(w)$;
     (b) $c'(\mathit{speaker}) \neq c(\mathit{speaker})$, $c'(\mathit{addressee}) \neq c(\mathit{addressee})$, and $c'(\mathit{speaker}) = c(\mathit{addressee})$;
     (c) if $X = \{w | w, g \models p, p \in Sub(\varphi) \cap \Phi\} \not\subseteq c(G)$, then $c'(G) = c(G) \cap X$, where $Sub(\varphi)$ is the set of subformulas of φ,

     where $\sigma \models \varphi$ iff $\sigma\, [\![\varphi]\!]\sigma$.

[6] This definition is the common knowledge by epistemic logic. See [10].

4 Applications to Some Dependent Questions

I will treat the significant cases, such as ∀-who/single answer, ∀-who/functional answer, ∀-exhaustive pair-list answer, who-who/single (pair list) answer (∀-who/single answer's invalidity), who-who/functional answer, and who-who/partial pair-list answer.

4.1 ∀-Who/Single Answers

(11a) is an example of the pair of ∀-who/single answer. (11b) is the logical form of (11a) in $DPL_{?¿}$.

(11) a. Who does every boy admire?; Agent Smith.
 b. $?\xi.\forall x.boy(x) \rightarrow person(\xi(x)) \wedge admire(x, \xi(x)) \wedge ¿\xi = Agent\_Smith \wedge \forall x.boy(x) \rightarrow person(\xi(x)) \wedge admire(x, \xi(x))$

The left conjunct of (11b) pushes the questioned state w where a Skolem function f assigns to each boy x a person who is admired by x. The questioned state w can be considered as the parametrized possible world $\langle w, f\rangle$, since for each boy $a \in D_w$, we can construct $w_{f(a)}$ where as for atomic proposition, only 'a admires b' holds.

The right conjunct of (11b) represents f as equivalent to $Agent\_Smith$, where $I(w)(Agent\_Smith)$ for any $w \in W_0$ satisfies the following condition:

(12) $\exists! d \in D_w.\forall x \in D_w.I(w)(Agent\_Smith)(x) = d$

I call the dependent answer which is denoted by a Skolem function symbol *proper-nominal* if it satisfies the condition similar with (12).

4.2 ∀-Who/Functional Answers

(13) a. Who does every boy admire?; His mother.
 b. $?\xi.\forall x.boy(x) \rightarrow person(\xi(x)) \wedge admire(x, \xi(x)) \wedge ¿\xi = mother \wedge \forall x.boy(x) \rightarrow person(\xi(x)) \wedge admire(x, \xi(x))$

In a similar way to (11b), the left conjunct of (14b) pushes the questioned state w where a Skolem function f assigns to each boy x a person who is admired by x, and $\langle w, f\rangle$ can be considered as a parametrized possible world. On the other hand, the right conjunct of (14a) represent 'His mother', since it can be considered as an abbreviated form of 'Every boy admires his mother'.

In this case, the Skolem function $I(w)(mother)$ for any $w \in W_0$ satisfies the following condition:

(14) $\forall x \in D_w.\exists! d \in D_w.I(w)(mother)(x) = d \;\&\; d \neq x$

I call the dependent answer which is denoted by a Skolem function symbol *functional* if it satisfies the condition similar with (14).

4.3 ∀-Who/Exhaustive Pair-List Answers

(15) a. Who does every boy admire?; Bill admires John, and Ken Sue.
b. $?\xi.\forall x.boy(x) \rightarrow person(\xi(x)) \wedge admire(x, \xi(x)) \wedge ¿\xi = f \wedge admire(Bill, John) \wedge ((\xi(Bill) = John \wedge \xi(Ken) = Sue) \equiv \forall x.boy(x) \rightarrow person(\xi(x)) \wedge admire(x, \xi(x)))$

Similarly with (11b), the left conjunct of (14b) pushes the questioned state w where a Skolem function f is assign to each boy x a person who is admired by x, and $\langle w, f \rangle$ can be considered as a parametrized possible world. On the other hand, the right conjunct of (15b) is the logical form of the answer in (15a). The logical form of exhaustive pair-list answer of (15a) can be (16).

(16) $admire(Bill, John) \wedge admire(Ken, Sue)$

But (16) does not imply that the answer denotes an exhaustive listing of the answers. Therefore, rather, I translate the exhaustive list answer in (15a) is (15b).

The exhaustiveness of the question is formalized in this way: For any contexts c, c', possible world stacks w, u, and variable assignments g, h, if

$$(c, w, g)\, [\![?\xi.\forall x.boy(x) \rightarrow person(\xi(x)) \wedge admire(x, \xi(x))]\!] (c', u, h),$$

then $u, h \models \exists\xi\forall x.boy(x) \rightarrow person(\xi(x)) \wedge admire(x, \xi(x))$. This means that for all boy $a \in D_{top(u)}$, there exists a person $b \in D_{top(u)}$ such that $\langle a, b \rangle \in I(top(u))(admire)$. The value of ξ, i.e., f assures this condition, since it is a Skolem function. Therefore, the exhaustiveness of the question is held in the questioned state.

On the other hand, the exhaustiveness of the answer is assured by the right condition of $\equiv$, i.e., $\forall x.boy(x) \rightarrow person(\xi(x)) \wedge admire(x, \xi(x)))$.

Therefore, I call a Skolem function *exhaustive listing* with respect to predicate $\pi(x)$ if for some formula $\psi(x)$, $\forall x.\pi(x) \rightarrow \psi(x, f(x))$.

4.4 Who-Who/Functional Answers

(17) a. Who admires who?; His mother.
b. $?\xi?^{\forall}x.person(x) \rightarrow person(\xi(x)) \wedge admires(x, \xi(x)) \wedge ¿\xi = mother \wedge \forall x.person(x) \rightarrow person(\xi(x)) \wedge admire(x, \xi(x))$

(17a) is a who-who/functional answer pair, and (17b) is its logical form. As we have seen in section 3.1, 'Who admires who?' in (17a) is translated to $?\xi?^{\forall}x.person(x) \rightarrow person(\xi(x)) \wedge admires(x, \xi(x))$, and as we have seen in section 4.2, 'His mother' in (17a) is translated to $\xi = mother \wedge \forall x.person(x) \rightarrow person(\xi(x)) \wedge admire(x, \xi(x))$.

The left conjunct of (17b) pushes the parametrized questioned state, say, $\langle w, f \rangle$ and assigns a Skolem function f' to ξ. Therefore, for all $a \in D_w$, $w_{f(a)}$, where a only admires $f'(a)$, is defined. The equation $f' = mother$ is represented by the right conjunct of (17b).

4.5 Who-Who/Partial Pair-List Answers

(18) a. Who admires who?; Bill admires John and Ken Sue.
b. $?\xi?^{\forall}x.person(x) \rightarrow thing(\xi(x)) \wedge admire(x, \xi(x)) \wedge ¿\xi = f \wedge admire(Bill, John) \wedge ((\forall x.person(x) \rightarrow person(\xi(x)) \wedge admire(x, \xi(x))) \rightarrow (\xi(Bill) = John \wedge \xi(Ken) = Sue))$

(18a) is a who-what/partial pair-list answer pair, and (18b) is its logical form. Why does $?\xi?^{\forall}x.person(x) \rightarrow thing(\xi(x)) \wedge admire(x, \xi(x))$ not constrain the exhaustivity? The formula pushes the parametrized questioned state, say, $\langle w, f \rangle$ and assigns a Skolem function f' to ξ. If f is a partial function, for some $a \in D_w$, $w_{f(a)}$ is undefined. I call this type of parametrized possible world a *partial parametrized possible world*.

Therefore, ∀-who dependent questions do not allow partial pair list answers, whereas who-who dependent questions allow partial pair list answers.

4.6 Who-Who/Single (Pair List) Answers

(19) a. Who admires who?; Bill admires John.
b. $?\xi?^{\forall}x.person(x) \rightarrow thing(\xi(x)) \wedge admire(x, \xi(x)) \wedge ¿\xi = f \wedge admire(Bill, John) \wedge ((\forall x.person(x) \rightarrow person(\xi(x)) \wedge admire(x, \xi(x))) \rightarrow (\xi(Bill) = John))$

Obviously, this is one case of who-who/partial pair list answers. This is the reason that ∀-who dependent questions do not allow single answers.

5 Conclusion

As we have seen, I have classified *dependent questions*, proposed $DPL_{?¿}$ by expanding DPL by introducing $?\bar{x}$-operator, $?^{\forall}\bar{x}$-operator, $?\bar{\xi}$-operator, ¿-operator, and a new model of question based on possible world stacks, parametrized possible worlds, and Skolem functions. As case studies, I have treated significant cases of dependent questions and their answers in $DPL_{?¿}$: ∀-who/single answer, ∀-who/functional answer, ∀-who/exhaustive pair list answer, who-who/functional answer, who-who/partial pair list answer, and who-who/single (pair list) answer. This is a new semantics, since a question is interpreted as an update function from a context to the questioned context, whereas an answer is interpreted as an update function from a questioned context to the context before the question. This is quite different from the Hamblin-Karttunen semantics or the Groenendijk-Stokhof semantics, and Groenendijk's update semantics of questions.

References

1. Maria Aloni. Questions under cover. In *Words, Proofs, and Diagrams*, pages 167–190. CSLI Publications, Stanford, 2002.
2. Maria Aloni and Robert van Rooy. The dynamics of questions and answers. In B. Jackson, editor, *Proceedings from SALT XII*, 2002.

3. Nicholas Asher and Alex Lascarides. Questions in dialogue. *Linguistics and Philosophy*, 21(3):273–309, 1998.
4. Nuel D. Belnap. Questions and answers in Montague Grammar. In S. Peters and E. Saarinen, editors, *Processes, Beliefs, and Questions*. D. Reidel Publishing Company, Dordrecht, 1982.
5. Nuel D. Belnap and Thomas B. Steel. *The Logic of Questions and Answers*. Yale University Press, New Haven, 1976.
6. Gennaro Chierchia. Functional wh and weak crossover. In *Proceedings of WCCFL 10*, pages 75–90, Stanford, 1991. CSLI.
7. Gennaro Chierchia. Questions with quantifiers. *Natural Language Semantics*, 1:181–234, 1993.
8. Ileana Comorovski. *Interrogative Phrases and the Syntax-Semantics Interface*. Kluwer Academic Publishers, Dordrecht, 1996.
9. Elisabet Engdahl. *Constituent Questions: The Syntax and Semantics of Questions with Special Reference to Swedish*. D. Reidel Publishing Company, Dordrecht, 1986.
10. Ronald Fagin, Joseph Y. Halpern, Yoram Moses, and V. Y. Vardi. *Reasoning about Knowledge*. The MIT Press, Cambridge, 1995.
11. Jelle Gerbrandy. Logic, language and computation. In Lawrence Cavedon, Patrick Blackburn, Nick Braisby, and Atsushi Shimojima, editors, *Logic, Language and Computation: vol.3*, pages 147–159. CSLI Publications, Stanford, 2000.
12. Jonathan Ginzburg. Resolving questions: Part I. *Linguistics and Philosophy*, 18(5):459–527, 1995.
13. Jonathan Ginzburg and Ivan A. Sag. *Interrogative Investigations: The Form, Meaning, and Use of English Interrogatives*. CSLI, Stanford, 2000.
14. Jeroen Groenendijk. The logic of interrogation. In Tanya Matthews and Devon Strolovitch, editors, *Proceedings from SALT IX, U.C. Santa Cruz*, pages 109–126, 1999.
15. Jeroen Groenendijk and Martin Stokhof. Type-shifting rules and the semantics of interrogatives. In Gennaro Chierchia, Barbara H. Partee, and Raymond Turner, editors, *Properties, Types, and Meanings*, volume 2. Kluwer Academic Publishers, Dordrecht, 1989.
16. Jeroen Groenendijk and Martin Stokhof. Dynamic predicate logic. *Linguistics and Philosophy*, 14:39–100, 1991.
17. Jeroen Groenendijk and Martin Stokhof. Questions. In Johan van Benthem and Alice ter Meulen, editors, *Handbook of Logic and Language*, pages 1057–1124. Elsevier Science B.V., Amsterdam, 1997.
18. C. L. Hamblin. Questions in Montague English. *Foundations of Language*, 10:41–53, 1973.
19. James Higginbotham. The semantics of questions. In Shalom Lappin, editor, *The Handbook of Contemporary Semantic Theory*, pages 361–383. Blackwell Publishers, Oxford, 1996.
20. James Higginbotham and Robert May. Questions, quantifiers and crossing. *Linguistic Review*, 1:41–79, 1981.
21. Nirit Kadmon. *Formal Pragmatics: Semantics, Pragmatics, Presupposition, and Focs*. Blackwell Publishers, Malden, 2001.
22. Lauri Karttunen. Syntax and semantics of questions. *Linguistics and Philosophy*, 1:3–44, 1977.
23. Lauri Karttunen and Stanley Peters. Interrogative quantifiers. In Christian Rohrer, editor, *Time, Tense, and Quantifiers*, pages 181–205. Niemeyers, Tübingen, 1980.
24. Susumu Kuno and Jane J. Robinson. Multiple *wh* questions. *Linguistic Inquiry*, 3:463–487, 1972.

25. Norihiro Ogata. What do you mean by "what do you mean'?: A formal representation and its dynamic semantics of meta-expressions in conversations. In Petr Sojka, Ivan Kopecek, and Karel Pala, editors, *Lecture Notes in Artificial Intelligence : Proceedings of Third International Workshop TSD2000, Brno, Czech Republic, September 13-16*, pages 393–398. Springer Verlag, Berlin, 2000.
26. Javier Gutiérrez Rexach. Questions and generalized quantifiers. In Anna Szabolsci, editor, *Ways of Scope Taking*, pages 409–452. Kluwer Academic Publishers, Dordrecht, 1997.
27. Anna Szabolsci. Quantifiers in pair-list readings. In Anna Szabolsci, editor, *Ways of Scope Taking*, pages 311–347. Kluwer Academic Publishers, Dordrecht, 1997.
28. Robert van Rooy. Exhaustivity and specificity: A parallelism between answers and pronouns. In Myriam Bras and Laure Vieu, editors, *Semantic and Pragmatic Issues in Discourse and Dialogue: Experimenting with Current Dynamic Theory*, pages 29–56. Elsevier, Amsterdam, 2001.
29. Robert van Rooy. Questioning to resolve dicision problems. *Linguistics and Philosophy*, 26:727–763, 2003.
30. Keiko Yoshida. Syntax and semantics of wh-quantifier interactions, 1993. Manuscript.

On Dependency and Quantification in Dynamic Semantics

Rick Nouwen*

School of Informatics, the University of Edinburgh, 2 Buccleuch Place, Edinburgh EH8 9LW, United Kingdom
rnouwen@inf.ed.ac.uk

Abstract. This paper considers how the interaction of quantification and dependent anaphora may be analysed in a dynamic semantics. It discusses a simple theory of the creation and accessibility of dependencies based on a dynamic semantics for distributivity and some basic assumptions on number agreement in discourse. This theory forms a partial defence of the line of semantics set out by van den Berg 1996. I argue, however, that it is essential to quantified anaphora to contextualise the labelling of antecedents.

1 Quantification, Accessibility and Dynamic Semantics

One of the main goals of the programme of *dynamic semantics* is to explain constraints on anaphora in model-theoretic terms. Under the slogan 'meaning is context-change-potential' the classical truth-conditional meaning of a form and this form's ability to change the context by introducing antecedents for future (pronominal) anaphora in the discourse are reconciled in a single framework. With respect to quantified statements, however, it is not at all straightforward how to model the dynamic semantics. This is due to the fact that plural pronouns with quantified antecedents do not in any trivial sense co-refer with their antecedent, nor are they bound by it, as was pointed out by Evans' famous example (1).

(1) Few senators admire Kennedy; and they are very junior. [Evans, 1980]
$\neq$ Few senators admire Kennedy and are very junior.
$\neq$ Few senators admire Kennedy and few senators are very junior.

Apart from the kind of anaphora exemplified in (1) quantification may furthermore give rise to dependent anaphora. In (2), the singular pronoun 'it' co-varies with the quantification over the subject pronoun 'they'.

(2) Every student wrote a paper and they each sent it to L&P.

* The work which lead to this paper was funded by a grant from the Niels Stensen Stichting, which I gratefully acknowledge. I furthermore wish to thank the workshop audience for interesting comments and discussion.

A. Sakurai et al. (Eds.): JSAI 2003/2004, LNAI 3609, pp. 383–393, 2007.

That is, the second sentence can only be understood to mean that each student sent the paper he or she wrote to L&P. The *dependency* between students and papers is preserved and accessed in discourse.

The discourse representation theory (DRT) of [Kamp and Reyle, 1993] gives an explicit model of the interaction between quantification and anaphora. Surprisingly, the analysis of this interaction relies on the assumption that quantifiers like 'every child' have no dynamics at all – they are themselves not (directly) responsible for making any antecedents for future anaphora available. To account for examples like (1) and (2), DRT stipulates a set of inference principles, which define in what way representations of a sentence may be manipulated to include representations of antecedents.

The inference principles posited by DRT go counter the ideas of dynamic semantics in two interrelated ways: (i) they necessarily involve a level of representation and (ii) they are independent of the semantics of quantificational expressions. One may clarify these shortcomings in terms of the notion of antecedent accessibility. Whether or not an antecedent is accessible in DRT depends not only on the result of interpreting some form, but moreover on the application of inference principles defined on representational structures.

DRT's analysis of the data is in sharp contrast to important assumptions behind *dynamic predicate logic* (DPL, [Groenendijk and Stokhof, 1991]). In a dynamic semantics based on DPL, the notion of the accessibility of an antecedent can be said to be completely semantic, obscuring the need for any representational level. Predicate logical forms in DPL are interpreted as relations between assignment functions. The values of the variables specified in these functions are the (potential values of) accessible antecedents. Resolved pronouns correspond to free variables, taking their value from a contextual assignment. Accessibility is thus reduced to a non-representational notion which is solely governed by semantics.

DRT's inference principles are incompatible with both DPL's lack of a representational level and its purely semantic notion of accessibility. Within DPL-style dynamic semantic theories, it has therefore been argued that examples like (1) and (2) ought to be accounted for without stipulating principles on top of the anaphora mechanisms that are directly available from the semantics (see especially [van den Berg, 1996] and [Krifka, 1996] but also, more recently, [Wang et al., 2003].)[1]

It is evident that the complexity of the semantic analysis will increase when turning to quantificational sentences. The example in (2), for one, shows that not only antecedents, but also relations between antecedents are stored in context. This undeniably calls for some sort of structured form of context (cf. [Krifka, 1996, van den Berg, 1996, Nouwen, 2003]). In this paper, I will review some design choices of dynamic analyses of quantification and argue for a semantics which is minimal in the sense that only one type of operation is responsible for introducing and accessing dependencies, roughly along the line of

[1] A different alternative is presented in [Ogata, 2002], where an abstraction procedure is made an explicit part of the (dynamic) semantics of quantificational sentences.

[van den Berg, 1996]. The van den Bergian notion of context as a set of partial assignment functions offers a direct and simple explanation of the distribution of dependent pronouns. However, I will argue that the interaction of anaphora with quantification (and especially with distributive predication) calls for a careful handling of variable names. I therefore present an incremental dynamic analysis of distributivity and dependency.

The structure of the paper is as follows. In section 2, I will introduce some assumptions on how number interacts with anaphora and review the accounts of dependency of [Asher and Wang, 2003] and [van den Berg, 1996]. Then, in section 3, an alternative to van den Berg's semantics will be presented within the framework of incremental dynamics [van Eijck, 2001]. Section 4 presents a application of the proposal to the problematic phenomenon called telescoping.

2 Dependency and Distributivity

In what follows I will make the following basic assumption concerning the role of number agreement in the process of pronominal anaphora resolution in discourse.

(3) a. Singular pronouns take antecedents which are both syntactically and semantically singular.
    b. Plural pronouns take antecedents which are syntactically or semantically plural (or both).

Whereas singular pronouns are strict about the number restrictions they impose on their antecedent, requiring both semantic and syntactic agreement,[2] plural pronouns may take all remaining kinds of antecedents.[3] The examples in (4) support this assumption.

(4) a. Most boys think they are/*he is wise.
    b. Most boys wrote a paper. They weren't/*It wasn't very well-written.
    c. Most boys wrote a paper. They/*he read a book as well.
    d. Every boy thinks *they are/he is wise.

[2] Note that a numeral like 'one book' allows for singular anaphora even though it is commonly accepted such numerals allow for an 'at least one' reading. In the semantics, the value of the variable will always be a singular one. In a form p like $\exists x(\text{book}(x) \wedge |x| = 1 \wedge P(x))$, the variable x ranges over groups containing just a single entity. The fact that p does not exclude the existence of larger groups of books that also have property P is irrelevant.

[3] In fact, it seems as if plural pronouns might actually be more loosely constrained then exposed here. One example is generally referred to as *gender bias*, where the gender-neutral plural pronoun is preferred over use of a singular pronoun which would enforce an explicit choice of a gender feature. (For instance, '*Every passenger is responsible for their luggage.*') In general, plural pronouns may be thought of as severely under-constrained with respect to the antecedent they take. What is important for what follows is the strict requirement of both semantic and syntactic agreement for singular pronouns. We can safely ignore extra-ordinary cases of plural pronoun use.

Since DPL has no way of representing pronouns – it is capable only of representing *resolved* utterances – the goal will be to devise a semantics which covers the attested accessibility patterns bearing in mind the above assumptions on agreement. So, '$\forall x(\varphi)(\psi)$' should not output a state in which x (necessarily) yields a singular value, since the corresponding 'Every N VP' does not allow for subsequent singular pronouns to pick up the antecedent introduced by the subject quantifier. (But see the discussion in section 4).

We will also use the agreement assumptions to make generalisations concerning dependent readings. Consider the following contrasts:

(5) Three students each wrote a paper.

a. #Together, they sent it to L&P.

b. They each sent it to L&P.

From (5-a) we may conclude that collectivity does not affect the plurality of the antecedent brought about by 'three students'. The continuation in (5-b), however, strongly suggests that distributive quantification makes accessible the atomic (singular) parts of the plurality in question. Apart from finding a semantics of (distributive) quantification which accounts for the introduction of plural antecedents, the goal is now to find a semantics of distributive quantification which entails the accessibility contrast in (5-a) and (5-b). The relationship between the semantic number of an antecedent and the mode of predication it occurs in is intuitively clear. Consider (6).

(6) Three students wrote a paper.

a. It wasn't very well-written.

b. They weren't very well-written.

Distributivity brings about a co-variation of the values for the object indefinite with the values in the restriction predicate. The result is a plurality of atomic values for the object indefinite each of which is related to an atomic part of the plurality of values brought about by the subject. In case (6) is understood collectively, no co-variation takes place. The result then is a single value for the indefinite object.

The mode of predication in a sentence is determined by a range of semantic and pragmatic factors. For instance, certain predicates may only take collections (e.g. 'be a good team'), while certain quantifiers enforce distributive quantification (hence '*Most students form a good team').

In their account of plurality in discourse, [Asher and Wang, 2003] propose that the mode of predication should be represented as a type of transition, where a collective transition passes on values in the form of collections and a distributive transition passes on values 'one at a time.' Moreover, a special kind of transition is used for the storage of dependencies in discourse. So, $\varphi;_{\text{dep}(x,y)}\,\psi$ expresses that respective atomic values for x and the corresponding depending values for y (as given by φ) are considered one at a time for the interpretation of ψ.

Asher and Wang's proposal is based on the argumentation that the two traditional lines of analysis of the distributivity/collectivity distinction cannot be maintained. The so-called term-ambiguity approach where the term predicated over is ambiguous is untenable since the same term may combine with a coordinated predication consisting of both a distributive and a collective part. Moreover, terms occurring in a distributive sentence may be picked up anaphorically and subsequently used collectively. If terms are really ambiguous, then such cases are hard to explain. The second approach Asher and Wang argue against, dubbed the predicate ambiguity approach, claims that it is the predication rather than the term which is ambiguous. This approach comprises analyses which make use of a distributivity operator (as in (one particular rendition of) [van den Berg, 1996] and the analysis to follow). Asher and Wang argue against such proposals on the basis of the persistence of modes of predication in discourse. In their own proposal, the mode of predication becomes a discourse-level interpretation mode. They argue that such an analysis is desired since distributivity, collectivity, dependency and cumulativity tend to persist in discourse. In a predicate ambiguity approach, the argument goes, this persistence is hard to explain. While I acknowledge the existence of a tendency towards persistence of modes of predication in discourse and appreciate the power of Asher and Wang's solution, I doubt whether it is actually an argument against the predicate ambiguity line of thinking. All that is wrong is that the persistence of a mode of predication does not follow from the semantics. That is, the predicate ambiguity approach needs an extra-semantic explanation about why distributive sentences seem to come in pairs. At the same time, however, Asher and Wang's approach needs an extra-semantic explanation of why discourse sometimes breaks with the persistence tendency. In fact, I believe there are reasons why, semantically, the persistence of a mode of predication is undesirable. We have seen from the examples in (6) that the mode of predication influences the way antecedents are introduced in discourse. For instance, dependencies may be seen as a side-effect of distributive predication. It has often been overlooked, however, that the mode of predication in which a pronoun is involved does not determine the pronoun's choice of antecedent. In (7), the pronoun 'they' is embedded in the explicitly distributive predication. Still, its antecedent is the collection John and Mary, not the varying atoms of that pair. (Compare with '*John and Mary both think they are being cheated on.*')

(7) John and Mary both think they make a great couple.

Cases comparable to (7) occur in discourse as well. Notice the contrast between (8-a) and (8-b).

(8) Three students each wrote a paper.
   a. They then each sent it to a (different) journal.
   b. They then each sent them to a (different) journal.

In both examples, both sentences are made explicitly distributive by using the distributor 'each'. The singular pronoun in (8-a) enforces the dependent reading.

The example in (8-b), however, has a reading were each of the three paper-writing students sent (copies of) the three written papers to a journal. (So, three packages containing the same three papers reach three (different) editors.)

This illustrates that the distributive reading of the sentence containing the pronoun does not enforce the pronoun to be interpreted in a dependent way. The kind of antecedent a pronoun in a distributive sentence takes is decided on by the resolution process. From a semantic perspective both collective and dependent antecedents are available. In Asher and Wang's proposal, however, the dependent transport of values in discourse (as caused by the 'dependent' first sentence) calls for a correction in the logical form once the resolution process decides to choose a collective antecedent for a pronoun, since the dependent transition renders the collective antecedent inaccessible.[4] In what follows, I intend to show that such a representation of anaphora at the level of logical form is unnecessary.

The contrast between (8-a) and (8-b) is not a counterexample against Asher and Wang's approach using modes of transition. But it does show that the choice between dependent readings and independent readings of pronouns is merely due to anaphora resolution. The 'mode' of the predication containing the pronoun or the antecedent is not a sufficient condition for whatever type of anaphora. This suggests that the fact that persistence does not follow directly from a predicate ambiguity approach might be a virtue rather than a vice.

In van den Berg's dynamic analysis of plurality [van den Berg, 1996], dependency is modelled in a rather straightforward way, yielding a simple but powerful formalism to describe the interaction of quantification and dependency. Let me abstract away from the details of van den Berg's analysis and present – what I feel – is the core of his proposal. Recall, first, that we concluded from (6) that distributivity is a *necessary* condition for both the storing and the accessing of a dependency, since without distributivity there is no source of the co-variation necessary for dependency effects. Dependency and the semantics of distributivity are therefore essentially linked in van den Berg's semantics.

Let X be a set of variables. An *information state* is a set of assignments defined on X. If F_X is such a state, then $F_X(x) = \{f(x)|f \in F_X\}$ (the collected values for x in F_X) and $F_X|_{x=d} = \{f \in F_X|f(x) = d\}$ (the substate of F_X where x's value is d.) In a state F_X, variable $y \in X$ is dependent on $x \in X$ if and only if:

$$\exists d, e \in F_X(x): \; F_X|_{x=d}(y) \neq F_X|_{x=e}(y)$$

Interpretation of predicate logical forms with respect to a state F_X may now proceed in either of two ways. The standard way results in collective interpretations across the board. $F_X[P(x)]G_Y$ is true if and only if $G_Y = F_X$ and

[4] This is most clear from examples where the antecedent is plural. Consider: "*Three students each wrote exactly two papers. They each sent them to L&P*". The second sentence is now truly ambiguous between a resolution for the pronoun to co-vary with the distributive quantification and one wherein the pronoun takes the collection of (six) written papers. In Asher and Wang's proposal the difference between these readings is not one between antecedent choice, but rather one between a combination of antecedent choice and a relating correction of logical form.

$F_X(x) \in I(P)$. What is considered here is the collection of values for x, $F_X(x)$. Dependency information (information at the level of substates) is not accessed. Using a distributivity operator δ_x the distributive reading may be obtained. $F_X[\delta_x(P(x))]G_Y$ is true if and only if for each atomic entity $d \in F_X(x)$ it holds that $F_X|_{x=d}[P(x)]G_Y|_{x=d}$.[5] That is, the predication $P(x)$ is evaluated step by step with respect to substates $F_X|_{x=d}$. In general for forms $\delta_x(\varphi)$ this has the following consequences. In case φ contains a variable dependent on x, then the values for this variable will co-vary with those of x during interpretation, since they are interpreted with respect to $F_X|_{x=d}$ (for some d) instead of F_X. Moreover, in case φ contains the introduction of a new discourse entity, then the values for that discourse entity will co-vary with those of x in the output state. In other words, accessed variables in φ are interpreted dependently and introduced variables in φ are stored dependently in van den Berg's proposal.

With respect to our discussion above, however, this does not suffice. While we saw that the storing of entities under distributivity inevitably leads to dependency, the accessing of antecedents within the scope of a distributivity operator could either co-vary with the running variable or not.

Given a state F_X with a dependency of $y \in X$ on $x \in X$, the independent reading of a pronoun indexed with y is simply the interpretation of y in the global state F_X. The dependent reading of such a pronoun is an interpretation with respect to a varying state $F_X|_{x=d}$, where d ranges over the atoms in $F_X(x)$. The problem is that both the dependent reading (interpretation with respect to a local perspective on context) and the independent reading (interpretation with respect to a global perspective on context) are associated with the same label (y, in this case). We observed, however, that the resolution process has a genuine choice between the two. So, *in context*, both perspectives should be present. Unfortunately, there is no straightforward way of combining F_X and $F_X|_{x=d}$ into one anaphoric resource pool, especially since the latter is a (most often proper) subset of the former. What is needed is a means of contextually determining with what label a certain anaphoric option is to be associated. The next section presents an account of distributivity and dependency in a formalism which fully contextualises the index of an antecedent.

3 Incremental Dynamics of Distributive Quantification

The framework of *incremental dynamics* (ID, [van Eijck, 2001, van Eijck, 2000]) differs from DPL-like formalisms in assuming that the introduction of a new discourse entity results in an incremented context, whereas in DPL it results in the random assignment of a value to a pre-determined variable. A *context* in ID is a stack. The action $\exists$ represents a push operation defined on stacks. The label of the newly introduced value depends on the size of the input stack. Performed on a stack of n entities, $\exists$ pushes a new entity to slot n (given that the first slot

[5] Unfortunately, this is a serious oversimplification, since it puts far too few restrictions on the output state G. An extra condition stating $F_X(x) = G_Y(x)$ should be added (cf. (9)). See [Nouwen, 2003] for discussion.

is labelled 0, the second 1, etc.) In ID, the labels of antecedents are not decided upon by the logical form, but rather by the context in which interpretation takes place. As a direct result, there is a straightforward way of combining contexts, namely the append operation, without the risk of a name clash. In the light of the discussion in the last section, this means that ID is a promising formalism for the simultaneous representation of global and local perspectives on context.

In [Nouwen, 2003, Nouwen, 2004], an incremental dynamic account of distributive quantification, dependency and plurality in discourse is presented. Let $s, s', \ldots$ range over stacks. Stacks are functions from $\{0, 1, 2, \ldots, n\}$ to the domain of entities D_e. The function $\{\langle 0, d_0\rangle, \langle 1, d_1\rangle, \langle 2, d_2\rangle, \ldots\}$ is written as $[d_0, d_1, d_2, \ldots]$. The size of a stack is the cardinality of its domain. We write $s^\wedge s'$ for appending stack s' to s, i.e. $s^\wedge s' = s \cup \{\langle i + |s|, d\rangle | s'(i) = d\}$. Let $S, S', \ldots$ range over sets of stacks of the same size. We write $S(i)$ for $\{s(i)|s \in S\}$ and $S|_{i=d}$ for $\{s \in S|s(i) = d\}$. Two sets of stacks may be combined as follows: $S^\sqcap S' = \{s^\wedge s'|s \in S \ \&\ s' \in S'\}$. The size of a set of stacks S, $|S|$, is the common size of the individual stacks.

The semantics is presented as a set of functions. Operators and predicates are represented as functions on indices and sets of stacks. This effectively does away with any meaningful representational level.

(9)
$$\begin{aligned}
\exists^\star &:= \lambda S.\lambda S'.\exists d \in D_e : S' = S^\sqcap\{[d]\}\\
P^\star &:= \lambda i.\lambda S.\lambda S'.S' = S \ \&\ M \models P(S(i))\\
\delta^\star_i &:= \lambda P.\lambda S.\lambda S'\, S(i) = S'(i)\ \&\\
&\qquad\qquad \forall d \in S(i) : S^\sqcap S'|_{i=d} \in P(|S| + i)(S^\sqcap S|_{i=d})\\
\varphi \cdot \psi &:= \lambda S. \cup \{\psi(S'')|S'' \in \varphi(S)\}
\end{aligned}$$

The functions in (9) have polymorphic types. '$\exists^\star$' is a state transition taking a set of stacks of some size n and returning sets of stacks of size n+1. The $\exists^\star$-quantifier is a push operator on sets of stacks. It increments the context with a single atomic individual. A predicate-function $P^\star$ takes an index and returns a test on states with respect to P and that index. The function $\delta^\star$ takes a predicate and an index to return a state transition. It quantifies over the set associated with the index. In the scope of a distributivity operator, the functional perspective on an input state ($S|_{i=d}$) and the global perspective (S) are combined. The set of accessible labels is thus doubled, but, since the label of an entity in a stack is determined by its position, there is no clash of labels. Outside the scope of $\delta^\star$, a global state S' remains. So, in a state S with some salient set of students in slot i, the following set of states represents the processing of a sentence 'each student wrote a paper.'

(10) $$(\delta^\star_i(\lambda n.\lambda T.(\exists^\star \cdot \text{paper}^\star(|T|) \cdot \text{write}^\star(n, |T|))(T)))(S)$$

States S' in the set described by (10) are such that $S'(i)$ returns the same set of salient students as $S(i)$ did. The plurality $S'(|S|)$ is a set of papers written by these students. Moreover, each substate $S'|_{i=d}$ for $d \in S(i)$ returns a single student d at i and a paper written by d at $|S|$. If we subsequently take states like S' as the input context for the processing of a sentence 'they each submitted PRO to L&P,' where PRO is some third person pronoun (say, either 'it' or 'them'), the resolution process has a choice between the following two sets of states (ignoring the choices for the subject pronoun).

(11) a. $(\delta_i^\star(\lambda n.\text{submit\_to\_L\&P}^\star(n, |S|)))(S')$
b. $(\delta_i^\star(\lambda n.\text{submit\_to\_L\&P}^\star(n, |S'| + |S|))(S')$

The set of states in (11-a) represents the reading where the set of papers written by students is accessed by the object pronoun (which will have to be plural in this case). The set in (11-b) collects output states for a reading where the pronoun is read dependently.

In sum, the incremental dynamics frameworks open up the possibility of describing dependent and independent anaphora using information states wherein the labels for antecedents are themselves contextualised.

4 Telescoping

There exists another phenomenon that crucially involves quantification and dependency. The infamous example (12) contrasts with the examples and the generalisation above in that it displays access to a dependency without the occurrence of a quantificational operator.

(12) Each chess set comes with a spare pawn. It is taped to the box.

This phenomenon is generally referred to as *telescoping* [Roberts, 1987]. With non-universal antecedent quantifiers this kind of reference is highly marked (in fact mostly infelicitous). Also, telescoping prefers the determiner 'each' over 'every' (see, for instance, [Wang et al., 2003]). Moreover, this kind of reference is only felicitous in case the resulting interpretation supports discourse coherence (see, for instance, [Roberts, 1987] and [Poesio and Zucchi, 1992]). I agree with [Wang et al., 2003], then, that a theory of telescoping should consist of two parts: (i) an explanation of the accessibility of singular reference to the running variable(s) of an antecedent quantificational sentence and (ii) an explanation of the marked nature of many such instances.

The goal of the present section is modest. I wish to show that van den Berg's δ-operator and its revised version, above, may help to offer a mechanism which explains the existence of examples like (12) and may offer a basis for a wider theory explaining the markedness of so many other examples.

The basic idea is the following. Say, T is a set of stacks in which slot j is dependent on i. Say, for instance, $T(i)$ is a set of chess sets and $T(j)$ a set of (corresponding) spare pawns. In other words, T is a state which could well be an output of the first sentence in (12). Say, now, that we want to communicate the following fact, presented as a test on information states:

(13) $\delta_i^\star(\ \text{taped\_to}^\star(|T| + j, \text{box\_of}^\star(i))\ )$

This tests whether each of the chess sets at i has its corresponding spare pawn taped to its box.[6] But to what in natural language will (13) correspond? One option is to choose a plural pronoun for i and a singular one for j. This obeys the two agreement rules we have stated in section 1. In T, i corresponds to a plurality and $|T| + j$ corresponds to an atomic value in the scope of $\delta_i^\star$. The values for i within the scope of $\delta_i^\star$, however, are singular (they are atomic chess sets). Given this, we might choose to use strict agreement with respect to the arguments of predication and use singular pronouns for both subject and object. The problem for the hearer with this choice is that she needs to assume that the sentence is distributive even though there is no plural subject. Inserting a distributivity operator when a plural subject is present is generally available, but inserting one otherwise will need to be backed up by some coherence-enhancing criteria.

The assumption that the insertion of a 'distributor' like $\delta^\star$ is free for syntactically plural predications readily explains the difference between (14-a) and (14-b) (from [Wang et al., 2003]):

(14) Each student in the syntax class was accused of cheating on the exam.
  a. # He had a Ph.D. in astrophysics.
  b. They had a Ph.D. in astrophysics.

Although incoherent, the plural pronoun in (14-b) is still understood as referring to the students in the syntax class. The incoherence of (14-a) fails to support the insertion of a distributivity operator in a syntactically singular predication.

Notice that the current proposal immediately explains why telescoping only occurs with singular determiners, since syntactically plural determiners block singular agreement.

In sum, a van den Bergian analysis of dependency in context allows for an explanation of the typical telescoping phenomena in terms of distributivity and agreement. Such a position would be right in-between the antecedent accommodation approach of Roberts [Roberts, 1987, Poesio and Zucchi, 1992] and the dynamic analysis of [Wang et al., 2003]. With the latter authors, I strongly believe quantification to be dynamic. The unmarked nature of plural anaphora in discourse makes any mechanism of ad-hoc antecedent creation unlikely.[7] With Roberts, however, I believe that telescoping is marked because it involves some operation which is not overtly present. Contra Roberts, however, I believe this operation to be the distributivity operator. The distributor may be freely inserted in plural predications. The argument slots in the resulting semantics, however, may be realised as singular pronouns, since they essentially involve atomic values.

---

[6] Since telescoping deals only with dependent readings, a van den Bergian analysis would work as well. The formula corresponding to (13) would be: $\delta_{x_i}(\ \text{taped\_to}(x_i, \text{box\_of}(x_j))\ )$.

[7] The account in [Ogata, 2003] presents a dynamic semantics for singular generics offering another dynamic alternative to Roberts-style antecedent accommodation.

5 Conclusions

DPL-style dynamic semantics offers tools for the modelling of the accessibility of antecedents for pronouns by interpreting natural language expressions as changes to anaphoric resource pools. Quantified noun phrases form a complication due to their non-straightforward referential behaviour. Quantified structures give rise to a number of accessibility patterns. Some of these patterns conflict with a simple-minded view of the way anaphoric resources are stored in context. I have defended the van den Bergian view on the representation of plurality in context. However, I have argued that a contextualised labelling of antecedents is called for and sketched how this may provide the basis for a minimalistic theory describing all cases of dependent and independent anaphora.

References

[Asher and Wang, 2003] Asher, N. and Wang, L. (2003). Ambiguity and Anaphora with Plurals in Discourse. In proceedings of: Semantics and Linguistic Theory 13 (SALT 13), University of Washington, Seattle, Washington.

[Evans, 1980] Evans, G. (1980). Pronouns. *Linguistic Inquiry*, 11(2):337–362.

[Groenendijk and Stokhof, 1991] Groenendijk, J. and Stokhof, M. (1991). Dynamic Predicate Logic. *Linguistics and Philosophy*, 14:39–100.

[Kamp and Reyle, 1993] Kamp, H. and Reyle, U. (1993). *From Discourse to Logic*. D. Reidel, Dordrecht.

[Krifka, 1996] Krifka, M. (1996). Parametrized sum individuals for plural reference and partitive quantification. *Linguistics and Philosophy*, 19:555–598.

[Nouwen, 2003] Nouwen, R. (2003). *Plural pronominal anaphora in context: dynamic aspects of quantification*. PhD thesis, UiL-OTS, Utrecht. LOT-series: No. 84, available from http://www.lot.nl/ or from author.

[Nouwen, 2004] Nouwen, R. (2004). On accessibility and plural pronouns. Ms. University of Edinburgh.

[Ogata, 2002] Ogata, N. (2002). Dynamic semantics of plurals DPL^*_Q. In *Electronic notes in theoretical computer science*, volume 67. Elsevier Science Publishers.

[Ogata, 2003] Ogata, N. (2003). Dynamic predicate logic with defeasible non-kind denoting singular generics. In *Workshop Proceedings: ICosS-4: Inference in computational semantics*, pages 157–170, Nancy, France. LORIA.

[Poesio and Zucchi, 1992] Poesio, M. and Zucchi, S. (1992). On telescoping. In *Proceedings of SALT 2*.

[Roberts, 1987] Roberts, C. (1987). *Modal Subordination, anaphora and distributivity*. PhD thesis, University of Massachussets, Amherst.

[van den Berg, 1996] van den Berg, M. (1996). *Some aspects of the internal structure of discourse: the dynamics of nominal anaphora*. PhD thesis, ILLC, Universiteit van Amsterdam.

[van Eijck, 2000] van Eijck, J. (2000). On the proper treatment of context. In *Proceedings of CLIN99*, Utrecht.

[van Eijck, 2001] van Eijck, J. (2001). Incremental dynamics. *Journal of Logic Language and Information*, 10(3):319–351.

[Wang et al., 2003] Wang, L., McCready, E., and Asher, N. (2003). Information Dependency in Quantificational Subordination. Paper presented at *Where Semantics Meets Pragmatics:*, Michigan.

Dynamic Interpretations and Interpretation Structures

Yasuo Nakayama

Graduate School of Human Sciences, Osaka University
1-2 Yamada-oka, Suita, Osaka, Japan
nakayama@hus.osaka-u.ac.jp

1 Introduction

Dynamic semantics has its roots in the research on anaphoric reference. It is well known that in general anaphoric relations are highly context sensitive and defeasible. However, most standard frameworks of dynamic semantics, such as Discourse Representation Theory (DRT) and Dynamic Predicate Logic, do not deal with problems of interpretation revision[1] (cf. [6]). Furthermore, these frameworks are strictly incremental in the sense that any old information is more entrenched than any new information.

Recently, several proposals have been made to overcome these shortcomings of Dynamic Semantics. Asher tried in [1] to combine DRT with belief revision. However, Asher kept the main features of DRT unchanged and only allowed limited revision of DRSs. At first, Asher reformulated DRSs as partial models and defined monotonic core logic of DRSs. Then, he completed this partial monotonic logic with a non-monotonic logic, where applications of the non-monotonic logic were strongly restricted. He called this method *local revision* and contrasted it to *global revision* in the AGM model that is a standard framework for formal treatment of belief revision (cf. [3]).

In this paper, I propose a flexible framework of dynamic semantics. I will show how to construct interpretations that can be revised. The Theory of Interpretation Structures (TIS) as proposed in this paper can be located not only in the tradition of dynamic semantics but also in the tradition of belief revision whose classical position is stated in [3]. TIS is a theory of global revision; one of its essential features lies in keeping revised information in the interpretation structure (IS). This makes cancellation of revision easily performable. In TIS, revision is realized by restructuring IS. This paper demonstrates TIS as a powerful framework for dynamic interpretations.

2 Natural Representation Language

NRL (Natural Representation Language) is a framework of dynamic semantics; it was proposed in [8]. In this section, NRL is briefly described by giving its axioms. NRL is a framework of extensional mereology (cf. Definition 1) combined with individuation by sortal predicates and with use of Skolem symbols (cf. Definition 2 and 3). Formulas of

[1] The theory of segmented DRSs (Discourse Representation Structures) proposed by Asher deals with structures of DRSs, but he uses the framework for describing discourse structures and not for analyzing interpretation revision (cf. [2]).

A. Sakurai et al. (Eds.): JSAI 2003/2004, LNAI 3609, pp. 394–404, 2007.

NRL are called *Discourse formulas* or *D-formulas*. NRL uses '{' and '}' as parentheses and comma ',' as conjunction.

Definition 1. *Extensional Mereology* (EM) consists of the following axioms, axiom schemas and definitions. Let $\{\psi \equiv \chi\}$ be an abbreviation of $\{\{\psi \rightarrow \chi\}, \{\chi \rightarrow \psi\}\}$.

MA1. Axioms for lattice theory.
MA2. Additional axioms for Boolean algebra.
MD1. $x \subset y \equiv \{x \cap y = x\}$.
MD2. $x \subsetneq y \equiv \{x \subset y, x \neq y\}$.
MD3. $x \subset_p y \equiv \{x \subset y, x \neq \emptyset\}$.
MA3. $\exists u\, \psi(u) \rightarrow \exists x\{\psi(x), \forall u\{\psi(u) \rightarrow u \subset_p x\}\}$.
MD4. $\{x = max(u)[\psi(u)]\} \equiv \{\{\psi(x), \forall u\{\psi(u) \rightarrow u \subset_p x\}\} \vee \{\forall u \neg\psi(u), x = \emptyset\}\}$.
MA4. For all Skolem function symbol $d_k : d_k(x \cup y) = d_k(x) \cup d_k(y)$.

$\subset_p$ in Definition 1 is *part-of* relation, which is the essential notion of mereology.

Definition 2. *Extensional Mereology with Sortal Individuation*(EMSI) is an axiom system with EM and the following axioms and definitions[2]. Predicate F that satisfies this axiom system is called "sortal predicate".

SA1. $\neg F(\emptyset)$.
SA2. $\{F(x), F(y)\} \rightarrow F(x \cup y)$.
SA3. $\{F(x \cup y), F(x), x \cap y = \emptyset\} \rightarrow F(y)$.
SD1. $x \subset_F y \equiv \{F(x), F(y), x \subset y\}$.
SD2. $atom_F(x) \equiv \{F(x), \forall u\{u \subset_F x \rightarrow u = x\}\}$.
SD3. $x \varepsilon_F y \equiv \{atom_F(x), x \subset_F y\}$.
SA4. $F(x) \rightarrow \{\forall u\{u \varepsilon_F x \equiv u \varepsilon_F y\} \rightarrow x = y\}$.
SA5. $F(x) \rightarrow \exists u\{u \varepsilon_F x\}$.
SD4. $x \subsetneq_F y \equiv \{F(x), F(y), x \subsetneq y\}$.
SD5. $\{x = sum_F(u)[\psi(u)]\} \equiv \forall u\{u \varepsilon_F x \equiv \{\psi(u), atom_F(u)\}\}$.
SD6. $\{x = max_F(u)[\psi(u)]\}\} \equiv \{\{\exists y\{F(y), \psi(y), \forall u\{\{F(u), \psi(u)\} \rightarrow u \subset_F y\}\} \rightarrow$
$\forall u\{\{F(u), \psi(u)\} \equiv u \subset_F x\}\}, \{\neg\exists y\{F(y), \psi(y), \forall u\{\{F(u), \psi(u)\} \rightarrow u \subset_F y\}\} \rightarrow x = \emptyset\}\}$.
SD7. $collective(x, \psi(u)) \equiv \{\psi(x), \forall u\{u \subsetneq x, u \neq \emptyset\} \rightarrow \neg\psi(u)\}\}$.
SD8. $collective_F(x, \psi(u)) \equiv \{\psi(x), \forall u\{u \subsetneq_F x \rightarrow \neg\psi(u)\}\}$.
SD9. $distributive(x, \psi(u)) \equiv \forall u\{u \subset_p x \rightarrow \psi(u)\}$.
SD10. $distributive_F(x, \psi(u)) \equiv \forall u\{u \subset_F x \rightarrow \psi(u)\}$.
SD11. When G is a unary predicate symbol, $non_F[G](x) \equiv \forall u\{u \varepsilon_F x \rightarrow \neg G(u)\}$.
SD12. A function symbol α of type $\langle thing, thing\rangle$ that satisfies the following condition is called "adjective": $\forall x\{\{F(x), F(\alpha(x))\} \rightarrow \{\alpha(x) \subset_F x\}\} \vee \{\exists x\{\alpha(x) \subset_p x\}, \forall x\{\alpha(x) \subset x\}\}$.

F in Definition 2 is intended as a sortal predicate that can be used for individuation of structured objects. For example, "human" and "animal" are sortal predicates. $\subset_F$ corresponds to IS-A relation with respect to F and ε_F expresses INSTANCE-OF relation with respect to F.

[2] EMSI is slightly changed from [8].

Definition 3. NRL is a theory of two-sorted logic with the following axioms and axiom schemas.
LA0. Axiom system EMSI.
LA1. Standard axioms for + .
LD1. $\{cd_F(x) = 1\} \equiv atom_F(x)$.
LA2. $\{x \cap y = \emptyset, cd_F(y) = 1\} \rightarrow \{cd_F(x) = n \equiv cd_F(x \cup y) = n + 1\}$.

cd_F means the *cardinality* of x with respect to F, i.e. the number of F-objects in x. In Definition 3, this notion is recursively defined.

The semantics of NRL is defined as follows:

Definition 4. Let $M = \langle\langle \mathbf{U}, \mathbf{N}\rangle, V\rangle$. Let $\mathbf{N}$ be the set of natural numbers and K be a D-formula.

1. M^* is a *Skolem expansion* of M with respect to K *iff*
   $[M^* = \langle\langle \mathbf{U}, \mathbf{N}\rangle, V^*\rangle]$ & $[V \subseteq V^*]$ &
   [For all Skolem constant symbols d_k, $V^*(d_k) \in \mathbf{U}$] &
   [For all n-ary Skolem function symbols d_k, $V^*(d_k)$ is a function from $\mathbf{U}^n$ into $\mathbf{U}$].
2. K is *true* with respect to M, β *iff*
   $\exists M^*([M^*$ is a Skolem expansion of M with respect to $K]$ &
   $[K$ is true with respect to $M^*, \beta])$
3. K is *true* with respect to M *iff*
   K is true with respect to M, β for all assignments β .

NRL is a framework of dynamic semantics that can be easily applied to represent the meaning of sentences with mass terms and sentences with plural terms. By using NRL, plural anaphora can be represented as simple as singular anaphora. This is a major advantage of NRL compared to DRT in [5].

3 Theory of Interpretation Structures

In this paper, the *Theory of Interpretation Structures* (TIS) is constructed in a similar way as the construction of the *Theory of Belief Structures* (TBS) that is a framework for belief revision (cf. [10]).

Definition 5

1. Let L be a language of NRL. Then, an *Interpretation Structure* ($IS = \langle \mathbf{S}, >, \sim\rangle$) is defined as follows:
   (a) $\mathbf{S}$ is a set of D-formulas in L.
   (b) $>$ is a partial ordering on $\mathbf{S}^2$.
   (c) $\sim$ is an equivalence relation on $\mathbf{S}^2$.
2. Given an interpretation structure IS. Then, the *intended content of IS*, denoted as $IC(IS)$, satisfies the following conditions:
   (a) $\forall X1, X2 \in \mathbf{S}(X1 > X2 \Rightarrow (X2 \in IC(IS) \Rightarrow X1 \in IC(IS))$.
   (b) $\forall X1, X2 \in \mathbf{S}(X1 \sim X2 \Rightarrow (X2 \in IC(IS) \Leftrightarrow X1 \in IC(IS))$.

IS and $IC(IS)$ can be seen as partial descriptions of interpreter's belief state. $K_1 > K_2$ means roughly that K_1 is *more firmly believed than* K_2. $K_1 \sim K_2$ means that K_1 is *as firmly believed as* K_2.

For change of IS and for interpretation of $IC(IS)$, we accept principles 1 - 3.

Interpretation Principles

1. For alteration of IS, there are the following three methods:
   (a) change of ordering,
   (b) addition of a new piece of information,
   (c) deletion of an old piece of information.
2. Principles for interpreters
   (a) An interpreter desires that his $IC(IS)$ is consistent. Therefore, he tries to avoid any contradiction when he finds one.
   (b) An interpreter desires that his $IC(IS)$ becomes richer in the long run.
   (c) Conservatism: The costs of interpretation revisions are high and therefore not desirable if not necessary.
3. Holistic principle for the intended content:
   NRL's interpretation is holistic, i.e. the total intended content is interpreted as a whole. This results from the interpretation method of Skolem symbols (cf. Definition 4).

By using the first group of principles, replacement of information can be defined by at first applying deletion (1c) to old information and then applying addition (1b) to new information. However, in TIS, replacement is rarely used. This is because the content replacement with respect to $IC(IS)$ can also be achieved by order change. Order change is preferable to replacement, because IS after order change still contains information that can be used in undoing this replacement in $IC(IS)$.

The idea behind TIS is similar to the approach of epistemic entrenchment proposed by Gärdenfors (cf. [3] Chap. 4). However, TIS is less formally characterized than the AGM model, so that more flexible treatment of problems becomes possible, while TIS has less formal results than the AGM model.

4 Dynamic Interpretations

4.1 Standard Interpretations

A standard interpretation of a discourse consists of the following two steps:

1. Constructing an interpretation structure (IS).
2. Identifying the intended content from IS ($IC(IS)$) by using interpretation principles in section 3.

In NRL, d_n and indexed pronouns, such as he_n, are used as Skolem symbols. In the following description, K_n describes the content of a given sentence and C_n describes a possible context for interpretation. In general, there are multiple possible context interpretations from which an appropriate one should be chosen.

Example 1. (Plural anaphora):
(1) Most farmers own a donkey. They are very cruel. They have a bad time.

K_1 : $\{d_1 = \text{FARMER}, d_2 = sum_{human}(u)[u\varepsilon_{human}d_1, d_3(u)\varepsilon_{animal}$ DONKEY, u owns $d_3(u)], Most_{human}(d_2, d_1)\}$,
where $Most_F(x, y) \equiv 2 \times cd_F(x) > cd_F(y)$, i.e. *more than half* of x are y.
K_2 : $\{cd_{human}(they_1) > 1, they_1$ are very cruel$\}$.
C_{2pn1} : $\{they_1 = d_1\}$.
C_{2pn2} : $\{they_1 = d_2\}$.
C_{2r1} : $\{distributive_{human}(u)[they_1, u$ is very cruel$]\}$.
C_{2r2} : $\{collective_{human}(u)[they_1, u$ are very cruel$]\}$.
K_3 : $\{cd_{animal}(they_2) > 1, they_2$ have a bad time$\}$.
C_{3pn1} : $\{they_2 = d_1\}$.
C_{3pn2} : $\{they_2 = d_2\}$.
C_{3pn3} : $\{they_2 = d_3(d_2)\}$.
C_{3r1} : $\{distributive_{animal}(u)[they_2, u$ has a bad time$]\}$.
C_{3r2} : $\{collective_{animal}(u)[they_2, u$ have a bad time$]\}$.

$d_3(d_2)$ means *the donkeys that at least one of the farmers own*. Axiom MA4 justifies this use of $d_3(d_2)$. The following tables and diagram show how an interpretation structure grows during the process of interpretation. At first, IS is determined. Then, $IC(IS)$ is calculated according to interpretation principles in section 3:

IS :	IS_1 :	$(\{K_1\}, \emptyset)$
	IS_2 :	$K_1 \sim K_2$ & $K_2 > C_{2pn2} > C_{2pn1}$ & $K_2 > C_{2r1} > C_{2r2}$
	IS_3 :	IS_2 & $(K_2 \sim K_3$ & $K_3 > C_{3pn3} > C_{3pn1}$ & $C_{3pn3} > C_{3pn2}$ & $K_3 > C_{3r1} > C_{3r2})$

IS_3:

K_1	
$\wr$ $> C_{2pn2}$	$> C_{2pn1}$
K_2	
$\wr$ $> C_{2r1}$	$> C_{2r2}$
$\wr$	$> C_{3pn1}$
$\wr$ $> C_{3pn3}$	$> C_{3pn2}$
K_3	
$> C_{3r1}$	$> C_{3r2}$

$IC(IS)$:	Stage 1	$IC(IS_1)$:	K_1
	Stage 2	$IC(IS_2)$:	$IC(IS_1) \cup (K_2 \cup C_{2pn2} \cup C_{2r1})$
	Stage 3	$IC(IS_3)$:	$IC(IS_2) \cup (K_3 \cup C_{3pn3} \cup C_{3r1})$

C_{2pn1} is incompatible with $K_1 \cup K_2 \cup C_{2pn2}$ and C_{2r2} is incompatible with $K_1 \cup K_2 \cup C_{2r1}$, so that they are omitted from $IC(IS_2)$. >From the same reason, C_{3pn1}, C_{3pn2}, and C_{3r2} are excluded from $IC(IS_3)$. The resulting interpretation of the discourse of (1) is given by $IC(IS_3)$, i.e., $K_1 \cup K_2 \cup C_{2pn2} \cup C_{2r1} \cup K_3 \cup C_{3pn3} \cup C_{3r1}$:

$\{d_1 = \text{FARMER}, d_2 = sum_{human}(u)[u\varepsilon_{human}d_1, d_3(u)\varepsilon_{animal}\ \text{DONKEY},$
$u \text{ owns } d_3(u)], Most_{human}(d_2, d_1)\}$
$\cup\{cd_{human}(they_1) > 1, they_1 \text{ are very cruel}\} \cup \{they_1 = d_2\}$
$\cup\{distributive_{human}(u)[they_1, u \text{ is very cruel}]\}$
$\cup\{cd_{animal}(they_2) > 1, they_2 \text{ have a bad time}\} \cup \{they_2 = d_3(d_2)\}$
$\cup\{distributive_{animal}(u)[they_2, u \text{ has a bad time}]\}.$

This interpretation corresponds to the following anaphora resolution:

Most farmers own a donkey. They, *the farmers who own a donkey*, are very cruel. They, *the donkeys that at least one of the farmers own*, have a bad time.

In TIS, not only the resulting interpretation but also interpretation structure is kept and used as a context for the ongoing interpretation process. This is a fundamental difference to systems proposed in [1], [11], and [12].

4.2 Revising Interpretation

Interpretation revision is carried out by restructuring IS. This can be achieved by changing the order of the old IS. After restructuring, the standard interpretation can be continued:

Step 1. Restructuring IS.
Step 2. Identifying the intended content from IS.

Example 2. (Revision):
(2) John has never read Russian novels. But Bill likes them.

$K_1 : \{d_1 = Russian(\text{NOVEL}), d_2 = John, d_3 = sum_{book}(u)[u\varepsilon_{book}d_1, d_1 \text{ has read } u], cd_{book}(d_3) = 0\}.$
$K_2 : \{d_4 = Bill, cd_{human}(them_1) > 1, d_4 \text{ likes } them_1\}.$
$C_A : \{them_1 = d_3\}.$
$C_B : \{them_1 = d_1\}.$

IS_1:

K_1

IS_2:

K_1	
$\wr$	
K_2	$> C_A > C_B$

Judgement of IS_2: $K_1 \cup K_2 \cup C_A$ has no model. Thus, IS_2 has to be restructured. Otherwise, anaphora resolution is impossible.

IS_{2R}:

K_1	
$\wr$	
$K_2 > C_B$	$> C_A$

The resulting interpretation of this discourse is given by $K_1 \cup K_2 \cup C_B$, which means: "John has never read Russian novels. But Bill likes *them* (= Russian novels)".

Kamp proposed a two-stage theory for the interpretation of presuppositions; these two stages consist of computation and judgment (cf. [4]). The process of revision described above is compatible with this two-stage theory. In order to compare TIS with DRT, we would like to consider examples (3a) and (3b).

Example 3a

(3a) Walter has a rabbit. His rabbit is white.

DRT interprets this example as follows (cf. [4] p. 231):

$$(\{s_0, w, y, t, s\},$$
$$\{n \subseteq s_0, t = n, t \subseteq s, Walter(w), rabbit(y), s_0 : have(w, y), s : white(y)\}).$$

In this DRS, s_0 and s are used as discourse referents for states. In TIS, the same example is interpreted as follows, where t_s represents the *speech time*:

$K_1 : \{d_1 = \text{Walter}, d_2 \varepsilon_{animal}\ \text{RABBIT}, s_0 : have(d_1, d_2), t_s \subset_p s_0\}.$
$C_1 : \{d_1 \varepsilon_{human} male(\text{HUMAN})\}.$
$K_2 : \{d_3 \varepsilon_{animal}\ \text{RABBIT}, s_1 : have(his_1, d_3), s_2 : white(d_3), t_s \subset_p s_1, t_s \subset_p s_2\}.$
$C_2 : \{his_1 = d_1, d_3 = d_2\}.$

$$IS_1 : \boxed{\begin{array}{l} K_1 > C_1 \\ \wr \\ K_2 > C_2 \end{array}}$$

$IC(IS_1) = K_1 \cup C_1 \cup K_2 \cup C_2.$
(Walter has a rabbit (d_2). *His* (= Walter) rabbit (d_2) is white.)

The content of $IC(IS_1)$ is essentially identical with the previous DRS. Their main difference lies in the redundancy of TIS formulation. However, this redundancy is not necessarily disadvantageous. To show this, let us consider the case that (3a) is continued as follows:

Example 3b

(3b) Walter has a rabbit. His rabbit is white. But Walter's is not white. It is brown.

In this example, it is likely "His" refers to a person who is not Walter. In such a case, it is quite difficult to accommodate the previous DRS to this change, because the change requires *ad hoc* addition of some discourse referents and corresponding changes in conditions of the DRS. In TIS, the change is accommodated by slightly restructuring *IS* and adding interpretation of new information, as follows:

$K_3 : \{d_4 = \text{Walter}, s_3 : have(d_4, d_5), s_4 : \neg white(d_5), t_s \subset_p s_3, t_s \subset_p s_4\}.$
$C_3 : \{d_4 = d_1, d_5 = d_2\}.$
$C_{2R} : \{his_1 \neq d_1\}.$ (an alternative interpretation of anaphora)
$K_4 : \{s_5 : brown(it_1)\}.$

$C_4 : \{it_1 = d_5, s_5 = s_4\}.$

$$IS_2: \begin{array}{|l|l} \hline K_1 \ > C_1 & \\ \ \wr & \\ K_2 > C_{2R} & > C_2 \\ \hline \ \wr & \\ K_3 \ > C_3 & \\ \ \wr & \\ K_4 \ > C_4 & \\ \hline \end{array}$$

$IC(IS_2) = K_1 \cup C_1 \cup K_2 \cup C_{2R} \cup K_3 \cup C_3 \cup K_4 \cup C_4.$
(Walter has a rabbit (d_2). *His* (≠ Walter) rabbit (d_3) is white. But Walter's (d_2) is not white. *It* (= d_2) is brown.)

This shows that accommodation of IS to interpretation change of anaphora resolution is straightforward. C_{2R}, which states *"His" does not refer to Walter*, is added to IS_1 and is preferred to the previous anaphora resolution described by C_2. Then, $IC(IS_2)$ is calculated from IS_2 according to the interpretation principles proposed in section 3.

4.3 Interpretation of Metonymy

TIS can be applied to various problems, such as interpretation of metonymy[3] and disambiguation of interpretation of expressions.

Example 4. (Presupposition and Metonymy):
(4) Plato is on the top shelf. It is bound in leather. He is a famous Greek philosopher.

$MB_0 : \{\psi(d_1)\}$, where $\psi(d_1)$ describes the presupposition about *the shelf*.
$K_1 : \{d_2 = Plato, d_1 \varepsilon_{furniture}$ SHELF, d_2 is on the top of $d_1\}$.
$K_{1R} : \{d_2 = Plato, d_1 \varepsilon_{furniture}$ SHELF, d_3 is on the top of $d_1\}$.
$C_{1R} : \{d_2$ wrote $d_3, d_3 \varepsilon_{thing}$BOOK$\}$.
$K_2 : \{it_1$ is bound in leather$\}$.
$C_{2A} : \{it_1 = d_1\}$.
$C_{2B} : \{it_1 = d_3\}$.
$K_3 : \{he_1 \varepsilon_{human} famous(Greek($PHILOSOPHER$))\}$.
$C_3 : \{he_1 = d_2\}$.

$$IS_3: \begin{array}{|c|l} \hline MB_0 > K_{1R} > C_{1R} & > K_1 \\ \hline \wr & \\ K_2 \quad > C_{2B} & > C_{2A} \\ \hline \wr & \\ K_3 \quad > C_3 & \\ \hline \end{array}$$

The resulting interpretation of this discourse is given by

$MB_0 \cup K_{1R} \cup C_{1R} \cup K_2 \cup C_{2B} \cup K_3 \cup C_3.$
(*Plato wrote a book* d_3. d_3 is on the top shelf. *It* (= d_3) is bound in leather. *He* (= Plato) is a famous Greek philosopher.)

[3] For interpretation of metonymy, see [9].

In the case of interpretation of metonymy, flexibility of context interpretation plays an essential role. The standard interpretation

$IC(IS_1) := MB_0 \cup K_1$
(Plato is on the top shelf.)

is rejected, because a human cannot be on the top shelf and the discourse therefore has no model. So, in the alternative interpretation, a new object d_3 is introduced into context C_{1R} and is related to Plato by an ordinary relation "writing". In general, an interpretation of a term D as metonymy introduces a Skolem symbol d_k with $R(D, d_k)$, where R expresses an ordinary relation.

This example also shows how to integrate presuppositions and contents of mutual beliefs into IS.

4.4 Disambiguation

From our last discussions, it must be obvious how to disambiguate the meaning of an expression. Suppose predicate F has two meanings, namely F_A and F_B.

$K_1 : \{\psi(d_1), F(d_1)\}$, where $F := \lambda x(F_A(x) \vee F_B(x))$.
$C_A : \{F_A(d_1)\}$.
$C_B : \{F_B(d_1)\}$.

In this case, disambiguation can be achieved by the following restructuring of IS:

Stage 1:

K_1	$> C_A$ $> C_B$

Stage 2: $K_1 \cup C_A$ is impossible. Thus, according to interpretation principle (2b), C_B should be preferred to C_A, in order to make $IC(IS)$ as rich as possible.

Stage 3:

$K_1 > C_B$	$> C_A$

When the meaning of F_A is inappropriate, its order is diminished and F_B becomes preferred. As a result, F_B is integrated in $IC(IS)$ (cf. Stage 3).

Generally, anaphoric resolution can be seen as a process of disambiguation (cf. [11], [12]), if there are several interpretation possibilities for an anaphoric relation. Cancellation of a disambiguation can be easily realized within TIS, while this is difficult for Reyle's approach, because his system removes information for alternative interpretations after disambiguation.

In some cases, *context identification* is meant by *disambiguation*. For example, many problems of *temporal underspecification* discussed in [12] can be interpreted as problems of *context identification* in the sense of this paper. Let us consider the following example (cf. [12] p. 264):

Example 5
(5a) John made a short trip to London. He visited the British Museum.
(5b) John made a short trip to London. He visited the Louvre.

The interpretation of (5a) within TIS is straightforward, because the standard strategy for identification of temporal contexts yields a consistent result:

K_{a1} : $\{d_1 = \text{John}, d_2 = \text{London}, s_1\text{: } \textit{make-a-short-trip}(d_1, d_2), s_1 < t_s\}$.
K_{a2} : $\{d_3 = \text{British Museum}, e_1 : \textit{visit}(he_1, d_3), e_1 < t_s\}$.
C_{a21} : $\{he_1 = d_1\}$.
C_{a22} : $\{e_1 \subset_p s_1\}$.

IS_{a2}:	
	K_{a1}
	$\wr \quad > C_{a21}$
	K_{a2}
	$\quad > C_{a22}$

$IC(IS_{a2}) = K_{a1} \cup K_{a2} \cup C_{a21} \cup C_{a22}$.
(John made a short trip to London. *He* (= John) visited the British Museum.)

The sentence (5b) seems unintelligible, if we interpret "He" as referring to John. In interpretation of (5a), we assumed that John's visit to the British Museum took place during his stay in London, but this kind of interpretation is impossible for (5b), because the Louvre is located totally outside of London. Thus, we need to revise this line of interpretation:

K_{b2} : $\{d_4 = \text{Louvre}, e_1 : \textit{visit}(he_2, d_4), e_2 < t_s\}$.
C_{b21} : $\{he_2 = d_1\}$.
C_{b22} : $\{e_2 \subset_p s_1\}$.
C_{b21R} : $\{he_2 \neq d_1\}$.
C_{b22R} : $\{e_2 \cap s_1 = \emptyset\}$.
C_{b2B} : $\{s_1 < e_2\}$. (Note that C_{b2B} implies C_{b22R}.)

IS_{b2A}:		
	K_{a1}	
	$\wr \quad > C_{b21R}$	$> C_{b21}$
	K_{b2}	
		$> C_{b22}$
		$> C_{b22R}$

$IC(IS_{b2A}) = K_{a1} \cup K_{b2} \cup C_{b21R} = K_{a1} \cup K_{b2} \cup \{he_2 \neq d_1\}$.
(John made a short trip to London. *Some other person* visited the Louvre.)

In (5b), we have no evidence for the temporal relation between John's trip to London and the visit to the Louvre by the other person. In TIS, this interpretation can be expressed by making temporal interpretation ambiguous. Thus, IS_{b2A} implies the information $K_{b2} > C_{b22}$ & $K_{b2} > C_{b22R}$ and $IC(IS_{b2A})$ does not contain any information on temporal relation between e_2 and s_1.

It is also possible to interpret the person referred to by "He" as John. In this case, his visit to the Louvre must take place after his stay in London:

IS_{b2B}:		
	K_{a1}	
	$\wr \quad > C_{b21}$	$> C_{b21R}$
	K_{b2}	
	$\quad > C_{b2B}$	$> C_{b22}$

$IC(IS_{b2B}) = K_{a1} \cup K_{b2} \cup C_{b21} \cup C_{b2B} = K_{a1} \cup K_{b2} \cup \{he_2 = d_1\} \cup \{s_1 < e_2\}$.
(John made a short trip to London. *Then John* visited the Louvre.)

$IC(IS_{b2A})$ and $IC(IS_{b2B})$ show two different context identifications for (5b). In this example, the first interpretation keeps the temporal underspecification, whereas the second interpretation yields a temporal specification.

5 Conclusions

TIS is a theory that combines NRL, a framework of dynamic semantics, with interpretation revision. By using TIS, interpretation processes can be described in detail. In this paper, I have sketched how to apply TIS to choice and revision of anaphoric references, representation of presuppositions, and interpretation of metonymies. TIS can also be applied to the distinction between the attributive and referential use of definite descriptions (cf. [7]).

References

1. Asher, N. (1991) "Discourse Representation Theory and Belief Dynamics," in: A. Fuhrmann and M. Morrau (eds.) *The Logic of Theory Change*, Springer, pp. 282-321.
2. Asher, N. (1993) *Reference to Abstract Objects in Discourse*, Kluwer Academic Pub.
3. Gärdenfors, P. (1988) *Knowledge in Flux*, The MIT Press.
4. Kamp, H. (2001) "The Importance of Presupposition," in: [13], pp. 207-254.
5. Kamp, H. and Reyle, U. (1993) *From Discourse to Logic*, Kluwer Academic Pub.
6. Muskens, R, van Benthem, J. and Visser, A. (1997) "Dynamics," in: J. van Benthem and A. ter Meulen (eds.) *Handbook of Logic and Language*, North-Holland, pp. 587 - 648.
7. Nakayama, Y. (1997) "Communication and Background Knowledge," in Japanese, *Kagakukisoron-Kenkyu*, (*Journal of the Japan Association for Philosophy of Science*), No. 88, Vol. 24 No. 2, pp. 67 - 74.
8. Nakayama, Y. (1999) "Mereological Ontology and Dynamic Semantics," *Annals of the Japan Association for Philosophy of Science*, Vol. 9 No. 4, pp. 29 - 42.
9. Nakayama, Y. (2000) "A Move towards a General Semantic Theory," *Proceedings of the 14th Pacific Asia Conference on Language, Information and Computation*, pp. 273-280.
10. Nakayama, Y. (2001) "Scientific Reasoning and Belief Structure," *Proceedings of the International Conference on Artificial Intelligence*, vol. II, CSREA Press, pp. 881-887.
11. Reyle, U. (1996) "Co-Indexing Labeled DRSs to Represent and Reason with Ambiguities," in: K. van Deemter and S. Peters (1996) *Semantic Ambiguity and Underspecification*, CSLI Pub., pp. 239-268.
12. Reyle, U. and Rossdeutscher, A. (2001) "Temporal Underspecification in Discourse," in: [13], pp. 207-254.
13. Rohrer, C. Rossdeutscher, A. and Kamp, H. (eds.) (2001) *Linguistic Form and its Computation*, CSLI Pub.

The Dynamics of a Japanese Reflexive Pronoun

Eric McCready

University of Texas at Austin
mccready@mail.utexas.edu

Abstract. This paper is concerned with the interpretation of the 'long-distance reflexive' *zibun*. I propose a new way to handle this item designed to account for intersentential binding facts. The account is couched within a version of dynamic semantics enriched with a notion of perspective. In the final part of the paper I show that this approach can be extended to account for intrasententially bound and indexical uses of *zibun*.

The present paper explores an approach to the 'long-distance reflexive' *zibun* that starts with different assumptions than most accounts to be found in the literature, which focus on the syntactic properties of this object.[1] No syntactic analysis has been able to capture all the phenomena; although each author has had valuable insights, the existence of *intersententially* bound *zibun* shows that a purely syntactic account cannot be possible. I will argue that the correct way of thinking about intersentential *zibun* is as an element that is *dynamically* bound by a salient object in the discourse context. This move allows us to not only analyze the phenomenon of intrasententially bound *zibun* that previous researchers have concentrated on, but also to analyze facts about intersentential binding and indexical uses of this object.

1 Background on *Zibun*

Zibun has been analyzed in the syntactic literature as polysemous, with realizations as an ordinary reflexive requiring a c-command relation (Kuroda, 1965), as an "empathetic logophor" which is sensitive to the 'camera angle' used in the discussion (Kuno, 1987), and as a *de se* reflexive which serves to self-ascribe a property to the speaker (Oshima 2002; see also the analysis of Chinese *ziji* of (Pan 1997)). One impetus behind this work is that on the assumptions of GB binding theory in the 1980s and early 1990s (cf. Chomsky (1981)), reflexive pronouns must be bound in their governing category, either a clause or a NP. This analysis turns out not to work for a large class of pronouns, however, the *long-distance reflexives*, discussed by various authors. Japanese *zibun* is one example of a long-distance reflexive; they are also found in Icelandic and other Scandinavian languages, Italian, Chinese, and Korean. The possibility of binding *zibun* from outside the clause is exhibited by (1):

[1] Thanks to the audience at LENLS for helpful comments and discussion, especially Rick Nouwen, Norihiro Ogata, and Katsuhiro Yabushita.

A. Sakurai et al. (Eds.): JSAI 2003/2004, LNAI 3609, pp. 405–415, 2007.

(1) Taro-wa Ziro-ga zibun-wa atama-ga ii to itta to
T-TOP Z-NOM self-NOM head-NOM good COMP say-PST COMP
itta
say-PST
'Taro$_i$ said that Ziro$_j$ said that he$_{i,j}$ is smart.'

Here Taro can bind *zibun* despite the fact that they appear in different clauses.

The most popular GB-style analyses of this problem involve covert movement at LF, where binding theory is sometimes taken to apply; a c-command relation between binder and bindee can be introduced by movement of one of these objects, allowing one to bind the other (Chomsky, 1981). These approaches successfully capture the general tendency of *zibun* to prefer binding by the subjects of clauses; this tendency has also played a role in analyses based on logophoricity. However, the proposed restriction is too strong; GB-based approaches incorrectly predict that *only* binding from subject position is possible; but *zibun* can be bound by objects within NPs or within adjunct clauses. Although there is a clear preference for binding by subjects, it is not difficult to find cases in which binding of *zibun* by elements in non-subject positions is possible, as in the following examples.

(2) Binding by antecedent in adjunct (Kameyama 1985):

Yamada-sensei-wa Taroo ni-totte zibun-no oya-no yoona
Prof. Yamada-TOP Taro for self-GEN parent-GEN existence
sonzai datta
COP-PST

'For Taro$_i$, Prof. Yamada$_j$ was like his$_{i/*j}$ own parent.'

(3) Binding by NP-internal antecedent (Iida 1996):

Hanako-no ziman-wa zibun-no musuko da
Hanako-GEN pride-TOP self-GEN son COP

'Hanako$_i$'s pride is her$_i$ son.'

Perhaps the most successful reanalysis to date that incorporates this other data has been that of Iida (1996), who uses a version of HPSG binding theory using notions of argument hierarchies and o-command to solve the problem (Pollard and Sag, 1994). The idea is essentially that an instance of *zibun* in a given grammatical role can be bound only by an argument in a higher position on the o-command hierarchy. Iida's theory pairs this notion with a set of pragmatic constraints, which together are very successful in accounting for the distribution of intrasententially bound *zibun*.

Intersentential binding of *zibun* is also possible in certain cases, as Iida notes, although her account doesn't extend to them due to its syntactic nature. The following examples are from Iida (1996); there she claims that the possibility of binding has to do with changes in perspective. For a coherent binding, one must maintain the perspective of the sentence in which the binder is introduced in the sentence that includes the instance of *zibun* to be bound:

(4) Ziro-wa totemo okkotteimasu. Taro-ga zibun-no tomodati-ni [Ziro-ga
Z-top very angry-is. T-nom self-gen friend-dat [Z-nom
kaita] e-o misete-simatta kara desu
drew] picture-acc show-perf-pst because cop.

'Ziro$_i$ is very angry. It's because Taro$_j$ showed his$_{i,j}$ friend a picture that Ziro drew.'

(5) Hanako-wa kinoo hidoku Taroo ni attattemasita. * Ziroo-ga
H-top yesterday badly T-dat bit-at Z-nom today
kyoo zibun-ni au koto-ni natteimasu
self-dat meet comp to is-planned

'Hanako$_i$ bit at Taroo$_j$ severely yesterday [Iida's gloss]. Ziroo$_k$ is supposed to see her$_i$ today.'

(6) Hanako-wa kinoo hidoku T-ni atattemasita. Ziro-ga zibun-o
H-top yesterday badly T-dat bit-at Z-nom self-acc
mattaku musisita kara desu
entirely ignored because cop

'Hanako$_i$ bit at T$_j$ severely yesterday. It is because Ziro$_k$ ignored her$_i$ completely.'

Iida comments that the reason for the coherence of these discourses is that the second sentence provides an *explanation* of the first, which allows the correct maintenance of perspective. Supporting this idea is the use of *kara* in the second sentences of the good discourses, which indicates the existence of a causal relation and the intention of the speaker to use the sentence as an explanation of what came before.

However, it seems that the situation of intersentential binding is more complicated than Iida suggests, in that a causal relation is not *required* for anaphora to be possible. Consider the following discourses ((8) is modeled on a Chinese example from Zhu (1997:89)):

(7) a. Taro-ga kore-o tabero to itta wake?
Taro-NOM this-ACC eat-IMP C said actually

'Taro said to eat this?'

b. soo da. zibun-ga kirai-na kuse-ni..
yes COP. ZIBUN-NOM dislike even-though

'Yes. Even though he doesn't like it himself.'

(8) Taro-wa hontoo-ni fuun-na hito da. Saikin okusan-ga
Taro-TOP really unlucky person COP. Recently wife-NOM
kuruma-ni hikareta. kondo-wa zibun-ga kubi ni natta.
car-DAT hit-PASS. next ZIBUN-NOM was-fired

'Taro is really an unlucky guy. Recently his wife got hit by a car. Now he got fired.'

No causal relations are to be found in the discourse in (8). Rather, it seems that the second and third sentence simply provide details about the proposition expressed by the first sentence. In the small dialogue in (7), the situation is even more clear: there cannot be a causal relation between the question and the answer given the semantic type of questions, which are standardly analyzed as sets of propositions. It is not clear how a set of propositions could itself be the cause of anything.[2]

Below I propose an account which is capable of providing a principled explanation for these cases. It is well known that discourses are not unstructured sequences of propositions, but have complex structures of their own (cf. Mann and Thompson 1986; Lascarides and Asher 1993; Webber et al. 2001; discourse segments (generally sentences or clauses) can be connected by various discourse relations. Recent work in Segmented Discourse Representation Theory (SDRT; Asher and Lascarides 2003) claims that discourse relations can be separated into two basic types, coordinating and subordinating relations. I claim that the possibility of using *zibun* intersententially crucially depends on the type of discourse relation that connects the sentences containing the (potential) antecedent and the instance of the reflexive: in particular, it must be a subordinating relation.

In SDRT, each discourse segment is treated as a speech act which introduces a label that marks its propositional content. Through a complex reasoning process involving nonmonotonic inference over discourse content, lexical information and world knowledge, binary discourse relations are inferred as holding between these labelled speech acts. The structure inferred is an acyclic graph which puts important constraints on anaphora: informally stated, for an anaphoric expression introduced in a given discourse segment K, only discourse referents introduced in segments which are connected to K by some (sequence of) discourse relations are available. Importantly, once a segment is attached to a node i in the discourse structure using a coordinating relation, nodes subordinate to i are no longer available for attachment. This definition has important consequences for anaphora. Coordinating relations can be understood as relating two speech act labels on the same level of a hierarchical graph, while subordinating relations introduce dependent nodes. Thus, discourse referents introduced by a speech act K connected to K' by a subordinating relation becomes unavailable for anaphoric expressions in a discourse segment K'' which is attached to K' by a coordinating relation.

Let us consider a concrete example.

(9) a. John had a wonderful evening last night. (π_1)
b. He had a great dinner. (π_2)
c. He ate salmon. (π_3)
d. He had a nice cheese. (π_4)
e. Then he won a dancing competition. (π_5)
f. ?It was a beautiful pink. (π_6)

[2] Note that I do not mean to claim that the *speech act* of asking a question cannot participate in causal relations, but that the semantic objects described by questions cannot be causal.

In SDRT, *Elaboration* is a subordinating relation inferred when one segment gives more detail on the information provided by a preceding one: thus we infer $Elaboration(\pi_1, \pi_2), Elaboration(\pi_2, \pi_3), Elaboration(\pi_2, \pi_4)$, and $Elaboration(\pi_1, \pi_5)$. The presence of the discourse marker *then* induces a monotonic inference of $Narration(?, \pi_5)$; world and lexical knowledge allow resolution of ? to π_2. *Narration* is a coordinating relation, and so 'closes off' the possibility of attaching π_6 to π_3, as would be needed to resolve *it* to the referent introduced by *salmon.* The result is that the discourse is incoherent.

I wish to suggest in the present paper that the distinction between coordinating and subordinating relations is important not only for anaphoric accessibility in later segments, as is now quite standardly accepted, but also has consequences for the *referent pool* available for certain kinds of anaphoric expressions. The idea is that subordinating relations are able to maintain a referent pool, but coordinating relations 'empty' it. *Zibun*, I argue, is one instance of an expression sensitive to this distinction.

2 *Zibun* Dynamically

One way to model the action of *zibun* is as an object which lacks inherent restrictions on its content; rather, a pool of referents exists that it can draw from that changes relative to what has been and what is being said. This pool of referents can be understood as a set of possible perspectives from which the situation described by a sentence or discourse can be viewed, where each perspective makes certain referents available for *zibun* in a way distinct from simple dynamic introduction of a discourse referent. Within the available referents, furthermore, one will be taken to be preferred to the others given the right circumstances. These circumstances include the use of inherently perspectival verbs or honorifics, and also facts about the world that enter into the inferencing mechanisms of the 'ordinary' language user.

I formalize this idea in the following way. taking standard DPL as a starting point (Groenendijk and Stokhof, 1991). I add a notion of a perspective and a mechanism for collecting antecedents for *zibun*; the set thus produced will be emptied with a change of perspective. I give the resulting logic the name 'RDPL'.[3]

Models for RDPL are defined just as in DPL, as are interpretation functions. Assignments look slightly different: we need a set of distinguished objects that can serve as antecedents for *zibun*, so instead of just assigning objects to variables, assignments in RDPL also have such a set associated with them. This set is notated P_n and is updated with processing of each bit of logical form in tandem with the standard assignment.

Definition 1. Assignments

- $f_{P_n}(x) \in D$

[3] 'Reflexive Dynamic Predicate Logic'.

What is the content of the set P_n? That depends on where we are in a discourse. I assume that the only semantic object capable of introducing elements into P_n is the existential quantifier. Thus the only clause of the DPL semantics that sees a serious change is for the existential quantifier. The rest have no effect on P_n. I use a DRT-style treatment of proper names, which often serve as antecedents for *zibun*, on which they are represented as existential quantifiers associated with a condition (e.g. *john*) which picks out a singleton set (cf. Kamp and Reyle 1993). Clauses for disjunction, implication, and universal quantification can be derived in the usual way.

Definition 2. Semantics of RDPL Formulas

- $[\![R(x_1,\ldots,x_n)]\!] = \{\langle g_{P_n}, h_{P_{n+1}}\rangle | h = g \wedge \langle [\![t_1]\!],\ldots,[\![t_n]\!]\rangle \in F(R) \wedge P_n = P_{n+1}\}$
- $[\![\neg\varphi]\!] = \{\langle g_{P_n}, h_{P_{n+1}}\rangle | h = g \wedge \neg\exists k : \langle h,k\rangle \in [\![\varphi]\!]\}$
- $[\![\varphi \wedge \psi]\!] = \{\langle g_{P_n}, h_{P_{n+1}}\rangle | \exists k : \langle g,k\rangle \in [\![\varphi]\!] \wedge \langle k,h\rangle \in [\![\psi]\!]\}$
- $[\![\exists x\varphi]\!] = \{\langle g_{P_n}, h_{P_{n+1}}\rangle | \exists k : k[x]g \wedge \langle k,h\rangle \in [\![\varphi]\!] \wedge P_{n+1} = P_n \cup \{x\}\}$

Note that since the existential quantifier may introduce objects into P_n, the universal quantifier is predicted by the semantics of DPL to do the same, for its nuclear scope. This prediction is borne out, as shown by the donkey sentence in (10):

(10) subete-no otoko-wa tomodati-o zibun-no heya-ni yonda
every man-TOP friend-ACC ZIBUN-GEN room-DAT invite-PST
'Every man$_i$ invited a friend to his$_i$ room.'

Now we move to the heart of the matter, the interpretation of *zibun* itself and the introduction of a dynamic notion of *perspective*. We have already done the groundwork by introducing the set P and its update mechanism; it remains to show how it is used by *zibun*. This is basically very simple. *Zibun* is assumed to freely choose any member of P_n as its referent; this is modelled by use of a choice function h. Which member is chosen will depend on various factors: world knowledge, grammatical elements, and what the interpreter perceives to be the intentions of the speaker; this last will be picked up on based on speech act used and so on discourse structure.

Perspectives are formalized in a simplistic way in this preliminary proposal, simply as a set of possible antecedents for *zibun*. Ultimately, of course, a perspective is more than this: ultimately, it should be considered as something more like an epistemic state, a set of propositions believed true by a perspective-holder, as well as a representation of the current state of a discourse.[4] I also treat the notion of perspective *change* in a simplified manner. To really do the notion justice we must understand exactly how a perspective is changed, and this seems to depend rather heavily on all the factors mentioned in the preceding paragraph. Since I do not at present have a comprehensive theory of perspectives, I allow arbitrary

[4] See Asher (1992) for a treatment of perspectives in the context of a formal treatment of the semantics of progressive aspect.

choice of points at which a perspective is changed, which I will model by use of a transition formula X, a 'blocker,' which affects output assignments by resetting the content of the antecedent set P_n but has no effect on the truth-conditional content of the discourse. Such a formula is stipulated to appear at the conclusion of each single sentence, except when it is connected to the following sentence by a subordinating discourse relation, as shown by the examples of intersentential binding above (cf. Asher and Lascarides 2003).

Definition 3. Semantics of *Zibun* and Perspective Changer

- $P_0 = \emptyset$
- $[\![zibun]\!] = h(P_n)$
- $[\![X]\!] = \{\langle g_{P_n}, h_{P_{n+1}}\rangle | g = h \wedge P_{n+1} = P_n \cap P_0\}$

Let's examine two simple examples to show how the system applies. We will use (6). The logical form of this discourse is shown in (11a), while (11b) shows the content of P at each stage of update. Since π and π' are connected by a subordinating relation, no blocking formulas are present, and so both Hanako (x) and Taro (y) are possible referents for *zibun* according to the present theory.

(11) a. $\pi : \exists x[h(x)]; \exists y[t(y)]; bite\_at(h,t); \pi' : \exists z[z(z)]; ignore(z, zibun);$
$Explanation(\pi', \pi)$

b. $\{x\}; \{x,y\}; \{x,y\}; \{x,y,z\}; \{x,y,z\}; \{x,y,z\}$

In the case of (5), however, the discourse relation is Narration, a coordinating relation; thus, blocking formulas appear after the content of each speech act, and intersentential binding is impossible, as shown in (12).

(12) a. $\pi : \exists x[h(x)]; \exists y[t(y)]; bite\_at(h,t); X; \pi' : \exists z[z(z)];$
$meet\_today(z, zibun); X; Narration(\pi', \pi)$

b. $\{x\}; \{x,y\}; \{x,y\}; \emptyset; \{z\}; \{z\}; \emptyset; \emptyset$

It should also be noted here that, given the way the update mechanism is defined, intrasentential binding of *zibun* is assumed to take place via a different mechanism. It is not clear whether a unified account is desirable, given that somewhat different constraints seem to operate on intra- vs. intersentential *zibun*. The next section considers possible ways to extend the present account to the intrasentential case, should one decide that doing so is the right way to go.

2.1 The Intrasentential Case

Extension to intrasentential binding is not simple, for a number of reasons. First, it is well known that the referent selected by *zibun* depends, not only on pragmatic and discourse facts, but also on such hard-to-capture things as the degree of empathy shown by a particular utterance containing *zibun* with the individuals described in the sentence, as expressed by the particular lexical items chosen. For instance, certain complex verbs express a degree of empathy with the agent or patient of the described event. Second, the way the RDPL formalism is set up causes difficulties for compositional interpretation: given the left-to-right nature of dynamic binding, it will be necessary to force possible antecedents to precede *zibun* in logical form regardless of the surface order of the clauses. This factor will cause problems in case *zibun* appears in a clause-initial adjunct, for instance. Of course, it is possible to construct a working system (for instance by allowing each proper name to introduce an existentially quantified variable which is attached at the leftmost point of the formula introduced by the clause), but it is difficult to construct one that does not appear *ad hoc*. For this reason, the logic introduced in the present paper applies only to the intersentential case. In this section I confine myself to some discussion on how the first problem could be overcome, given a satisfactory solution for the second.

An attractive possibility is to utilize Iida's (1996) analysis, which crucially involves the HPSG o-command hierarchy (on grammatical roles). This can be done by annotating discourse referents (in P) with grammatical role information at the time of their introduction. Concretely, one might make use of the f(unctional)-structures of Lexical-Functional Grammar (Bresnan, 2001). In this theory, constituent structure maps to a functional structure in which grammatical roles are mapped to unique objects, as shown in (13):

(13) a. John likes Bill.

b. $$\begin{bmatrix} \text{PRED} & \text{`LIKES}\langle\text{SUBJ,OBJ}\rangle\text{'} \\ \text{SUBJ} & \text{`John'} \\ \text{OBJ} & \text{`Bill'} \end{bmatrix}$$

LFG f-structures contain all necessary information for discourse referent annotation. A function-argument pair of the form ⟨SUBJ,'John'⟩ can be defined as introducing an object of the form j_{subj} into P_n.[5] Lexical entries for words/morphemes such as *morau/kureru* which indicate 'empathy' with one particular element then bias the interpretation of *zibun* toward the correct argument by selecting for an element with a particular subscript. If such an analysis is adopted, it becomes clear that we need not lose empirical coverage by switching to a dynamic interpretation of *zibun*, but rather the opposite. By adopting a dynamic account, we become able in principle to provide a unified account of intersentential and intrasentential binding of *zibun*.

Oshima (2002) and others note that *zibun* is associated with *de se* readings, readings of attitude-ascribing sentences in which a speaker self-ascribes an attitude. To see what this reading consists of, consider the following example due to Kaplan (1989):

(14) John believes that his pants are on fire.

a. John believes that John's pants are on fire.

b. John believes that his own pants are on fire.

(14) is ambiguous between the two readings in (14a) and (14b). Consider a situation in which John is looking at a mirror, which reflects an individual with

[5] The same trick would work in HPSG, of course.

his pants on fire. Imagine now that John does not realize that he is looking at his own reflection—that is, he thinks that the individual with burning pants is not him, but someone else. This reading is described by (14a). If John realizes that he is looking at himself, the sentence would have the reading in (14b). Assuming that John is completely amoral, only the second reading is one that would prompt him to try to put the fire out, for only this reading is 'self-directed' in the way that would cause him to act. The reading described by (14b) is the *de se* reading.

It turns out that when *zibun* is used the second reading is preferred; that is, *zibun* forces a *de se* reading. Within the current dynamic semantic setting, the different readings are distinguished by use of an *anchoring function* from discourse referents to objects: this function can be partial, to allow for the possibility that a speaker's mental representation of the discourse involves an object that actually doesn't exist (see Kamp et al. 2003 for discussion). *De se* readings are obtained by mapping the discourse referent representing the subject of the embedded proposition to the speaker by the anchoring function (Dekker, 2000). To handle the Japanese facts, anchoring functions can be further constrained so that they always map *zibun* to the attitude holder (in the case of potentially *de se* attitudes). Space considerations preclude spelling this analysis out in the present paper.

2.2 The Indexical Case

Zibun also has an indexical use: it can refer to the speaker of the utterance in which it appears, or to the interlocutor in certain special circumstances. In standard (Tokyo) Japanese and also in various dialects, an interpretation of *zibun* as an indexical referring to the speaker is available. Apparently this use comes from military jargon of the early 20th century, when members of the Japanese armed forces were required to refer to themselves as *zibun* when speaking to superiors (Martin, 1975); somehow this usage spread into everyday speech, and is now very common. As far as I can tell, this usage is available quite generally, although it is restricted to relatively colloquial speech. Here is an instance:

(15) zibun-wa hara het-teiru kedo.
self-TOP stomach empty-PROG though
'I'm hungry now anyway.'

The remaining indexical use exists only in certain dialects of Japanese. It's primarily associated with the dialect of the Kansai region (comprising, generally speaking, the region around Osaka and Kyoto), though it also appears in other Western dialects in Kyushu and the Chugoku region. *Zibun* in these dialects can also refer to the interlocutor; that is, it can also mean 'you,' as in (16):

(16) zibun-no suru koto-wa zenbu tadasii to omo-tteiru kara
self-GEN do thing-TOP all correct COMP think-PROG because
yaro
probably

'(That's) because you think everything you (I) do is right (isn't it).'

Note that *zibun* here can refer either to the speaker or to the hearer, though the second person interpretation is probably more natural. This second person use of reflexive pronouns turns out to be quite common cross-linguistically: similar facts can be found in Korean, Norwegian, and Kannada (Jeffrey Lidz, p.c.), and probably other languages as well.

We can account for indexical uses of *zibun* through a simple modification of the RDPL system described above. Above we assumed that the base set for update of possible *zibun*-antecedents, P_0, was the empty set. We need only assume that the base set is actually not empty to account for the possibility of *zibun* denoting the speaker, as follows:

- $P_0 = \{s\}$

For dialects of Japanese which permit reference to the interlocutor, we can add another element to this set (i). One might object to this move on the assumption that the second person use is simply a different lexical item than the first, for example with a different tone; if correct, this would motivate assuming the existence of two distinct base sets P_{0s} and P_{0i}. However, this is not the case, as McCready (2003) showed through phonetic experiments involving production data. Further, no phonetic difference is perceivable by interpreters (McCready, 2003), which also suggests that only one lexical item is involved.

Note that this semantics predicts that use of *zibun* is universally available to denote speaker or hearer. In fact, this prediction is probably not completely correct. First person indexical uses of *zibun* within the complements of attitude verbs seem to require focus-like intonational prominence. Also, the distribution of second person indexical *zibun* seems to be even more restricted in that it is felicitous primarily in the context of certain speech acts. I will not consider these complications further in the present paper.

3 Summary

In this paper, I proposed a new account of the Japanese long-distance reflexive *zibun* on which it is analyzed as a special sort of dynamically bound object associated with a set of potential binders. Each member of this set is taken to correspond to an individual from whose perspective the situation described by a sentence (or discourse) is viewed. Although the account is designed to handle cases of intersentential binding of *zibun*, I showed how it can be extended to cases of intrasentential binding and even indexical uses of *zibun*.

Bibliography

Asher, Nicholas. 1992. A default, truth-conditional semantics for the progressive. *Linguistics and Philosophy* 15:463–508.

Asher, Nicholas and Alex Lascarides. 2003. *Logics of Conversation*. Cambridge University Press.

Bresnan, Joan. 2001. *Lexical-Functional Syntax*. Blackwell.

Chomsky, Noam. 1981. *Lectures on Government and Binding*. Dordrecht: Foris.

Dekker, Paul. 2000. Coreference and representationalism. In K. von Heusinger and U. Egli, eds., *Reference and Anaphoric Relations*. Dordrecht: Kluwer.

Groenendijk, Jeroen and Martin Stokhof. 1991. Dynamic predicate logic. *Linguistics and Philosophy* 14:39–100.

Iida, Masayo. 1996. *Context and Binding in Japanese*. Stanford: CSLI. Ph.D Thesis, Stanford University, 1992.

Kameyama, Megumi. 1985. *Zero Anaphora: the Case of Japanese*. Ph.D. thesis, Stanford University.

Kamp, Hans and Uwe Reyle. 1993. *From Discourse to Logic*. Dordrecht, Reidel: Kluwer.

Kamp, Hans, Josef van Genabith, and Uwe Reyle. 2003. Discourse representation theory. Draft of article to appear in *Handbook of Philosophical Logic*.

Kaplan, David. 1989. Demonstratives. In J. Almog, J. Perry, and H. Wettstein, eds., *Themes from Kaplan*. Oxford University Press. Manuscript version from 1977.

Kuno, Susumu. 1987. *Functional Syntax: Anaphora, Discourse, and Empathy*. University of Chicago Press.

Kuroda, Sige-Yuki. 1965. *Generative Grammatical Studies in the Japanese Language*. Ph.D. thesis, MIT.

Lascarides, Alex and Nicholas Asher. 1993. Temporal interpretation, discourse relations and commonsense entailment. *Linguistics and Philosophy* 16:437–493.

Mann, William and Sandra Thompson. 1986. Relational propositions in discourse. *Discourse Processes* 9:57–90.

Martin, Samuel. 1975. *A Reference Grammar of Japanese*. New Haven: Yale University Press.

McCready, Eric. 2003. Prosodic contour and indexical interpretation in Osaka Japanese. unpublished ms.

Oshima, David Y. 2002. Logophoricity, empathy, and *de se* interpretation. Revised version of Tokyo University Masters' thesis.

Pollard, Carl and Ivan Sag. 1994. *Head-Driven Phrase Structure Grammar*. CSLI and University of Chicago Press.

Webber, Bonnie, Matthew Stone, Aravind Joshi, and Alastair Knott. 2001. Anaphora and discourse semantics. To appear in *Computational Linguistics*.

Zhu, Shensheng. 1997. *A Dynamic Semantic Theory of Chinese Anaphora*. Ph.D. thesis, University of Arizona.

Dynamic Semantics at Work

Rolf Schwitter and Marc Tilbrook

Centre for Language Technology, Macquarie University,
Sydney, NSW 2109, Australia
{schwitt, marct}@ics.mq.edu.au

Abstract. In this case study we show how an unambiguous semantic representation can be constructed dynamically in left-to-right order while a text is written in PENG, a controlled natural language designed for knowledge representation. PENG can be used in contexts where precise texts (e.g. software specifications, axioms for formal ontologies, legal documents) need to be composed. Texts written in PENG look seemingly informal and are easy to write and to read for humans but have first-order equivalent properties that make these texts computer-processable.

1 Introduction

Controlled natural languages are well-defined subsets of natural languages that have been restricted with respect to their grammar and their lexicon. Traditionally, controlled natural languages fall into two major groups: *human-oriented* and *machine-oriented* controlled natural languages [4].

PENG is a machine-oriented controlled natural language – with a restricted grammar and lexicon – that has been designed to write unambiguous and precise specifications for knowledge representation [11]. The writing process for a PENG text is supported by ECOLE, an intelligent text editor, that guides the user and guarantees compliance to the rules of the controlled language [12].

The language processor of the PENG system translates sentences dynamically into a flattened notational variant of Discourse Representation Theory (DRT) [6] [7] while the user writes the text. That means that after each word form that the user enters, the language processor creates semantic information that is analyzed in the context of a given discourse representation structure (DRS) and then either updates the current information state or delays this process if necessary. Additionally, the language processor generates look-ahead categories and sends them to the editor. These look-ahead categories provide syntactic hints that inform the user about choices on how to continue the current input string. The language processor is connected via a server with a theorem prover (OTTER; [9]) and a model builder (MACE; [8]) that run in parallel and allow for question answering and acceptability checking (i.e. consistency and informativeness checking) of a specified piece of knowledge.

In this paper, we will focus on how a DRS is dynamically constructed using a flattened notation for DRT conditions. Since PENG is a controlled natural language with well-defined syntactic properties, it allows for threading of information states in left-to-right order.

A. Sakurai et al. (Eds.): JSAI 2003/2004, LNAI 3609, pp. 416–426, 2007.

2 The Controlled Language PENG

PENG uses an unification-based phrase structure grammar as syntactic scaffolding for the DRS construction and a lexicon with (user-defined) content words and predefined function words.

2.1 The Grammar of PENG

The grammar of PENG currently consists of over 100 phrase structure rules that are processed by a chart parser which performs top-down parsing. The same rule formalism is used for building syntactic, semantic and pragmatic structures. This rule formalism uses an attribute-value notation to represent feature structures associated with the constituents. Below is a phrase structure rule that shows that a simple PENG sentence *s0* is composed of a noun phrase *n3* and a verb phrase *v3* followed by a full stop *fs*:

```
s0([crd:no,drs:D,para:P1-P4,tree:[s0,T1,T2,T3]]) -->
  n3([crd:_,arg:I,spec:Q,ana:A,drs:D,sco:S,para:P1-P2,tree:T1]),
  v3([crd:_,arg:I,drs:S,para:P2-P3,tree:T2]),
  fs([cat:fs,para:P3-P4,tree:T3]).
```

This rule encodes – among other things - the flow of semantic information and subject-verb agreement of the constituents. The attribute *arg* with the value *I* (i.e. a variable that stands for an entire feature bundle) constrains the argument structure of a sentence. The attribute *drs* guarantees that the semantic information *D* of the sentence is shared with the noun phrase (and eventually with its determiner). The attribute *sco* assures that the scope *S* of the noun phrase is passed to the verb phrase. The task of the attribute *para* is to construct a paraphrase for the input string and finally the attribute *tree* is used to build up a syntactic tree during processing time.

PENG sentences can be simple or complex and can be interrelated anaphorically with definite expressions. The main restrictions of the language are the use of present tense verbs and the control of plural constructions by explicit disambiguation markers [11]. Other restrictions that are important are the use of variables instead of personal pronouns, and the scope of quantifiers and negation. In PENG, a quantifier has always scope over all subsequent quantifiers in a sentence and a verb phrase negation has scope over the entire verb phrase. The order of quantifiers can be manipulated by constructors such as *there is a* and *for every*, for example *For every cyclist there is a race that is challenging.* These restrictions enable the scope of all quantifiers to be determined from the surface order.

2.2 The Lexicon of PENG

The lexicon of PENG is made up of content words and function words. Content words are defined by the user and can be added or modified during the writing process with the help of a lexical editor. The structure of content words can be simple (e.g. *cyclist*) or compound in various forms (e.g. *motorcyclist, tricyclist, champion cyclist*) like in full English. Semantically, compound words are treated as single units in the lexicon but their internal structure is not further analysed. In PENG, each content word can be associated with a set of (strict) synonyms and abbreviations. During processing time, synonyms and abbreviations are automatically replaced by the primary content word that has been specified in the lexicon and its corresponding logical representation. These replacements are reflected in a paraphrase that the PENG system generates for each sentence.

2.3 Writing in PENG

The writing process for a PENG text is supported by ECOLE, a look-ahead text editor [12]. After each word form that the user enters, the editor indicates what kind of syntactic structures can follow the current input string. Let us assume that the user is in the process of writing the noun phrase:

The American champion cyclist Lance Armstrong ...

After entering the word *champion*, the text editor displays the following kinds of look-ahead categories:

```
proper_noun | verb | auxiliary:[does] | noun:[cyclist] |
preposition:[of] | relative_pronoun:[who] | variable:[e.g. X1]
```

At this point, the user can employ all lexicalized content words and function words that belong to these categories to carry forward the input string. This way, the user is guided and does not need to learn and remember the restrictions of the controlled language. If the user enters a content word that is not in the lexicon, then the spelling checker of the PENG system fires up and checks whether that word is misspelled or not and suggests lexicalized alternatives. The user corrects the spelling of the word manually or selects an alternative from a menu. If the word is not yet in the lexicon, then the user may add the new word (in case of a content word) to the lexicon with the help of a lexical editor.

2.4 Anaphora Resolution in PENG

In PENG only definite noun phrases, proper nouns, and variables can be used as anaphoric expressions. Personal pronouns are not allowed in PENG but non-ambiguous variables can be used instead of pronouns, for example:

The cyclist A1 beats the cyclist A2. A1 is faster than A2.

Noun phrases must be accessible to be referred to by anaphoric expressions: indefinite noun phrases that are in the scope of a negation or a universal quantifier are not accessible for anaphoric reference from subsequent sentences in PENG. If the text contains – for example – the accessible noun phrase:

The American champion cyclist Lance Armstrong ...

then all nominal expressions that are part of this noun phrase (e.g. *The American champion* or *Lance*) can be used to corefer with the closest antecedent.

In the PENG system anaphora resolution is done dynamically while the user writes a text. That means whenever a complete noun phrase has been processed, the anaphora resolution algorithm checks if there exists a subsumption relation between the noun phrase and an accessible antecedent. If no antecedent can be found, then the noun phrase is treated as an indefinite noun phrase introducing a new discourse referent.

3 A Flattened Notation for DRT Conditions

Formally, a DRS is an ordered pair `<U,Con>` where `U` is a set of discourse referents and `Con` is a set of conditions [6]. For processing reasons, we represent a DRS as a term of the form `drs(U,Con)` consisting of a list `U` of discourse referents $[I_1, I_2, \ldots I_n]$ denoting entities and a list `Con` of conditions $[C_1, C_2, \ldots C_n]$ that describe properties or relations that these discourse referents must satisfy. DRSs can occur as constituents of larger (complex) DRSs. Complex DRS conditions are those involving implication, disjunction, and negation [7].

In our flattened notation for DRT conditions we treat concepts such as `cyclist(I)` as typed individuals `obj([cyclist],I)`. Concepts do not introduce predicate symbols anymore and can therefore be referred to by simple terms (see also [3]). The domain of discourse is divided into the domain of objects and the domain of eventualities (i.e. events and states). The domain of objects has a lattice-theoretic structure and is subdivided into groups, individuals and mass objects [10].

4 Dynamic Construction of DRSs

In Kamp's [6] *original* DRT the processing of a sequence of sentences S_1, S_2, ... S_n is carried out through a DRS construction algorithm that starts with the syntactic analysis of the first sentence S_1 and then transforms it with the help of DRS construction rules into a DRS K_1 which serves as the context for processing the second sentence S_2. This approach is sequential and does not emphasize the dynamic aspect of transforming information states while syntactic constituents are parsed.

Johnson and Klein's [5] reformulation of the DRS construction algorithm offers a solution here (see also [2]). In their approach each syntactic constituent of a sentence is related to an incoming and outgoing DRS. This relation is modeled by a difference list of the form *DrsIn-DrsOut*. This data structure threads the semantic information through the phrase structure rules of a (definite clause) grammar. The outgoing DRS is constructed from the incoming DRS (which contains information about available antecedents) plus conventional semantic information derived from the actual constituent. The meaning of a constituent can then be defined as the change in the DRS, after the constituent has been processed.

Threading of semantic information through a grammar can get complex, especially if a sentence consists of a number of quantifiers and constructors, since different scope-bearing elements of a sentence result in nested DRSs. As we will see in the next section, updating of a DRS sometimes needs to be delayed, especially if we want to deal with optional constituents such as prepositional and adverbial modifiers in a uniform way.

5 Implementation Issues

In this section, we will discuss how a DRS is dynamically constructed in PENG. To make this discussion self-contained, we will first present a pseudo-grammar that will later be used for threading difference lists through its structure. We will first focus on the crucial role that determiners play in this process and then describe in a stepwise manner how the sentence

The American champion cyclist Lance Armstrong wins no race in April.

is translated into a DRS. This sentence consists of a rather complex definite noun phrase *The American champion cyclist Lance Armstrong* in subject position and a verb phrase *wins no race* with a temporal modifier *in April.* The definite noun phrase is composed of an adjective (*American*), a compound common noun (*champion cyclist*), and a proper noun (*Lance Armstrong*) in appositive position. This noun phrase can be described by the following simplified phrase structure rules:

```
n3  --> d0, n2, { anaphora_resolution }.
n2  --> a2, n1.
n1  --> n0, ap.
ap  --> pn.
n0  --> xcn, ccn.
pn  --> xpn, cpn.
```

The first phrase structure rule indicates that anaphora resolution is done after the entire noun phrase has been processed. As the phrase structure rules foreshadow, compound words such as *champion cyclist* and *Lance Armstrong* are processed token by token as they are entered by the user (see Section 5.2 for details).

The verb phrase of the sentence is composed of a transitive verb *wins* that subcategorizes for a (negative) noun phrase *no race* whereas the prepositional phrase *in April* is syntactically attached to the verb. The simplified phrase structure rules have the following form in PENG:

```
v3  --> v2.
v2  --> v1, p2.
v1  --> v0, n3.
p2  --> p0, n3.
```

As we will see, these phrase structure rules will be expanded in a way that allows us to deal with, among other things, optional constituents such as prepositional and adverbial modifiers. This is important, since these optional constituents can open a new complex DRS space that embeds DRS conditions that have been derived from preceding constituents. As we will see in the following sections, the key question is to decide when to open a new DRS space and when to close it.

5.1 The Role of Determiners

The bulk of work that the DRS construction algorithm conducts is done in the rules for the determiners. Despite their minor syntactic role, determiners are the most important constituents for establishing the DRS of a sentence. Semantically, a determiner has two arguments: a restrictor and a scope. The restrictor consists of the conditions derived from the remaining noun phrase (= *n3 - d0*). The scope – itself probably consisting of a complex DRS – can be composed of the conditions outside the noun phrase (in our case semantic information derived from the verb phrase). Below is a (simplified) grammar rule for the definite determiner *the*:

```
d0([...,
    drs:D1-D3,res:[drs([],[])|D1]-D2,sco:D2-D3,
    ...]) -->
    { lexicon([lex:Determiner,...],[...]) },
    Determiner.
```

A definite determiner does not introduce a new DRS. However, since we do not known in advance whether the entire definite noun phrase is used anaphorically or not, we add the semantic information derived from the remaining noun phrase into an empty DRS *drs([],[])* that we place in front of the restrictor's incoming DRS *D1*. This DRS serves as a store that can be accessed by the anaphora resolution algorithm once the entire definite noun phrase has been processed. After anaphora resolution, the restrictor's outgoing DRS *D2* will contain the resolved DRS conditions that are then passed on to the scope's incoming DRS. After processing the verb phrase, the scope's outgoing DRS *D3* will contain the semantic information for the sentence as a whole.

Other determiners such as the negative determiner *no* manipulate the DRS in more complex ways:

```
d0([...,
    drs:D1-[drs(U1,[drs(U2,C2) -> drs([],[~drs(U3,C3)])|C1])|D3],
    res:[drs([],[])|D1]-D2,
    sco:[drs([],[])|D2]-[drs(U3,C3),drs(U2,C2),drs(U1,C1)|D3],
    ...]) -->
  { lexicon([lex:Determiner,...],[...]) },
  Determiner.
```

Here the restrictor pushes an empty DRS `drs([],[])` in front of the incoming DRS `D1` and makes this the active information space where all discourse referents and conditions for the remaining noun phrase are collected. The scope takes the restrictor's outgoing DRS `D2` and pushes a new empty DRS in front of it and makes this again a new active information space where all discourse referents and conditions outside the noun phrase are collected. The DRS for the restrictor `drs(U2,C2)` and the DRS for the scope `drs(U3,C3)` are then embedded into a complex outgoing DRS condition that consists of an implication `->` and a negation `~` representing the meaning of the negative determiner.

5.2 Processing the Restrictor

After processing the definite determiner, the restrictor `R1` is passed on to the constituent `n2` where the semantic information for the remaining noun phrase will be acquired:

```
n3([...,arg:[I|A],drs:D,sco:S,...]) -->
  d0([...,ana:yes,drs:D,res:R1-R3,sco:S,...]),
  n2([...,arg:[I|A],drs:R1-R2,...]),
  { anaphora_resolution(...,I,R2,R3,...) }.
```

Note that the remaining noun phrase consists of the nominal constituent *American champion cyclist Lance Armstrong*. The structure of the compound words will not be further analysed semantically but their semantics will be represented as a list of terms instead of a predicate (see Section 3). This form of representation makes it easy to add additional axioms later that specify the relation between these terms – if necessary.

The following (simplified) phrase structure rules are responsible for processing a compound noun such as *champion cyclist* dynamically:

```
n0([...,drs:D1-D2,...]) -->
  xcn([...,len:[]-Tokens,drs:D1-D2,...]),
  ccn([...,len:Tokens-[],...]).

xcn([...,arg:[ind:I|R],len:[]-Tokens,
     drs:[drs(U1,C1)|D1]-[drs([I|U1],[C3,C2|C1])|D1],...]) -->
  { lexicon([lex:[Token|Tokens],...],
            [...,arg:[ind:I|R],...,con:[C3,C2]]) },
  [Token].
```

Compound nouns are processed token by token. After the user enters the word *champion*, the first grammar rule `n0` triggers a lexicon lookup via `xcn` and retrieves the semantic information for the entire compound noun. This information (`I` for the discourse referent and `C3,C2` for the conditions) is then pushed on the outgoing DRS:

```
[drs([I|U1],[C3,C2|C1])|D1]
```

The remaining tokens *Tokens* of the compound noun are written on a list that is processed recursively as soon as these tokens are entered by the user. This way only one lexicon lookup is necessary for processing a compound noun.

After processing the entire noun phrase, the outgoing DRS *R2* of *n2* has the following form:

```
[drs([A],
     [prop([american],A),struc(A,atomic),
      obj([champion,cyclist],A),named([lance,armstrong],A)])|D1]
```

The anaphora resolution algorithm of *n3* checks now whether the acquired information in front of *D1* is accessible in *D1* or not.

5.3 Processing the Scope

The scope of the noun phrase will be realized by the semantic information derived from the verb phrase. The attribute *drs* is used to pass the scope *S* from the noun phrase into the verb phrase (see the grammar rule *s0* in Section 2.1).

The semantic information for the verb phrase is processed by the following simplified phrase structure rules:

```
v3([...,drs:D,...]) -->
  v2([...,drs:D,...]).

v2([...,drs:D,...]) -->
  v1([...,drs:D,sco:S,sco:hold:S1,...]),
  p2([...,drs:S,sco:S1,...]).

v1([...,drs:D,sco:S,sco:hold:S1,...]) -->
  v0([...,drs:S1,...]),
  n3([...,drs:D,sco:S,...]).
```

The important thing here is that in *v0* a new difference list *S1* with an incoming and outgoing DRS is constructed that collects the semantic information for the verbal event:

```
drs:[drs(U1,C1)|D]-[drs([E|U1],[C3,C2|C1])|D]
```

However, this information is not immediately used as scope for the subcategorized noun phrase *n3*, since it is not yet known at this stage of processing whether the verbal event will be modified later (as it is the case in our example) or not. Instead, the above-mentioned difference list *S1* is stored using the attribute *sco:hold* and passed on to the prepositional phrase *p2* via *v1*. In summary, the prepositional phrase *p2* receives the scope *S* of the noun phrase *n3* as DRS and the DRS of the verb *v0* as scope *S1*.

If the sentence would not contain a prepositional modifier, then the following additional rule would be used and guarantee that the scope of the noun phrase unifies with the semantic information for the verb stored on the holding list:

```
v2([...,drs:D,...]) -->
  v1([...,drs:D,sco:S,sco:hold:S,...]).
```

After processing the scope for the verb phrase, the outgoing DRS has the following final form:

```
[A,B]
prop([american],A),struc(A,atomic),obj([champion,cyclist],A),
named([lance,armstrong],A),struc(B,atomic),named([april],B)
  [C]
  obj([race],C),struc(C,atomic) ->
  ~ [D,E]
    prop(D,in,E,B),role(D,time),pred(E,[wins],A,C),evtl(E,event)
```

As the resulting DRS shows, the DRS conditions derived from the verb occur on the right hand side of the implication and fall under negation. This is a consequence of the quantificational potential of the negative determiner.

6 Check It or Prove It!

Texts written in PENG can be checked for consistency and informativeness after each new sentence (or paragraph) that the user enters. Apart from that, the user can also query a specification text. For example, the text in Fig. 1 is a reformulation of Lewis Carroll's *Grocer Puzzle* in PENG together with its DRS representation. The puzzle can be solved automatically by using the following question:

Is it true that no grocer is a cyclist?

or by a simplified version of that question:

Is no grocer a cyclist?

Both questions belong to the controlled language subset of PENG. However, before this can happen, the DRS in Fig. 1 needs first to be translated into a set of first-order formulas, since the third-party reasoning services that we employ can not process DRSs directly. Apart from this first-order representation, the reasoning services of PENG use additional lattice-theoretic (linguistic) axioms for the inference tasks (question answering in our case). For instance, the following axiom

```
(all X Y ((struc(X,atomic) & part_of(X,Y)) -> struc(Y,group))).
```

is used to relate a noun phrase (e.g. *every X*) that introduces an atomic object into the domain to a noun phrase (e.g. *all Xs*) that introduces a group. In our case, the answer to the puzzle can be deduced with the help of OTTER, a resolution-style theorem prover for first-order logic with equality [9].

OTTER automatically generates clauses for these input formulas, selects inference rules and strategies and solve the puzzle by deriving the empty clause.

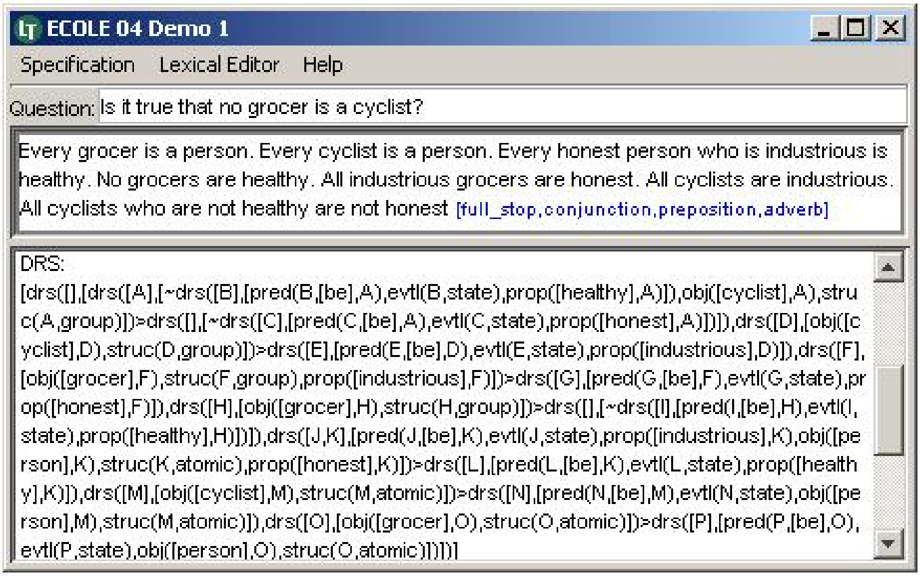


Fig. 1. The Text Editor ECOLE with Look-ahead Categories and DRS

7 Conclusions

PENG is a controlled natural language designed for knowledge representation and the PENG system provides a new kind of writing assistance to ensure that users write unambiguous texts that have the same formal properties as the underlying formal representation language (DRT).

In this paper, we focused on the dynamic aspects of constructing a DRS for texts written in PENG. For this purpose, we employed a flattened notation for DRT conditions and showed how DRSs can be constructed dynamically for the controlled natural language in left-to-right order while the texts are written. The presented approach can deal with, among other things, quantified noun phrases and very complex compound terms. As we have seen, embedding of new DRS conditions into a preceding DRS sometimes needs to be delayed in our incremental approach, since scope-bearing constituents may enforce a specific form of embedding in the final semantic representation that is not known at a certain point in time. However, the resulting formal representation is always an unambiguous reflection of the input text in controlled natural language.

Since texts written in PENG look seemingly informal, this controlled natural language can mediate between end-users and knowledge engineers. Moreover, it can support the knowledge acquisition process and increase the transparency of the knowledge representation for various kinds of application.

Currently, we are investigating if PENG can be used as a lingua franca for representing Semantic Web data and rules. The idea is to define a description logic equivalent subset of PENG and translate the resulting texts into Notation 3 or directly into RDF [1]. Also axioms of the RDFS and OWL layers could then be represented in a version of PENG – making these axioms writable, readable and understandable for man and machine.

Acknowledgments

The research reported here was supported by the Australian Research Council, Discovery Project DP0449928. The authors would also like to thank to all the folks at the Centre for Language Technology at Macquarie University for many discussions and valuable comments on the PENG project.

References

1. Berners-Lee, T.: Notation 3. An RDF language for the Semantic Web. Version 1.49, http://www.w3.org/DesignIssues/Notation.html, (2001)
2. Black, A.W.: Some Different Approaches to DRT. Draft. Centre for Cognitive Science, University of Edinburgh, June 27, (1997)
3. Hobbs, J.R.: Discourse and Inference. Draft. USC Information Science Institute, Marina del Rey, California, November 3 (2003)
4. Huijsen, W.O.: Controlled Language - An Introduction. In: Proceedings of CLAW 1998, Pittsburgh (1998) 1-15
5. Johnson, M., Klein, E.: Discourse, anaphora and parsing. In: Proceedings of the 11th International Conference on Computational Linguistics. Coling 86. Bonn (1986) 669–675
6. Kamp, H.: A theory of truth and semantic representation. In: J. Groenendijk et al. (eds.) Formal methods in the study of language. University of Amsterdam (1981) 277-322
7. Kamp, H., Reyle, U.: From Discourse to Logic, Dordrecht: Kluwer (1993)
8. McCune, W.: MACE 2.0 Reference Manual and Guide. ANL/MCS-TM-249. Mathematics and Computer Science Division, Technical Memorandum No. 249, Argonne National Laboratory, Argonne (2001)
9. McCune. W.: OTTER 3.3 Reference Manual. ANL/MCS-TM-263, Mathematics and Computer Science Division, Technical Memorandum No. 263, Argonne National Laboratory, Argonne (2003)
10. Schwertel, U.: Plural Semantics for Natural Language Understanding. A Computational Proof-Theoretic Approach. PhD Thesis, University of Zurich (2003)
11. Schwitter, R.: English as a Formal Specification Language. In: Proceedings of the Thirdteenth International Workshop on Database and Expert Systems Applications (DEXA 2002). Aix-en-Provence (2002) 228–232
12. Schwitter, R., Ljungberg, A., Hood, D.: ECOLE: A Look-ahead Editor for a Controlled Language, Proceedings of EAMT-CLAW03, Controlled Language Translation, May 15-17, Dublin City University (2003) 141–150

A Demonstrative Analysis of Anaphora in Hob-Nob Sentences

Takeo Kurafuji

University of the Ryukyus, Faculty of Education,
1 Senbaru, Nishihara-cho, Okinawa 903-0213, Japan
kurafuji@edu.u-ryukyu.ac.jp

Abstract. This paper discusses how to establish an anaphoric link between an indefinite NP and a pronoun in intensional contexts, in particular, contexts where different belief-holders are concerned, such as Hob-Nob sentences. It is shown that anaphoric relations in Hob-Nob sentences cannot be accounted for by dynamic binding or the E-type strategy, based on the evidence from the differences in distribution between overt and null pronouns in Japanese. To account for such cases, it is claimed, following Porter [15], that pronouns in intensional contexts should be treated as demonstratives and argued that as far as "the mode of presentations" of the antecedent NP and the pronoun anaphoric to it are consistent, the anaphoric relation between them can be established.

1 Introduction

The semantics of pronouns has long been one of the central issues in linguistics as well as philosophy of language. In the formal linguistics, there are two streams of investigation of pronouns, depending on the type of contexts where pronouns are used. Pronouns which are c-commanded by their antecedents have been analyzed from syntactic perspectives. In the formal semantics literature, on the other hand, much attention has been paid to pronouns anaphoric to non-c-commanding, non-referential, antecedents, which are known as donkey pronouns. Recently two approaches have been proposed to account for donkey pronouns: dynamic binding and the E-type strategy, and exciting issues on which is correct or whether we need both are going on.

Among various types of contexts where donkey pronouns are acceptable, this paper will investigate intensional contexts, in particular, cases where different attitude holders are involved, as in (1), which are often referred to as Hob-Nob sentences.

(1) Hob thinks a witch$_2$ blighted Bob's mare, and Nob thinks she$_2$ killed Cob's sow.

In (1), the pronoun in the second sentence can be anaphorically linked to the indefinite NP in the first sentence. The sequence of sentences in (1) can be felicitously used in several contexts, and the one which we are interested in is a scenario like (2).

(2) The Gotham City newspaper have reported that a witch, referred to as "Samantha", has been on quite a rampage. According to the article she has

A. Sakurai et al. (Eds.): JSAI 2003/2004, LNAI 3609, pp. 427 – 437, 2007.

been blighting farm animals and crops and throwing people down wells. In reality, there is no such person: the animals and crops all died of natural causes, and the people found at the well-bottoms had all stumbled in by accident in a drunken stupor. The news reporters simply assumed that a witch was responsible for all the mishaps, and dubbed her "Samantha." Hob and Nob both read the Gotham Star and, like most folks, they believe the stories about the witch. Hob thinks Samantha must have blighted Bob's mare, which took ill yesterday. Nob thinks Samantha killed his friend Cob's sow. (Edelberg [4], p2)

In this situation, the possible paraphrase of the second sentence of (1) is "Nob thinks that the same witch killed Cob's sow.' As Edelberg points out, the important fact is that the anaphoric link in this interpretation can be established even if there is no witch in the world of evaluation and furthermore Nob knows nothing about Hob or Bob's mare. As will be shown shortly, neither dynamic binding nor the E-type strategy can obtain the intended reading.

The present paper shows, based on the Japanese data, that neither dynamic binding nor the E-type strategy can account for Hob-Nob sentences, and claims, following Portner [15], that pronouns in intensional contexts should be treated as demonstratives, arguing that anaphoric links between pronouns and their antecedents are established as far as 'the modes of presentation' are consistent.

2 Dynamic Binding and the E-Type Strategy

This section discusses whether Hob-Nob sentences can be accounted for by dynamic binding and the E-type strategy. In this paper, I adopt Chierhcia's [2] version of dynamic binding and the E-type strategy. Chierchia's claim is that grammar has both dynamic binding and the E-type strategy, and they are complementary to each other.

Dynamic binding is roughly schematized as in (3), where the scope of the existential quantifier introduced in the first sentence goes beyond the sentence boundary, and binds the variable in the following sentence. Dynamic binding is thus possible only when the existential quantifier has the widest scope over the sentence in which it is.

(3) $\exists x[P(x)] \wedge Q(x) \implies \exists x[P(x) \wedge Q(x)]$

There are, however, cases where pronouns can be liked to the antecedents which do not have the widest scope. Well-known examples are paycheck sentences and bathroom sentences as given in (4) and (5), respectively.

(4) Everyone but John gave his paycheck to his wife. John gave it to his mistress.

(5) This building doesn't have a bathroom, or it's in a funny place.

Chierchia [2] argues that pronouns in these cases are interpreted via the E-type strategy, where a pronoun is interpreted as the most salient function in the given context. In (4), the pronoun is interpreted as a function from individuals into individuals, and the pronoun in (5) is interpreted as a function from the buildings into the bathrooms located in those buildings.

Now let us consider whether Hob-Nob sentences like (1) can be accounted for by dynamic binding or the E-type strategy. It is clear that dynamic binding cannot apply to (1), since the indefinite NP *a witch* is under the scope of the propositional attitude verb *think*. On the other hand, it is not clear how the E-type strategy applies to it. Suppose the most salient function in (1) is the one from individuals into the witches that those individuals think exist. Then there are two possibilities to be considered. One is the case where that function takes 'Hob.' In this case, the second sentence of (1) is interpreted as: Nob thinks that the witch that Hob thinks exists killed Cob's sow. This is not the interpretation that we want, since Nob doesn't know what Hob thinks exists. The other possibility is that the function takes 'Nob', and returns 'the witch that Nob thinks exists.' But nothing guarantees that the witch that Nob thinks exists is identical to the one that Hob thinks exists. Again, this is not the reading we want, and it can be concluded that the E-type strategy can't account for the reading obtained in (1).

Japanese provides interesting data that forces us to consider a third way of establishing anaphoric links in intensional contexts. The distributional differences between overt and null pronouns in Japanese roughly correspond to the division of labor between dynamic binding and the E-type strategy proposed by Chierchia. Generally, overt and null pronouns are interchangeable in Japanese, but in the contexts corresponding to (4) and (5), namely the contexts where dynamic binding is not available, null pronouns are perfectly OK, while the overt pronouns are less acceptable, if not totally ungrammatical.[1,2]

(6) John igai-no daremo-ga jibun-no kurejittokaado-o tsuma-ni
except-Gen everyone-Nom self-Gen credit.card-Acc wife-to
watashi-ta. John-wa **ø**/??**sore**-o aijin-ni watashi-ta.
give-Past -Top it-Acc mistress give-Past
'Everyone but John gave a credit card of his to his wife. John gave one of his to his mistress.'

(7) Kono tatemono-ni toire-ga na-i ka, **ø**/??**sore**-ga
this building-in bathroom-Nom Neg-Pres or it-Nom
henna tokoro-ni a-ru ka-no dochira-ka-dea-ru.
funny place-in exist-Pres or-Cop which-Q-Cop-Pres
'It is the case either that this building does not have a bathroom or that it is in a funny place.'

[1] Abbreviations; Top: topic, Nom: nominative, Acc: accusative, Gen: genitive, Pres: present tense, Past: past tense, Prog: progressive, Cop: copula, Q: question marker, Neg: negation, Comp: complementizer.

[2] For some speakers of Japanese, there is no difference between null and overt pronouns in these contexts. See Kurafuji [12] for an analysis of this type of judgment. The important point is, however, that there is no informant for whom overt pronouns are more acceptable than null pronouns.

The contrast between overt and null pronouns observed in (6) and (7) suggests that overt pronouns cannot be used when dynamic binding is not available. Based on these facts as well as many other pieces of evidence, Kurafuji [12] claims that overt pronouns in Japanese donkey contexts must be dynamically bound.

Interestingly enough, overt pronouns can be used in Japanese Hob-Nob sentences, a version of which is given in (8).

(8) John-wa [Mulder-ga eirian-o tsukamae-ta to] omot-tei-ru.
-Top -Nom alien-Acc catch-Past Comp think-Prog-Pres
Mary-wa [Scully-ga **sore**-o uchikoroshi-ta to] omot-tei-ru.
-Top -Nom it-Acc shoot.dead-Past Comp think-Prog-Pres
'John thinks that Mulder caught an alien. Mary thinks that Scully shot it dead.

The overt pronoun *sore* can be interpreted as 'the same alien that John thinks Mulder caught', but it can be felicitously used in the context where Mary doesn't know anything about John or Mulder, and no alien exists at all, just like (1). As briefly discussed above, Japanese donkey pronouns must be dynamically bound if it is overt. Hob-Nob sentences are contexts where dynamic binding is excluded, but nevertheless the overt pronoun is acceptable in (8). It has been also shown that the E-type strategy cannot account for the anaphoric link in Hob-Nob sentences. The question is then how the anaphora in Hob-Nob sentences should be analyzed without using dynamic binding or the E-type strategy.

3 The Semantics of Propositional Attitudes

Before discussing the demonstrative analysis of anaphora in intensional contexts, I introduce the semantics of propositional attitudes. In the traditional analysis such as Hinttika [8], the truth conditions of a sentence like *John believes that* ϕ are represented as in (9).

(9) $\|$John believes that $\phi\|$ is true in w iff $\|\phi\|$ is true in John's doxastic alternatives.

John's doxastic alternatives relative to w are the set of propositions given by a doxastic modal base. In this traditional view, propositional attitude verbs like *believe* are regarded as relations between attitude holders and propositions.

Quine [16], Kaplan [10], Lewis [14], and Cresswell and von Stechow [3] argue against such a traditional treatment of propositional attitudes, pointing out the problem concerning *de re* belief. Let us consider Quine's famous Ortcutt-example given in (10).

(10) Ralph believes that Ortcutt is a spy.

The scenario goes as follows. Ralph has glimpsed Ortcutt in a brown hat and he believes that the man in a brown hat that he has glimpsed is a spy. He also has glimpsed Ortcutt in a gray hat and he doesn't believe that the man in a gray hat that he glimpsed is a spy. We can felicitously report this situation by uttering (10) in spite of the fact that

Ralph believes both that Ortcutt is a spy and that Ortcutt is not a spy. In the traditional analysis, (10) is incorrectly interpreted as false, since there is no (John's belief) world in which both *Ortcutt is a spy* and *Ortcutt is not a spy* are true.

The above mentioned authors' approach to this problem is as follows. First, propositional attitude verbs are not relations between attitude holders and propositions, but rather they are relations between attitude holders, *res*, and properties. (10) is thus understood as in (11).

(11) Ralph ascribes to Ortcutt the spy-property.

Second, there is a special relation between attitude holders and *res*. Following Lewis [14], Cresswell and von Stechow [3] characterize this relation as follows.

> ... if *a* is to have a belief about *b* then *b* must stand in a relation to *a* which puts him into cognitive contact with *a*. We shall say (using the terminology of David Lewis in [14]) that such a relation ξ is *suitable* for a person *a* and a *res b* in a world w.

The suitable relation is also called an acquaintance relation. In the scenario for (10), there are two acquaintance relations, R and R'; R is the relation 'x glimpsed y in a brown hat on a certain occasion O,' and R' is the relation 'x glimpsed y in a gray hat on a certain other occasion O'.' Due to the different relations, for Ralph, the man in a brown hat is a different individual than the man in a gray hat, and hence his belief worlds contain no contradiction.

Such an analysis of propositional attitudes is sometimes called the *De Se* analysis, and it will play a crucial role in Portner's approach in accounting for anaphoric links in intensional contexts.

4 Hob-Nob Anaphoras as Demonstratives

In this section, I will outline Portner's [15] approach to anaphoric links in intensional contexts. He claims that pronouns in these contexts should be treated as referential. So first we will introduce the semantics of demonstratives that he adopts, and then we will see how it will be related to the analysis of pronouns in intensional contexts.

4.1 Demonstratives

The referent of a demonstrative expression is determined in several ways. For example, in cases like (12), the referent of a demonstrative is provided by a pointing gesture by the speaker, which is, probably, the easiest case.

(12) [That dog]$_2$ eats apples.

Now suppose that *that dog*$_2$ refers to Shelby. Suppose further that Carol is giving some food to Shelby. Then I can say (13).

(13) She$_3$ is giving him some apples.

In this case, no demonstrating gesture like pointing is necessary. If we are acquainted with her by the appearance she presents as she gives Shelby some food, we can link *she₃* to Carol. Following Heim [7], Portner calls a mode of presentation 'a guise', and he labels the pointing gesture in (12) **can** and the mode of presentation in (13) **car**. The relation between the speaker s, a guise, and the reference is represented by a function F as given in (14), and as Portner notes, this is essentially Kaplan's [10] semantics of demonstratives.

(14) a. F(s)(**can**) = Shelby
b. F(s)(**car**) = Carol

Based on the idea that indices are the bearers of reference, he further introduces a function H, a function from indices onto guises. So, H(2) = **can** and H(3) = **car**. With F and H, the assignment function g is defined as in (15), where g maps indices of directly referential terms onto individuals.

(15) If α_i is a directly referential term, then g(i) = F(s)(H(i))

4.2 Acquaintance Relations as Guises

The semantics of demonstratives given in the last subsection is assimilated to the acquaintance relation between the content of attitude and the attitude holder. In order to understand this, let us begin with the simple *de re* reading of (16).

(16) John believes that $Mary_2$ walked in.

As introduced above, in Quine-Kaplan-Lewis-Cresswell&von Stechow's analysis of *de re* readings, (16) is paraphrased as follows: There is an acquaintance relation R such that (i) John bears R to Mary, and (ii) John believes that whoever he bears R to walked in. In Portner's [15] approach, the truth conditions of (16) are represented as in (17).

(17) $\|believe'(John, Mary_2, \lambda x[x\ walk\ in])\|^{w, g}$ is true iff for some acquaintance relationship R,
$R(\|John\|^{w, g}, \|Mary_2\|^{w, g})$, and
for all $y \in Dox(\|John\|^{w, g})$, $\|\lambda x[x\ walk\ in]\|^{w_y, g}(\iota z(R(y, z)))$.

He follows Lewis's [13] theory, in which individuals are treated as existing in only one world and correspondences between individuals are established by counterpart relations. The accessibility relation Dox in (17) is one of the counterpart relations. John exists only in the world of evaluation w, while in worlds other than w, John's counterparts y exist. 'w_y' stands for the world in which y exists.

The truth conditions of a belief sentence with more than one relation like *John believes $Mary_2$ introduced $Bill_3$ to Sue_4* are generalized as in (18) (cf. Cresswell and von Stechow [3]).

(18) $\|believe'(a, <b_1, ..., b_n>, \lambda x_1 ...\lambda x_n[\phi])\|^{w, g}$ is true iff for acquaintance relationships $R_1, ..., R_n$,
$R_1(\|a\|^{w, g}, \|b_1\|^{w, g}), ... ,$ and $R_n(\|a\|^{w, g}, \|b_n\|^{w, g})$, and
for all $y \in Dox(\|a\|^{w, g})$, $\|\lambda x_1 ...\lambda x_n[\phi]\|^{wy, g} (\iota z(R_n(y, z)))...(\iota z(R_1(y, z)))$.

Instead of n-many relations as in (18), we can assume a single acquaintance relation, which applies to indices and individuals (= attitude holders) and returns individuals. For example, R(John)(2) = Mary, R(John)(3) = Bill, R(John)(4) = Sue. This is generalized as in (19).

(19) $\|believe'(a, <b_1, ..., b_n>, \lambda x_1 ...\lambda x_n[\phi])\|^{w, g}$ is true iff for some acquaintance relationship R,
$R(\|a\|^{w, g})(1) = \|b_1\|^{w, g}, ...,$ and $R(\|a\|^{w, g})(n) = \|b_n\|^{w, g}$, and
for all $y \in Dox(\|a\|^{w, g})$, $\|\lambda x_1 ...\lambda x_n[\phi]\|^{wy, g} (R(y)(n))...(R(y)(1))$.

Suppose that John is acquainted with Mary by seeing her on his left. Then the function R means that John stands in the 'seen-on-his-left' relation to Mary. This is exactly the same as a guise. Let us call Mary's guise **mar**. Then F(John)(**mar**) = R(John)(2) = Mary. So we get (20) (Portner's (84)).

(20) For any individual y and index i, F(y)(H(i)) = R(y)(i).

R is now defined in terms of F and H as in (21).

(21) $R =_{def} \lambda y\lambda i[F(y)(H(i))]$

With R(y), (19) is represented as in (22), where the indexed variables in $\|\phi\|$ are interpreted with respect to R(y).

(22) $\|believe'(a, <b_1, ..., b_n>, \lambda x_1 ...\lambda x_n[\phi])\|^{w, g}$ is true iff for some acquaintance relationship R,
$R(\|a\|^{w, g})(1) = \|b_1\|^{w, g}, ... ,$ and $R(\|a\|^{w, g})(n) = \|b_n\|^{w, g}$, and
for all $y \in Dox(\|a\|^{w, g})$, $\| \phi \|^{wy, R(y)}$.

Now let us consider a sequence of sentences like (23).

(23) John believed that [a woman]$_1$ was in the room. He believed that she$_1$ was happy.

The *de re* analysis in (22) seems not to apply to (23), since there is no unicorn with which John is acquainted. But Portner [15] argues that all of John's doxastic counterparts are acquainted with unicorns. Suppose John looked at a sign on the door saying "Occupied" and guessed that there was a woman inside but actually there was nobody inside. In this situation, the first sentence of (23) is interpreted as true, and the indefinite NP receives the *de dicto*-reading. Portner suggests that the sentence should be analyzed as *de re* about the room and proposes that the women in their respective worlds are presented to John's counterparts under the guise 'in the room whose door I am looking at.' Let us call this guise **wom**, then the indefinite NP is interpreted as follows.

(24) For any individual y, F(y)(**wom**) is the individual in the room whose door y is looking at.

The truth conditions of the first sentence of (23) is represented as in (25), which says that the first sentence of (23) is true iff all of John's doxastic counterparts looked at the door of the room, in which a woman x_1 existed.

(25) $\|$believe'(John, [woman'(x_1) $\wedge$ in-the-room'(x_1)]$\|^{w,\, g}$ is true iff for some acquaintance relationship R, R(y)(i) = the thing presented to y under the i-th guise provided in w, and
for all y $\in$ Dox($\|$John$\|^{w,\, g}$), $\|$woman'(x_1) $\wedge$ in-the-room'(x_1)$\|^{wy,\, R(y)}$.

Contrary to (22), (25) does not have a condition like R($\|$John$\|^{w,g}$)(1) = $\|$a woman$\|^{w,g}$, since there is no woman. Instead, R encodes John's belief as guise like the one in (24).[3]

The truth conditions of the second sentence of (23) are given in (26).

(26) $\|$believe'(John, happy'(x_1)$\|^{w,\, g}$ is true iff for some acquaintance relationship R, R(y)(i) = the thing presented to y under the i-th guise provided in w, and
for all y $\in$ Dox($\|$John$\|^{w,\, g}$), $\|$happy'(x_1)$\|^{wy,\, R(y)}$.

According to the truth conditions in (26), the second sentence of (26) is true iff all of John's doxastic counterparts looked at the door of the room, in which the same woman was happy.

The intuition behind this analysis is that as long as an attitude holder keeps his attitude to a proposition constant, the acquaintance relations in the first sentence and that in the second sentence remain the same, and hence the pronoun refers to the same individual denoted by the antecedent.

4.3 Hob-Nob Sentences

The apparent difficulty of applying Portner's approach to Hob-Nob sentences like (1) comes from the fact that different attitude holders are concerned. In order to obtain the identity of the witch in question, the equality between (27a) and (27b) must be guaranteed.

(27) a. F(Hob)(H(2))
b. F(Nob)(H(2))

[3] Norihiro Ogata (personal communication) suggested to me that the epsilon operator be used for the analysis of indefinite NPs in intensional contexts. This alternative sounds very attractive, but I would like to leave it open which approach is better.

Let us consider if the same guise is provided to Hob and Nob. As described in the scenario in (2), they got acquainted with Samantha by reading *The Gotham Star*. This means that reading the paper serves as a guise. Let us call this guise **sam**. The function in (27) is thus represented as in (28).

(28) a. F(Hob)(**sam**)
b. F(Nob)(**sam**)

Recall that guises are modes of presenting individuals to attitude holders. So, if two attitude holders are presented something in the same mode, then the presented objects should be identical. Hence, in general, (29) holds (cf. (24)).

(29) For any individual y, F(y)(**sam**) is the individual whose article in the Gotham Star was read by y.

Now, the truth conditions of (1) are straightforward, as given in (30) for the first sentence and (31) for the second.

(30) $||$think'(Hob, [witch'(x_2) $\wedge$ blight'(Bob's mare)(x_2)]$||^{w, g}$ is true iff for some acquaintance relationship R, R(y)(i) = the thing presented to y under the i-th guise provided in w, and
for all y $\in$ Dox($||$Hob$||^{w, g}$), $||$witch'(x_2) $\wedge$ blight'(Bob's mare)(x_2)$||^{wy, R(y)}$.

(31) $||$think'(Nob, [kill'(Cob's sow)(x_2)]$||^{w, g}$ is true iff for some acquaintance relationship R, R(y)(i) = the thing presented to y under the i-th guise provided in w, and
for all y $\in$ Dox($||$Nob$||^{w, g}$), $||$kill'(Cob's sow)(x_2)$||^{wy, R(y)}$.

When the acquaintance relation R in (30) is the same as that in (31), the individual with index 2 in (30) is identical to that in (31). In other words, the Hob-Nob type anaphoric dependency is acceptable only when we can set up a context that is specific enough to provide the identical acquaintance relationship to different attitude holders, as the scenario in (2). With an identical acquaintance relationship, the referent of the indefinite noun phrase for one attitude holder can be identical to the referent of the pronoun for the other. Like the *imagine* case, this analysis regards *a witch* and *she* as referential as if such an individual actually existed.

5 Consequences

Japanese overt pronouns discussed in this paper such as *sore* are called *So*-series demonstratives and analyzed as part of the demonstrative paradigm in Japanese traditional grammar. This fact suggests that Portner's analysis is on the right truck. And the anti-E-type property of overt pronouns observed in (7) and (8) should be related to the fact that they are *So*-series demonstratives. As shown in (32) and (33), English demonstratives are also not acceptable in paycheck or bathroom sentences.

(32) Every man but John gave his $paycheck_i$ to his mistress.
*John gave that $paycheck_i$ /$that_i$ to his mistress.

(33) *This building doesn't have a $bathroom_i$ or that $bathroom_i$ /$that_i$ is in a funny place.

These data show that demonstratives are subject to accessibility conditions. The ungrammaticality of (32)-(33) can be accounted for by assuming that demonstratives are identity maps, which are represented as follows.

(34) a. that $paycheck_i$ ==> $\iota y\ [paycheck'(y) \wedge y = x_i]$
b. $that_i$ ==> $\iota y\ [C(y) \wedge y = x_i]$
'C' is some contextually specified property.

A demonstrative is interpretable if x_i is dynamically bound. If unbound, it is not interpreted. In (32) and (33) dynamic binding is not available, so the demonstratives in these examples are not interpreted. The accessibility requirement of demonstratives thus can be accounted for. An indexical interpretation of the demonstrative obtains if x_i is associated with an act of demonstration like pointing. With Portner's notion, this is represented as in (35).

(35) $\iota y\ [paycheck'(y) \wedge y = F(s)(H(i))]$,
where H is a function from indices onto guises.

However, no appropriate guise is supplied in the contexts of (32) and (33), which also accounts for the ungrammaticality of these sentences. Thus the translations given in (35) account for the ungrammaticality of demonstratives in paycheck sentences and bathroom sentences both in English and Japanese.

Acknowledgement

I would like to thank the participants at the international workshop of Logic and Engineering of Natural Language Semantics, in particular, Norihiro Ogata and Yasuo Nakayama for their helpful comments and suggestions. I am also grateful to an anonymous reviewer for informative suggestions. All errors are mine.

References

1. Asher, N.: Reference to Abstract to Objects in Discourse. Kluwer Academic (1993).
2. Chierchia, G.: Dynamics of Meaning: Anaphora, Presupposition, and the Theory of Grammar. The University of Chicago Press, Chicago (1995).
3. Cresswell, M J., von Stechow. A.: *De re* Belief Generalized. Linguistics and Philosophy 5 (1982) 503-535.

4. Edelberg, W.: A New Puzzle about Intentional Identity. Journal of Philosophical Logic 15 (1986) 1-25.
5. Geach, P.: Intensional Identity. Journal of Philosophy 64 (1967) 627-632
6. Heim, I.: Presupposition Projection and the Semantics of Attitude Verbs. Journal of Semantics 9 (1992) 183-221.
7. Heim, I.: Anaphora and Semantic Interpretation: A Reinterpretation of Reinhart's Approach. In: Sauerland. U., and Percus, O. (eds.): The Interpretive Tract, MIT Working Papers in Linguistics 25. MIT, Cambridge (1998) 205-246.
8. Hintikka, J.: Knowledge and Belief. Cornell University Press, Ithaca (1962).
9. Kamp, H.: Prolegomena to a Structural Theory of Belief and Other Attitudes. In C. A. Anderson, C. and Owens, J. (eds.) The Role of Content in Language, Logic and Mind, CSLI (1990) 27-91.
10. Kaplan, D.: Quantifying In. In Davidson, D. and Hintikka, J. (eds.) Words and Objections. Reidel, Dordrecht (1969).
11. Kaplan, D.: Demonstratives: An Essay on the Semantics, Logic, Metaphysics, and Epistemology of Demonstratives and Other Indexicals. (1977): also reprinted in Harnish R., (ed.) Basic Topics in the Philosophy of Language. Harvester Wheatsheaf, London (1994) 275-319.
12. Kurafuji, T.: Japanese Pronouns in Dynamic Semantics: The Null/Overt Contrast. Ph.D. Thesis, Rutgers University (1999).
13. Lewis, D.: Counterpart Theory and Quantified Modal Logic. The Journal of Philosophy 65 (1968) 113-26.
14. Lewis, D.: Attitude De Dicto and De Se. The Philosophical Review 88 (1979) 513-543
15. Portner, P.: Modal Discourse Referents and De Se Attitudes. ms. Georgetown University (1996).
16. Quine, W.V.O.: Quantifiers and Propositional Attitudes. The Journal of Philosophy 53 (1956) 177-187.

Temporal Dynamic Semantics of Factual Counterfactuals

Sumiyo Nishiguchi

Osaka University Graduate School of Letters Department of English Linguistics
1-5 Machikaneyama, Toyonaka 560-0043, Japan
UCLA Department of Linguistics
3125 Campbell Hall, Box 951543, Los Angeles, CA 90095-1543
snishigu@ucla.edu

Abstract. This paper discusses the fake past tense morphology used for present state in Japanese (Teramura 1984; Iatridou 2000). Unlike Korean and other languages, the past tense marker "ta" can express an unexpected finding or remembrance at the time of speech (Inoue and Ubukoshi 1997). I claim that this construction corresponds to subjunctive conditionals with a covert negative antecedent, and that counterfactuality is involved in such non-past past tense, even though the proposition expressed is factual. It is comparable to the counterfactual analysis of factive emotive predicates such as "sorry"and "glad"(Heim 1992; Giorgi and Pianesi 1997). I show that non-past past construction updates the information state, and revises the context.

1 Fake Past: Unexpectancy and Remembrance

Peculiarly, in Japanese, the past tense marker "ta" may express the present tense, when used mainly with stative predicates such as "aru" or "iru" (be/exist). For example, when the speaker has been looking for today's newspaper, a possible utterance at finding it would be:

(1) Koko-ni at-ta.
here-Loc be-Past
'(It) was here'

Although the speaker is aware of the existence of the newspaper at the time of speaking, which is not in the past, the past tense marker is more appropriate than the present tense in this situation. The literal meaning which (1) expresses is (2), not (3):

(2) $\exists x \exists e \exists t' [newspaper'(x) \wedge be-here'(x,e,t') \wedge t' \supseteq t]$
(where t is the speech time)

(3) $\exists x \exists e' \exists t' [newspaper'(x) \wedge be-here'(x,e',t') \wedge t' \prec t]$

A. Sakurai et al. (Eds.): JSAI 2003/2004, LNAI 3609, pp. 438–448, 2007.

What the speaker expresses is that the presense of the newspaper includes the speech time.

Therefore, the interpretation of the non-past past tense is, following Enç's (1996) descriptions:

(4) Where w is a possible world and I, I' are intervals, the denotation of $NON-PAST\ PAST$ at $\langle w, I\rangle$ is an interval I' such that $I' = I$.

Why do Japanese speakers use the fake past tense, instead of the present tense (5), although the presence of the newspaper belongs to the speech time (Iatridou 2000)?

(5) Koko-ni aru.
here-Loc be.
'It is here'

(5) is not used in the above described situation, since the present tense does not express unexpectancy. This "ta" has been called '"ta" of finding' by Japanese traditional grammarians (Mikami 1972, Machida 1989, among others). In contrast, in English, the past tense in the above context is also possible but better with "all along" (6a), and the present tense equally expresses surprise (6b):

(6) a. It was here (all along).
b. It is here.

Another example with the fake past in Japanese would be:

(7) Kyo-wa tanjobi dat-ta.
today-Top birthday be-Past
'Today was my birthday'

The context of (7) is that e.g., the speaker has forgotten his birthday until he looks at the calendar. Although his birthday is the same as the time of speaking, the past marker "ta" is used. This "ta" has been called '"ta" of remembrance.' What is the difference between (7) and its present counterpart (8)?

(8) Kyo-wa tanjobi da.
today-Top birthday be
'Today is my birthday'

(8) is only a statement of the truth that the day of the speech time is the birthday. It does not indicate that the speaker has realized this. The characteristics of these past sentences, which are actually non-past, is that the speaker has not known the event attached "ta" until the speech time. "I did not know that there was a newspaper on the table" is the presupposition for (1) (Machida 1989).

Another example would be:

(9) Jerry-wa sutaffu dat-ta.
Jerry-Top staff be-Past
'Jerry was a member of staff'

"Staffu dat-ta" in (9) does not mean that Jerry's being a member of the staff was in the past. Rather, it is a fact which the speaker had not been aware of until the speech time.

This peculiar "ta" is not limited to stative predicates. However, the fake "ta" in non-stative predicates does not necessarily express surprise or unexpectancy:

(10) A, basu-ga ki-ta.
Oh, bus-Nom come-Past
'Oh, the bus came'

Unless there is any emphatic stress or an exclamation mark at the end of the sentence, (10) does not indicate surprise. The speaker has expected that the bus was to come according to the bus schedule. (10) does not express any more surprise or unexpectancy than the corresponding present tense (11):

(11) A, basu-ga kuru.
Oh, bus-Nom come
'Oh, the bus comes'

In section 2.1, I claim that this fake past construction corresponds to counterfactual conditionals in other languages in the morphological behavior. This "ta" construction covertly presupposes negative conditional antecedent clause which contains the past tense morpheme - "when I have thought not p, it has been p." Section 2.2 explains that, although the proposition p is found to be factual, counterfactuality is involved. Section 2.3 points out that the degree of counterfactuality differs between stative and eventive predicates.

In section 3, we will discuss the interaction between tense and information. In particular, this non-past "ta" construction updates the information states. It is in line with Takagi who analyzes these "ta" as perfective, rather than the past tense, which denotes the transition of the viewpoint from the stage without information to the stage with information (Takagi 1993; Kinsui 2000).

2 Subjunctive Conditionals with Factual Counterfactuality

This section explains that the non-past past sentence presupposes the existence of a negative antecedent of conditionals. In other words, the Japanese fake "ta" construction has an elided *if*-clause in which the speaker holds the contrary proposition. The schema is as follows:

(12) (When I have assumed not-p), it was p all along.

2.1 Elided Subjunctive Conditionals

I claim that the fake past tense construction in Japanese has an elided antecedent of the subjunctive conditional. Although there exists no subjunctive morphology in Japanese as in other languages like German or Italian, the morphological behavior corresponds to English subjunctive in "contrary-to-fact" conditionals in Binnick's (1991) terminology.

Counterfactual conditionals have the subjunctive mood expressed with subjuctive morphology in Italian and other languages. Correspondingly, in English, when the falsity of the *if*-clause is presupposed, both the protasis and the apodosis take past tense:

(13) a. If I were rich, I'd travel around the world (but unfortunately I am not).
b. If John had arrived on time, we would have gone to cinema. (Giorgi and Pianesi 1997)

In contrast, indicative conditionals do not presuppose anything about the truth of the protasis (Giorgi and Pianesi 1997):

(14) a. If I am rich, it is because I worked hard. (modified from Binnick 1991)
b. If it rains, I will cancel the appointment.

The Japanese fake past "ta" construction also presupposes the contrary-to-fact protatis. For example, the utterance (1), "koko-ni at-ta ('it was here')," presupposes the antecedent, "although I have thought it was not here":

(15) Koko-ni-*nai*-to omot-te-i-tara, at-ta.
Here-loc-Neg-Comp think-Ger(und)-be-Cond be-Past
'Although (I) have not expected that (it) was here, (it) was'

"Kyo-wa tanjobi dat-ta ('Today was the birthday')" in (7) also presupposes that the speaker expected the contrary:

(16) Kyo-wa tanjobi-de-*nai*-to omot-te-i-tara, so dat-ta.
Today-Top birthday-be-Neg-Comp think-Ger-be-Cond so be-Past
'While I did not think it was my birthday today, it was'

The following examples further show that non-past past sentences can combine with the contrary propositions (a) because "ta" in the consequent requires unexpectancy and finding. On the other hand, the affirmative presupposition cannot be taken as an antecedent (b):

(17) a. Jerry-wa *gakusei*-ka-to omot-te-i-tara sutaffu dat-ta.
Jerry-Top student-Q-Comp think-Ger-be-Cond staff be-Past
'(I) have thought Jerry to be a student, but (he) was a staff'
b#Jerry-wa *sutaffu*-ka-to omot-te-i-tara, sutaffu dat-ta.
Jerry-Top staff-Q-Comp think-Ger-be-Cond staff be-Past
'(I) have thought Jerry to be a staff, but (he) was a staff'

(18) a. Kare-wa *oki*-te-iru-no-ka-to omot-te-i-tara, ne-te-i-ta.
he-Top awake-be-fact-Q-Comp think-Ger-be-Cond sleep-Ger-be-Past
'(I) have thought that he is awake, but (he) was sleeping'
b#Kare-wa *ne*-te-iru-no-ka-to omot-te-i-tara yappari
he-Top asleep-Ger-be-fact-Q-Comp think-Ger-be-Cond as-I-expected
ne-te-i-ta.
sleep-Ger-be-Past
'(I) thought that he was sleeping, but (he) was sleeping'

When p is already presupposed in the antecedent, the non-past p construction cannot adequately follow.

Notably, non-past past "ta" construction cannot coincide with other conditional connectives such as "to", "nara" or "ba," which do not contain the past tense morpheme:

(19) a#Kyo-wa tanjobi-de-nai-to omot-te-iru-*to*, so dat-ta.
Today-Top birthday-be-Neg-Comp think-Ger-be-Cond so be-Past
'While I did not think it was my birthday today, it was'
b.*Kyo-wa tanjobi de-nai-to omot-te-iru-*nara*, so dat-ta.
Today-Top birthday be-Neg-Comp think-Ger-be-Cond so be-Past
'While I did not think it was my birthday today, it was'
c.*Kyo-wa tanjobi-de-nai-to omot-te-ire-*ba*, so dat-ta.
Today-Top birthday-be-Neg-Comp think-Ger-be-Cond so be-Past
'While I did not think it was my birthday today, it was'

Getting rid of the past tense marker from the main clause makes (19) grammatical:

(20) a. Taro-ga kuru-*to*, Hanako-ga yorokobu.
Taro-Nom come-Cond Hanako-Nom rejoice
'If Taro comes, Hanako is happy'
b. Taro-ga kuru-*nara*, Hanako-ga yorokobu.
Taro-Nom come-Cond Hanako-Nom rejoice
'If Taro comes, Hanako is happy'
c. Taro-ga kure-*ba*, Hanako-ga yorokobu.
Taro-Nom come-Cond Hanako-Nom rejoice
'If Taro comes, Hanako is happy'

The reason why only "tara" is compatible with the "ta" construction is that "tara" contains the past morpheme "ta" in itself. Therefore, my claim is that this is the subjunctive construction "...*ta*ra, ...*ta*" with shifted tense in both the *if*-clause and the main clause. This "ta" construction corresponds to subjunctives in other languages.

(21) < ...*ta*-ra>, ...*ta*.

In the next section, I show how counterfactuality is involved with this construction, even though the proposition expressed is in fact factual.

2.2 Factual Counterfactuality

Counterfactuality is related to the Japanese fake past construction, in the sense that factive emotives also involve counterfactuality (Heim 1992; Giorgi and Pianesi 1997; von Fintel 1999). I claim that the counterfactuality shifts the tense backward, so that the factual predicates which are also grammatical with the present tense contain the past tense morpheme "ta." The analysis is based on the forward branching $T \times W model$ (Thomason 1984; Condoravdi 2002).

Even though the proposition expressed is factual, counterfactuality is related to the construction discussed above. In the examples below, as in (1) and (7), the presence of the newspaper in (22) and the fact that "today" is the speaker's birthday in (23) are factual, not counterfactual:

(22) Koko-ni at-ta.
Here-Loc be-Past
'(It) was here'

(23) Kyo-wa tanjobi dat-ta.
Today-Top birthday be-Past
'Today was my birthday'

How could the counterfactual be factual? Heim (1992) suggests that factive desire predicates such as "be glad" and "be sorry" involve counterfactuality. α *is glad that* p and α i*s sorry that* p presuppose that α believes that p. The attitude expressed by such a sentence compares the world as α believes it to be with the world that α believes it would be if p were not true (see von Fintel 1999). Thus, factive "be glad" and "be happy" possess the element of counterfactuality.

Similarly, the speaker of (22) compares the actual world with the world which he had believed to be true. The finding of the newspaper had not been expected, so the reality is contrary to his past belief. The actual world w_3 at the utterance time t_0 is foreign to the counterfactual world w_2 which was compatible to his belief at a time prior to t_0. Following Thomason's (1984) world-time model, I assume $T \times W$ frames (Condoravdi 2002). In this model, the worlds are complete histories of the actual world and the alternative worlds branched forth:

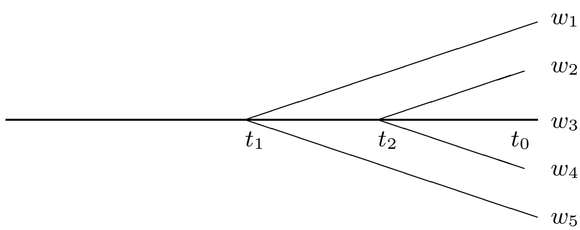


Fig. 1. $t_1 \prec t_2 \prec t_0$ and t_0 is the speech time

Suppose that the actual world is w_3 at t_0 in Fig. 1. According to the shared belief between the speaker and the hearer at t_2, the actual world at t_0 was supposed to be the counterfactual world w_2, where the newspaper is not on the table, instead of w_3 where it is on the table, as represented in (24) and (25):

(24) $PAST(BELIEVE(speaker, \sim being'(x))) \wedge$
$PRES(BELIEVE(speaker, \sim being'(x)))$

(25) $\exists x \exists t_o \exists w_3 [newspaper'(x) \wedge know'(speaker, being'(x))$ in w_3 at $t_o] \wedge$
$\forall w' \forall t' [w'$ is compatible with what the speaker believes in w_3 at $t' \wedge$
$t' \prec t_o \rightarrow believe'(speaker, \sim being'(x))$ in w' at $t']$

The speaker's perspective shifts backward to t_2 and corrects the path of the past actual world from the one leading to w_2 to the one leading to w_3 in the utterance (22). Since the present world w_3 is counterfactual to the addresser and the addressee, the speaker corrects their erroneous past belief by the past marker. Non-past past tense expresses the counterfactual factuality of the present world:

(26) Where s is a state, τ is a function yielding the temporal trace of stativity in a given world, $[t, \_)$ designates an interval with t as an initial subinterval and extending to the end of time, and $\simeq_t$ be the modal base,
$\lambda w \exists w' \exists t' [t' \prec t_0 \wedge w \simeq_{t'} w' \wedge$
$\exists s \exists x [[newspaper'(x) \wedge \sim being-here'(x))](w')(s) \wedge \tau(s, w') \subseteq [t', \_)]]$
(modified from Condoravdi 2002)

2.3 Degree of Counterfactuality

This section discusses that the degree of counterfactuality differs between stative and non-stative predicates.

(27) a. A, basu-ga ki-ta.
Oh, bus-Nom come-Past
'Oh, the bus came'
b. A, hon-ga at-ta.
Oh, book-Nom be-Past
'Oh, here was a book'

There is a significant difference between the degree of surprise of the two sentences. The speaker of (27a) is more surprised to find the book than the speaker of (27b) who had not expected the book to be there. On the contrary, in (27b), the speaker had expected the arrival of the next bus at the bus stop. The appearance of bus was presumed. With non-stative predicates like (27b), the speaker had not expected the finding of the book. In other words, the presupposed beliefs of the speakers are different:

(28) a. $HAVE(\sim BELIEVE(speaker, \exists x(bus'(x) \wedge come'(x))))$
b. $HAVE(BELIEVE(speaker, \sim \exists x(book'(x) \wedge be-here'(x))))$

The negation takes wide scope over the *BELIEVE* operator in (28a), since it is not the case that the speaker of (27a) expected the bus to come. On the other hand, negation takes narrow scope in (28b), because the speaker of (27b) had believed that the book should not have been there. In other words, it is not that the speaker of (27a) had believed that the bus would never come. If so, he would not have been waiting at the bus stop. When the speaker sees the bus, he says (27a) in order to draw the attention of the hearer to the approaching bus. However, the speaker of (27a) had a belief in the absence of the book. The speaker's disbelief in the absence of the book is much stronger than the one in the coming of the bus.

Therefore, there is less counterfactuality involved with the non-stative predicates in the fake past tense. The degree of counterfactuality differs between stative and non-stative predicates.

The "ta" of finding with stative predicates presupposes the following in (29) and (30):

(29) $PAST(BELIEVE(speaker, \sim \phi) \wedge NOW(BELIEVE(speaker, \phi))$

(30) $\forall t \forall t' (t \prec t' \wedge t' \prec now \wedge BELIEVE(speaker, \sim \phi))$ at t $\wedge$ $BELIEVE(speaker, \phi)$ at t')

3 Dynamic Semantics

This section reveals that the non-past past construction in Japanese updates the information states of the speaker and the hearer with added belief. The fake past tense changes the context, but the corresponding present tense does not.

3.1 Added Belief

Within the framework of dynamic semantics, the sentence (27b) adds the following information to the shared belief between the speaker and the hearer, while the corresponding sentence in the present tense (6) does not:

(31) $PAST(BELIEVE(speaker, being'(x)))$

As $PAST$ takes scope over $BELIEVE(speaker, being'(x))$, the perspective of the knowledge is not the time of utterance, but of the past. This is because the speaker now realizes that what he had not believed to be there had existed there. The speaker corrects the information about the past history of the world, as well as the present belief. The following information is also added:

(32) $PRESENT(KNOW(speaker, being'(x)))$

Furthermore, the speaker now knows that the counterfactual world which he had believed to be true had not been so. In line with Condoravdi (2002), MB designates the modal base a modal depends on for its interpretation. MB is taken to be a temporary determined function from world-time pairs to sets of worlds:

(33) $I = \langle W, T, \prec, g \rangle$, where the following constraints hold: W is a set of worlds. T is a set of time ordered by the relation $\prec$. For t', t in T, $t' \prec t$ is to be read "t' precedes t." g is an assignment function.
Non-past past knowledge:
$NON - PASTPAST(KNOW(speaker, \phi))$ is true at $\langle w, t \rangle$ iff there exists w' such that $w' \subseteq MB\ (w, t')$ where $t' \prec t$, and the speaker knows that ϕ is false at $\langle w', t \rangle$.

3.2 Update Semantics

Veltman's update semantics (Veltman 1996), defined as in (34)-(36), clearly distinguishes between the fake past (37a) and the present tense (37b):

(34) Update semantics for language L (Veltman 1996):
Σ: a set of relevant information states
[ ]: a function that assigns to each sentence ϕ an operation $[\phi]$ to Σ
$\langle L, \Sigma, [\,] \rangle$: update system
σ: state
ϕ: sentence
$\sigma[\phi]$: the result of updating σ with ϕ

(35) ϕ is accepted in σ, i.e., $\sigma[\phi] = \sigma \Leftrightarrow \sigma \Vdash \phi$

(36) σ is strengthened, i.e., $\sigma \leq \sigma[\phi] \Leftrightarrow \sigma \nVdash \phi$

(37) a. Koko-ni aru.
here-Loc be
'(It) is here'
b. Koko-ni at-ta.
Here-Loc be-Past
'(It) is here'

In (37a), the following condition holds, i.e., the input states of (37a) are accepted:

(38) $\sigma[PRES(KNOW(speaker, being(x)))] = \sigma$

In (37b), the following condition holds, i.e., the input states of (37b) are strengthened:

(39) $\sigma \leq \sigma[PAST(KNOW(speaker, being'(x)))]$

The existence of the newspaper is the new information, because the speaker of (37b) had not known the existence of the newspaper until the time of the utterance. On the contrary, (37a) does not add any information to the speaker's knowledge.

3.3 Context Change Potential

Here we adopt the view that the meaning of a sentence is its context change potential (CCP) (Heim 1992). The CCP of the fake past morphology revises the context, while the corresponding present tense does not. The fake past morpheme shows that the actual world was not one of the doxastically accessible worlds until the speaker revised the context (i.e., the state of information) when finding the truthfulness of the proposition of the non-past past sentence.

Dox is a function from worlds to sets of worlds, which conforms to the speaker a's belief in w at time t':

(40) For any $w \in W, Dox(a, w, t') = \{w' \in W : w'$ conforms to what a believes in w at $t'\}$ (modified from Heim 1992)

Prior to the time of utterance, the interpretation of ϕ had not been a subset of the doxastically accessible worlds for the speaker a:

(41) Where $t' \prec now, [\![\phi]\!] \not\subseteq Dox(a, w, t')$

However, at t'', the contexts are revised:

(42) Let $\langle W, T, \prec, Dox, A \rangle$ be a model where W is a set of possible worlds, T a set of time, $\prec \subseteq T \times T$, $Dox : A \times W \times T \rightarrow pow(W)$, $a \in A$, the speaker, $t_0 \in T$, the current time, $c \subseteq W \times T \times G$, and G a set of variable assignments:
$[\![ta]\!]^c = c(w, t_0, g) \in Dox(a, w, t')$ is false at t' where $t' \prec t_0$.
Where $rev(c)$ is the revision of c, $w' \in Dox(a, w, t'')$, for all t'' such that $t' \prec t'' \prec t_0$, $rev(c) \in Dox(a, w, t'')$

In contrast, the interpretation of ϕ with present tense has been a subset of the doxastic worlds since t'. Thus, it does not revise the context.

4 Conclusion

The present paper presented an analysis of the non-past past construction as a subjunctive construction. It presupposes an elided antecedent clause, whose proposition is contrary to the main clause. Even though the proposition expressed by this type of sentence is factual, counterfactuality motivates the speaker to use the past marker in order to express remoteness from the assumed present actual world. The past tense updates the information state and revises the context, while the corresponding present sentence does not.

Acknowledgement

I thank Norihiro Ogata, Philippe Schlenker, Terry Parsons, Paul Elbourne and the audience at UCLA for valuable comments and suggestions on the earlier versions of this paper. I also thank Paul A. Harvey and Ben Keil for stylistic improvement, and Andrei Antonenko for the help with LaTeX.

References

Binnick, Robert I.: Time and the Verb: A Guide to Tense and Aspect. Oxford University Press, New York Oxford (1991)

Condoravdi, Cleo: Temporal Interpretation of Modals: Modals for the Present and for the Past. In: D. I. Beaver et al. (eds.): The Construction of Meaning. CSLI (2002) 59-87

Enç, Murvet: Tense and Modality. In: Shalom Lappin (ed.): The Handbook of Contemporary Semantic Theory. Blackwell, Oxford (1996) 345-358

Farkas, Donka: On the Semantics of Subjunctive Complements. In: P. Hirschbuhler (ed.) Romance Languages and Modern Linguistic Theory. John Benjamins, Amsterdam (1992) 69-105

von Fintel, Kai: NPI Licensing, Strawson Entailment, and Context Dependency. In: Journal of Semantics 16 (1999) 97-148

Giorgi, Alessandra and Fabio Pianesi, Tense and Aspect: From Semantics to Morphosyntax. Oxford University Press, New York Oxford (1997)

Groenendijk, Jeroen and Martin Stokhof: Dynamic Predicate Logic. In: Linguistics and Philosophy 14 (1991) 39-100

Heim, Irene: Presupposition Projection and the Semantics of Attitude Verbs. In: Journal of Semantics 9 (1992) 183-221

Iatridou, Sabine: The Grammatical Ingredients of Counterfactuality. In: Linguistic Inquiry 31 (2000) 231-270

Inoue, Masaru: Gendai Nihongo no Ta, “Ta” no Gengogaku, Hitsuji Shobo. Tokyo (2001) 97-163

Inoue, Masaru and Ubukoshi Naoki: Kakokei no Shiyo ni Kakawaru Goyoronteki Yoin: Nihongo to Chosengo no Baai. In: Kokuritsu Kokugo Kenkyujo (ed.): Nihongo Kagaku 1. Tosho Kanko Kai, Tokyo (1997)

Kinsui, Satoshi: Toki no Hyogen. In: Toki, Hitei to Toritate. Iwanami, Tokyo (2000) 3-92

Machida, Ken: Nihongo no Jisei to Asupekuto. Arc, Tokyo (1989)

Mikami, Akira: Gendai Goho Josetsu. Kuroshio, Tokyo (1972)

Mo, Sejong: Hakken, Omoidashi ni okeru ‘ru’ kei to ‘ta’ kei. In: Nihongogaku 12 (1992) 88-97

Stalnaker, Robert: A Theory of Conditionals. In: Studies in Logical Theory, American Philosophical Quarterly, Monograph, 2 (1968)

Takagi, Kazuhiro: Ninshiki to Hatsuwa no Katei o Koryo shita Imi Kijutsu no Kokoromi: Nihongo no Bunmatsu Hyogen o Rei ni. M.A. thesis. Kobe University of Foreign Studies (1993)

Takubo, Yukinori: Danwa Kanri Riron-niyoru Nihongo-no Hanjijitsu Joken Bun. In: Takashi Masuoka (ed.): Nihongo no Joken Hyogen. Kuroshio, Tokyo (1993)

Teramura, Hideo: Nihongo no Shintakusu to Imi 2. Kuroshio, Tokyo (1984)

Thomason, Richmond H.: Combinations of Tense and Modality. In: Dov Gabbay and Franz Guenthner (eds.): Handbook of Philosophical Logic: Extensions of Classical Logic. Reidel, Dordrecht (1984) 135-165

Veltman, Frank: Defaults in Update Semantics. In: Journal of Philosophical Logic 25 (1996) 221-261

English Present Perfect Revisited: A Unified Semantics as a Tense and a Lower-Level Ambiguity Represented in DRT*

Yoko Mizuta

National Institute of Informatics, Research Information Research Division
2-1-2 Hitotsubashi, Chiyoda-ku, Tokyo 101-8430, Japan
ymizuta@nii.ac.jp

Abstract. In this paper, I propose a systematic account of the semantics of English present perfect (PresPerf) from both empirical and formal perspectives. I incorporate the insights of an extended now (XN) analysis ([1]) into the dynamic semantics framework, DRT ([2]). I demonstrate that PresPerf is licensed by its relation to an XN interval and that different readings are attributed to the specific way that the embedded eventuality interacts with an XN. I provide the semantics of PresPerf compositionally in terms of a general component attributed to *have* + *en* and specific conditions attributed to adverbials and verb semantics. In addition, I characterize PresPerf as a tense by showing that the XN interval serves as "temporal location" ([3]), a determinant of tense. An alleged connection between PresPerf and a perfect state is only entailed in the case of events.

1 Introduction

English present perfect (PresPerf) has been the focus of a great deal of discussion over the years, with different views being presented. Yet the following major issues still remain controvertial: 1) the semantics of different readings and their relation to each other, and 2) the status of the construction in the English tense/aspect system. I will approach these issues from both empirical and formal perspectives and shed new light on the semantics of PresPerf.

The following sentences illustrate major types of reading of PresPerf.

(1) a. John *has lost* his keys.
    b. John *has left*.

(2) John *has lived* in NY before.

* I am grateful to the following people for helpful comments on earlier versions of this paper; Anastasia Giannakidou (University of Chicago), Norihiro Ogata (Osaka University), Akira Ishikawa (Sophia University), the audience of the LENLS workshop (May 31, 2004, Kanazawa, Japan) and of Sophia University Linguistics Colloquium (July 15, 2004, Tokyo, Japan), Katie Gruber (University of Chicago), and an anonymous reviewer in the LNCS committee. Shortcomings are of course solely mine.

A. Sakurai et al. (Eds.): JSAI 2003/2004, LNAI 3609, pp. 449–459, 2007.

(3) a. John *has lived* in NY since 2000.
b. We*'ve known* each other for a long time.

Sentences (1a) - (1b) yield the resultative reading, (2) the experiential reading, and (3a) - (3b) the continuative reading. I will refer to the continuative reading as the U- reading ("universal reading") and other readings as E-reading ("existential reading").[1] The U-reading occurs with stative verbs or a stative version of eventive verbs, but not vice versa.

(4) John *has lived* in NY for three years.

For example, (4) is ambiguous between the E-reading (i.e. John once lived in NY for three years) and the U-reading (i.e. John lives in NY, and the duration of his stay there now amounts to three years). A proper account of PresPerf should be able to handle all types of reading including those illustrated above.

A common practice has been to relate PresPerf to a perfect state (i.e. a state after the culmination of the eventuality at issue) and therefore to regard PresPerf as an aspect. I call this the "Perfect state (PS) account". Examples (1a) - (2) apparently fit this view. But examples (3a) and (3b) present a critical problem, for the state mentioned by the sentence (e.g. John's living in NY) still obtains at the speech time. The PS account therefore does not apply to sentences with the U-reading, at least not straightforwardly.

Another line of analysis is the "Extended now (XN) analysis" ([1], [5], [6], [7]). This view relates PresPerf to an XN interval, an temporal interval which extends from the speech time into the past. In this view, (1a) asserts that the John-loosing-keys state holds during a certain XN interval, while (3a) asserts that the John-living-in-NY state holds during an XN interval introduced by the *since*-phrase. Thus, the XN analysis applies to both the E- and the U-readings in a consistent manner. This is favorable.

Now, formal dynamic semantics frameworks are available for a semantic analysis of sentences in a discourse. These include DRT (Discourse Representation Theory, [2]) and DAT (Dynamic Aspect Tree, [8]). Given the sophisticated environment they provide, there are good reasons for analyzing the semantics of PresPerf in these frameworks. Incidentally, both Kamp and Reyle ([2]) and ter Meulen ([8]) take the PS account in their analysis of PresPerf. However, these frameworks are in themselves open to alternative approaches, and in fact the insights of the XN analysis provided in the descriptive framework should not be dismissed.

In what follows, 1) I discuss the shortcomings of the PS account and defend the insights of the XN analysis, providing further empirical evidence to support it (Sec. 2); 2) I incorporate the insights of an XN analysis into DRT, and anlyze a set of PresPerf sentences with major types of reading (Sec. 3), and 3) I generalize my analysis and provide the semantics of PresPerf in DRT, which represents the core semantics of PresPerf as well as the ambiguity between different readings (Sec. 4).[2]

[1] For the variety of readings and terminology, see [4].
[2] This paper focuses on DRT but the same line of approach applies also to DAT.

2 Overview of Previous Analysis

I will discuss the shortcomings of the PS account and defend the insights of the XN analysis by providing more empirical evidence.

2.1 The Problems with the PS Account

Kamp and Ryle's ([2]) analysis of (1b) is illustrated in the Discrouse Representation Structure (DRS) (5. It follows the lines of the PS account. The DRS reads as follows. Discourse elements *t*, *s*, *x*, *e* are introduced under the conditions specified: state *s* overlaps with the speech time *n*, and it abuts John's leaving event *e* ("e ⊃⊂ s"). The abutting relation between an event *e* and the current state *s* plays a central role in the PS account. Other types of E-reading are treated in a similar fashion. Note that this abutting relation is defined in purely temporal terms, and therefore it holds forever, once an event culminates.

(5) DRSs for (1b) with the E-reading and for (4) with the U-reading along the lines of Kamp and Ryle ([2])

a.

t,s,x,e
t = n
s ∘ t
John(x)
e ⊃⊂ s
e: x leave

b.

n, t, s, x, y, s', e, mt
t = n
s ∘ t
John(x)
NY(y)
e = beg(s')
e ⊃⊂ s
3-years(mt)
duration(s') = mt
s': x live in y
¬ [*e'* ; e' ≤ n ; e' = end(s')]

The application of the PS account to the U-reading is not trivial, as I noted above. Kamp and Reyle ([2]) propose analyzing (4) as illustrated in (5b). First, they stipulate a start-of-the-state event (e.g. John started to live in NY), and they treat the current state as a perfect state of that starting event ("e = beg(s')" and "e ⊃⊂ s "). Also, they stipulate the condition that the state has not yet come to an end (the final part in the DRS, i.e. the embedded DRS with a negation "¬").

The PS account suffers from the following problems. First, the stipulations for the U-reading are counter-intuitive. The start-of-the-state event would be expressed by 'start to live' etc. (e.g. John *started to* live in NY; We *got to* know each other) rather than by 'live' by itself (e.g. John *lived* in NY). This suggests that the state in question is identified directly, not via the starting event. This point is more evident in PresPerf progressive sentences such as (6a):

(6) a. John *has been playing* the piano since this morning.
    b. John *is playing* the piano now.

Intuitively, (6a) is a PresPerf version of (6b). Sentence (6a) refers to a wider range of time than its simple present counterpart (6b) does. In Kamp and Reyle's analysis, however, (6a) should be analyzed as describing "a perfect state of the starting event of a progressive state". This involves too much complication and thus it is counter-intuitive. As McCoard ([1]) mentions, the PS account applied to the U-reading as above blurs the fundamental aspectual distinction between completive and continuative statuses: In essence, it reduces a continuative status to a completive status. This is a problem.

The PS account has another problem in the treatment of the resultative reading as illustrated in (5a). Given that the abutting relation is defined in purely temporal terms, as I noted above, (5a) only represents that the event precedes the current time. No element actually represents a resultative state (i.e. John being gone). Thus, the resultative and the mere perfect state readings (i.e. *John has once left*) would receive the same analysis. But apparently, these two readings are semantically ambiguous, as traditional ambiguity tests suggest (e.g. John *has left*, so *has* Mary).

The following data ("iterative perfect") provide further counter-evidence to the PS account.

(7) John *has visited* NY *twice* since 2000.

Sentence (7) means that there were two occurrences of a John-visiting-New-York event during the interval introduced by the *since*-clause. However, assuming that the scope of PresPerf includes the adverbial *twice*, the PS account predicts the wrong reading: John-visiting-NY-twice events were completed by 2000 and he has been in the perfect state since 2000 with respect to these events. Diagrammatically ("•" stands for an event of John's visiting New York):

(8) a. ——————-2000——•——•——now : the right reading
    b. ——•——•——2000——————now : the wrong reading

Furthermore, as we observe it in (5a) and (5b), the semantics provided for PresPerf sentences comes out quite differently between the E- and the U-readings, and therefore it is hard to see the shared, core meaning of PresPerf, if there is one. These points all demonstrate shortcomings of the PS account.

2.2 The Extended-Now (XN) Analysis

McCoard ([1]) examines the four major theories which he identifies in the literature and eventually defends the XN theory. He illustrates how the notion of an XN (i.e. the temporal interval extending from the speech time to some time in the past) is comprehensively able to account for empirical data. This view follows the lines of Comrie's ([9]) argument that PresPerf concerns the temporal domain, 'anterior present', preceding the present, and that it is accordingly characterized as a tense rather than as an aspect.

More recent literature also follows this line of analysis. Dowty ([5]) defines an XN as follows:

(9) An "extended now"
XN(t,t') $\stackrel{\text{def}}{=}$ *t'* is a final subinterval of *t*
Specifically,
XN(t,n) means that *n* is a final subinterval of *t*,
i.e. *t* in XN(t,n) is an "extended now" interval

In an XN analysis in the form of traditional logic ([5], [6]), examples (10a) and (10b) are analyzed as (11a) - (11c), where *e* and *s* stand respectively for an event and a state.

(10) a. John *has left.* (E-reading: resultative state)
b. John *has lived* in NY for three years. (E- and U-readings)

(11) a. (10a): $\exists t \exists e[\ XN(t,n)\ \wedge\ e \subset t\ \wedge\ e : leave(John)\ ]$
b. (10b), E-reading
$\exists t \exists e[\ XN(t,n)\ \wedge\ s \subset t\ \wedge\ s : live(John, in\ NY)\ \wedge$
$duration(s) = 3\ years\ ]$
c. (10b), U-reading ("s ∘ t" stands for a temporal overlap.)
$\exists t \exists e[\ XN(t,n)\ \wedge\ s \circ t\ \wedge\ s : live(John, in\ NY)\ \wedge$
$duration(s) = 3\ years\ ]$

We observe in the above that this line of analysis is transparent between an event and a state and between the E- and U-readings. Also, sentence (12) with the iterative reading can be analyzed as (13).

(12) John *has visited* NY twice.

(13) The XN analysis of (12)
$\exists t \exists e_i (i = 1, 2)[\ XN(t,n)\ \wedge\ (e_i \subset t)\ \wedge\ e_i : visit(John, NY)\ ]$

(13) follows the lines of (11a), which means that the XN analysis is also transparent between the iterative reading and the resultative reading. This is also favored. These facts demonstrate the strengths of the XN analysis.

3 Incorporating an XN into DRT

I acknowledge the strengths of an XN analysis through the discussion in the last section and propose incorporating its insights into the DRT framework. I will illustrate an XN analysis of representative PresPerf sentences in DRT.

3.1 Preliminaries: Basic Temporal Notions

In order to investigate the tense and aspect systems in English and Japanese in discourse semantics frameworks, Mizuta ([3]) employs the following basic temporal notions, which I adopt here.

(14) The basic temporal notions ([3])
- E (Event time): in Reichenbach's (1947) sense
- TP (Temporal perspective point in the sense of Kamp and Ryle ([2]), also Reference time in Reichenbach's sense):
  The time from which the eventuality is seen.
- R_{pt} (Reference point in Kamp and Reyle's 1993 sense):
  The perspective point which explains narrative progression.
- TLoc (Temporal location: introduced here):
  The temporal interval in which the eventuality is seen.

TLoc is a specific way of framing time: it includes or overlaps with E. TLoc plays a crucial role in my DRT analysis of PresPerf. It also has general significance in the DRT analysis of other constructions ([3]). In what follows, I will analyze PresPerf sentences TP and TLoc, in particular.

3.2 E-Reading (Resultative State Reading)

Sentences (1a) and (1b) fit the same scheme and are analyzed respectively as (15a) and (15b).

(15) DRSs for (1a) and (1b)

a.

n, t, e, e-time, x, y
TP = n
XN(t, n)
TLoc = t
at(e-time, e)
LB(e-time, TLoc)
John(x)
keys (y)
belong-to(y,x)
e: x loose y

b.

n, t, e, e-time, x
TP = n
XN(t, n)
TLoc = t
at(e-time, e)
LB(e-time, TLoc)
John(x)
e: x leave

The so-far duration of the result state introduces an XN interval *t*, and licenses the use of PresPerf. The time of the triggering event constitutes the left boundary of TLoc ("LB(e-time, TLoc)"). The 'result' is not limited to a concrete one but is more flexible. DRS (15a) represents the following: The temporal perspective is located at the speech time ("TP=n"); an XN interval *t* is introduced ("XN(t, n)");

the temporal location is on *t* (”TLoc=t”); and *t* has as its left boundary the time of John-loosing-his-keys event (”LB(e-time, TLoc)”). Most importantly, an XN interval *t* constitutes TLoc, and *e-time* constitutes its left boundary. I consider that the absence of adverbials introduces a default local context and triggers the resultative state reading. In *John has lost his keys before*, in contrast, the adverbial *before* introduces a global context and triggers an experiential reading. In light of this, I acknowledge a ’zero adverbial’ which triggers the resultative state reading.

3.3 E-Reading (Iterative and Experiential Reading)

Our next focus is the iterative reading. Experiential reading is regarded as a special case of iterative reading.

(16) a. John *has visited* NY twice.
     b. John *has visited* NY before.

(17) DRSs for (16a) and (16b)

a.

$n,\ t,\ e_i,\ e\text{-}time_i,\ x,\ y$
TP = n
XN(t, n)
TLoc = t
at(e-time$_i$, e$_i$)
e-time$_i$ $\subset$ TLoc
$\mid i \mid = 2$
John (x)
NY(y)
e_i: x visit y

b.

$n,\ t,\ e_i,\ e\text{-}time_i,\ x,\ y$
TP = n
XN(t, n)
TLoc = t
at(e-time$_i$, e$_i$)
e-time$_i$ $\subset$ TLoc
$\mid i \mid \geq 1$
John (x)
NY(y)
e_i: x visit y

(17a) represents the following: there are two occurrences of John-visiting-New-York events, e$_i$ (i=1,2), within an XN interval *t*. As in (15a) and (15b), TLoc is located at the XN interval. Unlike the resultative reading, however, the event time is flexible within TLoc: TLoc provides the domain of the event time. With a *since*-clause such as *since 2000*, the left boundary of the XN interval is specified. In that case, the following conditions are added: ”t’ $\subset$ 2000, LB(t’, t)”. The proposed scheme applies also to the experiential reading, as in (2): the number of occurence is one or more ($\mid i \mid \geq 1$). The DRSs for *John has lived in NY twice* fit the same scheme as (17a): Bound states are treated like events.

3.4 U-Reading

I now analyze an example with the U-reading in relation to its simple present version.

(18) a. John *has lived* in NY. (with the U-reading)
b. John *lives* in NY.

(19) DRSs for (18a) and (18b)

a.

n, t, s, s-time, x, y
TP = n
XN(t, n)
TLoc = t
at(s-time, s)
s-time ⊃ TLoc
John (x)
NY(y)
s: x live in y

b.

n, t, s, s-time, x, y
TP = n
TLoc = n
at(s-time, s)
s-time ⊃ TLoc
John (x)
NY(y)
s: x live in y

The PresPerf version is only different from the simple present version in the way the state of John living in NY is framed. The simple present version focuses on the speech time even though the state at issue was actually initiated some years ago. This point is illustrated in the DRSs above as the difference in the location of TLoc. In (19a) for the PresPerf version, TLoc is located at the XN interval t, whereas in (19b), TLoc is located at the speech time n. In (19a), the condition that the eventuality time includes TLoc, and therefore the XN interval, entails that the state holds at the speech time. I consider this condition to be essential to the U-reading. The U-reading is characterized by a total overlap of the eventuality time with the XN interval. Since the state is unbound, this comes out as an inclusion relation (i.e. "s-time ⊃ TLoc").

3.5 U-Reading and E-Reading with a *for*-Phrase

Sentence (20a) with the E- and the U-readings are analyzed as (21a). A related example (20b) with only the U-reading is analyzed as (21b).

(20) a. John *has lived* in NY *for three years.*
b. John *has lived* in NY *for these three years.*

Unlike previous work ([2], [6]), I posit the same scheme for both the E- and the U-readings of (20a). Sentence (20a) asserts that there is an occurrence of John living in NY for three years in a certain XN interval. Even with the U-reading, the sentence concerns a bound state, which can be treated like an event. The difference between the E- and the U-readings depends on the location of the three-year period. In the case of the E-reading, *s-time* is a true subset of TLoc ("s-time ⊂ TLoc"): it does not include the speech time. In the case of the U-reading, *s-time* coincides with TLoc ("s-time = TLoc") and it includes the speech time. The two cases are incorporated into an inclusion relation ("s-time ⊆ TLoc"). In contrast, sentence (20b) with a locating adverbial *for these*

(21) DRSs for (20a) and (20b)

a.

$n,\ t,\ s,\ s\text{-}time,\ x,\ y$
TP = n
XN(t, n)
TLoc = t
at(s-time, s)
s-time $\subseteq$ TLoc
dur(s-time) = 3 yrs
John (x)
NY(y)
s: x live in y

b.

$n,\ t,\ s,\ s\text{-}time,\ x,\ y$
TP = n
XN(t, n)
TLoc = t
at(s-time, s)
s-time $\supset$ TLoc
dur(s-time) = 3 yrs
John (x)
NY(y)
s: x live in y

three years introduces an XN, which includes the speech time. The sentence thus triggers only the U-reading. Notice that DRS (21b) follows the lines of (19a).

4 The Semantics of Present Perfect

Based on the analysis of PresPerf sentences illustrated above, I now provide the semantics of the PresPerf construction. Both the core part and the semantics of different readings are represented.

4.1 General Scheme

The scheme consists of a general component to be applied across readings and specific conditions embedded in it, which trigger different readings. The general component illustrated in (23) on the next page introduces an XN interval and TLoc is located on it. For practical reasons, I use a single label *e* (eventuality) and *e-time* (eventuality time) for both an event and a state.

(22) Specific conditions
a. iterative/experiential reading e.g. (17a), (17b), (21a)
e-time $\subseteq$ TLoc (i.e. TLoc is the temporal domain in which *e* happens.)
b. Resultative reading e.g. (15a), (15b)
LB(e-time, TLoc) (i.e. The event time is the left boundary of TLoc.)
c. Continuative reading (U-reading) e.g. (19a), (21b)
e-time $\supset$ TLoc (i.e. The eventuality holds throughout TLoc.)

Events (as opposed to states) fit into (22a) and (22b). In the case of (22a), the event time should not include the speech time. (If it does, the event is in a progressive state.) These two cases entail a perfect state at the speech time. Thus, the relation to a perfect state is not inherent to PresPerf but is attributed to the verb semantics. The PS account focuses on examples of eventive verbs and mistakenly incorporates this condition into the semantics of PresPerf.

(23) General DRS for PresPerf sentences

n, t, e, e-time, ...
TP = n
XN(t, n)
TLoc = t
at(e-time, e)
...
Specific conditions
...
e: ......

4.2 Construction Rules

DRS construction rules for PresPerf are now provided as (24).

(24) The proposed DRS construction rules for PresPerf
a. General: by virtue of the morphological device *have* + *en*
Triggering configurations:
$\gamma \subseteq \gamma' \in \mathrm{Con}_K$: $[_{S(s)}$ u [VP'(s) $[_{VP2}$ [HAVE $\mathrm{VP}_{[STAT=\alpha]}$]]]]
Introduce in U_K: new discourse referents
t (an XN interval), *e* (an event if α = -, or a state, if α = +), e-time
Introduce in Con_K: new conditions
XN(t,n), TLoc = t, at(e-time, e)
Replace γ by: $[_{S(e)}$ u [VP'(e) [$\mathrm{VP}_{[STAT=\alpha]}$](e)]]]
b. Specification: by virtue of adverbials and verb semantics

- The iterative reading triggered by counting or frequency adverbials and others (e.g. *before*, *ever*): Introduce: e-time $\subseteq$ TLoc
- The resultative state reading triggered by a zero adverbial and an accomplishment/ achievement/ activity VP representing a telic eventuality: Introduce: LB(e-time, TLoc)
- The U-reading triggered by a stative VP: Introduce: s-time $\supset$ TLoc

The construction rules consist of a general component pertaining to the morphological level and a specification component. This conforms to the above-proposed model of the semantics of PresPerf represented in (23) and (22).

The proposed model provides a systematic view of PresPerf. It represents that there is one PresPerf construction, which pertains to the morphological device (*have* + *en*) and that it takes as its parameter adverbials and verb semantics, which lead to different readings.[3]

5 Theoretical Implications

Cross-linguistically, Japanese encodes both the U-reading and its simple present version as "nonpast stative", using *tei-ru* (e.g. *sum-dei-ru*, live-stative-nonpast,

[3] Reyle and Rossdeutscher ([10]) proivdes insights into underspecification.

'has lived'/'lives') On the other hand, the resultative reading and its simple past version are both encoded as "past eventive", using *ta* (e.g. *nakushi-ta*, loose-eventive past, 'has lost'/'lost'). The PS account would work for Japanese *ta*, but English PresPerf has a wider coverage. The tense view of PresPerf can accommodate the wider variety that English PresPerf covers.

6 Conclusion

Providing further empirical evidence, I defended the insights of an XN analysis and incorporated them, with some modifications, into a dynamic semantics framework, DRT. I conclude that: 1) PresPerf is licensed by its relation to an XN interval; 2) different readings are attributed to the specific way that the embedded eventuality interacts with an XN interval and to the contributions of adverbials and the verb semantics; and 3) the semantics of PresPerf is provided compositionally in terms of a general component and specific conditions. The compositional model of PresPerf I provide sheds light on both the differences between PresPerf sentences and the core part shared by them.

I characterize PresPerf as a tense. An alleged connection between PresPerf and a perfect state, which the PS account is based on, is not inherent to the construction itself but is only entailed in the case of events.

References

1. McCoard, R.: The English Present Perfect: Tense Choice and Pragmatic Inference. Amsterdam: North-Holland (1978)
2. Kamp, H., Reyle, U.: From Discourse to Logic. Kluwer Academic Publishers (1993)
3. Mizuta, Y.: A Discourse-semantic Analysis of Tense and Aspect in English and Japanese. PhD thesis, University of Chicago (2002)
4. Depraetere, I.: On the resultative character of present perfect sentences. Journal of Pragmatics **29** (1998) 597–613
5. Dowty, D.: Word Meaning and Montague Grammar. Dordrecht: Reidel (1979)
6. von Stechow, A.: Remarks on kamp-reyle's (1993) analysis of the english perfect. (ms)
7. Giannakidou, A.: A puzzle about until and the present perfect. In Alexiadou, A.e.a., ed.: Perfect Explorations. Mounton de Gruyter (2003)
8. ter Meulen, A.: Representing Time in Natural Language: the Dynamic Interpretation of Tense and Aspect. MIT Press (1995)
9. Comrie, B.: Tense. Cambridge University Press (1985)
10. Reyle, U., Rossdeutscher, A.: Temporal underspecification in discourse. In Rohrer, C., Rossdeutscher, A., Kamp, H., eds.: Linguistic Form and its Computation. CSLI Publications (2001)

Workshop on Active Mining (AM-2004)

The workshop on Active Mining (AM-2004) was held on June 1, 2004 at Ishikawa Kousei Nenkin Kaikan in Kanazawa City, Japan, as a part of the Eighteenth Annual Conference of the Japanese Society for Artificial Intelligence (JSAI-2004). This is the third workshop that focuses on Active Mining; the first one was held on December 9, 2002, as a part of the Second IEEE International Conference on Data Mining (ICDM'02), and the second one was held on October 28, 2003 as a part of the 14th International Symposium on Methodologies for Intelligent Systems, both at Maebashi TERRASA, Maebashi City, Japan.

Active mining is a new direction in the knowledge discovery process for real-world applications handling various kinds of data with actual user need.

Our ability to collect data has been increasing at a dramatic rate, which we call *information flood*. However, our ability to analyze and understand massive data lags far behind our ability to collect them. The value of data is no longer in *how much of it we have.* Rather, the value is in how quickly and effectively can the data be reduced, explored, manipulated and managed.

For this purpose, Knowledge Discovery in Databases (KDD) emerges as a technique that extracts implicit, previously unknown, and potentially useful information (or patterns) from data. However, recent extensive studies and real-world applications show that the following requirements are indispensable to overcome *information flood:* 1) identifying and collecting the relevant data from a huge information search space (active information collection), 2) mining useful knowledge from different forms of massive data efficiently and effectively (user-centered active data mining), and 3) promptly reacting to situation changes and giving necessary feedback to both data collection and mining steps (active user reaction).

Active mining is proposed as a solution to these requirements, which collectively achieves various mining needs. By *collectively achieving* we mean that the total effect outperforms the simple add-sum effect that each effort can bring.

The objective of this workshop was to gather researchers as well as practitioners who are working on various research fields of active mining, share hard-learned experiences, and shed light on future development of active mining. This workshop addressed many aspects of active mining ranging from theories, methodologies, algorithms, to their applications. Through this workshop, we produced modern solutions facilitating data collection, processing and knowledge discovery and created synergy among different branches.

This book contains 8 papers selected from 11 papers presented in the workshop. We would like to thank all the authors who have submitted papers, with whom the workshop was possible.

July 2005

Masayuki Numao, Takahira Yamaguchi,
Shusaku Tsumoto and Hiroshi Motoda

A. Sakurai et al. (Eds.): JSAI 2003/2004, LNAI 3609, p. 463, 2007.

Spiral Discovery of a Separate Prediction Model from Chronic Hepatitis Data

Masatoshi Jumi[1], Einoshin Suzuki[1], Muneaki Ohshima[2], Ning Zhong[2], Hideto Yokoi[3], and Katsuhiko Takabayashi[3]

[1] Electrical and Computer Engineering, Yokohama National University, Japan
jumi@slab.dnj.ynu.ac.jp, suzuki@ynu.ac.jp
[2] Faculty of Engineering, Maebashi Institute of Technology, Japan
ohshima@wi-lab.com,zhong@maebashi-it.ac.jp
[3] Division of Medical Informatics, Chiba University Hospital, Japan
yokoih@telemed.ho.chiba-u.ac.jp, takaba@ho.chiba-u.ac.jp

Abstract. In this paper, we summarize our endeavor for spiral discovery of a separate prediction model from chronic hepatitis data. We have initially proposed various learning/discovery methods including time-series decision tree, PrototypeLines, and peculiarity-oriented mining method for mining the data. This experience has motivated us to model physicians as considering typical cases with the specific disease and ruling out clearly exceptional cases. We have developed a spiral discovery system which learns a prediction model for each type of cases, and obtained promising results from experiments.

1 Introduction

Medical data are challenging to data miners since they show problems related to data quantity, data quality, and data form [5]. Chronic hepatitis data [1], which show high diversities in various aspects of cases, satisfy this nature and necessitate us to develop novel data mining methods. For the chronic hepatitis data, providers presented several objectives including progress of chronic hepatitis (difference between type B and type C) and effect of interferon therapy. Among them, LC (liver cirrhosis) prediction from blood test data can be regarded as one of the most important objectives since it can potentially substitute routine tests for a biopsy, which represents a highly invasive test.

We have been working on knowledge discovery from the chronic hepatitis data including the LC prediction. Our methods include a decision tree learner for time-series classification problem [5], a visualization method for medical test data [4], and a peculiarity-oriented mining method [6]. The methods have proved to be effective for various purposes including detection of exceptional cases.

From the endeavor, we are motivated to model physicians as starting their differential diagnoses by distinguishing typical cases from apparently exceptional cases. In this paper, we propose a discovery method which distinguishes the two types of cases in its separate prediction model based on the abovementioned data mining methods.

A. Sakurai et al. (Eds.): JSAI 2003/2004, LNAI 3609, pp. 464–473, 2007.

2 Knowledge Discovery from the Chronic Hepatitis Data

2.1 Liver Cirrhosis Prediction

Chronic hepatitis represents a disease in which liver cells become inflamed and harmed by virus infection. In case the inflammation lasts a long period, the disease comes to an end which is called a liver cirrhosis (LC). During the process to an LC, the degree of fibrosis, which consists of five stages ranging from F0 (no fibrosis) to F4 (LC), represents an index of the progress. The degree of fibrosis can be inspected by biopsy which picks liver tissue by inserting an instrument directly into liver. A biopsy, however, cannot be frequently performed since it requires a short-term admission to a hospital and involves danger such as hemorrhage. Therefore, if we can predict the degree of fibrosis with a conventional medical test such as a blood test, it would be highly beneficial in medicine.

A time sequence $\boldsymbol{A}$ represents a list of values $\alpha_1, \alpha_2, \cdots, \alpha_I$ sorted in chronological order. A data set D consists of n examples $e_1,\ e_2,\ \cdots,\ e_n$, and each example e_i is described by m attributes $a_1,\ a_2,\ \cdots,\ a_m$ and a class attribute c. We assume that an attribute a_j originally represents a time-series attribute which takes a time sequence as its value. The class attribute c represents a nominal attribute and its value is called a class. We show an example of a data set which consists of time-series attributes in Figure 1. The data set consists of examples 84, 85, 930, each of which is described with time-series attributes GPT, ALB, PLT, and a class.

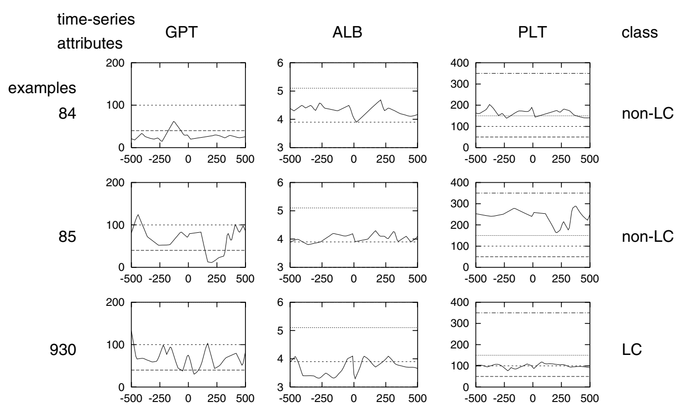


Fig. 1. Data set which consists of time-series attributes

In classification from time-series data, the objective represents induction of a classifier, which predicts the class of an example e, given a training data set D. In this paper, we assume liver cirrhosis (LC) and non-liver cirrhosis (non-LC) as classes.

2.2 Our Endeavor for the Chronic Hepatitis Data

Our time-series decision tree was proposed since physicians requested to use time sequences which exist in data in a classifier [5]. We applied the method to the LC prediction problem and the results attracted their interests. The physicians commented that the obtained decision tree is highly valid although we used medical knowledge only for selecting medical tests in the data [5]. Moreover, most of the cases who are mispredicted by the decision tree were recognized as exceptions by the physicians.

Our PrototypeLines represents a visualization method which is considered to enable discovery of interesting knowledge without extensive training [4]. Application of PrototypeLines to the chronic hepatitis data revealed interesting characteristics. A student in computer science discovered at least two exceptions which were both recognized as missing diseases in the original data by a domain expert. Moreover, a physician who was introduced PrototypeLines for the first time discovered an exceptional condition of a case after a 5-minute explanation.

Our peculiarity-oriented mining method obtains tendencies among examples with peculiar data [6]. The peculiar data are detected based on a distance measure, and the number of the peculiar data for an example is called the number of peculiar attributes. The method has successfully discovered exceptional cases in terms of interferon therapy.

3 Discovery of a Separate Prediction Model

3.1 Overall Architecture

Based on the motivation presented in Section 1, we propose a discovery method which learns a separate prediction model. Our method employs a naive Bayes learner and a 1-NN (1-nearest neighbor) classification method for typical cases and exceptional cases respectively, and refines the separate prediction model by a spiral interaction with a medical expert. The term "spiral" is employed since it represents iterative refinement of discovered knowledge. In the interaction, exceptional cases are identified with the data mining methods mentioned in Section 2.2 and a likelihood-based method to be proposed in the next Section.

Figure 2 shows the architecture of our method. In our method, detection of exceptional cases with our previous methods as well as update of a naive Bayes learner for typical cases are iterated in a spiral manner with interaction with an expert. A naive Bayes classifier predicts a class $\hat{c}_{NBayes,i}$ of an example e_i assuming that each attribute a_j is independent. Here v_{ij} represents the value for attribute a_j of p.

$$\hat{c}_{NBayes,i} = \text{argmax}_c \Pr(c) \prod_{j=1}^{m} \Pr(a_j = v_{ij} \mid c) \quad (1)$$

As the result, we obtain a list of typical cases, conditional probabilities for typical cases, and a list of exceptional cases. Classification of a new case begins by detection of exceptional cases based on a 1-NN method, which will be presented in Section 3.3. If a case who is close to the new case is found, our method outputs the class of the former case. Otherwise, the class of the new case is predicted with a naive Bayes classifier.

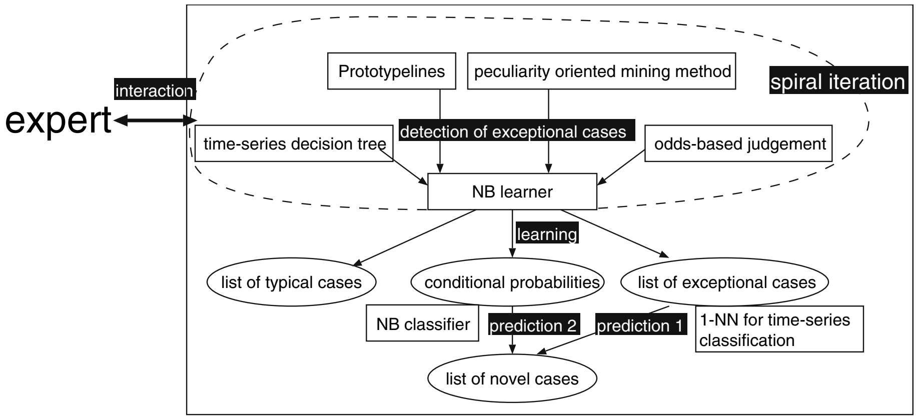


Fig. 2. Architecture of the proposed method

3.2 Likelihood-Based Judgment for a Degree of Exception

In a classification problem, an example can be intuitively regarded as "typical" or "atypical". The typicalness $\Phi(p)$, which is based on the right-hand side of Eq. (1), of a case p represents a degree to which p belongs to its class c compared with the other class $\overline{c}$. Here v_{ij} represents the value for an attribute a_j of p.

$$\Phi(p) = \frac{\Pr(c)\prod_{j=1}^{m}\Pr(a_j = v_{ij} \mid c)}{\Pr(\overline{c})\prod_{j=1}^{m}\Pr(a_j = v_{ij} \mid \overline{c})} \tag{2}$$

The larger $\Phi(p)$ is, the more certain that p belongs to its class c.

If a naive Bayes classifier is relatively accurate for the class c, the typicalness $\Phi(p)$ tends to be large and vice versa. Thus the typicalness $\Phi(p)$ is relative since it depends on the preciseness of a naive Bayes classifier in terms of the class c. The degree how a naive Bayes classifier is precise for c can be measured by the degree of correct classification and the degree of incorrect classification. For c and $\overline{c}$, we represent the number of correctly-predicted examples by the naive Bayes method μ and ν respectively. Likewise, for c and $\overline{c}$, we represent the number of incorrectly-predicted examples by the naive Bayes method μ_e and ν_e respectively. The preciseness $\Psi(\hat{c})$ of a naive Bayes classifier of the estimated class $\hat{c}$ represents the ratio of the precision for c and the precision for $\overline{c}$.

$$\Psi(\hat{c}) = \frac{(\mu+1)(\nu+\nu_\mathrm{e}+2)}{(\mu+\mu_\mathrm{e}+2)(\nu_\mathrm{e}+1)} \tag{3}$$

For our LC prediction problem, we define a degree of exception $E(p, \hat{c})$ for a case p and his/her estimated class $\hat{c}$ as follows in order to discriminate exceptional cases from typical cases.

$$E(p, \hat{c}) = \lceil -\log_{\Psi(\hat{c})} \Phi(p) \rceil \quad (4)$$

Intuitively, $E(p, \hat{c})$ represents an evaluation index which is equal to the number of upvaluated digits below the decimal point when we measure the typicalness $\Phi(p)$ of a case p in terms of the preciseness $\Psi(\hat{c})$ of a naive Bayes classifier of the estimated class $\hat{c}$. When prediction of the naive Bayes method is accurate, $\Psi(\hat{c})$ tends to be large, and the absolute value of $E(p, \hat{c})$ is relatively small even if the absolute value of $\Phi(p)$ is large. This fits our intuition that certain information rarely leads to an extreme degree of exception.

In each loop, cases whose degrees of exception are no less than a user-specified threshold are detected as exceptional cases. The loop continues until prediction accuracy estimated with 10-fold cross validation either decreases or reaches 100 %.

3.3 Similarity Between a Pair of Cases Based on Dynamic Time Warping

Dynamic time warping (DTW) represents an index of dissimilarity between a pair of time sequences [3]. Unlike Euclidean distance, DTW can allow distortion along the time axis since a point in a time sequence can correspond to multiple points in the other sequence.

A 1-NN method based on DTW constantly showed high accuracy in our experiments for time-series classification including the LC prediction problem [5]. Due to this good result, we have chosen this method for detecting exceptional cases in test data. Following [5], the window width [3] of DTW was settled to 10 % of the length of the time sequence, and a dissimilarity measure $H(e_i, e_j)$ between a pair of examples e_i, e_j has been employed. This measure normalizes the results $G(e_i(a_k), e_j(a_k))$ of DTW for e_i and e_j in terms of a time-series attribute a_k with the maximum value $q(a_i)$, where $q(a_k) \geq \forall i \forall j G(e_i(a_k), e_j(a_k))$.

$$H(e_i, e_j) = \sum_{k=1}^{m} \frac{G(e_i(a_k), e_j(a_k))}{q(a_k)} \quad (5)$$

4 Experiments

4.1 Application of Conventional Methods

In the experiments, we used data from 180 days before the first biopsy to the day of the first biopsy following advice of medical experts. As the result, the numbers of LC and non-LC cases are 159 and 112 respectively. The blood tests that we use are GOT (glutamic oxaloacetic transaminase = AST (asparate aminotransferase)), GPT (glutamic pyruvic transaminase = ALT (alanine aminotransferase)), TTT (thymol turbidity test), ZTT (Zinc sulfate turbidity test), D-BIL (direct bilirubin), I-BIL (indirect bilirubin), T-BIL (total bilirubin), ALB (albumin), CHE (cholineesterase), TP (total protein), T-CHO (total cholesterol), WBC (white blood cell), PLT (platelet), and HGB (hemoglobin). In our experiment, we used 46 LC cases and 55 non-LC cases each of whom has test values for all of these blood tests.

For the classifier, we first averaged each time sequence then discretized each value as we show in Table 1, where we use for each attribute value U: extremely high, V: very high, H: high, N: normal, L: low, v: very low, and u: extremely low. Each conditional probability is estimated using Laplace correction in order to cope with the 0-occurrence problem [2]. A missing value is ignored both in estimating probabilities and in classifying an example.

Table 1. Blood tests in the chronic hepatitis data

attribute	intuitive explanation	discretization
GOT	amount of broken	$N \leq 40 < H \leq 100 < V \leq 200 < U$
GPT	liver cells	$N \leq 40 < H \leq 100 < V \leq 200 < U$
TTT	degree of	$N \leq 5 < H \leq 10 < V \leq 15 < U$
ZTT	immune activity	$N \leq 12 < H \leq 24 < V \leq 36 < U$
D-BIL	disorder of	$N \leq 0.3 < H \leq 0.6 < V \leq 0.9 < U$
I-BIL	bile excretion	$N \leq 0.9 < H \leq 1.8 < V \leq 2.7 < U$
T-BIL		$N \leq 1.2 < H \leq 2.4 < V \leq 3.6 < U$
ALB	decrease of	$v \leq 3.0 < L \leq 3.9 < N \leq 5.1 < H \leq 6.0 < V$
CHE	protein generation	$v \leq 100 < L \leq 180 < N \leq 430 < H \leq 510 < V$
TP		$v \leq 5.5 < L \leq 6.5 < N \leq 8.2 < H \leq 9.2 < V$
T-CHO		$v \leq 90 < L \leq 125 < N \leq 220 < H \leq 255 < V$
WBC	good order	$u \leq 2.0 < v \leq 3.0 < L \leq 4.0 < N \leq 9.0 < H$
PLT	of a liver	$u \leq 50 < v \leq 100 < L \leq 150 < N \leq 350 < H$
HGB	hemoglobin	$L \leq 12 < N \leq 18 < H$

First, the medical expert investigated display result of PrototypeLines from 500 days before the first biopsy to 500 days after the first biopsy. As the result, cases 380, 336, 928, 251, 602, 913 are recognized as candidates of exceptional cases. Inspecting time sequences for these cases, he concluded that case 336 is recognized as an exception since his/her PLT is very low and his/her ALB, CHE, TP, WBC are low although s/he belongs to non-LC. Moreover a fatty liver is suspected since his/her GPT is much higher than GOT. Case 928 is suspected to show acute aggravation or an uninterpretable disease thus is removed.

Similarly, we applied our peculiarity-oriented mining method to the data, and showed cases with no less than four peculiar attributes to the medical expert. As the result, cases 611 and 903 show low ALB and CHE though their TP are high and T-CHO are normal. Since these four blood tests are closely related to each other and typically synchronize, these cases are recognized as exceptions.

Case 916 showed a peculiar sequence for T-CHO compared with his/her ALB, CHE, TP thus is recognized as exception.

The above cases are considered to be peculiar in the whole data. In order to detect peculiar cases in LC cases or in non-LC cases, we applied this method to data with the corresponding class. From non-LC data, several cases are detected including the abovementioned cases 615 and 160, of which WBC and PLT only are low. At this moment, the medical expert decided to postpone decisions on "borderline cases" each of whom is similar to cases in the other class, and to remove only apparently abnormal cases. For instance, decision on case 596 is postponed although s/he looks very differently from the other cases who belong to 596's degree of index F2. Blood tests ALB, CHE, TP, T-CHO are closely related to each other as we described earlier, and a case who shows low values for one or two of them is called a monocytopenia or a bicytopenia respectively. While the medical expert was inspecting case 755, he discovered that such borderline cases can be classified into either bicytopenia or monocytopenia. In these results, cases 160, 918, 596, 755 are judged bicytopenia, and cases 615, 925, 596, 755 are judged monocytopenia. He inferred that removing these borderline cases might reduce predictive accuracy of similar kinds of cases, and this effect should be investigated.

Similarly, interesting cases are also detected from LC data. We confirmed removal of case 611, who was detected from the whole data, since s/he is suspected to suffer hemorrhage right after the biopsy. Case 903 was judged to stay in typical cases since s/he belongs to "partial LC" who shows partial aggravation of blood tests. We postponed removal of case 916, who was also detected from the whole data. Case 943 is judged to be removed since s/he is suspected to suffer from constitutional hyperbilirubinemia (as a complication) since his/her TTT and ZTT are high but no other results suggest LC.

The medical expert then inspected misclassified cases from the time-series decision tree. Case 737 shows good blood test results although s/he belongs to LC, thus is judged as a compensatory LC. Case 615 shows good blood test results except for T-CHO thus is misclassified. This is due to the nature of a decision tree which makes prediction based on a relatively small number of attributes. Case 755 shows bad results although s/he belongs to non-LC. We have concluded that misclassified cases from the time-series decision tree represent borderline cases.

4.2 Application of the Proposed Method

Following the policy in the previous Section, we first defined exceptional cases as abnormal cases who are suspected to suffer from different diseases and tried to obtain the separate prediction model. After careful examination, cases 916, 928, 943 were first disregarded as exceptions. The naive Bayes learner was applied to the remaining data, and the likelihood-based method in Section 3.2 was applied in order to detect candidates of exceptional cases, in which we used 3 as the value of the threshold. As the result, 14 cases showed degrees of exception no less than 3. The cases identifiers are 203, 206, 737, 758 (LC, degree 5); 236, 245 (LC, degree 4); 184, 244, 251, 553, 699, 913 (LC, degree 3); and 160 (non-LC, degree 3). The medical experts inspected their time sequences but none of them were recognized as abnormal thus the result at this moment was judged as final.

We then defined exceptional cases as borderline cases who resemble to cases in the other class. After careful examination, cases who are removed before the first spiral are grouped in terms of their causes. The following non-LC cases, though they belong to either F1 or F2, resemble to LC cases.

1. Blood cell decrease: 336, 160, 596, 755 (only PLT), 918, 926 (only PLT).
2. ALB decrease: 336, 596, 755, 918, 926.
3. CHE decrease: 336, 160, 596, 926.
4. T-CHO decrease: 615, 925.

The following LC cases, though they belong to F4, resemble to non-LC cases.

1. Blood cell non-decrease: 611, 903, 602 (only WBC), 203 (only WBC), 404, 737, 244 (only PLT).
2. ALB non-decrease: 737, 244, 913.
3. CHE non-decrease: 203, 737, 244, 913, 404.
4. T-CHO non-decrease: 611, 203, 244, 251, 404, 737, 903.

We define that a cycle consists of application of the naive Bayes method, detection of candidates of exceptional cases by the likelihood-based method, and determination of exceptional cases by the medical expert. Below we show exceptional cases detected in each cycle and their causes. All of them belong to LC.

1. Second cycle: 184 (ALB, CHE, T-CHO, WBC), 236 (ALB, CHE, T-CHO, WBC), 206 (ALB, CHE, T-CHO, WBC, PLT), 245 (ALB, T-CHO), 758 (ALB, CHE, TP, T-CHO), 553(ALB, CHE, TP, T-CHO, WBC).
2. Third cycle: 166 (ALB, T-CHO, CHE), 699 (ALB, CHE, TP, T-CHO, WBC, PLT).
3. Fourth cycle: 623 (ALB, CHE, T-CHO, WBC).

We show the final conditional probabilities of the naive Bayes classifier in Table 2. In the Table, for each blood test a and a category v, $\hat{\Pr}(a{=}v|\text{non-LC})$ $(n(a{=}v|\text{non-LC}))$ | $\hat{\Pr}(a{=}v|\text{LC})$ $(n(a{=}v|\text{LC}))$ are shown, where $n(\cdot)$ represents the corresponding number of examples in the data set. For instance, there are 6 non-LC cases and 1 LC cases for ZTT=N thus the probabilities are obtained using Laplace correction since there are 47 non-LC cases and 26 LC cases in the data set which corresponds to the Table. In the Table, we emphasize categories each of which shows more than 3 times of difference and no smaller than 10 % with boldface and with underline for non-LC predominant and LC predominant respectively. It should be noted that order information on attribute values is considered in this process.

According to the medical expert, this kind of conditional probabilities represent useful information in building a medical expert system. Compared with its first probabilities, the number of categories that are effective in discriminating LC cases from non-LC cases increases. This signifies that the removed cases

Table 2. Conditional probabilities (%) and numbers of examples for typical cases, where each categories shows $\hat{\Pr}(a{=}v|\text{non-LC})$ $(n(a{=}v|\text{non-LC}))$ | $\hat{\Pr}(a{=}v|\text{LC})$ $(n(a{=}v|\text{LC}))$

<table>
<tr><td>GOT</td><td>N: 35.3(17) | 3.3(0)</td><td>H: 39.2(19) |43.3(12)</td><td>V: 21.6(10) |46.7(13)</td><td>U: 3.9(1) | 6.7(1)</td></tr>
<tr><td>GPT</td><td>N: 15.7(7) | 6.7(1)</td><td>H: 51.0(25) |23.3(6)</td><td>V: 21.6(10) |60.0(17)</td><td>U: 11.8(5) |10.0(2)</td></tr>
<tr><td>TTT</td><td>N: 39.2(19) |26.7(7)</td><td>H: 27.5(13) |46.7(13)</td><td>V: 29.4(14) | 6.7(1)</td><td>U: 3.9(1) |20.0(5)</td></tr>
<tr><td>ZTT</td><td>N: 13.7(6) | 6.7(1)</td><td>H: 68.6(34) |73.3(21)</td><td>V: 13.7(6) |16.7(4)</td><td>U: 3.9(1) | 3.3(0)</td></tr>
<tr><td>D-BIL</td><td>N: 88.2(44) |36.7(10)</td><td>H: 7.8(3) |46.7(13)</td><td>V: 2.0(0) |10.0(2)</td><td>U: 2.0(0) | 6.7(1)</td></tr>
<tr><td>I-BIL</td><td>N: 96.0(47) |75.9(21)</td><td>H: 2.0(0) |17.2(4)</td><td>V: 2.0(0) | 6.9(1)</td><td></td></tr>
<tr><td>T-BIL</td><td>N: 96.0(47) |69.0(19)</td><td>H: 2.0(0) |24.1(6)</td><td>V: 2.0(0) | 6.9(1)</td><td></td></tr>
<tr><td>ALB</td><td>L: 12.2(5) |60.7(16)</td><td>N: 87.8(42) |39.3(10)</td><td></td><td></td></tr>
<tr><td>CHE</td><td>v: 2.0(0) |10.0(2)</td><td>L: 2.0(0) |53.3(15)</td><td>N: 92.2(46) |33.3(9)</td><td>H: 3.9(1) | 3.3(0)</td></tr>
<tr><td>TP</td><td>L: 4.0(1) | 3.4(0)</td><td>N: 92.0(45) |93.1(26)</td><td>H: 4.0(1) | 3.4(0)</td><td></td></tr>
<tr><td>T-CHO</td><td>L: 2.0(0) |16.7(4)</td><td>N: 88.0(43) |76.7(22)</td><td>H: 6.0(2) | 3.3(0)</td><td>V: 4.0(1) | 3.3(0)</td></tr>
<tr><td>WBC</td><td>u: 1.9(0) | 6.5(1)</td><td>v: 1.9(0) | 6.5(1)</td><td>L: 9.6(4) | 12.9(3)</td><td>N: 82.7(42) |71.0(21)
H: 3.8(1) | 3.2(0)</td></tr>
<tr><td>PLT</td><td>u: 2.0(0) |10.0(2)</td><td>v: 5.9(2) |50.0(14)</td><td>L: 27.5(13) |26.7(7)</td><td>N: 64.7(32) |13.3(3)</td></tr>
<tr><td>HGB</td><td>L: 4.1(1) |21.4(5)</td><td>N: 95.9(46) |78.6(21)</td><td></td><td></td></tr>
</table>

actually represent borderline cases. It has been also observed that the prediction accuracy of the naive Bayes method increases.

4.3 Analysis of Experimental Results

For the experimental results in the previous Section, the accuracy of the separate prediction model is 73.5 %. More precisely, the accuracies for exceptional cases and typical cases were 52.4 % and 79.2 % respectively. The overall accuracy is similar to that of a conventional naive Bayes classifier, but it should be noted that the predictive accuracy for LC cases is higher than the predictive accuracy obtained by a conventional naive Bayes classifier. The accuracies for LC and non-LC cases were 77.3 % and 70.4 % respectively mostly because the correctly predicted cases with the 1-NN method were all LC. Our separate prediction model, which first applies the 1-NN method for predicting the class of exceptional cases, regards LC cases important since it is adequate in predicting exceptional LC cases due to the nature of its dissimilarity measure[1]. This fits the nature of the LC prediction problem in which overlooking of LC cases costs more than misprediction of non-LC cases.

In data mining, discovered knowledge is typically more important than high accuracy. Our experiments have revealed that detection of exceptional LC cases could be done by searching for asynchronism in blood tests ALB, CHE, T-CHO, WBC, PLT; and HGB might be safely ignored. Although it is impossible to detect an exceptional LC case who shows good results for all blood tests, our 1-NN method could detect exceptional LC cases relatively accurately. A partial LC case who shows partial aggravation of blood tests was known among medical experts only empirically, but we have succeeded in detecting several of them. Such cases might suffer from genetic problems, and detailed inspection can be expected to reveal their true causes.

[1] Exceptional LC cases are relatively stable in their time sequences and are more easily predicted with the 1-NN method.

5 Conclusions

In this paper, we have described our endeavor with our separate prediction model. The motivation is based from our previous endeavor, from which we are motivated to model physicians as starting their differential diagnoses by distinguishing typical cases from clearly exceptional cases.

Among conventional prediction methods for LC cases, ALB such as Child Pugh classification is the most frequently used and PLT is also known a good indicator. Blood chemistry and complete blood count tests such as CHE, T-CHO, WBC were known to decrease as liver cirrhosis progresses, but various factors have prohibited their quantitative evaluation. We obtained comments that our endeavor with the separate prediction model is expected to contribute to such analysis. Our future work concerns building an effective data mining method as well as contributing to analysis of the chronic hepatitis data.

Acknowledgement

This work was partially supported by the grant-in-aid for scientific research on priority area "Active Mining" from the Japanese Ministry of Education, Culture, Sports, Science and Technology.

References

1. P. Berka: ECML/PKDD 2002 Discovery Challenge, Download Data about Hepatitis, http://lisp.vse.cz/challenge/ecmlpkdd2002/ (current September 28th, 2002).
2. P. Domingos and M. Pazzani: On the Optimality of the Simple Bayesian Classifier under Zero-One Loss, *Machine Learning*, Vol. 29, No. 2/3, pp. 103–130 (1997).
3. E. J. Keogh: Mining and Indexing Time Series Data, *Tutorial at the 2001 IEEE International Conference on Data Mining (ICDM)*, http://www.cs.ucr.edu/%7Eeamonn/tutorial_on_time_series.ppt (2001).
4. E. Suzuki, T. Watanabe, H. Yokoi, and K. Takabayashi: Detecting Interesting Exceptions from Medical Test Data with Visual Summarization, *Proc. Third IEEE International Conference on Data Mining (ICDM)*, pp. 315-322 (2003).
5. Y. Yamada, E. Suzuki, H. Yokoi, and K. Takabayashi: Decision-tree Induction from Time-series Data Based on a Standard-example Split Test, *Proc. Twentieth International Conference on Machine Learning (ICML)*, pp. 840-847 (erratum http://www.slab.dnj.ynu.ac.jp/erratumicml2003.pdf) (2003).
6. N. Zhong, Y. Y. Yao, and M. Ohshima: Peculiarity Oriented Multi-Database Mining, *IEEE Transaction on Knowledge and Data Engineering*, Vol. 15, No. 4, pp. 952-960 (2003).

Process to Discovering Iron Decrease as Chance to Use Interferon to Hepatitis B

Yukio Ohsawa[1,4], Hajime Fujie[2], Akio Saiura[3], Naoaki Okazaki[4], and Naohiro Matsumura[5]

[1] Graduate School of Business Sciences, University of Tsukuba
osawa@gssm.otsuka.tsukuba.ac.jp
[2] Department of Gastroenterology, The University of Tokyo Hospital
[3] Department of Digestive Surgery, Cancer Institute Hospital, Tokyo
[4] Graduate School of Information Science and Technology, The University of Tokyo
[5] Faculty of Economics, Osaka University

Abstract. Chance discovery is the process of human interaction with the environment for discovering events significant for making a decision. We executed the *double helix* process of chance discovery, on the blood-test data for hepatitis B, for obtaining scenarios telling when and how symptoms essential for treatment appear. In the double-helical process of chance discovery, the presented scenario maps are evaluated and fed back to the following cycles, to obtain novel and potentially useful knowledge for treatment. Due to the combination of the objective facts in the data and the subjective focus of the hepatologists' concerns in this process, the relation between the changes of iron quantities due to iron-carrying proteins and the cure of hepatitis B with interferon, has got clarified visually.

1 Introduction: Scenarios in the Basis of Critical Decisions

A scenario is defined as a sequence of events sharing a context. According to the definition of "chance" in [Ohsawa and McBurney 2003], i.e., an event or a situation significant for decision making, a chance occurs at the cross point of multiple scenarios because a decision is to select one scenario in the future. Generally speaking, a set of scenarios form a basis of decision making, in domains where the choice of a sequence of events affects the future significantly. For example, let us stand on the position of a surgeon looking at a time course or clinical course of symptoms observed in an individual patient. This surgeon should provide the patient with proper treatment at the right time. If he does so, the patient's disease may be cured. Otherwise the patient's status might be worsened radically. The problem is to choose one from multiple scenarios. For example, suppose states 4 and 5 in Eq. (1) mean two opposite situations.

Scenario 1 = {state1 -> state2 -> state3 -> state4 (a normal condition)}.
Scenario 2 = {state 0 -> state2 -> state5 (a fatal condition)}. (1)

Each event-sequence in Eq.(1) is called a *scenario* if the events in it share some common context. For example, Scenario 1 is a scenario in the context of cure, and

A. Sakurai et al. (Eds.): JSAI 2003/2004, LNAI 3609, pp. 474 – 484, 2007.

Scenario 2 is a scenario of the context of disease progress. The surgeon should choose an effective action at the time of state 2, in order to turn this patient to state 3 and state 4 rather than to state 5, if possible. Such a state as state 2, essential for making a decision, is a c*hance* in this case.

Detecting an event at a crossover point among multiple scenarios, as state 2 above, and selecting the most valuable scenario at such a cross point means a chance discovery. Discovering a chance and taking it into consideration is required for making valuable useful scenarios, but the discovery should come after a number of possible scenarios are proposed.

2 Scenario “Emergence” in the Mind of Experts

Scenarios presented from the viewpoint of each participant’s environment, are bridged via ambiguous pieces of information about the different mental worlds they belong to. From these bridges, each participant recognizes situations or events which may work as “chances” to import others’ scenarios to get combined with one’s own. In the example of Eq.(1), a surgeon who almost gave up (paying attention only to Scenario 2) may obtain a new hope in Scenario 1 proposed by his colleague who noticed that state 2 is common to both scenarios – only if it is still before or at the time of state 2. Here, state 2 is uncertain in that its future can potentially proceed into either of two directions, and this uncertainty can make a chance, an opportunity not only a risk.

In this paper the authors applied a method for aiding scenario emergence, by means of the interaction with real data using two tools of chance discovery, KeyGraph in [Ohsawa 2003b] and Polaris [Okazaki and Ohsawa 2003]. Here, KeyGraph with additional causal directions in the co-occurrence relations between values of variables in blood-test data of hepatitis patients (let us call this a *scenario map*), and functions of Polaris helps in dealing with data matching with the concern of experts, i.e. hepatologists here.

These tools help in obtaining useful scenarios of a chrolonological course of hepatitis with/without treatment, reasonably restricted to an understandable type of patience, from the complex real data taken from the mixture of various scenarios. The

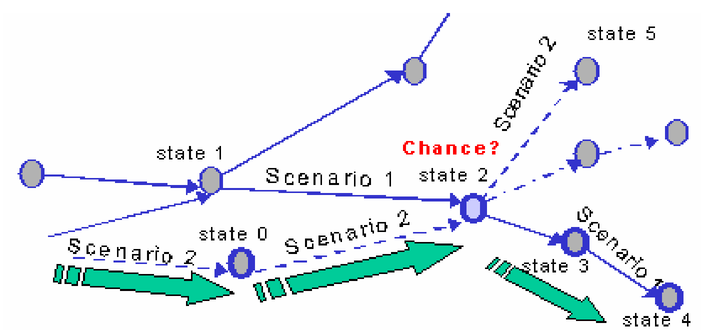


Fig. 1. A chance existing at the cross point of scenarios. The scenario in the thick arrows emerged from Scenario 1 and Scenario 2.

scenarios obtained for hepatitis were evaluated by two hepatologists, a surgeon and a physician, as useful in finding a good chance to treat hepatitis patients. We can say our discovery process worked quite well under the hard condition that it is very rare that the full scenarios of really critical worsening or exceptionally successful treatment occur.

3 Tools for Accelerating the Process of Chance Discovery

3.1 The Double Helix: The Process Model of Chance Discovery

In the studies on chance discovery, the discovery process has been supposed to follow the Double Helix (DH) model [Ohsawa 2003a] as in Fig. 2. The DH process starts from a state of user's mind concerned with catching a new chance (we use "concern" for including the meaning of worrying, i.e., concern with risks not only with opportunities, because a "chance" includes the meaning of a risk). This *concern* is reflected to acquiring *object-data* to be analyzed by data-mining tools specifically designed for chance discovery. Looking at the visualized result of this analysis, possible scenarios and their values rise in each user's mind. Then users get to be participants of a co-working group for chance discovery, sharing the same visual result. Then, words corresponding to the bridges among the various daily contexts of participants are visualized in the next step of visual data mining applied to the *subject-data*, i.e., the text data recording the thoughts and opinions in the discussion. Via the participants' understanding of these bridges, the islands get connected and form novel scenarios. By this time, the participants may have discovered chances on

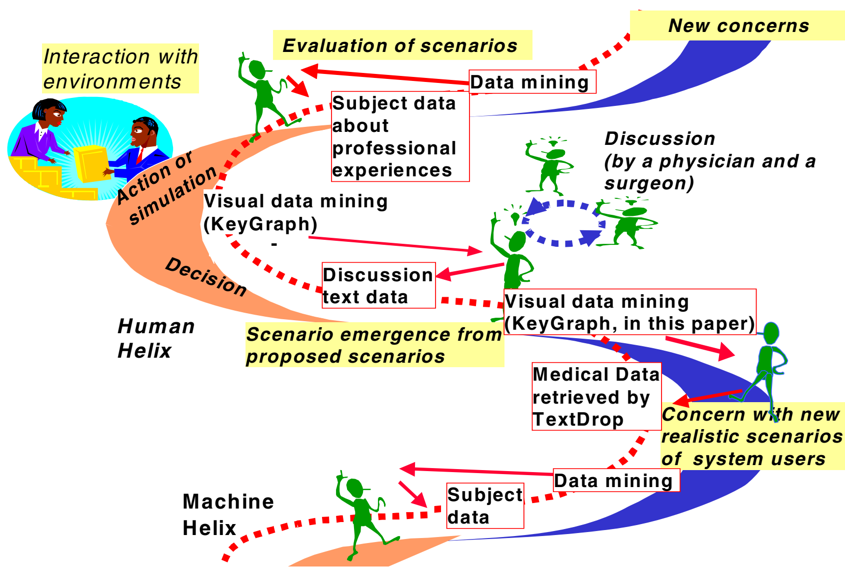


Fig. 2. The DH Model: A process model of chance discovery

the bridges, because each visualized island corresponds to a certain scenario familiar to some of the participants and a bridge means a cross-point of those familiar scenarios. Based on these chances, the user(s) take actions, or simulate actions in a virtual environment, and obtain concerns with new chances – the helical process returns to the initial step of the next cycle.

DH is embodied in this paper, in the application to obtain hepatitis scenarios. Users watch and discuss on KeyGraph [Ohsawa 2003b], working on Polaris, an interactive visual interface which accelerates the DH process [Okazaki and Ohsawa 2003], applied to the subject-data and the object-data in the process. They think and talk about scenarios the diagram may imply, looking at the visual output of KeyGraph.

3.2 KeyGraph for Visualizing Scenario Map

KeyGraph is a computer-aided tool for visualizing the map of event relations in the environment, in order to aid the process of chance discovery. If the environment represents a discussion, an event may represent a word by a participant. By visualizing the map where the words appear connected in a graph, one can see the overview of participants' interest. Suppose a text (string-sequence) *D* is given, describing an event-sequence sorted by time, with periods (``.") inserted at the parts corresponding to the moments of major changes. For example, let text *D* be:

D = "*Mr. A: In the market of general construction, the customers decreased.*
Mr. B: Yes… Our company, building from concrete and steel, is in this bad trend.
Mr. C: This state of the market induces a further decrease of customers. Our company may have to introduce restructuring for satisfying customers.
Mr. B: Then the company can reduce the price of concrete, steel, and construction.
M .D: But that may reduce us the power of this company." (2)

In the case of a document as in Eq.(2), periods are put at the end of each sentence. In the case of a sales (Position Of Sales: POS) data, periods are put in the end of each basket. *KeyGraph*, of the following steps, is applied to *D* ([Ohsawa 2003b] for details).

***KeyGraph-Step* 1:** Clusters of co-occurring frequent items (words in a document, or events in a sequence) are obtained as basic clusters, called *islands*. That is, items appearing many times in the data (e.g., the word "market" in Eq.(2)) are depicted with black nodes, and each pair of these items occurring often in the same sequence unit (a *sentence* in a document, a bought set of items in each basket in sales data, etc) is linked to each other, e.g., `` market - customers - decrease" for Eq.(2) with a solid line. Each connected graph obtained here forms one island, implying the existence of a common context underlying the belonging items.

***KeyGraph-Step* 2:** Items which may not be so frequent as the black nodes in islands but co-occurring with multiple islands, e.g., ``restructuring" in Eq.(2), are obtained as *hubs*. A path of links connecting islands via hubs is called a *bridge*. If a hub is rarer than black nodes, it is colored in a different color (e.g. red). We can regard such a new hub as a candidate of *chance*, i.e., items significant (assertions in a document, or latent demand in a POS data) with respect to the structure of item-relations.

In the example of Fig. 3, the result of KeyGraph on Polaris, the island {customers} means the context asserting that the importance of customers is established, and the island of {steel, concrete, company} shows the established business context of a company. The bridge "restructuring" shows the company may introduce restructuring, e.g. firing employees, for winning the good feeling of customers. "Restructuring" might be rare in the communication of the company staffs, but this expresses the concern of the employees.

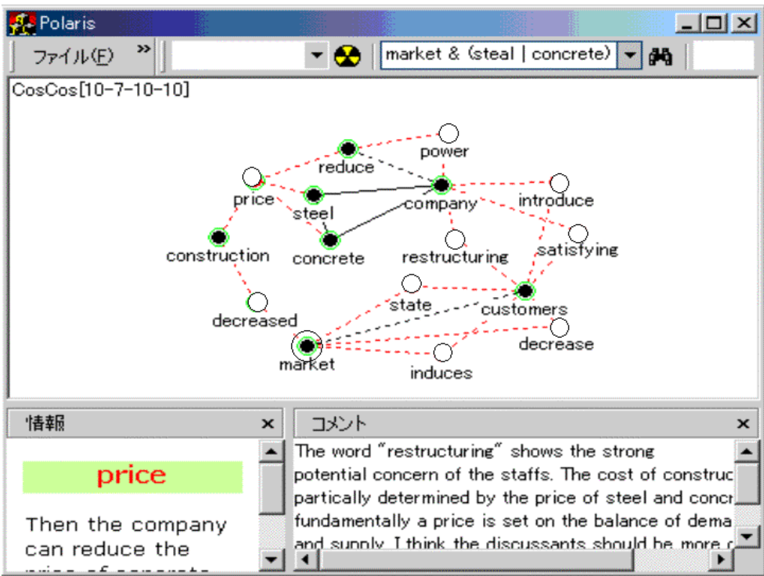


Fig. 3. An example of *KeyGraph on Polaris*: Islands are obtained from *D* in Eq.(2), each including event-set {market}, {steel, concrete, company}, {customers} etc. The double-circled nodes and white nodes show frequent and rare words respectively, forming hubs of bridges.

On Polaris, user can feedback new concern, i.e., the interest in new object-data by entering a new search query as "market & (concrete OR steel)" in the tool bar shown in Fig.3, as mentioned in 4.2. In this case, the user is concerned about the market of concrete and steel, in the chat among Mr. A, Mr. B, Mr. C, and Mr. D. Further more, if user becomes concerned with a topic shown by a node in the graph, he/she can click on the node to read the sentences including words corresponding to the node (the lower left window in Fig.3). The set of extracted sentences becomes the input to KeyGraph, if user selects going to the next cycle of the DH process.

3.3 Retrieving Data Relevant to User's Concern

Polaris has a function for Boolean-selection of the part of data corresponding to users' concern described in a Boolean formula, e.g,

concern ="(product A | product B) & product C & !product D". (3)

For this Boolean expression of user's concern, Polaris obtains a focused data made of baskets including product A or product B, and product C, but not including product D. This becomes a revised input to KeyGraph, at the step where user acquires a new concern and seeks a corresponding data, in the DH process. This is useful if the user can express his/her own concern in Boolean formula as in (2). The concern of a user/users might be more ambiguous, especially in the beginning of the DH process. In such a case, the user is supposed to enter the formula specifically representing one's own concern. Having KeyGraph, query-search function, and the clickable nodes, Polaris supports the user can follow the procedure below to realize a sped up DH process.

Table 1. The DH process supported by KeyGraph and Text Drop

Step 1) Extract a part of the object-data on Polaris, corresponding to user's concern with events or with the combination of events expressed in a Boolean formula.

Step 2) Apply KeyGraph to the data in Step 1) in order to visualize the map representing the relations between events, and attach causal arrows as much as possible, with the help of experts of the domain.

Step 3) Manipulate KeyGraph in a group work with domain experts, as follows:
3-1) Move/remove nodes and links in KeyGraph, considering their importance.
3-2) Write down comments about scenarios, proposed on KeyGraph.

Step 4) Read or visualize (with KeyGraph) the subject-data, i.e., the participants' comments obtained in 3-2), and choose noteworthy and realistic scenarios.

Step 5) Execute or simulate (draw concrete images of the future) the scenarios obtained in Step 4), and, based on this experience, refine the statement of the new concern in concrete words. Go to Step 1).

4 Results for the Diagnosis Data of Hepatitis

4.1 The Hepatitis Data

The following shows the style of data obtained from blood-tests of hepatitis cases. Each event represents a pair, of a variable and its observed value. That is, an event put as "a_b" means the value of variable a was b. For example, T-CHO_high (T-CHO_low) means T-CHO (total cholesterol) was higher (lower) than the predetermined upper (lower) bound of normal range. Note that the lower (higher) bound of each variable was set higher (lower) than values defined in hospitals, in order to be sensitive to the moments the variable takes an unusual value. Each line delimited by '.' represents the sequence of blood-test results for one patient. As in

Eq.(3), we regard one patient as a unit of co-occurrence of events. As a result, the scenario of a typical chrolonological course is expected to appear as a connected path in the scenario map obtained with KeyGraph.

$$\begin{aligned} &\text{Case1} = \{\text{event1, event2, ….., event m1}\}. \\ &\text{Case2} = \{\text{event 2, event 3, …., event m2}\}. \\ &\text{Case3} = \{\text{event 1, event 5, …, event m3}\}. \end{aligned} \tag{3}$$

For example, suppose we have the data in Table 2, where each event means a value of a certain attribute of blood, e.g. GPT_high means the status of a patient whose value of GPT exceeded its upper bound of normal range. Each period ('.') represents the end of one patient's case. If the doctor is interested in patients having experiences of both GTP_high and TP_low, then the doctor can enter "GTP_high & TP_low" to Polaris in Step 1) in Table 1 and get the italic lines as an input to KeyGraph in Step 2).

By applying KeyGraph to this data, the following components are obtained:

- *Islands of events*: A group of events co-occurring, i.e. occurring often to the same patients. The doctor is expected to know a patient's status corresponding to each island, because events in an island are frequent.
- *Bridges across islands*: A patient may switch from one island to another, in the progress of the disease and in the treatment.

Table 2. An Example of Blood Test Data for KeyGraph

GPT_high TP_low TP_low GPT_high TP_low GPT_high TP_low.
ALP_low F-ALB_low GOT_high GPT_high HBD_low LAP_high LDH_low TTT_high ZTT_high ALP_low CHE_high D-BIL_high F-ALB_low F-B_GL_low.
GOT_high GPT_high LAP_high LDH_low TTT_high ZTT_high F-ALB_low F-B_GL_low G_GL_high GOT_high GPT_high I-BIL_high LAP_high LDH_low TTT_high ZTT_high GOT_high GPT_high LAP_high LDH_low TP_low TTT_high ZTT_high B-type CAH2A
D-BIL_high F-CHO_high GOT_high GPT_high K_high LAP_high LDH_low T-CHO_high TP_low UN_high T-BIL_high ALP_high D-BIL_high GOT_high GPT_high I-BIL_high LDH_high T-BIL_high B-type CAH2B.

The data dealt with here was 771 cases, taken from 1981 through 2001. Fig. 4 is a KeyGraph obtained, for cases of progressive hepatitis B. The causal arrows in Step 2) of the DH Process, which does not appear in the original KeyGraph of Ohsawa (2003), depict approximate causations. If the direction from X to Y is apparent for an expert, the expert puts the arrow in the graph. If this direction is not apparent, the two results of KeyGraph are compared, one for the data retrieved for entry "X" with Polaris, and the other for the data retrieved for entry "Y." If the expert judges the former includes more causal events than the latter, X is regarded as a preceding event of Y in a scenario.

Sometimes, the order of causality and the order of occurrence time are opposite. For example, the upper threshold of ZTT may be set low and easy to exceed than that of G_GL, which makes ZTT_high appear before G_GL_high even though ZTT_high is a result of G_GL_high. In such a case, we compare the results of KeyGraph, one for data including G_GL_high and the other for data including ZTT_high. Then, if the latter includes F1, an early stage of fibrosis, and the former includes F2, a later stage, we can understand G_GL_high was really preceding ZTT_high. Let us call a KeyGraph with the arrows made in this way, a *scenario map*.

4.2 The Double Helix Process Executed for Hepatitis B

In the preliminary cycle of DH process, we had understood the scenario map became a mixture of cases according to the subject-data (comments) from hepatologists. Thus we separated the data into each scenario, i.e., AH (acute hepatitis), CAH (chronic aggressive hepatitis) etc., by spotlighting events typical to each scenario. Below we present how we reached a significant discovery in the continued process.

1) The obtained KeyGraph and how users understood it: For the cases of AH, the obtained scenario map matched with the background knowledge of the hepatologists. For CAH, some of their tacit experiences were externalized. For example, a quick sub-process from LDH_high to LDH_low (LDH: lactate dehydrogenase) shown in the scenario map had been sometimes observed in the introductory steps of fulminant hepatitis B, but has not been published because of the rareness.

2) Deepened concerns, the focused object data, and new scenarios: The results in 1) drove us to see more into the details of the progress of fibrosis denoted by F1, F2, F3, and F4 (or LC: liver cirrhosis). Fig. 4 is the scenario map for hepatitis B, in the spotlights of F1, F2, F3, and F4 (LC), i.e., for data extracted for entry "type-B & (F1 | F2 | F3 | F4 | LC)"with Polaris. This result shows a novel connection among the basic state transitions in fibrosis listed from a. to e. below, useful for understanding the status of a patient at an arbitrary time.

a. A chronic active hepatitis sometimes turns into a severe progressive hepatitis and then to cirrhosis or cancer, in the case of hepatitis B.
b. The final states of critical cirrhosis co-occur with kidney troubles, and grows to malignant tumors (cancer) with the deficiencies of white blood cells.
c. Recovery is possible from the earlier steps of fibrosis.
d. LDH_low after high LDH_high can be a sign of fulminant hepatitis.
e. The low Fe (iron) level with cirrhosis can be a turning point to the recovery of liver.

In Fig.4, the appearance of FE_low, on the *only* bridge from cirrhosis to recovery, can be useful for finding the optimal timing to treat a patient of hepatitis B, and seems relevant to [Rubin et al 1995] having suggested iron reduction may improve the response of chronic hepatitis to interferon. However, Fig.4 does not include "interferon." A possible interpretation of this figure is that iron-reduction has been applied to patients. An iron reduction is to take blood out of the body, in order clean the iron pool stored too much in the liver. Hayashi et al (1995) also showed iron reduction improves the condition of a patience of hepatitis C. However, the problem is that doctors rarely use iron reduction for hepatitis B, so the FE_low in Fig.4 does not seem to mean iron reduction.

3) The subject-data from the discussions by users: Hepatologists made a discussion looking at Fig.4, from which the words were regarded as new subject-data in the DH process and visualized with KeyGraph. From the result, shown in Fig.5, we can summarize the comments of hepatologists. First, the iron reduction for hepatitis of type B is not realistic. They explained transferrin, the protein denoted by F_B-GL in figures, carries iron to/from the liver and arranges the quantity of iron. Assuming iron reduction is not used, the discussion focused attention to "I-BIL_low" on the way to the "recovery" in Fig.4. This decrease in I-BIL (indirect bilirubin) seems to mean hem, a substance from hemoglobin to increase bilirubin, is carried back to the liver as in the normal status. This carrying is done by the protein F_A2-GL. Therefore, we can guess proteins such as F_A2-GL are both working actively on the way to recovery.

The discussion then paid attention to “FE_high” near “FE_low” in Fig.4, and suggested it is possible that FE_low and FE_high means ASC (asymptomatic carriers). In the case of an ASC, a clear symptom of progress is hardly observed but the patient may suffer from sudden worsening, and sometimes the wrong progress and recovery interleaves to lead to a fatal liver condition. However, sometimes a patient of ASC recovers. Because of missing symptoms, it is important to discover events significant for deciding to execute a proposed treatment, i.e. the chance to drive into the better scenario.

4) The effect of interferon, extracted as a result: Reflecting the new concern, we extracted the blood-test data of patients whose history includes “type-B & FE_low & FE_high,” and regarded each time of test as one sentence for KeyGraph. The result of KeyGraph for this extracted data is shown in Fig.6. We find the cycle where FE_high and FE_low appear in turns, and the events below occur in turns. That is, lipoprotein metabolism is weakened, with the decrease in hemoglobin (HBD_low). Then, iron and platelets in the blood decreases. Following this, kidney troubles and the decrease in amylase (AMY_low), implies liver cirrhosis. Finally, cancer and anomalous conditions of coagulation/fibrinolysis appear, and treatment operations or the

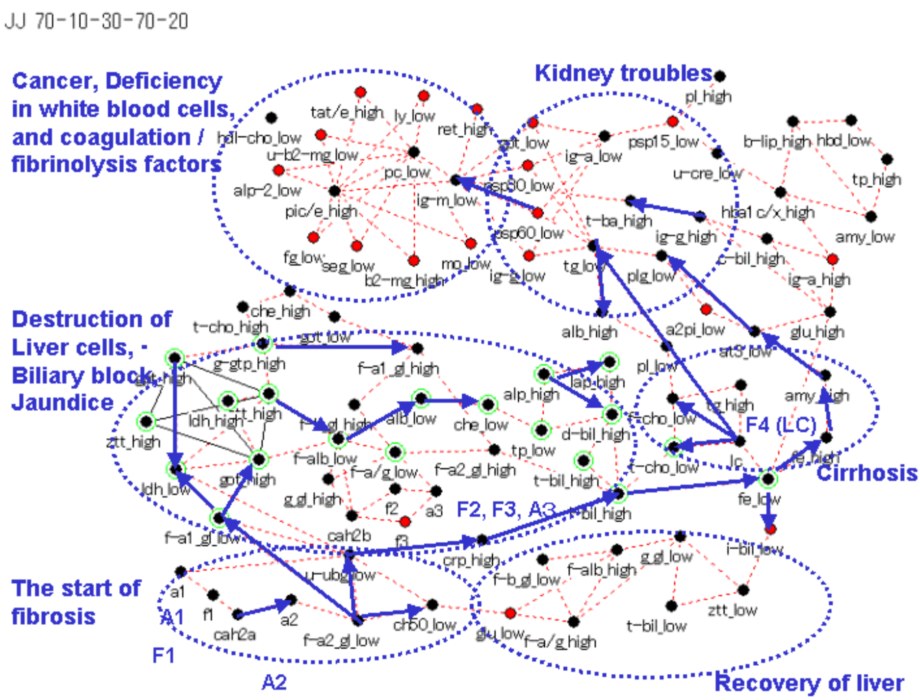


Fig. 4. The scenario map for hepatitis B, for the spotlights of F1, F2, F3, and F4 (LC)

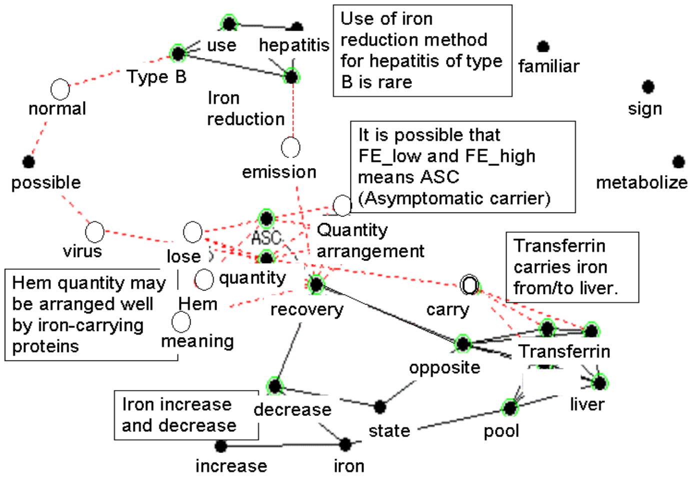


Fig. 5. KeyGraph for the comments of hepatologists looking at Fig.6

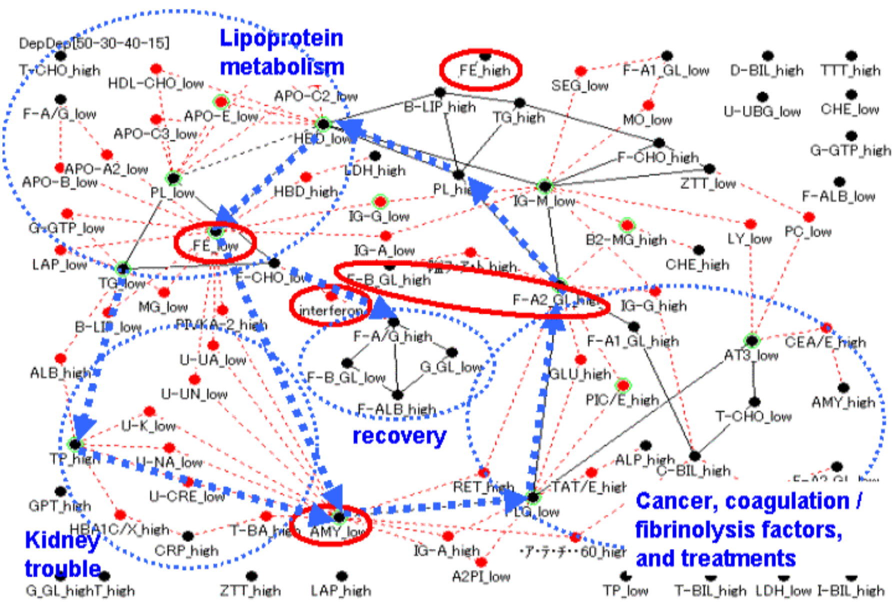


Fig. 6. The latest scenario map for hepatitis B

existence of cancer is implied by the high value of amylase (AMY_high) [Miyagawa et al 1996, Chougle 1992]. Then, following the activation of F_A2-GL and F_B-GL, iron in the blood increases, and the patient's hemoglobin decreases back. This cyclic process means the scenario of progress come after treatment. In Fig.6, we find that the recovery via FE_low, as was shown in Fig.4, had the use of interferon. This corresponds to the result of [Rubin et al 1995], and we can conclude the path to recovery in Fig.4 and Fig.6, was caused by interferon at the best timing i.e., at the time of iron decrease.

5 Conclusions

Scenario of the progress and the cure of hepatitis B were discovered. The relevance of iron and the disease has been implied in previous results, and this paper clarified that the decrease in iron shows a good timing for using interferon to hepatitis B. These are the effects of double helix (DH) process accelerated by using Polaris.

References

Chougle A; Hussain S; Singh PP; Shrimali R., 1992, Estimation of serum amylase levels in patients of cancer head and neck and cervix treated by radiotherapy, *Journal of Clinical Radiotherapy and Oncology*. 1992 Sept; 7(2): 24-26

Hayashi, H., T. Takikawa, N. Nishimura, M. Yano, T. Isomura, and N. Sakamoto. 1994. Improvement of serum aminotransferase levels after phlebotomy in patients with chronic active hepatitis C and excess hepatic iron, *American Journal of Gastroenterol.* 89: 986-988

Miyagawa S, Makuuchi M, Kawasaki S, Kakazu T, Hayashi K, and Kasai H., 1996, Serum Amylase elevation following hepatic resection in patients with chronic liver disease., *Am. J. Surg.* 1996 Feb;171(2):235-238

Ohsawa Y and McBurney P. eds, 2003, Chance Discovery, Springer Verlag

Ohsawa Y., 2003a, Modeling the Process of Chance Discovery, Ohsawa, Y. and McBurney eds, *Chance Discovery*, Springer Verlag pp.2—15 (2003)

Ohsawa Y, 2003b, KeyGraph: Visualized Structure Among Event Clusters, in Ohsawa Y and McBurney P. eds, 2003, *Chance Discovery*, Springer Verlag: 262-275

Okazaki N and Ohsawa Y, 2003, Polaris: An Integrated Data Miner for Chance Discovery, In *Proceedings of The Third International Workshop on Chance Discovery*, Crete, Greece

Rubin RB, Barton AL, Banner BF, Bonkovsky HL., 1995, Iron and chronic viral hepatitis: emerging evidence for an important interaction. in Digestive Diseases

A Novel Hybrid Approach for Interestingness Analysis of Classification Rules

Tolga Aydın and Halil Altay Güvenir

Department of Computer Engineering, Bilkent University,
06800 Ankara, Turkey
{atolga, guvenir}@cs.bilkent.edu.tr

Abstract. Data mining is the efficient discovery of patterns in large databases, and classification rules are perhaps the most important type of patterns in data mining applications. However, the number of such classification rules is generally very big that selection of interesting ones among all discovered rules becomes an important task. In this paper, factors related to the interestingness of a rule are investigated and some new factors are proposed. Following this, an interactive rule interestingness-learning algorithm (IRIL) is developed to automatically label the classification rules either as "interesting" or "uninteresting" with limited user participation. In our study, VFP (Voting Feature Projections), a feature projection based incremental classification learning algorithm, is also developed in the framework of IRIL. The concept description learned by the VFP algorithm constitutes a novel hybrid approach for interestingness analysis of classification rules.

1 Introduction

Data mining is the efficient discovery of patterns, as opposed to data itself, in large databases [4]. Patterns in the data can be represented in many different forms, including classification rules, association rules, clusters, sequential patterns, time series, contingency tables, and others [5]. However, the number of discovered patterns is usually very big and the user analyzing the patterns is generally interested in a subset of them. Therefore, selection of interesting patterns is an important research topic.

In this paper, we concentrate on the patterns represented by the classification rules and develop an interactive rule interestingness-learning algorithm (IRIL) to automatically classify these rules as interesting or uninteresting, with limited user participation. In our study, VFP (Voting Feature Projections), a feature projection based incremental classification-learning algorithm, was also developed in the framework of IRIL. Being specific to our concerns, VFP takes the rule interestingness factors as features and is used to learn the rule interestingness concept and to classify the newly learned classification rules. The concept description learned by the VFP algorithm constitutes a novel hybrid approach for interestingness analysis of the classification rules.

Section 2 describes the interestingness issue of patterns. Section 3 is devoted to the knowledge representation used in our study. Section 4 and 5 are related to the training

A. Sakurai et al. (Eds.): JSAI 2003/2004, LNAI 3609, pp. 485–496, 2007.

and classifying phases of the VFP algorithm. IRIL is explained in the following section. Giving the experimental results in Section 7, paper is concluded.

2 Interestingness Issue of Patterns

The interestingness issue has been important ever since the beginning of data mining research [1]. There are many factors contributing to the interestingness of a discovered pattern [1, 2, 3]. Some of them are coverage, confidence, completeness, action ability and unexpectedness. The first three factors are objective, action ability is subjective and unexpectedness is sometimes regarded as subjective [7, 8, 9] and sometimes as objective [10, 11]. Objective interestingness factors can be measured independently of the user and domain knowledge. However, subjective interestingness factors are not user and domain knowledge independent. The measurement of a subjective interestingness factor may vary among users analyzing a particular domain, may vary among different domains that a particular user is analyzing and may vary even for the same user analyzing the same domain at different times.

An objective interestingness measure is constructed by combining a proper subset of the objective interestingness factors in a suitable way. For example, objective interestingness factor x can be multiplied by the square of another objective interestingness factor y to obtain an objective interestingness measure of the form xy^2. It is also possible to use an objective interestingness factor x alone as an objective interestingness measure (e.g. *Confidence*). Discovered patterns having *Confidence* ≥ *threshold* are regarded as "interesting". Although the user determines the threshold, this is regarded as small user intervention and the interestingness measure is still assumed to be an objective one.

The existing subjective interestingness measures in the literature are constructed upon unexpectedness and action ability factors. Assuming the discovered pattern to be a set of rules induced from a domain, the user gives her knowledge about the domain in terms of fuzzy rules [9], general impressions [8] or rule templates [7]. The induced rules are then compared with user's existing domain knowledge to determine subjectively unexpected and/or actionable rules.

Both types of interestingness measures have some drawbacks. A particular objective interestingness measure is not sufficient by itself [9]. They are generally used as a filtering mechanism before applying a subjective measure. On the other hand, subjective measures are sometimes used without prior usage of an objective one. In the case of subjective interestingness measures, user may not be well in expressing her domain knowledge at the beginning of the interestingness analysis. It'd be better to automatically learn this knowledge based on her classification of some presented rules as "interesting" or "uninteresting". Another drawback of a subjective measure is that the induced rules are compared with the domain knowledge that addresses the unexpectedness and/or action ability issues. Interestingness is assumed to depend on these two issues. That is, if a rule is found to be unexpected, it is automatically regarded as an interesting rule. However, it would be better if we learned a concept description that dealt with the interestingness issue directly and if we benefited from unexpectedness and action ability as two of the factors used to express the concept description. That is, interestingness of a pattern may depend on factors other than unexpectedness and action ability issues.

The idea of a concept description that is automatically determined and directly related with the interestingness issue motivated us to design IRIL algorithm. The concept description learned by the VFP algorithm, which was also developed in this framework, constitutes a novel hybrid approach for interestingness analysis of classification rules.

To ensure that the concept description is directly related to the rule interestingness issue, some existing and newly developed interestingness factors that have the capability to determine the interestingness of rules were used instead of the original attributes of the data set. Current implementation of IRIL does not incorporate unexpectedness and action ability factors, requiring no need for domain knowledge. Although all the interestingness factors are of type objective in the current version of IRIL, the thresholds of the objective factors are learned automatically rather than expressing them manually at the beginning. The values of these thresholds are based upon the user's classification results of some presented rules. So, although in the literature subjectivity is highly related to the domain knowledge, IRIL differs from them. IRIL's subjectivity is not related with the domain knowledge. IRIL makes use of objective factors (actually the current version makes use of only objective factors) but for each such factor, it subjectively learns what ranges of factor values (what thresholds) lead to interesting or uninteresting rule classifications if only that factor is used for classification purposes. That is, IRIL presents a hybrid interestingness measure.

IRIL proceeds interactively. An input rule is labeled if the learned concept descripion can label the rule with high certainty. If the labeling or classification certainty factor is not of sufficient strength, user is asked to classify the rule manually. The user looks at the values of the interestingness factors and labels the rule accordingly. In IRIL, concept description is learned or updated incrementally by using the interestingness labels of the rules that are on demand given either as "interesting" or "uninteresting" by the user.

3 Knowledge Representation

The aim of the study presented in this paper is to label a set of classification rules as interesting or uninteresting. This labeling problem is modeled as a new classification problem and a *rule set* is produced for the given rules, which are previously learned by applying a rule induction algorithm on a data set. Each instance of the rule set is represented by a vector whose components are the interestingness label and the interestingness factor values having the potential to determine the interestingness of the corresponding rule.

The classification rules used in the study are probabilistic and have the following general structure:

If (A_1 *op value*$_1$) AND (A_2 *op value*$_2$) AND …AND (A_n *op value*$_n$) THEN
($Class_1$: $probability_1$, $Class_2$: $probability_2$,…,$Class_k$: $probability_k$)

In the above structure, A_i's are the features, $Class_i$'s are the classes and $op \in \{=, \neq, <, \leq, >, \geq\}$.

The instances corresponding to probabilistic classification rules have either "interesting" or "uninteresting" as the interestingness label, and the interestingness factors shown in Table 1. In this new classification problem, these factors are treated as determining features, and interestingness label is treated as the target feature of the rule set.

Table 1. Features of the rule set

Feature	**Short description and/or formula**
Major Class	$Class_i$ that has the highest probability
Major Class Frequency	Ratio of the instances having $Class_i$ as the class label in the data set
Rule Size	Number of conditions in the antecedent part of the rule
Confidence with respect to Major Class	$\|Antecedent$ & $Class_i\|$ / $\|Antecedent\|$
Coverage	$\|Antecedent\|$ / $\|N\|$
Completeness with respect to Major Class	$\|Antecedent$ & $Class_i\|$ / $\|Class_i\|$
Number of Classes with Zero Probability	Number of classes having zero probability
Standard Deviation of Class Probabilities	Standard deviation of the class probabilities
Major Class Probability	Maximum probability value
Minor Class Probability	Minimum probability value
Decisive	True if Std.Dev.of Class Probabilities > s_{min}

Each feature carries information about a specific property of the corresponding rule. For example, if we let $Class_i$ to take the highest probability, it then becomes the *Major Class* of that classification rule. If we shorten the representation of any rule as "If *Antecedent* THEN $Class_i$" and assume the data set to consist of N instances, we can define *Confidence*, *Coverage* and *Completeness* as in Table 1. Furthermore, a rule is decisive if the standard deviation of the class probabilities is greater than s_{min}, whose definition is given in the following equation:

$$s_{min} = \frac{1}{(Class\ Count - 1)\sqrt{Class\ Count}} \quad (1)$$

If all the classes have equal probability in a rule, then the standard deviation of the probabilities becomes zero and the rule becomes extremely indecisive. This is the worst distribution that can happen. The next worst distribution is obtained if exactly one class has a zero probability, and the remaining classes have equal probability. The standard deviation of the probability values in such a situation is called s_{min}.

4 Training in VFP Algorithm

VFP (Voting Feature Projections) is a feature projection based classification-learning algorithm developed in our study. It is used to learn the rule interestingness concept and to classify the unlabeled rules in the context of modeling rule interestingness problem as a new classification problem.

The training phase of VFP, given in Figure 3, is achieved incrementally. On a nominal feature, concept description is shown as the set of points along with the numbers of instances of each class falling into those points. On the other hand, on a numeric feature, concept description is shown as the normal (gaussian) probability density functions for each possible class. Training can better be explained by looking at the sample data set in Figure 1, and the associated learned concept description in Figure 2.

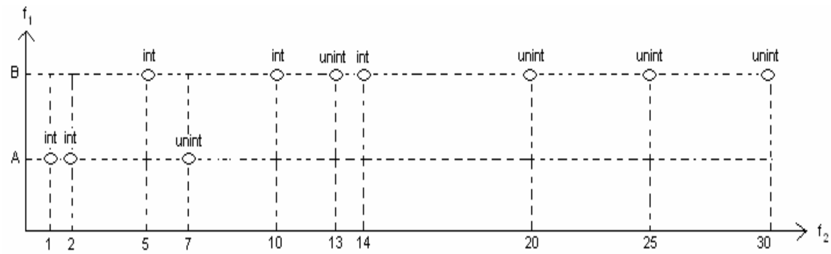


Fig. 1. Sample data set

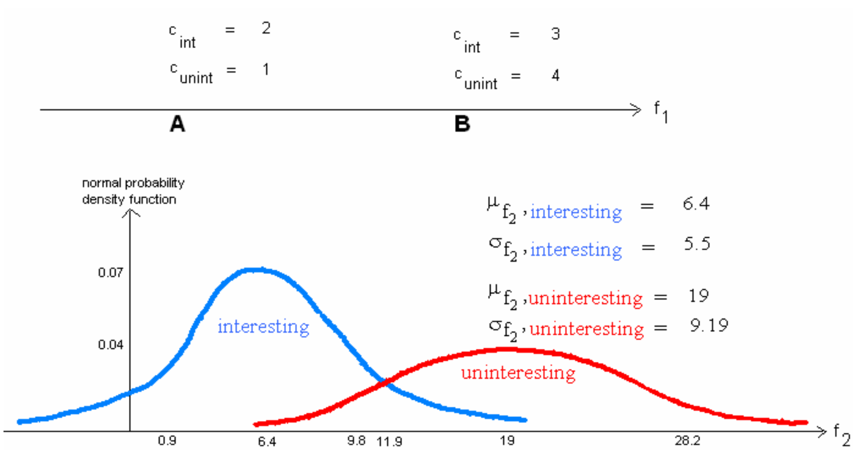


Fig. 2. Concept description learned for the sample data set

The example data set consists of 10 training instances, having nominal f_1 and numeric f_2 features. f_1 takes two values: "A" and "B", whereas f_2 takes some integer values. There are two possible classes: "interesting" and "uninteresting". f_2 is assumed to have gaussian probability density functions for both classes.

```
VFP_train (t)        /* t: newly added training instance */
begin
   let c be the class of t
   let others be the remaining classes other than c
   if training set = {t}
      for each class s
         class_count[s] = 0

   class_count[c]++

   for each feature f
      if f is nominal
         p = find_point(f,t_f)
         if such a p exists
        /* if t_f value exists in the training set */
            point_class_count [f,p,c] ++
         else  /* add new point for f */
            add a new p' point
            point_class_count [f,p',c] = 1
            point_class_count [f,p',others] = 0
      else if f is numeric
        if training set = {t}
           μ_f,c = t_f          ,       μ_f,others = 0
           μ²_f,c = t_f²        ,       μ²_f,others = 0
           σ_f,c = Undefined
           norm_density_func._f,c = Undefined
        else
           n = class_count[c]
           μ_f,c = (μ_f,c * (n-1) + t_f) / n        /*update*/
           μ²_f,c = (μ²_f,c * (n-1) + t_f²) / n   /*update*/
```

$$\sigma_{f,c} = \sqrt{\frac{n}{n-1}\left(\mu^2_{f,c} - (\mu_{f,c})^2\right)}$$

$$norm\_density\_func._{f,c} = \frac{1}{\sigma_{f,c}\sqrt{2\pi}} e^{-\frac{(x-\mu_{f,c})^2}{2\sigma_{f,c}^2}}$$

```
   return  { For numeric features:
           {   norm_density_func._f,c (∀f, c)
           { For nominal features:
           {   point_class_count[f, p, c] (∀f, p, c)
end.
```

Fig. 3. Incremental training in VFP

In Figure 3 for a nominal feature *f*, *find_point* (f, t_f) procedure tries to find (t_f), the new training instance's value at feature *f*, in the *f* projection. If t_f is found at a point *p*, then *point_class_count* [*f*, *p*, *c*] is incremented, assuming that the training instance is of class *c*. If t_f is not found, then a new point *p'* is constructed and *point_class_count* [*f*, *p'*, *class*] is initialized to 1 for *class* = *c*, and to 0 for all other classes. In our study, features used in VFP are the interestingness factor values computed for the classification rules, and we have only "interesting" and "uninteresting" as the classes.

For a numeric feature *f*, if a new training instance *t* of class *c* is examined, we let the previous training instances of class *c* to construct a set *P* and let $\mu_{f,c}$ and $\sigma_{f,c}$ to be the mean and the standard deviation of the *f* feature projection values of the instances in *P*, respectively. The previous training instances' values on *f* need not be stored anywhere, so $\mu_{f,c}$ and $\sigma_{f,c}$ are updated incrementally. Updating $\sigma_{f,c}$ incrementally requires $\mu^2_{f,c}$ to be updated incrementally, as well.

5 Classification in VFP Algorithm

Classification phase of VFP is shown in Figure 4. The query instance is projected on all features and each feature gives votes for each class. If a feature is not ready for classification process, it gives zero, otherwise gives normalized votes. Normalization ensures that each feature has the same weight in classifying the query instances. However, if a feature is not ready, it is not involved in the classification process, therefore need not give normalized votes. For a feature to be ready for the classification process, it should have at least two different values for each class.

The classification starts by giving zero votes to classes on each feature projection. The features that are not ready do not participate in the classification process. The participating features are handled accordingly. For a nominal feature *f*, *find_point* (f, q_f) procedure is used to search whether q_f exists in the *f* projection. If q_f is found at a point *p*, feature *f* gives votes for each class as shown in the equation below, and then these votes are normalized to ensure equal voting power among features.

$$feature\_vote\,[f, c] = \frac{point\_class\_count\,[f, p, c]}{class\_count\,[c]} \tag{2}$$

In the above equation, we divide the number of class *c* instances on point *p* of feature projection *f* by the total number of class *c* instances to find the class conditional probability of falling into the *p* point. For a linear feature *f*, each class gets the vote given in equation 3. Normal probability density function values are used as the vote values. These votes are then normalized, too.

$$feature\_vote\,[f, c] = \lim_{\Delta x \to 0} \int_{q_f}^{q_f + \Delta x} \frac{1}{\sigma_{f,c}\sqrt{2\pi}}\, e^{-\frac{(q_f - \mu_{f,c})^2}{2\sigma_{f,c}^2}}\, dx \tag{3}$$

```
VFP_query(q)            /* q: query instance*/
begin

   for each feature f
      for each class c
         feature_vote[f,c] = 0

      if feature_ready_for_query_process(f)

         if f is nominal
            p = find_point(f,q_f)
            if such a p exists
            /* if q_f value exists in the training set */
               for each class c
                  feature_vote [f,c] = point_class_count[f,p,c] / class_count[c]
               normalize_feature_votes (f)
               /* such that Σ_c feature_vote[f,c] = 1 */

         else if f is numeric
            for each class c
```

$$g = \frac{1}{\sigma_{f,c}\sqrt{2\pi}} e^{-\frac{(q_f - \mu_{f,c})^2}{2\sigma_{f,c}^2}}$$

$$feature\_vote\,[f,c] = \lim_{\Delta x \to 0} \int_{q_f}^{q_f+\Delta x} g\,dx$$

```
            normalize_feature_votes (f)

   for each class  c
```

$$final\_vote\,[c] = \sum_{f=1}^{\#Features} feature\_vote[f,c]$$

```
   for each class  c
```

$$\text{if } \min_{i=1}^{\#Classes} final\_vote[i] < final\_vote\,[c] = \max_{i=1}^{\#Classes} final\_vote[i]$$

```
      classify q as "c" with a certainty factor C_f
      return C_f
    else
      C_f = -1
      return C_f
end.
```

Fig. 4. Classification in VFP

Final vote for any class *c* is the sum of all votes given by the features. If there exists a class *c* that gets the highest vote and there also exists at least one other class that gets a lower vote than *c*, then class *c* is predicted to be the class of the query instance. The certainty factor of the classification (C_f) is computed as follows:

$$C_f = \frac{final\_vote\,[c]}{\sum_{i=1}^{\#Classes} final\_vote\,[i]} \tag{4}$$

If no prediction is made, certainty factor is taken as "–1" to indicate this situation.

6 IRIL Algorithm

IRIL algorithm, shown in Figure 5, needs two input parameters: *R* (The set of classification rules) and $MinC_t$ (Minimum Certainty Threshold). It tries to classify the rules in *R*. If $C_f \geq MinC_t$ for a query rule *r*, this rule is inserted into the successfully classified rules set (R_s). Otherwise, two situations are possible: either the concept description is not able to classify *r* (C_f = -1), or the concept description's classification (prediction of *r*'s interestingness label) is not of sufficient strength. If $C_f < MinC_t$, rule *r* is presented, along with its computed eleven interestingness factor values such as *Coverage*, *Rule Size*, *Decisive* etc., to the user for classification. This rule or actually the instance holding the interestingness factor values and the recently

```
IRIL ( R, MinC_t)
begin
   R_t ←∅,    R_s ← ∅
   repeat
      for each rule r ∈ R
         C_f ←VFP_query (r)
            if  C_f < MinC_t
               ask the user to classify r
               set C_f of  this classification to 1
               insert r into R_t
               VFP_train (r)
            else
               add r into R_s
            remove r from R

      for each rule r ∈ R_s
         C_f ←VFP_query (r)
            if  C_f  < MinC_t
               remove r from R_s
               add r into R
   until R is empty
   output rules in R_s
end.
```

Fig. 5. IRIL algorithm

determined interestingness label of this rule is then inserted into the training rule set R_t and the concept description is reconstructed incrementally.

All the rules in R are labeled either automatically by the classification algorithm, or manually by the user. User participation leads rule interestingness learning process to be an interactive one. When the number of instances in the training rule set increases, the concept description learned tends to be more powerful and reliable. When the labeling of the rules ends, the rules in R_s are relabeled by the latest version of the concept description. Because there may exist some rule r that was classified as "interesting" with a sufficient certainty factor by a weak version of the concept description, but now labeled as "interesting" or "uninteresting" with an insufficient certainty factor by the latest and the most reliable version of the concept description. Such rules called as R_{exc} are excluded from R_s and inserted into R. Therefore, we have $R = R_{exc}$ and $R_s = R_s - R_{exc}$. The cycle is repeated until R gets empty and IRIL concludes by presenting the labeled rules in R_s. It is guaranteed that the number of cycles is not infinite and R eventually gets empty. Proof is as follows:

At the end of any cycle, if $R_{exc} = \{\}$ then we are done. If $R_{exc} \neq \{\}$, then at least one rule will be classified by the user and then added into the R_t since the current version of the concept description could not classify the rules in R_{exc} with sufficient certainty. Unless $R_{exc} = \{\}$, at the end of each cycle R_t will expand by at least one element. Therefore, the cycle will be repeated $|R|$ times at most.

7 Experimental Results

IRIL algorithm was tested to classify 184 classification rules induced from a financial distress domain using a benefit maximizing feature projection based rule learner proposed in [6]. The data set of the financial distress domain is a comprehensive set consisting of 25632 data instances and 164 determining features (159 numeric, 5 nominal). There are two classes: "Profit" and "Loss". The data set includes some financial information about 3000 companies collected during 10 years and the class feature states whether the company made a profit or loss in a particular year. Domain expert previously labeled all the 184 induced rules to make accuracy measurement possible. The expert labeled 50 rules (27.17%) as "interesting" and 134 rules (72.83%) as "uninteresting".

The results for $MinC_t = 60\%$ shows that the user classifies 54 rules with 100% certainty, and 130 rules are classified automatically with $C_f > MinC_t$. User participation is 29% in the classification process. While labeling the rules, user participation increases in proportion to the $MinC_t$ as expected. In the classification process, it is always desired that rules are generally classified automatically, and user participation is low.

If we look at the accuracy results for $MinC_t = 60\%$, they are measured as 80%, 94.87% and 73.63% for the rules in R_s (overall accuracy), for the actually interesting rules in R_s (accuracy among interesting rules) and for the actually uninteresting rules in R_s (accuracy among uninteresting rules), respectively. It is important to keep the three accuracy values close to each other. For instance, if the above three accuracy values were 65%, 20% and 75%, respectively, we would easily claim that IRIL made

Table 1. Results for IRIL

	$MinC_t$ 60%	$MinC_t$ 65%	$MinC_t$ 75%
Number of rules	184	184	184
Number of rules classified automatically with high certainty	130	108	90
Number of rules classified by user	54	76	94
User participation	29%	41%	51%
Overall Accuracy	80%	87.04%	90%
Accuracy among interesting rules	94.87%	95.45%	95.12%
Accuracy among uninteresting rules	73.63%	81.25%	85.71%

biased classifications in favor of "uninteresting" class. Because, accuracy among uninteresting rules is too high, whereas accuracy among interesting rules is too low. Furthermore, user herself labels 134 of the rules (72.83%) as uninteresting, so we could label all the rules as "uninteresting" without using IRIL that would result in an accuracy value of 72.83%, which is very close to the overall accuracy of 65%. Fortunately, IRIL makes unbiased classifications since the three accuracy values are balanced. The accuracy values generally increase in proportion to the $MinC_t$. Because the higher the $MinC_t$, is, the higher the user participation is. And higher user participation leads to learning a more powerful and predictive concept description.

8 Conclusion

(IRIL feature projection based interactive rule interestingness learning algorithm) was developed and gave promising experimental results. The concept description learned by the VFP algorithm, also developed in the framework of IRIL, constitutes a novel hybrid approach for interestingness analysis of classification rules. The concept description differs among the users analyzing the same domain. That is, IRIL determines the important rule interestingness factors for a given domain subjectively, by making use of objective factors.

As future work, other classification learning algorithms, which need not be feature projection based, can be used in the framework of IRIL. On the other hand, other objective and subjective interestingness factors, especially unexpectedness, may be used as the features of the rule sets.

References

1. Frawely, W.J., Piatetsky-Shapiro, G., and Matheus, C.J., "Knowledge discovery in databases: an overview" *Knowledge Discovery in Databases*, AAAI/MIT Press, 1991, 1-27
2. Major, J.A., and Mangano, J.J., "Selecting among rules induced from a hurricane database" *Proceedings of AAAI Workshop on Knowledge Discovery in Databases*, 1993, 30-31

3. Piatetsky-Shapiro, G., and Matheus, C.J., “The interestingness of deviations” *Proceedings of AAAI Workshop on Knowledge Discovery in Databases*, 1994, 25-36
4. Fayyad, U., Shapiro, G., and Smyth, P., “From data mining to knowledge discovery in databases” *AI Magazine 17(3)*, 1996, 37-54
5. Hilderman, R.J., and Hamilton, H.J., “Knowledge discovery and interestingness measures: a survey” *Technical Report*, Department of Computer Science, University of Regina, 1999
6. Güvenir, H.A., “Benefit Maximization in Classification on Feature Projections” *Proceedings of the 3rd IASTED International Conference on Artificial Intelligence and Applications* (AIA’03), 2003, 424-429.
7. Klemettinen, M., Mannila, H., Ronkainen, P., Toivonen, H., and Verkamo, A.I., “Finding interesting rules from large sets of discovered association rules” *Proceedings of the 3rd Int. Conf. on Information and Knowledge Management*, 1994, 401-407.
8. Liu, B., Hsu, W., and Chen, S., “Using general impressions to analyze discovered classification rules” *Proceedings of the 3rd Int. Conf. on KDD*, 1997, 31-36.
9. Liu, B., and Hsu, W., “Post-analysis of learned rules”, *AAAI*, 1996, 828-834.
10. Hussain, F., Liu, H., Suzuki, E., and Lu, H., “Exception rule mining with a relative interestingness measure” *Proceedings of the Pacific-Asia Conference on Knowledge Discovery and Data Mining*, 2000, 86-97.
11. Dong, G., and Li, J., “Interestingness of discovered association rules in terms of neighborhood-based unexpectedness” *Proceedings of the 2nd Pacific-Asia Conference on Knowledge Discovery and Data Mining*, 1998, 72-86.

Preliminary Evaluation of Discovered-Rule-Filtering Methods

Yasuhiko Kitamura[1], Akira Iida[2], and Keunsik Park[3]

[1] School of Science and Technology, Kwansei Gakuin University,
2-1 Gakuen, Sanda, Hyogo 669-1337, Japan
ykitamura@ksc.kwansei.ac.jp
http://ist.ksc.kwansei.ac.jp/~kitamura/
[2] Graduate School of Engineering, Osaka City University,
3-3-138, Sugimoto, Sumiyoshi-ku, Osaka, 558-8585
{iida, tatsumi}@kdel.info.eng.osaka-cu.ac.jp
[3] Graduate School of Medicine, Osaka City University,
1-4-3, Asahi-Machi, Abeno-ku, Osaka, 545-8585
kspark@msic.med.osaka-cu.ac.jp

Abstract. Data mining systems semi-automatically discover knowledge by examining large volumes of data, but the knowledge so discovered is not always novel to users. We introduce a discovered-rule-filtering approach that uses information retrieval results from the Internet to assess rules discovered by data mining and find those that are novel to the user. To implement this approach, we create 2 methods: the micro view method and the macro view method. In the micro view method, we extract keywords from a discovered rule and rank the rule referring to the number of hits returned when the keywords are submitted to an appropriate database. In the macro view method, we first retrieve documents by submitting every pair of extracted keywords and then form keyword clusters according to the results. We evaluated the methods by sending out a questionnaire to medical students and using the MEDLINE database as our Internet source. The evaluation indicates that the macro view method is promising.

1 Introduction

Active mining [1] is a new approach to data mining; it tries to discover "high quality" knowledge that meets the user's demand in an efficient manner by integrating information gathering, data mining, and user reaction technologies. This paper introduces a discovered-rule-filtering method [3,4,5] that extracts from the large number of rules output by a data mining system a small number of novel rules by using information retrieved from the Internet.

Data mining is an automated method to discover useful knowledge by analyzing large volumes of data mechanically [6]. Generally speaking, conventional data mining methods try to discover statistically significant patterns from a large volume of raw data contained in a given database. Unfortunately, considering only statistically significant features yields a large number of rules, most of which are already known to the user. To cope with this problem, our discovered-rule-filtering approach winnows the large number of rules returned by a data mining system to find the small number of rules that are novel to the user. To judge whether a rule is novel or not, we utilize

A. Sakurai et al. (Eds.): JSAI 2003/2004, LNAI 3609, pp. 497 – 506, 2007.

an information source on the Internet and judge its novelty according to the number of retrieved documents that relate to the rule.

This paper introduces two implementations of the proposed approach: the micro view method and the macro view method. We evaluate the methods by questioning medical students. We describe the concept and the process of discovered rule filtering using the example of clinical data mining in Section 2. We then show the micro view and the macro view methods in Section 3 and evaluate them in Section 4 by questionnaire. Finally we conclude this paper with our future work in Section 5.

2 Discovered Rule Filtering

The target of our active mining project is a clinical examination database of hepatitis patients, which is offered by the Medical School of Chiba University; on which 10 research groups cooperatively work as a common data source [7]. Several groups have already discovered some sets of rules. For example, Yamaguchi et al. in Shizuoka University analyzed sequential trends between GPT (Glutamic Pyruvic Transaminase), which represents the progress of hepatitis, and other blood test data, and has discovered a number of rules, one of which is shown in Fig. 1 [8].

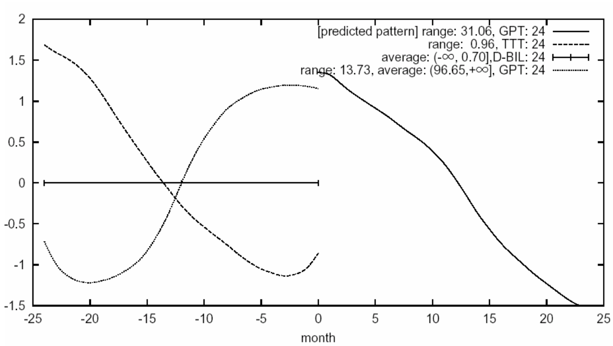


Fig. 1. An example of discovered rule [8]

This rule shows a relation among GPT, TTT (Thymol Turbidity Test), and D-BIL (Direct Bilirubin) and can be expressed as “If, for the past 24 months, D-BIL stays unchanged, TTT decreases, and GPT increases, then GPT will decrease for the following 24 months.” A data mining system can semi-automatically discover a large number of rules by analyzing a set of given data, but the discovered rules are likely to include many that are well known to the user. Showing all the discovered rules without exception does not help the user. We need to be able to identify and shows only those rules that are novel to the user. We propose to judge whether a rule is novel or not by utilizing information sources on the Internet; novelty is inversely proportional to the number of retrieved documents related to the discovered rule.

When a set of discovered rules is output by a data mining system, the discovered-rule-filtering system first retrieves information related to the rules from the Internet and then filters the rules based on the information retrieval results. In our study, we are interested in discovering knowledge in a hepatitis clinical database, and it is not easy to gather information related to hepatitis from the Internet by using a naïve search engine because Internet information sources generally contain a huge amount of various and noisy information. We use the MEDLINE (MEDlars on LINE) database as the information source since it is the largest bibliographical database in the medical and biological domain. PubMed (http://www.ncbi.nlm.nih.gov/entrez/query.fcgi) is a free MEDLINE search service on the Internet run by NCBI (National Center for Biotechnology Information). By using Pubmed, we can retrieve MEDLINE documents by submitting a set of keywords just like a normal search engine. In addition, we can retrieve documents according to the year of publication and/or the category of documents. These functions are not provided by normal search engines.

The discovered-rule-filtering process takes the following steps.

Step 1: Extracting keywords from a discovered rule
We need to develop a set of proper keywords to retrieve MEDLINE documents that relate to a discovered rule. Such keywords are extracted from the rule and the domain of data mining as follows.

- **Keywords extracted from a discovered rule.** These keywords represent attributes in a discovered rule. For example, keywords that can be extracted directly from the discovered rule shown in Fig. 1 are GPT, TTT, and D-BIL because they explicitly appear in the rule. If any abbreviation is not acceptable to Pubmed, it is translated into its full normal name. For example, TTT and GPT are translated into "thymol turbidity test" and "glutamic pyruvic transaminase", respectively.
- **Keywords related to the mining domain.** These keywords represent the purpose or the domain of the data mining task. Together with keywords extracted from the rule, they are submitted to the Pubmed as the common keywords to improve the quality of retrieved documents. For hepatitis data mining, "hepatitis" is a domain keyword. The domain keywords are implicit keywords, and we do not directly refer to such keywords hereafter.

The rule shown in Fig.1 includes information not only about relations among attributes but also about how the attributes change, but it is difficult to represent the latter bit of information in a sequence of keywords. This problem is left as future work.

Step 2: Filtering the Discovered Rules
We filter the discovered rules by using the results of MEDLINE document retrieval. We have created the micro view method and the macro view method for this purpose. The methods are detailed in the following section.

3 Two Methods for Discovered Rule Filtering

How to filter discovered rules according to the search result of MEDLINE document retrieval is a most important issue of this work. We have two methods; the micro view method and the macro view method, to filter discovered rules [5].

3.1 Micro View Method

The micro view method retrieves documents directly related to a discovered rule. It utilizes the document retrieval results not only to filter the discovered rules, and shows the rules and the documents to the user. This allows the user to expand her insights on the rule and the data mining task [3]. The micro view method is quite simple and is based on the following hypotheses.

[Hypotheses] (Micro View Method)
1. The number of documents related to a known rule is large.
2. The number of documents related to an unknown rule is small.
3. The number of documents related to a garbage rule is zero

We hypothesize that known rules have been the subject of many papers. On the other hand, unknown rules have been the subject of only a few papers. Nobody has any interest in garbage or nonsense rules, and so no papers are related to them.

The first filtering stage is to eliminate garbage rules. Since the border between known rules and unknown ones is vague, the second stage ranks rules using the number of related documents.

We note that the performance of the micro view method strongly depends on the performance of document retrieval, and that existing keyword-based document retrieval techniques do not well support the retrieval of appropriate documents related to any particular rule. Generally speaking, when a rule is simple with a small number of attributes, the Pubmed system returns a large number of unrelated noisy documents. When a rule is complicated with a large number of attributes, it returns only few documents.

3.2 Macro View Method

The macro view method tries to roughly observe the research trends implicit in each discovered rule. Given a rule, it submits every pair of keywords extracted from the rule, not the whole sequence of the keywords, to the Pubmed system, and integrates the results in the form of a keyword co-occurrence graph to judge the novelty of the rule.

Figures 2, 3, and 4 show keyword co-occurrence graphs. In each graph, a node represents a keyword and edge length represents the inverse of the frequency of co-occurrences of the keywords connected by the edge. The score attached to the edge represents the frequency of co-occurrence. Hence, the more documents related to a pair keywords are retrieved from Pubmed, the closer the keywords are located in the graph.

For example, Fig. 2 shows that the relation between all pairs formed from ALB, GPT, and T-CHO is strong. Fig. 3 shows that the relation between T-CHO and GPT is strong, but that between chyle and either of T-CHO or GPT is rather weak. Fig. 4 shows that the relations among GPT, female, and G-GTP are strong, but the relation between hemolysis and G-GTP and those between "blood group a" and the other keywords are weak.

We then form clusters of keywords by using the Hierarchical Clustering Scheme [9]. As a strategy to form clusters, we adopt the complete linkage clustering method

(CLINK). In the method, the distance between clusters A and B is defined as the longest among the distances of every pair of keywords in cluster A and a keyword in cluster B. The method initially forms a cluster for each keyword. It then repeatedly merges clusters within a threshold length into one or more clusters.

We consider that cluster number is strongly related to research activity as follows.

[Hypothesis] (Macro View Method)
1. The number of clusters concerning a known rule is 1.
2. The number of clusters concerning an unknown rule is 2.
3. The number of clusters concerning a garbage rule is more than 3.

A rule with only one cluster is regarded as a known rule because a large number of papers that use all pairs of keywords in the rule have been published. A rule with two clusters is regarded as an unknown rule. This is because each cluster represents a lot research activity, but little cross-cluster research has been done. A rule with more than two clusters is regarded as a garbage rule. Such a rule is too complex to understand because the keywords are partitioned into many clusters and the rule consists of many unknown factors.

For example, if we set the threshold of CLINK to 1 (the frequency of co-occurrences is 1), the rule in Fig. 2 is regarded as a known rule because all the

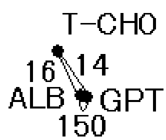


Fig. 2. The keyword co-occurrence graph of rule including GPT, ABL, and T-CHO

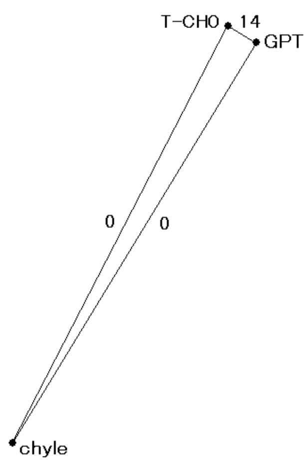


Fig. 3. The keyword co-occurrence graph of rule including GPT, T-CHO, and chyle

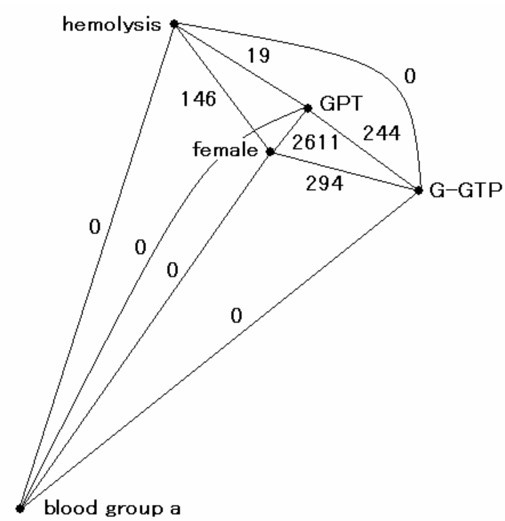


Fig. 4. The keyword co-occurrence graph of rule including GPT, G-GTP, hemolysis, female and "blood group a"

keywords are merged into a single cluster. Keywords in Fig. 3 are merged into two clusters; one cluster consists of GPT and T-CHO and another consists of chyle only. Hence, the rule is judged to be unknown. Keywords in Fig. 4 are merged into 3 clusters as GPT, G-GTP, and female form a cluster and each of hemolysis and "blood group a" forms a different cluster.

4 Evaluation of Discovered-Rule-Filtering Methods

4.1 Questionnaire

We evaluated our discovered-rule-filtering methods by the questionnaire method. The intent was to verify the hypotheses underlying the micro view and the macro view methods. We first made a questionnaire containing two questions as shown in Fig. 5. The 20 items in Q1 came from rules discovered by the data mining group in Shizuoka University [8] by extracting keywords from the rules. The reason why we did not show the discovered rules to the subjects is because we would like to evaluate the effectiveness of showing only attribute keywords extracted from the rules. If we show the discovered rules directly to the subjects, the subjects might judge them by considering more than just the relation among attribute keywords, ex. how the attributes change, which would invalidate the evaluation.

The 20 items in Q2 were randomly chosen from keywords in the discovered rules. The purpose of Q2 was to show how the number of retrieved documents correlates to the evaluation of medical students when we limit the number of submitted keywords to 2, as discussed in Section 4.2.

We sent out the questionnaire to 47 medical students in Osaka City University. The students were soon going to take the state examination to become a medical doctor, so we could assume that they were well briefed about the latest medical knowledge.

Q1: How do you guess the result when you submit the following keywords to the Pubmed system? Choose one among A, B, and C.

A (Known): Documents about a fact that I know will be retrieved.
B (Unknown): Documents about a fact that I do not know will be retrieved.
C (Garbage): No document will be retrieved.

(1) [A B C] ALT and TTT
(2) [A B C] TTT, Direct-Bilirubin, and ALT
(3) [A B C] ALT, Total-Cholesterol, and Hepatitis C
(4) ….

Q2: Choose one among four choices about the relation between the following items.

A: The items have a strong relation with each other.
B: The items have a medium relation with each other.
C: The items have a weak relation with each other.
D: The items have no relation with each other.

(1)[A B C D] ALT, Total-Bilirubin
(2)[A B C D] ALT, Total-Cholesterol
(3)[A B C D] ALB, Total-Cholesterol
(4) …

Fig. 5. Questionnaire sent to medical students

4.2 Evaluation of the Micro View Method

We here evaluate the micro view method by analyzing the relation between the number of documents hit by the keywords and the ratio of choices in Q1 made by medical students. Fig. 6 shows the result. We plot the relation between the ratio of choice and the number of retrieved documents for each item in Q1. We also add regression lines to show the relation more clearly. We assessed the significance by using the t-test method at the risk level of 5%, but could not find any significant relation.

The reason why the micro view method, in which all the keywords extracted from a rule are directly submitted to Pubmed, does not work well is because the number of hits seems to depend so strongly on the number of keywords submitted to Pubmed. Generally speaking, increasing the number of submitted keywords, decreases the number of retrieved documents.

If we limit the number of submitted keywords to 2, the method shows better performance. Fig. 7 shows the relation between the number of documents and the evaluation of medical students obtained from Q2. The number of keywords used in Q2 was fixed at 2. We scored choices A, B, C, and D as 3, 2, 1, and 0, respectively. The averaged score and the number of documents are significantly correlated since the correlation coefficient is 0.54. Hence, if we limit the number of submitted keywords to 2, the result reflects the evaluation of medical students.

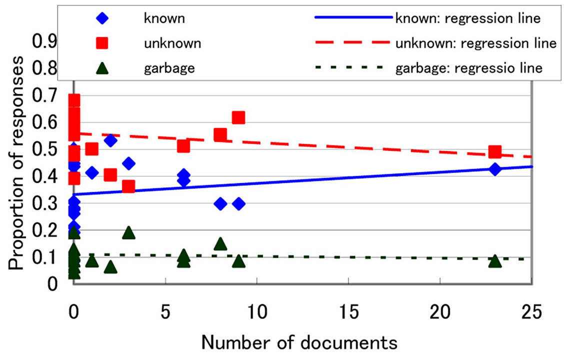


Fig. 6. The relation between the ratio of choice and the number of documents

4.3 Evaluation of the Macro View Method

We verified the hypotheses of the macro view method by using the result of Q1 of the questionnaire. We show the relation between the number of clusters and the average ratio of choice in Fig. 8. The threshold of CLINK is 1. At the risk level of 5%, the graph shows two significant relations.

- As the number of clusters increases, the average ratio of "unknown" increases.
- As the number of clusters increases, the average ratio of "known" decreases.

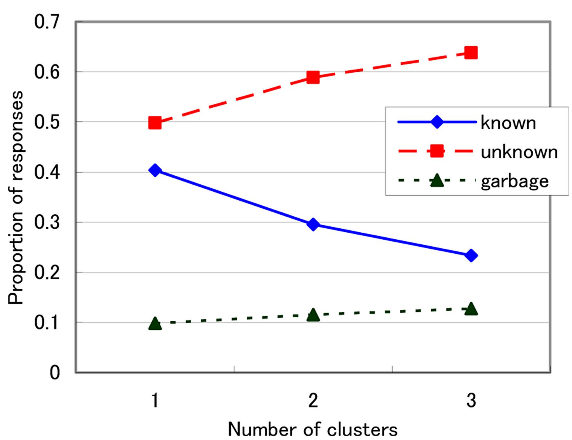


Fig. 7. The relation between the number of documents and the evaluation of medical students when the number of submitted keywords is 2

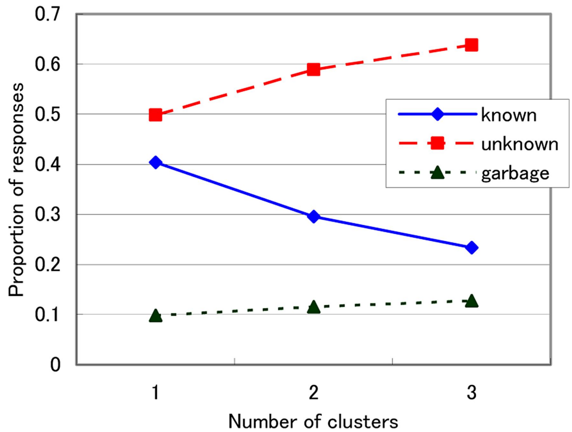


Fig. 8. The relation between the number of clusters and the evaluation of medical students

The result does not show any significant relation about "garbage" because the number of students who chose "garbage" was relatively small compared to those making other choices and does not depend on the number of clusters. We suppose that the medical students hesitated to judge a rule as being garbage.

The hypotheses of the macro view method are partly supported by this evaluation. The maximum number of clusters in this examination was 3. We still need to examine how medical students or experts judge rules with more than 4 clusters.

5 Summary

We proposed two discovered-rule-filtering methods, the micro view and the macro view methods, to identify novel rules from among those returned by a data mining system; both methods use information retrieved from the Internet. We evaluated the methods by using the questionnaire method. The results indicate that the output of the macro view method reflects the evaluation of medical students.

Our future work is summarized as follows.

- We need to improve the performance of the information retrieval technique which is based on naïve keyword search. We plan to improve the performance by applying natural language processing techniques [10].
- We will apply the discovered-rule-filtering methods to a practical application domain such as hepatitis data mining, and evaluate its feasibility.

Acknowledgement

This work is supported by a grant-in-aid for scientific research on priority area by the Japanese Ministry of Education, Science, Culture, Sports and Technology.

References

1. H. Motoda (Ed.), Active Mining: New Directions of Data Mining, IOS Press, Amsterdam, 2002.
2. R. Baeza-Yates and B. Ribeiro-Neto, Modern Information Retrieval, Addison Wesley, 1999.
3. Y. Kitamura, K. Park, A. Iida, and S. Tatsumi. Discovered Rule Filtering Using Information Retrieval Technique. Proceedings of International Workshop on Active Mining, pp. 80-84, 2002.
4. Y. Kitamura, A. Iida, K. Park, and S. Tatsumi, Discovered Rule Filtering System Using MEDLINE Information Retrieval, JSAI Technical Report, SIG-A2-KBS60/FAI52-J11, 2003.
5. Y. Kitamura, A. Iida, K. Park, and S. Tatsumi, Micro View and Macro View Approaches to Discovered Rule Filtering. Proceedings of 2nd International Workshop on Active Mining, pp.14-21, 2003.
6. U. M. Fayyad, G. P. Shapiro, P. Smyth, and R. Uthurusamy (Eds.), Advances in Knowledge Discovery and Data Mining, AAAI Press, 1996.
7. H. Yokoi, S. Hirano, K. Takabayashi, S. Tsumoto, Y. Satomura, Active Mining in Medicine: A Chronic Hepatitis Case – Towards Knowledge Discovery in Hospital Information Systems -, Journal of the Japanese Society for Artificial Intelligence, Vol.17, No.5, pp.622-628, 2002. (in Japanese)
8. M. Ohsaki, Y. Sato, H. Yokoi, and T. Yamaguchi, A Rule Discovery Support System for Sequential Medical Data – In the Case Study of a Chronic Hepatitis Dataset -, Proceedings of International Workshop on Active Mining, pp. 97-102, 2002.
9. S. C. Johnson, Hierarchical Clustering Schemes, Psychometrika, Vol.32, pp.241-254, 1967.
10. T. Yamasaki, M. Shimbo, and Y. Matsumoto: A MEDLINE document search system using section information, JSAI, SIG-KBS-A301-05, 2003.

Proposal of Relevance Feedback Based on Interactive Keyword Map

Yasufumi Takama and Tomoki Kajinami

Tokyo Metropolitan Institute of Technology
6-6 Asahigaoka, Hino Tokyo 191-0065, Japan
ytakama@cc.tmit.ac.jp

Abstract. The relevance feedback based on a keyword map is proposed so that a Web interface can be more interactive. There exists vast amount of information in the Web, from which users usually gather information without definite information needs. Therefore, it is difficult for users to organize and understand what they have gathered from the Web. From this viewpoint, we have proposed the concept of RBA-based interaction, in which analysis operation aims to assist users in understanding the context of their web interaction. However, the currently developed interface focuses on the information flow from the interface to users. As the first step for realizing relevance feedback (RF) based on interactive keyword map, this paper proposes the algorithm for extracting the pair of keywords that reflects a user's interest from the keyword map. Experimental results are given for showing how the algorithm works on the keyword map that is modified by the user, and for discussing the difference between the RF based on keyword map and conventional RF methods.

1 Introduction

A Web interaction is defined as users' activities for viewing and collecting web pages with using search engines and Web browsers. There exists vast amount of information in the Web, from which users usually gather information without definite information needs. Therefore, it is difficult for users to organize and understand what they have gathered from the Web. We have proposed the concept of RBA-based interaction, in which analysis operation aims to assist users in understanding the context of their web interaction. The Web interface that supports RBA-based interaction employs both keyword map visualization and document clustering, which respectively present users the topic distribution and document clusters within gathered document set[13]. However, the interface currently focuses on the information flow from the interface to users. In this paper, the relevance feedback based on a keyword map is proposed so that the interface can be more interactive. As the first step for realizing keyword map-based RF, this paper proposes the algorithm for extracting the pair of keywords that reflects a user's interest from the keyword map. Experimental results are given for

A. Sakurai et al. (Eds.): JSAI 2003/2004, LNAI 3609, pp. 507–516, 2007.

showing how the algorithm works on the keyword map that is modified by the user, and for discussing the difference between the RF based on keyword map and conventional RF methods.

2 Related Work

2.1 Concept of Retrieval, Browsing, Analysis (RBA)-Based Interaction

One of the essential properties of our activities in the Web is that we do not always have the predetermined target topics while surfing on the Web. Therefore, not only submitting relevant queries, but also evaluating the relevance of web pages is difficult for us. Through the interaction with the Web, We find the topics of interest, and acquire the background knowledge about the topics, based on which the relevance of pages is evaluated.

Considering the commercial success of web search engines, it is rational that we assume the following steps for locating and gathering information in the Web:

Retrieval. Obtains a set of pages by submitting tentative query to a search engine.

Browsing. Starting from individual documents in the retrieved results, browses their neighboring pages and collect (save) the relevant ones.

We call the interaction based on these two steps RB-based interaction. It should be noted that a user cannot always evaluate the relevance of pages correctly, and the evaluation criteria frequently changes while interacting with the Web. In other words, the context that affects the evaluation criteria is composed of the pages that have been gathered so far. Therefore, we claim that the "analysis" step should be combined with RB-based interaction. We call the interaction based on these three steps RBA-based interaction. Although Gershon[4] has already denoted the importance of the analysis step, in which the properties within a single page is analyzed. Our focus is on analyzing the set of gathered documents.

From this viewpoint, conventional information visualization systems[1, 2, 5–7, 15–17] contribute for supporting RBA-based interaction to some extent. However, they put the analysis step inside retrieval and browsing steps. That is, the visualized space by browsing support systems is mainly used for users to browse the hyperspace. The space visualized by clustering-based information visualization systems helps user explore retrieved results. On the other hand, we have proposed to visualize the set of documents that is gathered as a result of the user's RB-based interaction[13].

Document clustering-based visualization is suitable for our aims, because it is assumed that a user usually gathers the pages of interest from various Web sites, and most documents have no direct hyperlink to others. In particular, this assumption becomes valid in retrieval step.

In order for users to understand context information from the visualized results, presenting only document clusters is not enough, but the relationship among clusters should also be presented. The SOM-based visualization systems can satisfy this to some extent, but the obtained structure seems to be fixed, even if users can manipulate the visualized space with fisheye or fractal operation[16]. Furthermore, we think that the obtained document clusters should be presented to users as lists, because Web users are familiar with the document lists that are returned by most of search engines.

Therefore, we have proposed to visualize both of document and keyword space. Document clusters are presented to users as lists, while keyword space is visualized so that the relationship among document clusters can be reflected. For visualizing the keyword space, we employed the **keyword map** [12], on which the keywords extracted from documents are arranged so that the pair of keywords that frequently appears in the same documents can be arranged closely to each other.

The point is how to relate the keyword map with document space, and we have proposed a landmark-based approach, called plastic clustering method[11].

A prototype interface has been developed based on server-side programming technique. A user can interact with the Web with ordinary Web browsers as usual. The interface displays a small control panel on a separate browser window, which provide users with several assist for collecting pages as well as for analyzing the topic information over collected document set.

2.2 Relevance Feedback

Interaction should be bidirectional. That is, interactive interface should not only provide users with information in understandable manner, but also get their intentions and preferences. Relevance feedback (RF) is one of major approaches for implicitly obtaining the users' preferences.

Conventional RF algorithms[3, 8] modify a profile (query) vector based on user's judgment (relevant or irrelevant) on the retrieved documents. In this case, the user's intention is estimated indirectly from the document space. The FISH View system[10] extracts the user's viewpoint from the diagram, in which the user groups documents hierarchically. There also exists the system that supports the user's query generation by presenting the related keywords[9]. However, it is not RF approach in the sense the user has to generate the Boolean query manually.

As noted in section 2.1, it is rational that the interaction between humans and the Web involves existing search engines. Although the conventional relevance feedback technique is basically based on vector space model (VSM), it should be combined with widely-used search engines such as Google. That is, a query vector as a result of relevance feedback should be converted to a set of keywords, which can be submitted as a query to usual search engines. An easy solution for that is to select a couple of keywords that have higher weights in the query vector than others. However, the conversion from a query vector to a set of keywords is indirect approach for inferring a user's intention or preference, because a user has to select documents in spite of what he finally wants a set of query keywords.

In this paper, we propose keyword map-based RF, which infers the user's intention from the keyword space. This approach is more direct than the conventional RF algorithms applied to document space. It can also be said that keyword map-based RF is more flexible than conventional RF, as the latter cannot obtain arbitrary combination of keywords. Finally, keyword arrangement can reflect user's intention more implicitly than the diagram used by the FISH View system.

3 Relevance Feedback Based on Keyword Map

3.1 Keyword Map Visualization System

A keyword map-based information visualization system is developed for visualizing the topic distribution within a document set[12]. The developed system called TMIT (Topic Map Idea Tool) employs the spring model[14] to arrange keywords on 2D space. Although a number of information visualization systems employ the 3D graphics, they seem to be suitable for the facilities such as museum, where visitors use the systems. We claim that the system that can be in daily use should be simple. Therefore, we employ the 2D graphics. The basic algorithm of TMIT is as follows.

1. Define the distance l_{ij} between keyword i and j based on their similarity $R_{ij}(\in [-1,1]$[1]$)$ by Eq. (1) (m is positive constant).

$$l_{ij} = m\,(1 - R_{ij})\,. \tag{1}$$

2. The moving distance of keyword i in each step, $(\delta_{xi}, \delta_{yi})$, is calculated by Eq. (2).

$$(\delta_{xi}, \delta_{yi}) = \left(c\frac{\partial E}{\partial x_i}, c\frac{\partial E}{\partial y_i}\right), \tag{2}$$

$$E = \sum_i \sum_j \frac{1}{2} k_{ij} \left(d_{ij} - l_{ij}\right)^2, \tag{3}$$

$$d_{ij} = \sqrt{\left(x_i - x_j\right)^2 + \left(y_i - y_j\right)^2}. \tag{4}$$

3. In each step, the center of gravity is adjusted to the center of 2D space.

In addition to this basic algorithm, an arrangement priority based on **spring constant** is introduced [14][2]. It can be understood from Eq. (3) that the influence of strong spring (with large spring constant) is greater than that of weak ones. Here, the springs connecting to focused keywords (such as landmarks[11]) are given larger spring constant than others, so that they can have priority than other keywords in terms of arrangement.

[1] In the current keyword map system, $R_{ij} \in (0, 1]$ when keywords co-occur within documents, and R_{ij}=–1 when they do not appear within a document.

[2] Another arrangement priority based on frictional force is also introduced for considering the topic stream, which is out of scope and omitted in this paper.

3.2 Keyword-Pair Extraction for Relevance Feedback

The keyword map system currently implemented considers the information flow from the system to a user. In this subsection, the information flow from a user to the system is considered.

When a keyword map is presented to a user, he usually finds the difference between the keyword arrangement on the map and his background knowledge. Therefore, he wants to modify the arrangement, as he likes. If the system can infer the user's intention from the keyword map modified by him, relevance feedback can be available.

Let us consider the following cases:

1. A user rearranges the keyword A close to keyword B, which were initially arranged far away from each other.
2. A user moves apart keyword A and B, which were initially arranged close to each other.

In the first case, the user estimates the relationship between keyword A and B closer than the initial keyword map. Therefore, collecting new document that contain both keywords should satisfy the user's interest. The latter case might be more complicated, and there will be several possibilities. For example, a user might simply want documents containing keyword A but B (i.e., a query might be "A AND NOT B".). As for another possibility, the user might want to divide the topic represented by keyword A and B into two detailed topics. In this case, finding new keywords that bridge keyword A and B will be useful for the user.

In this paper, we proposed a method for extracting such keyword pairs as discussed above, from a user's modification on a keyword map. Extracting such keyword pairs is expected to be a fundamental process for realizing keyword map-based RF.

In the following algorithm, an input data file for keyword map (KData) and the data file for keywords' coordinates in the modified map (XYData) are given. KData stores the similarity S_{Ki} $(= R_{lm} \in [-1, 1])$ for every keyword pair p_i $(w_l$, $w_m)$, and XYData stores the coordinates $(x_i,\ y_i)$ of every keyword w_i on the map after the user's modification.

1. Calculate similarities S_{Xi} for each keyword pair p_i, based on the distance d_i between the keywords, from XYData. The d_M is the maximum distance among all keyword pairs.

$$S_{Xi} = 1 - (d_i/d_M)\,. \tag{5}$$

2. Translate S_{Ki}s in KData into value within [0,1] by Eq. (6).

$$S'_{Ki} = \max\left(S_{Ki}, 0\right). \tag{6}$$

3. For each keyword pair p_i, calculate the degree of "farness" ($\mathrm{Far}(S'_{Ki})$) and "nearness" ($\mathrm{Near}(S'_{Ki})$) in KData, and those ($\mathrm{Far}(S_{Xi})$ and $\mathrm{Near}(S_{Xi})$) in

XYData by Eq. (7) and (8), respectively.

$$\mathrm{Far}\,(x) = \max\left(-\frac{x}{t} + 1, 0\right), \tag{7}$$

$$\mathrm{Near}\,(x) = \max\left(\frac{x-t}{1-t}, 0\right). \tag{8}$$

4. Extract the keyword pairs having high values calculated by Eq. (9) as "Far2Near" pairs.

$$V_i^{F2N} = \max\left(\mathrm{Near}(S_{Xi}), \mathrm{Far}(S'_{Ki})\right) \ldots \mathrm{Near}(S_{Xi}), \mathrm{Far}(S'_{Ki}) > 0,$$
$$0 \ldots \text{otherwise.} \tag{9}$$

5. Extract the keyword pairs having high values calculated by Eq. (10) as "Near2Far" pairs.

$$V_i^{N2F} = \max\left(\mathrm{Near}(S'_{Ki}), \mathrm{Far}(S_{Xi})\right) \ldots \mathrm{Near}(S'_{Ki}), \mathrm{Far}(S_{Xi}) > 0,$$
$$0 \ldots \text{otherwise.} \tag{10}$$

4 Experiments on Keyword Map-Based Relevance Feedback

The experiments on keyword map-based relevance feedback are performed with using the prototype interface shown in Section 2.1, combined with the algorithm described in Section 3.2. Currently, the algorithm has not been yet implemented inside the prototype interface, and it is difficult to perform experiments with test subjects. Therefore, the section shows the examples how the proposed algorithm works on a keyword map actually generated from the retrieval results of existing search engine. Furthermore, the advantage of the keyword map-based RF against conventional RF is also discussed based on the examples. It should also be noted that the experiments are performed on Japanese Web pages, and results are translated from Japanese into English hereafter.

When applying the algorithm to a keyword map, the parameter t in Eq. (7) and (8) should be given. In the experiments, R_{lm} is given based on Jaccard coefficient regarding co-occurrence of keywords l and m within a document set. Therefore, $t = 0.25$ for farness and $t = 0.5$ for nearness are used for the KData, because two keywords are assumed to be highly related if R_{lm} exceeds 0.5. The value for farness is determined empirically, so that the number of Far2Near keyword pairs can be limited. As for the XYData, $t = 0.5$ for farness and $t = 0.9$ for nearness are used, because we assume that users will arrange the keywords that they want to discriminate with more than half distance of the max distance among keywords. The value for nearness is determined empirically, in order to reduce the number of Far2Near keyword pairs.

A query "Kanazawa[3] AND Sightseeing (Kanko)" is submitted to the Google, and top 10 pages are collected as an initial page set, from retrieved result. Figure1

[3] Kanazawa is the name of the city where AM2004 workshop is held.

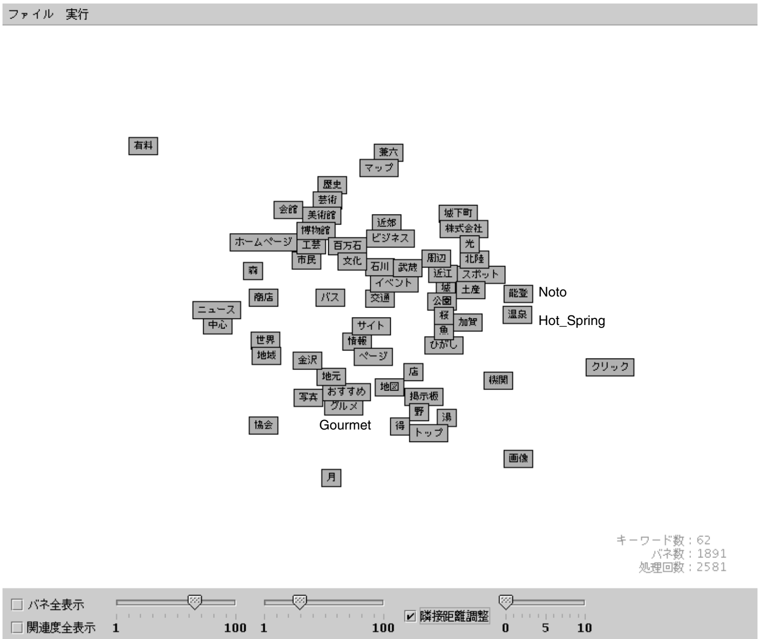


Fig. 1. Keyword Map Generated from Initial Document Set

shows the keyword map that the prototype interface generates from the initial page set.

In Fig. 1, the word “Noto” and “Hot Spring” (Onsen) are close, from which “Gourmet” is far away. In fact, “Noto” and “Hot Spring” appear in the same set of documents, and “Gourmet” co-occurs with those keywords in only a single document.

The keyword map system used in the experiments is improved to be interactive, so that the relevance feedback can be available. In particular, a user can drag any keyword to arbitrary position, and other keywords are arranged automatically based on spring model. After the modification, the system can output the XYData as noted in Section 3.2.

Figure 2 and 3 show the maps that are modified by a user from the initial arrangement. In Fig. 2, “Noto” and “Gourmet” are close, and “Hot Spring” is far from those keywords. On the other hand, “Hot Spring” and “Gourmet” are close, and “Noto” is far from those in Fig. 3.

From Fig. 2, three pairs are extracted as Far2Near, whereas 5 pairs are extracted as Near2Far with the proposed algorithm. Among them, “Gourmet AND Noto” can obtain the highest Far2Near score, and “Hot Spring AND Noto” can obtain the highest Near2Far score.

On the other hand, from Fig. 3, 19 Far2Near pairs and a single Near2Far pair, “Hot Spring AND Noto” are respectively extracted. Among Far2Near pairs, “Gourmet AND Hot Spring” has the highest score. These results show that

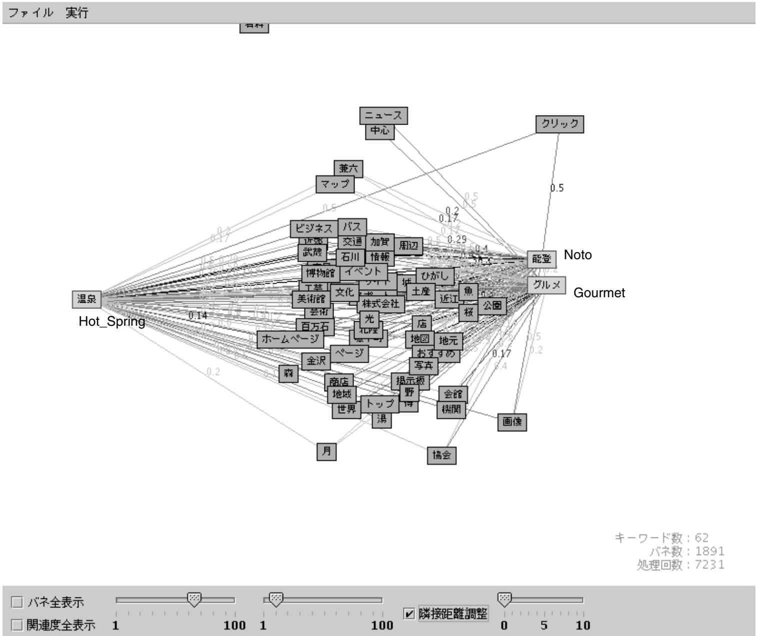


Fig. 2. Keyword Map Edited by User (1)

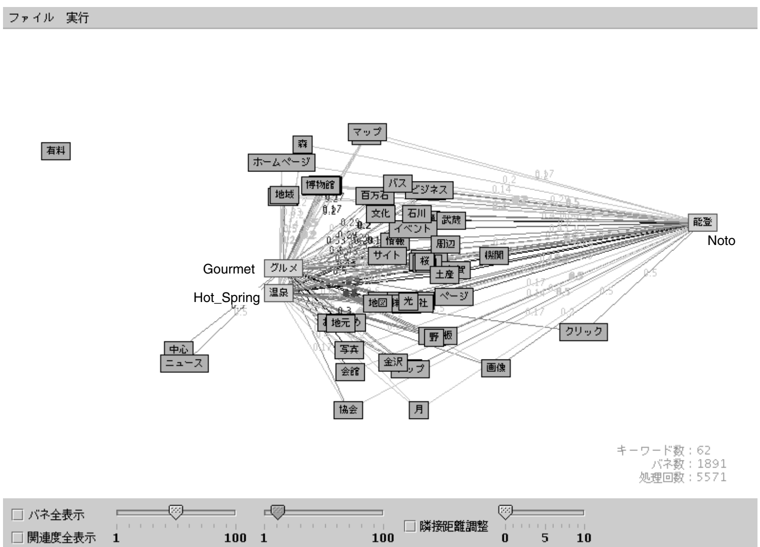


Fig. 3. Keyword Map Edited by User (2)

the proposed algorithm can extract appropriate keyword pairs based on the difference between the initial data set and modified keyword map.

Furthermore, the difference between the result of Fig. 2 and that of Fig. 3 clearly shows the advantage of keyword map-based RF against conventional RF. As noted above, “Noto” and “Hot Spring” appear within the same documents, d_1, d_2, and d_3 among 10 documents. On the other hand, “Gourmet” is contained in 3 documents, d_1, d_5, d_6. Therefore, only a single document, d_1, contains all three keywords.

When Rocchio-based RF[3] with TFIDF weighting (i.e. conventional RF) are performed with d_1, d_2, d_3, d_5, d_6 as positive examples and other 5 documents as negative ones, those three keywords obtain high positive weights. Whereas, when only d_1 is given as positive, “Noto” and “Gourmet” obtain positive weights, and “Hot Spring” obtain negative weight, which corresponds to the result of Fig. 2. This result can be the query “Gourmet AND Noto AND NOT Hot Spring”.

Then, how to obtain the query “Gourmet AND Hot Spring AND NOT Noto” with conventional RF? It seems difficult to obtain such a query, because “Noto” and “Hot Spring” appears within the same set of documents. In this example, such a query cannot be obtained by removing either d_2 or d_3 from positive document set. This result show the keyword map-based RF is more flexible than conventional RF when used in combination with existing search engines.

5 Conclusion

The relevance feedback based on interactive keyword map system is proposed. For the first step towards realizing keyword map-based RF, an algorithm is proposed for extracting the pair of keywords that reflects a user’s interest from the keyword map modified by the user. Experimental results show how the algorithm works on the keyword map that is modified by the user, and discuss the difference between the RF based on keyword map and conventional RF methods.

We have already proposed the concept of Retrieval, Browsing, and Analysis (RBA)-based interaction. The prototype interface employs the keyword map visualization system so that users can easily understand the context of their interaction with the Web. Combination of the prototype interface with the algorithm proposed in this paper will realize bidirectional web interaction between users and the Web.

References

1. Ackerman, M. et al., “Learning Probabilistic User Profiles,” AI Magazine, Vol. 18, No. 2, pp. 47—56, 1997.
2. Armstrong, R., Freitag, D., Joachims, T., Mitchell, T., “WebWatcher: A Learning Apprentice for the World Wide Web,” AAAI Spring Symposium on Information Gathering from Heterogeneous, Distributed Environments, 1995.
3. Baeza-Yates, R. and Ribeiro-Neto, B., “5. Query Operations” in Modern Information Retrieval, Addison Wasley, 1999.

4. Gershon, N., LeVasseur, J., Winstead, J., Croall, J., Pernick, A., Ruh, W., "Case Study: Visualizing Internet Resources," Proc. Information Visualization (INFOVIS'95), pp. 122—128, 1995.
5. Hearst, M. A. and Pedersen, J. O., "Reexaming the Cluster Hypothesis: Scatter/Gather on Retrieval Results," Proc. Of 19th Int'l ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'96), pp. 76-84, 1996.
6. Lieberman, H., "Letizia: An Agent That Assists Web Browsing," Proc. 14th Int'l Joint Conf. on Artificial Intelligence (IJCAI95), pp. 924—929, 1995.
7. Mukherjea, S., Hara, Y., "Visualizing World-Wide Web Search Engine Results," Int'l Conf. on Information Visualization, p.400—405, 1999.
8. Onoda, T., Murata, H. and Yamada, S., "Document Retrieval based on Relevance Feedback with Active Learning," SIG-KBS-A301 (JSAI), pp.13-18, 2003.
9. Sunayama, W., Ohsawa, Y. and Yachida, M. "A Search Interface with Supplying Search Keywords by Using Structure of User Interest," J. of Japan Society for Artificial Intelligence, Vol. 15, No. 6, pp. 1117-1124, 2000.
10. Takama, Y. and Ishizuka, M., "FISH VIEW System: A Document Ordering Support System Employing Concept-structure-based Viewpoint Extraction," J. of Information Processing Society of Japan, Vol. 41, No. 7, pp.1976-1986, 2000.
11. Takama, Y. and Hirota, K., "Web Information Visualization Method Employing Immune Network Model for Finding Topic Stream from Document-Set Sequence," J. of New Generation Computing, Vol. 21, No. 1, pp. 49-59, 2003.
12. Takama, Y. and Tetsuya, H., "Application of Immune Network Metaphor to Keyword Map-based Topic Stream Visualization," Proc. 2003 IEEE Int'l Symp. on Computational Intelligence in Robotics and Automation (CIRA2003), pp. 770-775, 2003.
13. Y. Takama, "Intelligent Interface based on Retrieval, Browsing, Analysis Operations," *4th International Conference on Intelligent Technologies (InTech'03)*, pp. 806–811, 2003.
14. Takasugi, K. and Kunifuji, S., "A Thinking Support System for Idea Inspiration Using Spring Model," J. of Japanese Society for Artificial Intelligence, Vol. 14, No. 3, pp. 495—503, 1999 (written in Japanese).
15. Teraoka, T. and Maruyama, M., "Research Report: Adaptive Information Visualization Based on the User's Multiple Viewpoints –Interactive 3D Visualization of the WWW –," Proc. IEEE Symposium on Information Visualization (InfoVis'97), pp. 25—28, 1997.
16. Yang, C. C., Chen, H., Hong, K., "Internet Browsing: Visualizing Category Map by Fisheye and Fractal Views," Proc. Int'l Conf. On Information Technology: Coding and Computing (ITCC'02), pp. 34—39, 2002.
17. Zamir, O. and Etzioni, O., "Grouper: A Dynamic Clustering Interface to Web Search Results," Proc. 8th International WWW Conference, 1999.

A Correlation-Based Approach to Attribute Selection in Chemical Graph Mining

Takashi Okada

Department of Informatics, Kwansei Gakuin University, 2-1 Gakuen, Sanda-shi, Hyogo 669-1337, Japan
okada-office@ksc.kwansei.ac.jp

Abstract. The huge number of descriptive features is often a problem in data mining. We analyzed structure activity data for dopamine antagonists, which involves selecting useful features from numerous fragments extracted from their chemical structures. Correlation coefficients among categorical variables were used to select attributes. Chemists evaluated the rules obtained by the cascade model, and the importance of attribute selection was confirmed.

1 Introduction

One of the challenging problems in data mining is to cope with the vast numbers of attributes. A typical example involves finding important genes from among millions of single nucleotide polymorphisms (SNPs) that explain the cause of a given disease. We analyzed structure-activity relationships using linear fragments derived from chemical graphs. There were 2,000~3,000 meaningful fragments. This is much smaller than the SNP problem, but we cannot obtain valuable knowledge unless we can overcome this problem.

Association rule mining has been used to solve numerous attributes problems [1]. It can detect frequent itemsets in customers' baskets selected from the thousands of items sold in a supermarket. However, its success depends on the sparseness of the data. That is, the method thinks of the few items in the basket, and it does not consider items that do not appear in the basket. When we treat a dense dataset, a huge number of itemsets appears, resulting in a combinatorial explosion of the itemset lattice. Our cascade model also constructs an itemset lattice [2, 3]. It can handle a dense dataset, as it detects a useful rule from a single link located at a shallow level in the lattice. Nevertheless, it is limited to 100~150 attributes. Improvements are necessary in order to treat a dataset with numerous attributes.

A regression analysis usually uses attribute selection procedures to avoid the overfitting and instability of the model that arises from collinearity among the explanatory variables. Attribute selection is also useful in the decision tree approach, when a dataset contains more than a few dozen attributes.

This paper reports an attempt to introduce attribute selection to the mining of SAR from chemical graphs. The next section gives a brief overview of the analysis, including a basic introduction to the mining method and the problems encountered. The categorical attributes are selected using correlation coefficients, which are

A. Sakurai et al. (Eds.): JSAI 2003/2004, LNAI 3609, pp. 517 – 526, 2007.

defined in Section 3. The results of applying the method to chemical graph mining are shown in Section 4, where the effects of attribute selection are discussed referring to the quality of rules judged from a chemist's perspective.

2 Mining Chemical Graphs Using the Cascade Model

2.1 Overview of the Dopamine D2 Antagonists Analysis

Dopamine is a neurotransmitter in the brain. Neural signals are transmitted via the interaction between dopamine and proteins known as dopamine receptors. There are five different receptor proteins, D1 – D5, each of which has a different biological function. Certain chemicals act as antagonists for these receptors. An antagonist binds to a receptor without functioning as a neurotransmitter, thereby blocking the dopamine molecule.

We used the MDDR database developed by Prous Science and MDL as the data source [4]. It contains 1,349 chemical structures that describe dopamine (D1, D2, D3, and D4) antagonist activities. The sample problem used in this paper is to discover the structural characteristics responsible for D2 antagonist activity, which is the hardest problem among the four antagonist activities.

Figure 1 shows a brief scheme of the analysis. All of the structural formulae of the chemicals were stored in an SDF file, a common data exchange format that is used in computer chemistry. Then, molecular orbital calculations using MM-AM1-Geo software were used to derive three electronic properties: HOMO, LUMO, and Dipole. LogP values were calculated using the program ClogP to give the hydrophobic

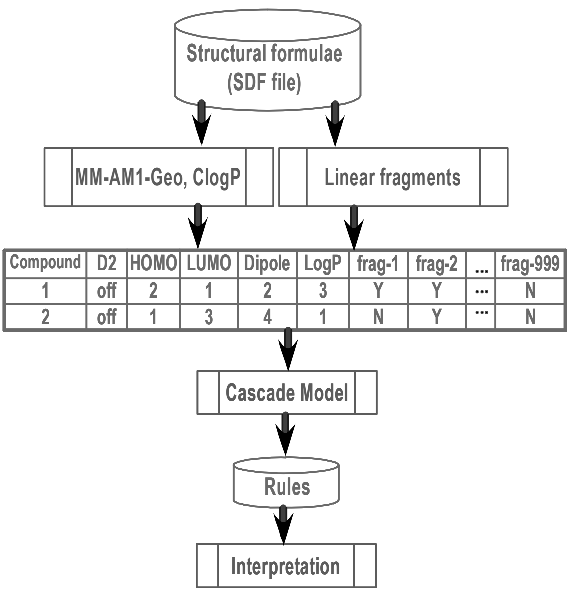


Fig. 1. Flow chart of chemical graph mining with the cascade model

property of each molecule. We also extracted many linear fragments contained in the chemical graphs, and used the presence/absence of these fragments in a molecule as another attribute. Klopman first introduced this type of linear fragment [5], and further developments were made by Okada [6] and Kramer *et al.* [7]. Linear fragments are expressed by constituent elements and bond types like "c3H:c3---C4H-N3", and they are used as attribute names. Publication of our fragment-generation algorithm is pending [8].

Obviously, the number of all possible fragments is too large. Therefore, the length of linear fragments was limited to those shorter than 9, and one of the terminal atoms of a fragment was restricted to be a heteroatom or a carbon involved in a double or triple bond. Consequently, we obtained 8,041 fragments, which was still too many for analysis using the current implementation of the cascade model. Therefore, we selected 114 fragments that appeared in 15 to 85% of the compounds.

Application of the cascade model to the table generated rules that characterized the structures of D2 antagonists. Chemists interpreted and evaluated the resulting rules.

2.2 The Cascade Model and the Datascape Survey

The cascade model can be considered an extension of association rule mining [1]. The method creates an itemset lattice in which an [attribute: value] pair is used as an item to constitute itemsets. Links in the lattice are selected and interpreted as rules. That is, we observe the distribution of the right hand side (RHS) attribute values along all links. If a distinct change in the distribution appears along some link, then we focus on the two terminal nodes of that link. The itemset at the upper end of a link is {[A: y]} and item [B: n] is added along the link. If a marked activity change occurs along this link, we can write a rule:

```
Cases: 200 ==> 50 BSS=12.5
IF [B: n] added on [A: y]
THEN [Activity]:    .80 .20  ==>  .30 .70 (y n)
THEN [C]:           .50 .50  ==>  .94 .06 (y n)
```

where the added item [B: n] is the main condition of the rule, and the items at the upper end of link {[A: y]} are considered preconditions. The main condition changes the ratio of active compounds from 0.8 to 0.3, while the number of supporting instances decreases from 200 to 50. *BSS* is the between-groups sum of squares, which is derived from the decomposition of the sum of squares for a categorical variable. Its value can be used as a measure of the strength of a rule. The second "THEN" clause indicates that the distribution of the values of attribute [C] also changes sharply with the application of the main condition. This description is called the *collateral correlation*.

Recently, we modified the method of conducting a *datascape survey* in order to reduce the number of rules [9], and to denote the details of the data distribution specified by a rule [10]. These functions facilitated the interpretation of rules.

2.3 Attribute Selection Problem

There is no reason to justify the selection of the 114 fragments that appeared in 15-85% of the compounds. In fact, by browsing the structural formulae, the chemists noticed other important fragments that contribute to D2 activity. However, if we use more attributes, the combinatorial explosion in the lattice size prohibits analysis. Past experience suggests that the upper limit of attributes is 100~150.

Analysts often encounter a pair of fragments with the same number of supporting compounds, such as O1=S4-c3:c3H and S4-c3:c3H. The support for the latter must always be greater than or equal to that of the former, since the latter is a substructure of the former. If the support of the two is equal, then they appear in exactly the same compounds, and the selection of both fragments as attributes is redundant. That is, the correlation coefficient between these two attributes is 1.0.

Omission of an attribute from such pairs should enable analysis using more attributes with lower support. Furthermore, attribute pairs do not need to be correlated completely. We can omit an attribute if it is in a highly correlated pair. Therefore, we decided to introduce a correlation coefficient between pairs of attributes, and use it as a criterion to omit/keep attributes.

3 Correlation Coefficient Between Categorical Variables

The correlation coefficient is a well-known concept in the world of numerical attributes. Recently, we introduced generalized covariance using a vector expression for the value difference [11], and a uniform treatment of covariance became possible for numerical and categorical variables. Here, we briefly mention a special case that defines the correlation coefficient between a pair of binary attributes.

Gini successfully defined the variance of categorical data [12]. He first showed that the following equality holds for the variance of a numerical variable x_i.

$$V_{ii} = \left(\sum_a (x_{ia} - \bar{x}_i)^2 \right) \Big/ n = \frac{1}{2n^2} \sum_a \sum_b (x_{ia} - x_{ib})^2 \quad , \tag{1}$$

where V_{ii} is the variance of the i-th variable, x_{ia} is the value of x_i for the a-th instance, and n is the number of instances.

Then, he introduced the distance definition (2) into the value differences in (1), and obtained the categorical variance expression (3), which is known as the Gini-index.

$$x_{ia} - x_{ib} \begin{cases} = 1 & \text{if } x_{ia} \neq x_{ib} \\ = 0 & \text{if } x_{ia} = x_{ib} \end{cases} \quad , \tag{2}$$

$$V_{ii} = \frac{1}{2n^2} \sum_a \sum_b (x_{ia} - x_{ib})^2 = \frac{1}{2}\left(1 - \sum_r p_i(r)^2 \right) \quad . \tag{3}$$

Extension of this definition to the covariance fails, if we simply change $(x_{ia} - x_{ib})^2$ to $(x_{ia} - x_{ib})(x_{ja} - x_{jb})$. We used a regular simplex expression for the value of the categorical variable, and used the vector expression $\overrightarrow{x_{ia}x_{ib}}$, instead of the scalar $x_{ia} - x_{ib}$, in the definition of the variance. We proposed that V_{ij} be defined as the sum of the inner products of $\overrightarrow{x_{ia}x_{ib}}$ and $\overrightarrow{x_{ja}x_{jb}}$, where two regular simplexes for x_i and x_j are rotated to give the maximum value for V_{ij}. It is defined using the formulae:

$$V_{ij} = \max\left(Q_{ij}(L)\right) \quad , \tag{4}$$

$$Q_{ij}(L) = \frac{1}{2n^2} \sum_a \sum_b \left\langle \overrightarrow{x_{ia}x_{ib}} \middle| L \middle| \overrightarrow{x_{ja}x_{jb}} \right\rangle \quad . \tag{5}$$

Here, ***L*** is an orthonormal transformation applicable to the value space. The bracket notation <***e***|***L***|***f***> is evaluated as the scalar product of vectors ***e*** and ***Lf*** (or $\boldsymbol{L}^{-1}\boldsymbol{e}$ and ***f***). If the vectors ***e*** and ***f*** are of unequal lengths, zeros are first added to the shorter vector.

		x_j		
		u	v	
x_i	r	n_{ru}	n_{rv}	$n_{r\cdot}$
	s	n_{su}	n_{sv}	$n_{s\cdot}$
		$n_{\cdot u}$	$n_{\cdot v}$	n

We applied this definition to the simplest 2 × 2 contingency table shown to the left, where $n_{r\cdot}$ *and* $n_{\cdot u}$ have marginal distributions. Straightforward application of (5) to this table gives the following expressions for V_{ii}, V_{jj} and V_{ij}, and the correlation coefficient R_{ij} is given by (9).

$$V_{ii} = n_{r\cdot}n_{s\cdot}/n^2 = \tfrac{1}{2}\left(1-(n_{r\cdot}/n)^2-(n_{s\cdot}/n)^2\right) \quad . \tag{6}$$

$$V_{jj} = n_{\cdot u}n_{\cdot v}/n^2 = \tfrac{1}{2}\left(1-(n_{\cdot u}/n)^2-(n_{\cdot v}/n)^2\right) \quad . \tag{7}$$

$$V_{ij} = \frac{\left|n_{ru}n_{sv} - n_{rv}n_{su}\right|}{n^2} \quad . \tag{8}$$

$$R_{ij} = \frac{V_{ij}}{\sqrt{V_{ii}V_{jj}}} \quad . \tag{9}$$

The numerator in (8) is the critical term used to represent the extent of the dependency between two variables. In fact, the correlation coefficient is 1.0 (0.0) for completely dependent (independent) data, respectively.

4 Results and Discussion

We applied the attribute selection scheme to the dopamine D2 antagonist problem. In all, 8,041 fragments were generated. First, we selected a fragment as an attribute, if the probability of its appearance satisfied the following condition:

$$edge < P(\text{fragment}) < 1.0 - edge \quad . \tag{10}$$

When *edge* was set to 0.01, 0.02, 0.03, 0.05, 0.10, and 0.15, the number of fragments selected was 1,698, 1,056, 730, 377, 176, and 114, respectively. We used the presence/absence of these fragments as the initial attribute set $\{x\}$.

4.1 Attribute Selection Using Correlation Coefficients

The attribute selection procedure is as follows:

1. Calculate the correlation coefficients among all attribute pairs x_i and x_j, and add the pair to the list *pairs*, if it satisfies the condition: $R_{ij} > min\text{-}R_{ij}$.
2. Sort *pairs* in descending order of R_{ij}.
3. Pop *pairs*, and get pair: x_i and x_j.
4. Delete an attribute (x_i or x_j) from $\{x\}$, if both attributes are members of $\{x\}$.
5. Repeat steps 3 and 4 until every pair in *pairs* has been examined.

When we delete an attribute from a correlated attribute pair at step 4, the longer fragment is kept in the attribute set because an analyst can get more information from the longer attribute when it appears in a rule.

Figure 2 shows the numbers of attributes selected on a log scale for six *edge* values, as the value of $min\text{-}R_{ij}$ changes to 1.0, 0.99, 0.97, 0.95, 0.90, 0.85, 0.80, 0.75, and 0.70. Here, no attribute selection is carried out at $min\text{-}R_{ij}$ = 1.0, and attributes in perfectly correlated pairs are omitted at $min\text{-}R_{ij}$ = 0.99.

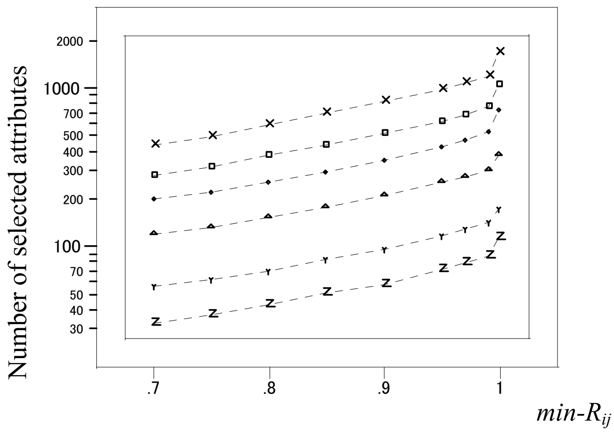


Fig. 2. Number of selected attributes for different values of $min\text{-}R_{ij}$

The shapes of the curves in the figure do not depend on the *edge* values. Another interesting point is the steep slopes seen at the right ends of the plots. This means that 20~30% of the attributes in the initial attribute sets are completely correlated in the chemical graph mining process using linear fragments. Roughly speaking, about half of the attributes are omitted at $min\text{-}R_{ij}$ = 0.90. Therefore, we conclude that the attribute selection scheme using correlation coefficients works well at reducing the number of attributes.

4.2 Effects of Lattice Size

Lattice expansion in the cascade model is controlled by the parameter *thres*. The smaller the value of *thres*, the more nodes in the lattice we examine. In this application, the value of *thres* was usually in the range 0.15-0.2, and the number of nodes in the lattice ranged from 5,000-30,000 using 100-150 attributes. We examined lattice size with changes in *edge* and *min-*R_{ij} values, for three *thres* values: 0.15, 0.175, and 0.20.

Figure 3 shows rough contour maps of the number of nodes (*#nodes*) in the lattice, where the y-axis is *min-*R_{ij} and the x-axis is the number of selected attributes (*#attributes*) in (A) and *edge* in (B). In the figure, the calculated points are shown using '+', while the points that resulted in the combinatorial explosion of the lattice are not depicted. The lowest contour line (*#nodes* = 3,000) is indicated by arrows.

The contour lines in (A) are all more or less parallel to the y-axis for large *min-*R_{ij} values, while they dip to the bottom right corner for small *min-*R_{ij} values. This shows that the lattice size does not change sharply when we use more uncorrelated attributes. In fact, we could use 400-500 attributes selected from more than 1,000 attributes.

The contour lines in Figure 3B are drawn from the upper right to the bottom left corners. The meaning of this is seen by inspecting the two ⊕ points and the two + points near the gray contour in the top right map. The data for these four points are summarized in Table 1.

Table 1. Calaculated results for 4 points near a gray contour (*thres*=0.15)

Point	*edge*	*min-*R_{ij}	*#attributes*	*#nodes*	*#detected*	*#rules*	*score*
P1	0.02	0.70	287	4992	23	6 (3)	3
P2	0.05	0.80	155	5983	39	8 (4)	3
P3	0.10	0.90	130	5223	72	9 (4)	2
P4	0.15	0.99	88	6265	97	14 (5)	2

This table shows that a similar number of nodes emerge from a wide range of *#attributes* (88 – 287). That is, a correlated attributes set increases the lattice size, while an uncorrelated set depresses it. As a result, attribute selection using a lower value of *min-*R_{ij} proves very useful in reducing the lattice size.

4.3 Evaluating Rules

The quality of the rules produced is more important than the number of attributes selected or the lattice size. The column *#detected* in Table 1 shows the number of links detected with large *BSS* values, where optimization of a rule starts. The column *#rules* shows the number of resulting rules, with the number of principal rules after the rules organization step shown in parentheses [9]. These numbers tend to increase with the value of *min-*R_{ij}. The appearance of many rules does not produce useful knowledge. For example, there are many highly correlated attributes in the calculation at P4, which might cause redundant rules. In fact, the increase in the number of principal rules is very limited.

Here, we introduce a scheme for evaluating rules. After considering various rules, the analysts noticed three important substructures relevant to D2 antagonist activity: an aromatic ether, a tertiary amine separated from an aromatic ring by three single bonds, and a CO group bonded to an amine. The appearance of these features in rules

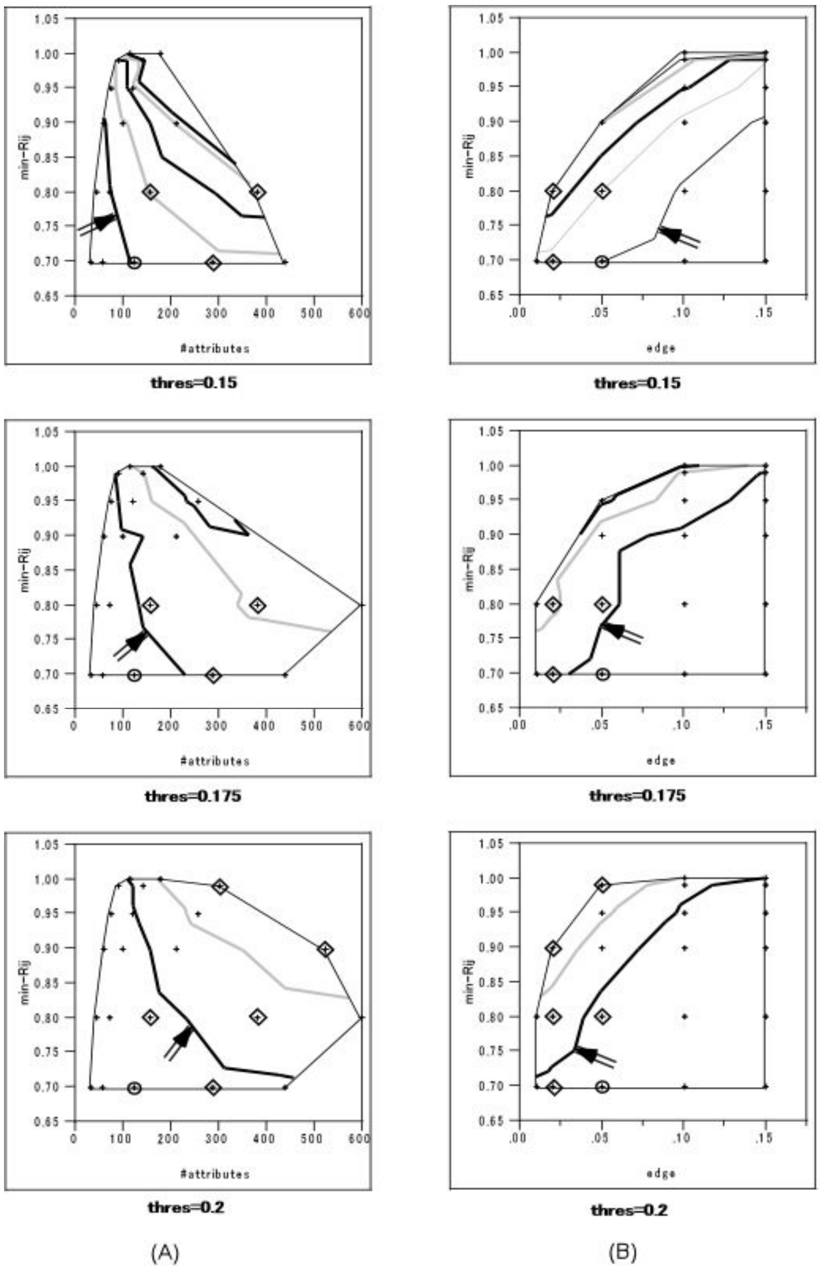
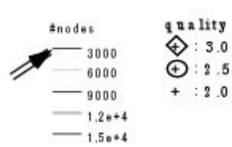


Fig. 3. Contour maps for the number of nodes in the lattice



was used to judge their quality. That is, we searched for these three features in the main condition of the principal rules, and the number of features found was used to *score* the rule set. When a feature appears in a relative rule only, we counted it as 0.5. Note that the appearance of a feature is counted only once. Therefore, the highest *score* of the resulting rules is 3. This evaluation scheme is rough, as the true mechanism of D2 antagonist activity is unknown. Nevertheless, we expect that this *score* will serve as a guide to judge the quality of rule sets.

The last column in Table 1 shows this *score* for four calculations. Note that the number of rules has no meaning from this viewpoint. The next problem is to find adequate values leading to a good rule set for the three parameters: *edge*, $min\text{-}R_{ij}$ and *thres*.

In Figure 3, the calculated points with *score*=3, 2.5, 2 are shown by ◈, ⊕, and +, respectively. The distribution of high score points in Figure 3A indicates that neither the number of attributes nor the size of the lattice is directly related to the *score* of a rule set. By contrast, Figure 3B shows that rule sets with high scores result from calculations at *edge* = 0.02 and 0.05. Attribute selection using $min\text{-}R_{ij}$ = 0.8 seems to give better rule sets.

Therefore, the suggested plan for mining is to use relatively smaller *edge* values, and then to select attributes using $min\text{-}R_{ij} \cong 0.8$. The effect of *thres* seems to be limited, given that the objective of mining is to comprehend the rough characteristics of chemical graphs.

5 Concluding Remarks

The attribute selection scheme introduced in this paper is essentially a method to cope with collinearity among explanation attributes. Many studies have attempted to solve this problem in the field of regression analysis. They include various attribute selection schemes, the canonical regression method, and partial least squares.

Of the mining methods used for categorical data, the reduct concept for the rough set clearly solves this problem [13]. However, its implementation cannot treat thousands of attributes. Another approach from the mining community is the closed itemset concept in association rule mining [14, 15]. It is used to compute long frequent itemsets quickly, and it is also used to filter rules to omit redundant ones. However, this method is useful only when a pair of attributes correlates completely. Even if the correlation coefficient is larger than 0.99, the method cannot be applied to data with a noise instance.

The cascade model also encounters the collinearity problem. The method first incorporates collateral correlations in a rule expression. It illustrates attributes with high correlations to the main condition, and greatly helps an analyst to interpret rules [3]. Furthermore, correlated attributes result in the generation of a pair of rules that overlap considerably. This problem is solved by organizing rules into principal and relative rules [9]. The attribute selection introduced here is useful for reducing the lattice size. Moreover, the omission of a correlated attribute cuts self-evident collateral correlations, and it also reduces the number of relative rules, thereby reducing the load faced by an analyst. All these functions work for partially correlated attributes, and this method offers a superior framework than those given by the closed itemset.

Comprehensive analysis of ligands for dopamine receptor proteins is now underway using the proposed system, which should discriminate not only antagonists but also agonists. We are also investigating the factors that distinguish antagonists and agonists. The results will serve as a model of work in the field of qualitative SAR analysis.

Acknowledgments

The author thanks Dr. Masumi Yamakawa and Dr. Hirotaka Niitsuma of Kwansei Gakuin University for their valuable discussions.

References

1. Agrawal, R., Srikant, R.: Fast Algorithms for Mining Association Rules. Proc. VLDB (1994) 487-499
2. Okada, T.: Rule Induction in Cascade Model based on Sum of Squares Decomposition. Principles of Data Mining and Knowledge Discovery (Proc. PKDD'99), LNAI 1704, Springer-Verlag (1999) 468-475
3. Okada, T.: Efficient Detection of Local Interactions in the Cascade Model. In: Terano, T. et al. (eds.) Knowledge Discovery and Data Mining PAKDD-2000. LNAI 1805, Springer-Verlag (2000) 193-203
4. MDL Inc.: http://www.mdl.com/products/knowledge/drug_data_report/index.jsp
5. Klopman, G.: Artificial Intelligence Approach to Structure-Activity Studies. J. Amer. Chem. Soc. 106 (1984) 7315-7321
6. Okada, T.: Discovery of Structure Activity Relationships using the Cascade Model: The Mutagenicity of Aromatic Nitro Compounds. J. Computer Aided Chemistry, 2 (2001) 79-86
7. Kramer, S., De Raedt, L., Helma, C.: Molecular Feature Mining in HIV data. In: Proc. of the Seventh ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD-01) (2001) 136-143
8. Okada, T., Yamakawa M., Niitsuma, H.: Spiral Mining using Comprehensible Attributes Generated from Molecular Structures. Submitted to post-congress proceedings of Active Mining 2003
9. Okada, T.: Datascape Survey using the Cascade Model. In: Satoh, K. et al. (eds.) Discovery Science 2002. LNCS 2534, Springer-Verlag (2002) 233-246
10. Okada, T.: Topographical Expression of a Rule for Active Mining. In: Motoda, H. (ed.) Active Mining. IOS Press, (2002) 247-257
11. Okada, T.: A Note on Covariances for Categorical Data. In: Leung, K.S. et al. (eds.) Intelligent Data Engineering and Automated Learning - IDEAL 2000. LNCS 1983, Springer-Verlag (2000) 150-157
12. Gini, C.W.: Variability and Mutability, contribution to the Study of Statistical Distributions and Relations, Studi Economico-Giuridici della R. Universita de Cagliari (1912). Reviewed in: Light, R.J., Margolin, B.H.: An Analysis of Variance for Categorical Data. J. Amer. Stat. Assoc. 66 (1971) 534-544
13. Pawlak Z.: Rough sets: Theoretical Aspects of Reasoning about Data. Dordrecht: Kluwer (1991)
14. Pasquier, N., Bastide, Y., Taouil, R., Lakhal, L.: Efficient Mining of Association Rules Using Closed Itemset Lattices. Information Systems, 24 (1) (1999) 25-46
15. Zaki, M.J., Hsiao, C.J.: CHARM: An Efficient Algorithm for Closed Itemset Mining. In: Proc. SDM’02, SIAM (2002) 457-473

Improving Multiclass ILP by Combining Partial Rules with Winnow Algorithm: Results on Classification of Dopamine Antagonist Molecules

Sukree Sinthupinyo[1], Cholwich Nattee[1], Masayuki Numao[1], Takashi Okada[2], and Boonserm Kijsirikul[3]

[1] Department of Architecture for Intelligence, The Institute of Scientific and Industrial Research, Osaka University,
8-1 Mihogaoka, Ibaraki, Osaka, 567-0047, Japan
{sukree,cholwich,numao}@ai.sanken.osaka-u.ac.jp

[2] Center for Information & Media Studies, Kwansei Gakuin University
okada@kwansei.ac.jp

[3] Department of Computer Engineering, Chulalongkorn University
boonserm.k@chula.ac.th

Abstract. In this paper, we propose an approach which can improve Inductive Logic Programming in multiclass problems. This approach is based on the idea that if a whole rule cannot be applied to an example, some partial matches of the rule can be useful. The most suitable class should be the class whose important partial matches cover the example more than those from other classes. Hence, the partial matches of the rule, called *partial rules*, are first extracted from the original rules. Then, we utilize the idea of Winnow algorithm to weigh each partial rule. Finally, the partial rules and the weights are combined and used to classify new examples. The weights of partial rules show another aspect of the knowledge which can be discovered from the data set. In the experiments, we apply our approach to a multiclass real-world problem, classification of dopamine antagonist molecules. The experimental results show that the proposed method gives the improvement over the original rules and yields 88.58% accuracy by running 10-fold cross validation.

1 Introduction

In recent years, Inductive Logic Programming (ILP) has been widely applied to various real-world applications [1,2]. Standard ILP systems are usually two-class classifier (positive and negative classes). A test example which matches with some rules is classified as positive class, while the example which does not match with any rule is classified as negative class. This causes some troubles when we need to use ILP in multiclass problems. In such problems, when a test example does not match with any rule or matches with some rules from more than two classes, we cannot determine which class is most suitable for the example.

A. Sakurai et al. (Eds.): JSAI 2003/2004, LNAI 3609, pp. 527–537, 2007.

In this paper, we propose an approach which can utilize the standard ILP's rules in multiclass problems. Our approach is based on the idea that if a whole rule cannot be applied to an example, some parts of rule may match with that example. Thus, we can make use of these matches to determine the class for the example. The most suitable class should be the class whose number of important matches is higher than those of other classes. Thus, in our approach, we first extract some part of rule which will be used as *partial rule*. Then, all partial rules are given the importance in term of weights using Winnow-based approach [3]. Finally, the partial rules and the weights are combined and used to classify new examples. Moreover, the weights assigned to the partial rules also show another aspect of the characteristic of data set that is very useful in knowledge discovery fashion.

We apply our approach to a real-world problem, classification of dopamine antagonist molecules. Dopamine antagonist is a kind of molecules which can block the binding between dopamines and dopamine receptors in the signal transfer process in human brain. The excessive levels of the dopamine have been implicated in schizophrenia. Hence, in the medical treatment of schizophrenic patients, the dopamine antagonist molecules are used to decrease the signal transfer level which can limit the effect of the high density of dopamines. The knowledge discovered from this domain may be useful for schizophrenic drug development.

1.1 Related Work

Srinivasan and King [4] proposed the work using ILP to discover new attributes or *features* which are then used by linear regression to predict chemical activity. Our work is different in that we focus on using ordinary ILP's rules in multiclass problems; the extracted features in our work are the parts of the original rules and aimed to be combined with other method to classify new examples in multiclass fashion. The following are the works proposed to help ILP in multiclass problems. Dietterich and Bakiri [5] proposed the method which employs the error correcting code to represent the class of examples and tries to predict the most suitable class for test examples. Round Robin Rule Learning proposed by Fürnkranz [6] focuses on training examples rearrangement. The training examples from each class are used to train the learner several times. A test example is tested with all trained classifiers. The most winning class is selected as the class of the test example. Eineborg and Bostr [7] proposed the method for selecting the class for the uncovered examples, Rule Stretching. The method aims to deal with an uncovered example by generalising the original rules to cover the example and select the most accurate rule as the rule which best matches with the example.

1.2 Paper Outline

The paper is organized as follows. In the next section, we present a concept of ILP and the obstacles when ILP is applied to multiclass problems. The partial rule extraction strategy and the weight adjustment are expressed in Section 3 and

Section 4, respectively. The details of the experiments are presented in Section 5. The paper ends with the conclusion in Section 6.

2 Using ILP in Multiclass Problems

ILP is the Machine Learning technique which is originally proposed as a two-class classifier. ILP aims to construct a rule set that covers all positive examples and none of the negatives. The output of ILP is the first-order rules which will be used to classify new examples. This causes some troubles when we need to use ILP in multiclass problems, i.e. (1) how to construct the rule for each class, and (2) how to select the class for each example. In the former case, as mentioned earlier, ILP systems search for the rules which cover positive examples, however, in multiclass problems, we need to construct the rules for each class. Hence, the additional techniques must be used to help ILP to construct the rules, such as one-against-all, round robin rule learning [6], and loss-based decoding [8]. Nevertheless, in this work, we emphasize on the latter case. Thus, we employ the common method, one-against-all, to construct the rules for each class.

In the one-against-all algorithm, a k-class problem is reduced to k two-class problems. To generate the rules for class i, the training examples are organized by using the training examples of class i as positive examples and using the training examples of class j where $j = 1, ..., k$ and $j \neq i$ as negative examples. For example, our data set contains 4 classes, i.e. D1, D2, D3, and D4 (as will be described in Section 5). We use the training examples of class D1 as the positive examples and use those of classes D2, D3, and D4 as the negatives for learning rules of class D1. Using this strategy, the obtained rules are unordered.

The problem of class selection arises when an example does not exactly match with any rule or matches with some rules from more than two classes, especially in case of unordered rules. In case of ordered rules, the class selection is not complicated, the example which does not exactly match with any rule can be classified as the default class, while the example which matches with multiple rules from different classes can be classified as the class of the higher order rule. However, in case of unordered rules, as constructed in this work, ILP's rules alone cannot select the class of the example which does not match with any rule or matches with multiple rules from different classes. Hence, we propose an approach which is based on the idea that if the whole rule cannot be applied to the example, we can utilize some partial matches of the rule to determine the most appropriate class.

In our experiments, we employ an ILP system, Aleph [9], to construct the rules for each class. The rule construction of Aleph starts with building the most specific clause, called *bottom clause*. Then, to seek for the best generalized clause, Aleph provides many search algorithms which users can select the most suitable one for their domain. In our experiments, we selected the randomized search method using an altered form of the GSAT algorithm [10] that was originally proposed for solving propositional satisfiability problems. The GSAT algorithm provided by Aleph is modified to suit the clause searching process in ILP fashion.

3 Partial Rules

As described in the previous section, our approach is based on the idea that some partial matches in the rule can be used to classify the unclassifiable examples. Hence, several parts of a rule or *partial rules* are first extracted from the original rules. Then, they are used to classify unseen examples collaboratively. The following describes our partial rule extraction algorithm.

A *partial rule* is *a rule whose body contains a valid sequence of the literals, from the body of the original rule, which starts with the literal consuming the input variables in the head of the rule*. The partial rule extraction algorithm is based on the idea of the newly introduced variables, similar idea as the feature extraction in BANNAR [11]. As shown in Fig. 1, each input variable in a literal is introduced as a new variable in some preceding literals. Thus, we group the literal which consumes the new variable and the literal which introduces that variable into the same sequence. For example, in Fig. 1, the new variable `D` introduced in literal `link(A, B, C, D)` is used as the input variable of literal `D=2.7`. Thus, we group these two literals into the same sequence, `link(A, B, C, D), D=2.7`.

Our partial rule extraction strategy is described below:

- Given the original rule $l_0 \leftarrow l_1, l_2, ..., l_n$ where l_0 is the literal in the head and l_i when $i = 1...n$ is the literal in the body of the rule.
- Construct all possible primitive partial rules $p_0 \leftarrow p_1, p_2, ..., p_m$ where p_0 is l_0; p_i when $i = 1...m$ is the literal selected from l_j when $j = 1...n$; p_{i+1} consumes the variable(s) newly introduced in p_i; p_m does not introduce new variable or there is no l_j consuming new variable introduced in p_m.
- Make all possible combinations of the primitive partial rules, constructed from the previous step, which have the common variables not occurring in the head l_0.

For example, from the original rule, as shown in Fig. 1:

```
molecule(A) :- atm(A, B, C, D, E, F), C=n, E=2.8, bond(A, G, B, H,
               I, J), gteq(J, 1.5).
```

The primitive partial rules are:

```
molecule(A) :- atm(A, B, C, D, E, F), C=n.
molecule(A) :- atm(A, B, C, D, E, F), E=2.8.
molecule(A) :- atm(A, B, C, D, E, F), bond(A, G, B, H, I, J),
               gteq(J, 1.5).
```

All combinations of the primitive partial rules which have the common variables not occurring in the head of the rule are:

```
molecule(A) :- atm(A, B, C, D, E, F), C=n, E=2.8.
molecule(A) :- atm(A, B, C, D, E, F), C=n, bond(A, G, B, H,
               I, J), gteq(J, 1.5).
```

```
molecule(A) :- atm(A, B, C, D, E, F), E=2.8, bond(A, G, B, H,
               I, J), gteq(J, 1.5).
molecule(A) :- atm(A, B, C, D, E, F), C=n, E=2.8, bond(A, G,
               B, H, I, J), gteq(J, 1.5).
```

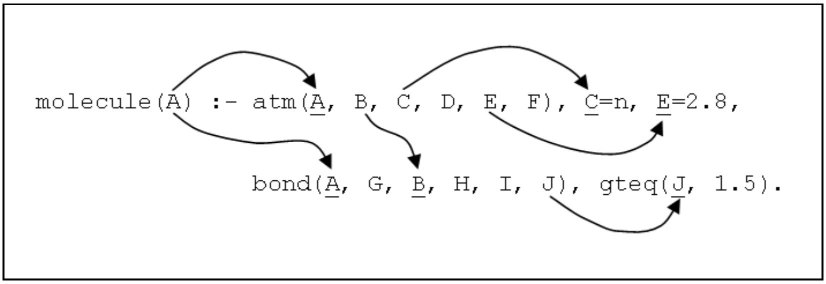


Fig. 1. New variables consumption. The underlined characters show the input variables of literals.

4 Weight Adjustment Using Winnow-Based Approach

As described earlier, our approach is based on the idea that some partial matches can be used to classify new examples. Thus, we extract the partial rules from the original rules and use them collaboratively for classifying examples. The idea of using many partial rules to classify an example is that the partial rules are assigned the importance in form of the weights of each class and all applicable partial rules are combined with their weights for determining the class of the example. When we need to classify an example, we determine the summation of the weight of each class of all partial rules which match with the example. The class which has the highest summation of weights is selected as the class of the example.

The Winnow algorithm [3] is originally proposed as a linear threshold algorithm. For an input vector x, weight vector w, promotion factor $\alpha > 1$, and threshold $\theta > 0$, the algorithm predicts 1 if $w \cdot x \geq \theta$. Intuitively speaking, the Winnow algorithm activates the output if the input x is high enough. If $w \cdot x$ is too low, the weight vector w is increased by updating $w_i \leftarrow \alpha^{x_i} w_i$. On the other hand, if $w \cdot x$ is too high, the weight vector w is decreased by updating $w_i \leftarrow \alpha^{-x_i} w_i$. However, in our approach, the concept of prediction scheme is different. In our class prediction, instead of comparing the summation to the threshold we need only the highest summation of the weight from each class, so that we can make use of the Winnow algorithm by employing the following strategy.

Given a problem with n partial rules, m classes, and promotion factor α. P is a vector of length n, where element p_i of P is a partial rule. W_i is a vector of length m, where element $w_{i,j}$ of W_i is the weight of class j of partial rule p_i. V is a summation vector of length m, where v_i of V is the summation of the weights of class i. The weight vector W_i are updated by using the following procedure.

- Initialize all $w_{i,j} = 1$
- Until termination condition is met, Do
  - For each training example e, Do
    - Initialize all $v_i = 0$ and c as the class of e
    - For all partial rules p_i which match with e, add corresponding W_i to V,

$$V = V + W_i$$

    - Let v_k be the maximum element in V, predict the example e as class k
    - If $c = k$, no update is required; otherwise the weight w_i corresponding to p_i which matches with e is updated by,

$$w_{i,j} = \begin{cases} \alpha w_{i,j} & \text{if } j = k, \\ \\ \alpha^{-1} w_{i,j} & \text{if } j = c. \end{cases}$$

Each partial rule is weighed by a weight vector whose elements are for each class and we classify an example as the class which has the highest summation of the weights of the applicable partial rules. When an example is incorrectly classified, the output class is different from the target class. This means the summation of the weight of the output class of all applicable partial rules is higher than that of the target class. Thus, we decrease the weights of the output class of all applicable partial rules by using Winnow algorithm's weight updating equation, $w_{i,j} = \alpha^{-1} w_{i,j}$ and increase the weights of the target class of all applicable partial rules by using promotion factor, $w_{i,j} = \alpha w_{i,j}$.

To classify an unseen example e, we use the following strategy.

- Initialize all $v_i = 0$
- For all partial rules p_i which match with e, add corresponding W_i to V,

$$V = V + W_i$$

- Let v_k be the maximum element in V, classify the example e as class k

5 Experiments

The data set used in the experiments contained 1366 molecules of dopamine antagonist molecules of 4 classes, D1, D2, D3, and D4. Information of the molecules was originally described in term of the position in three dimension space of atoms, types of atoms, types of bonds, and dopamine antagonist activity of molecules. However, the position in three dimension space was not useful for discriminating examples because a molecule could rotate or move to other positions in the space. Hence, we converted the position of atoms to the relation between atoms and bonds. We instead represented the information of atoms, bonds, and distances between atoms in term of three predicates, `atm/6`, `bond/6`, and `link/4`, respectively. The details of these three predicates are described below:

- `atm(A, B, C, D, E, F)` represents that the atom `B` is in molecule `A`, is type `C`, forms a bond with oxygen atom if `D` is 1, otherwise it does not link to any oxygen atom, has distance `E` to the nearest oxygen atom, and has distance `F` to the nearest nitrogen atom.
- `bond(A, B, C, D, E, F)` represents that the bond `B` is in molecule `A`, has atoms `C` and `D` on each end, is type `E`, and has length `F`.
- `link(A, B, C, D)` represents that in the molecule `A`, the distance between atoms `B` and `C` is `D`.

The following is an example of rules obtained from the experiments.

```
molecule(A,d1) :- link(A, B, C, D), bond(A, E, F, C, G, H), D=6.9,
    H=1.4, bond(A, I, J, F, K, H), bond(A, L, M, J, G, H),
    bond(A, N, B, O, G, P).
[Positive cover = 53 Negative cover = 5]

molecule(A,d2) :- atm(A, B, C, D, E, F), C=n, bond(A, G, B, H, I,
    J), gteq(J, 1.5), atm(A, L, M, D, N, O), N=5.1, O=1.5.
[Positive cover = 42 Negative cover = 1]

molecule(A,d3) :- link(A, B, C, D), D=4.1, atm(A, B, E, F, G, H),
    H=4.1, bond(A, I, B, J, K, L), bond(A, M, C, N, K, L).
[Positive cover = 56 Negative cover = 1]

molecule(A,d4) :- link(A, B, C, D), D=4.4, bond(A, E, C, F, G, H),
    bond(A, I, F, J, G, H), bond(A, K, L, C, G, H), bond(A, M, J,
    N, G, H).
[Positive cover = 130 Negative cover = 8]
```

We compared our approach with two other approaches, i.e. Majority Class [12,13] and Decision Tree Learning. As described in Section 2, ILP's rules alone cannot classify the examples which match with multiple rules from different classes or do not match with any rule, so that we make the rules be fairly compared to other methods by using the Majority Class in such cases.

In the Majority Class method, we selected the class which had the maximum number of examples in training set as the default class. An example which matched with only rule(s) from one class was classified as that class, while an example which could not match with any rule was classified as the default class. In case of the examples which matched with the rules from two or more classes, we selected the class of which the matched rules covered maximum number of examples.

Another method compared in our experiment is the Decision Tree Learning (DTL) algorithm. DTL is a well-known propositional Machine Learning technique which employs the Information Theory to guide in searching for the best theories. The decision tree learner used in our experiments is C4.5 system [14]. To compare our method to C4.5, we used the truth values obtained by comparing the partial rules with examples as the attributes of the examples. By using this attribute set, the examples originally represented as first order logic were transformed into propositional representation. Finally, these transformed examples are used as the training examples for C4.5. For example, assume that we have 8 partial rules. When we compare the example, `molecule(m06497)`, with all partial rules, the second and fifth partial rules are *true* while the other partial rules are *false*. The training examples for C4.5 of `molecule(m06497)` will be $< false, true, false, false, true, false, false, false >$.

We ran 10-fold cross validation experiment using three methods, the original ILP system with the Majority Class method (ILP+Majority Class), Partial Rules and DTL (PR+DTL), and our approach, Partial Rules and Winnow algorithm (PR+Winnow).

The accuracy shown in Table 1 was separately evaluated when the rules were used as in two-class fashion. The covered examples were classified as positive, while the uncovered examples were classified as negative. The accuracy of each class was obtained from the test set consisting of only the examples from the test set of that class. The accuracy in Table 1 shows the accuracy of the rules from each class. The average accuracy of all classes is 76.42%. Furthermore, this percentage of accuracy also shows the coverage ratio on the test set.

Table 2 shows the accuracy of each approach in classifying test examples in multi-class fashion. The accuracy of ILP+Majority Class approach is 79.11%. This shows that only the Majority Class method can slightly improve the accuracy of the original rules. The accuracy of PR+DTL is 85.71%, higher than ILP+Majority with 99.5% confidence level using the standard paired t-test method. The accuracy of PR+Winnow is 88.65%, higher than ILP+Majority and PR+DTL methods with 99.5% and 99.0% confidence level respectively using the same comparing method.

Table 3 shows the ratio between the number of examples correctly classified and the number of examples for each portion. Exactly Covered column indicates the number of the examples covered by the rule(s) from only one class, Multiple Covered column indicates the number of the examples covered by the rules from different classes, and Uncovered column indicates the number of the examples which are not covered by any rule. The results show that our approach remarkably improved the accuracy in Multiple Covered and Uncovered portion. In Multiple Covered portion, PR+Winnow correctly classified 77 of 97 examples, whereas only 49 examples were correctly classified by the Majority Class method. For the uncovered examples, PR+Winnow correctly classified 171 of 220 examples, while only 68 examples were correctly classified by the Majority Class method.

An example of some partial rules which are highly weighed is shown below.

```
molecule(A):- atm(A, E, F, G, H, I), bond(A, N, E, O, P, M),
              atm(A, O, F, G, Q, R), H=2.4.
              [0.1999, 33.5451, 0.2812, 0.5303]
```

Table 1. The accuracy of the output rules used to classify only the positive examples of each class

Class	Accuracy (%)
D1	77.42
D2	70.30
D3	74.80
D4	83.16
Average	76.42

Table 2. The accuracy of the compared methods

Method	Accuracy (%)
ILP+Majority Class	79.11±4.37
PR+DTL	85.71±3.41
PR+Winnow	88.65±3.85

Table 3. Improvements over the original rules with Majority Class method, reported according to exactly covered examples, multiple covered examples, and uncovered examples

Method	Exactly Covered	Multiple Covered	Uncovered
ILP+Majority Class	965/1049	49/97	68/220
PR+Winnow	962/1049	77/97	171/220

```
[The original rule is
molecule(A):- link(A, B, C, D), atm(A, E, F, G, H, I), D=5.6,
              H=2.4, gteq(I, 3.8), bond(A, J, B, K, L, M),
              bond(A, N, E, O, P, M), atm(A, O, F, G, Q, R),
              lteq(Q, 2.9), lteq(M, 1.4), bond(A, S, C, T, P, U).]

molecule(A):- atm(A, B, C, D, E, F), bond(A, G, B, H, I, J),
              atm(A, H, K, D, L, M), L=6.5.
              [0.9524, 0.1566, 35.2224, 0.1904]
[The original rule is
molecule(A):- atm(A, B, C, D, E, F), bond(A, G, B, H, I, J),
              atm(A, H, K, D, L, M), link(A, H, N, O),
              atm(A, N, K, D, L, M), atm(A, H, K, D, L, M),
              L=6.5, F=1.3.]
```

The weights in the above example show another advantage of our approach. We can see that when an example matches with these highly weighed partial rules, the example has the high probability of being classified as the class whose weight is very high. This provides us some knowledge which can be discovered from the dataset, different from the original rules which sometimes are too specific and not useful. Our approach can seek for some pieces of knowledge which are more important than the others in the original rule. For example, the second partial rule in the above example shows that if an unseen example matches with this partial rule, that example has the high probability of being classified as class 3 which is very highly weighed.

6 Conclusion

We have proposed an approach that can improve ILP in multiclass problem. Our method is based on the idea that the unequally important partial rule matching with an example can be useful for classifying the example. The partial rules are extracted from the original rules and are assigned the importance in term of the weights obtained from Winnow-based approach. The experimental results on classifying the activity of the dopamine antagonist molecules show that our approach was successfully applied to such domain by yielding 88.58% accuracy. The accuracy obtained from the experiments also shows that using only the matching of the partial rules and an attribute learner C4.5 improved the accuracy over using the original rules with the majority class method, and the accuracy was much more improved when using the proposed method. Furthermore, the weights of the partial rules also show some pieces of knowledge which are previously hidden in the original rules.

References

1. Enot, D.P., King, R.D.: Application of Inductive Logic Programming to Structure-based Drug Design. In Lavrac, N., Gamberger, D., Todorovski, L., Blockeel, H., eds.: Proc. 7th European Conf. On Principles and Practice of Knowledge Discovery in Databases., Springer-Verlag (2003)
2. Quiniou, R., Cordier, M.O., Carrault, G., Wang, F.: Application of ILP to cardiac arrhythmia characterization for chronicle recognition. Lecture Notes in Computer Science **2157** (2001) 220–227
3. Littlestone, N.: Learning quickly when irrelevant attributes abound: A new linearthreshold algorithm. Machine Learning **2** (1988) 285–318
4. Srinivasan, A., King, R.: Feature construction with inductive logic programming: A study of quantitative predictions of biological activity aided by structural attributes. In Muggleton, S., ed.: Proceedings of the 6th International Workshop on Inductive Logic Programming, Stockholm University, Royal Institute of Technology (1996) 352–367
5. Dietterich, T.G., Bakiri, G.: Solving multiclass learning problems via error-correcting output codes. Journal of Artificial Intelligence Research **2** (1995) 263–286
6. Fürnkranz, J.: Round robin rule learning. In Brodley, C.E., Danyluk, A.P., eds.: Proceedings of the 18th International Conference on Machine Learning (ICML-01), Williamstown, MA, Morgan Kaufmann Publishers (2001) 146–153
7. Eineborg, M., Boström, H.: Classifying uncovered examples by rule stretching. Lecture Notes in Computer Science **2157** (2001)

8. Allwein, E.L., Schapire, R.E., Singer, Y.: Reducing multiclass to binary: A unifying approach for margin classifiers. In: Proc. 17th International Conf. on Machine Learning, Morgan Kaufmann, San Francisco, CA (2000) 9–16
9. Srinivasan, A.: The Aleph Manual (2001)
10. Selman, B., Levesque, H.J., Mitchell, D.: A New Method for Solving Hard Satisfiability Problems. In: Proc. 10th National Conference on Artificial Intelligence, AAAI Press (1992) 440–446
11. Kijsirikul, B., Sinthupinyo, S., Chongkasemwongse, K.: Approximate match of rules using backpropagation neural networks. Machine Learning **44** (2001) 273–299
12. Clark, P., Boswell, R.: Rule induction with CN2: Some recent improvements. In: Proc. Fifth European Working Session on Learning, Berlin, Springer (1991) 151–163
13. Laer, W.V., Raedt, L.D., Dzeroski, S.: On multi-class problems and discretization in inductive logic programming. In: International Syposium on Methodologies for Intelligent Systems. (1997) 277–286
14. Quinlan, J.: C4.5: Programs for machine learning. Morgan Kaufmann Publishers (1993)

Mining of Three-Dimensional Structural Fragments in Drug Molecules

Hiroaki Kato, Takashi Koshika, Yoshimasa Takahashi, and Hidetsugu Abe

Department of Knowledge-based Information Engineering,
Toyohashi University of Technology,
1-1 Hibarigaoka Tempaku-cho, Toyohashi, 441-8580 Japan
{hiro, kosika}@cilab.tutkie.tut.ac.jp,
taka@mis.tutkie.tut.ac.jp, abe@cilab.tutkie.tut.ac.jp

Abstract. This paper describes an approach to three-dimensional structure data mining of drug molecules. The approach is based on reduced graph representation of molecular structures and 3D substructure searching. The procedure was implemented as a software tool, called FragSearch. The tool allows us to actively find meaningful 3D structural features that appear in a particular class of drug molecules. Usage of the approach is discussed with some computational trials using a real dataset of drug molecules.

1 Introduction

An understanding of the structural features of molecules is essential for solving many problems in chemistry. In particular, a substructural analysis and a functional group analysis are very important to structure-activity (or property) studies and rational molecular design based on them. Besides, it is well known that molecular properties, including biological functions, closely related not only to the atomic connectivity but also to the three-dimensional (3D) spacial arrangement of the atoms [1]. Such structural feature analyses could be done manually for a small set of molecules. However, the works are quite tedious and very time consuming for a large set of molecules, even if they are handled with the topological or the 2D representation. For the reason, some computerized approach is required of a systematic data analysis on the structural features for the case.

There are a variety of techniques for molecular data mining including inductive logic programming and inductive databases. For example, De Raedt et al. proposed the level-wise version space algorithm that forms the basis of the inductive query and database system MolFea (Molecular Feature Miner), but it was restricted to linear molecular fragments as patterns [2]. AGM (Apriori-based Graph Mining) approach was proposed by Inokuchi et al. to overcome their limitations [3]. They also applied it to the 3D graph structured data using multiple labeled edges [4]. However, their approach for mining the frequent 3D subgraphs was dependent on the information of the original labels of edges (i.e. bond types in a molecular structure) and virtual links (path lengths, or topological distances between the non-adjacency atoms).

For knowledge discovery based on 3D common structural feature analysis of organic molecules, Takahashi et al. reported a computer program, named COMPASS

A. Sakurai et al. (Eds.): JSAI 2003/2004, LNAI 3609, pp. 538 – 545, 2007.

(COMmon geometric PAttern Search System) [5]. In the work, each molecule was treated as a set of points that correspond to its constituent atoms in the 3D space. The set of points was described by a matrix representation, of which each element involves the inter-atomic distance within the molecule. Thus, it allows us to represent the detail structural information of a molecule, including its 3D geometry, with a weighted graph of which the nodes and edges correspond to atoms and the inter-atomic distances between them, respectively. On the basis of this representation, a maximal common subgraph matching algorithm was employed for the searching of the geometric patterns which are shared into the molecules [6]. However, the COMPASS was implemented to search for just the largest common fragments that are shared with all the molecules as "greatest common divisor". Therefore, it is often difficult to get meaningful results when some of the molecules to be analyzed have quite different chemical structures from the others [7]. Besides, only the maximal common substructural feature(s) are not always important for the structure-activity problem. To this point, the COMPASS couldn't find any other smaller fragments excepting the maximal common subgraph(s) for the searching.

In this paper, on the basis of the COMPASS approach we have developed an alternative fragment search system for more flexible structural feature analysis. For a given set of molecules, the system allows us to enumerate all the fragments which contain more than the specified number of atoms, and it also allows us to search for the fragments appeared with a specified value of the frequency or more (Fig. 1).

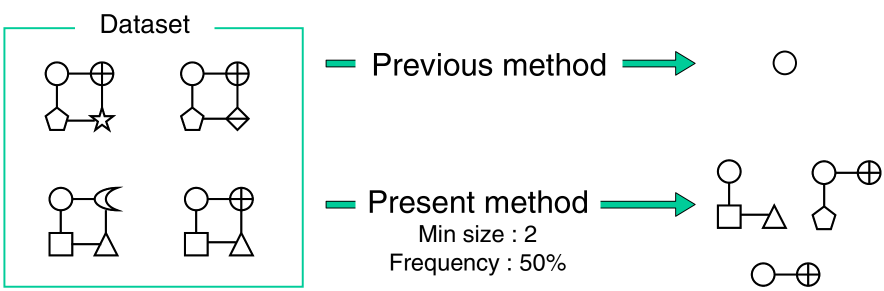


Fig. 1. Basic concept of the frequent fragment pattern search in the present work

In the present work, simply the 3D geometric pattern of atoms is used as the primary information at the feature mining. Then, atomic type, bond type, and/or other additional attributes of molecules are also available for filtering the searching results. Alternatively, several reduced graph representation of 3D molecular structure can also be employed for the more efficient analysis.

2 Methods

2.1 Basic Algorithm for Finding Frequent Fragment Patterns

At first, all distinct pairs of atoms (size-2 fragments) of the first molecule in a chemical structure database are generated with their inter-atomic distances. For the other molecules, the 3D substructure searching is carried out using each fragment generated in the above as a query, and the fragments that satisfied user-defined search conditions survive. For keeping the information whether a size-2 fragment satisfies the conditions specified, a candidate matrix whose elements correspond to the fragments is prepared. When the fragment consisting of *i*-th and *j*-th atoms of the molecule satisfies the conditions, the element c(*i*, *j*) of the matrix is set to 1, otherwise set to 0. The same process is applied for the second molecule, third, and so on. It is obvious that a larger fragment made by extending the size-2 fragment which does not satisfy the conditions never satisfies them, either. Therefore, this matrix can be used for pruning the unnecessary search procedures.

Next, the fragments of size N+1 are generated from a size-N fragment with the candidate matrix for the reference molecule. For each generated fragment, a strict validation is made using the result for size-N fragments. Then 3D substructure search is done to test the conditions of survival. For efficiently exploring, the database searching is terminated when it becomes clear that the fragment in question doesn't satisfy the conditions specified in advance. The procedures are repeated until further extension of the fragment fails (Fig. 2).

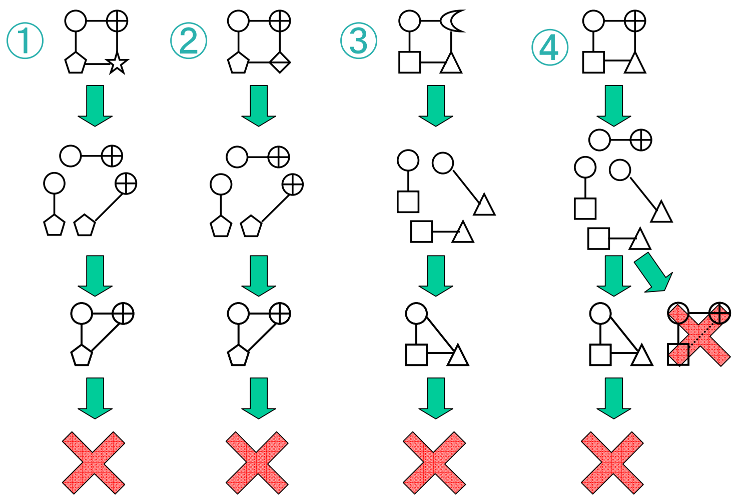


Fig. 2. Illustrative example of an exploration of the frequent fragments among four small graphs. In this case, a threshold value of the frequency was set to 50%.

Finally, all information about the fragments whose sizes are equal to the user-defined threshold value or more is saved into the fragment dictionary. To avoid redundancy, the smaller fragments that succeeded in the further extension are discarded. Then, all the entries of the dictionary are sorted according to their frequencies. When two or more fragments are regarded as identical within the allowance of the distance predefined, the fragments with the lower frequency are removed from the dictionary. If they have the same frequency, the fragment originated from the parent molecule that has the smallest registry number is chosen as a representative. Furthermore, if the parent molecules are the same, i.e. plural similar fragments are generated from an identical parent molecule, then the fragment which consisted of the smaller indexes of atoms (in the dictionary order) is prioritized.

2.2 User Defined Parameters for Fragment Searching

The user can specify the searching conditions [8] as the following:

(1) A threshold value of the frequency of appearance (or minimum support [3]); For the given set of molecules, the fragment patterns with the specified value or more are considered as the candidates for the larger fragments. The value is specified by percentage point (%).
(2) The minimum number of atoms for the fragments to be explored.
(3) The allowance at testing the equivalency for the inter-atomic distances; The inter atomic distances are regarded as equivalent when the difference is smaller than this allowance value. The input value is specified by the unit of Å.

In addition to these conditions, some other optional attributes to be matched can be specified such as atomic type, atomic charge, and/or enantiomer geometries.

2.3 Reduced Representation of Molecular Structure

It is expected that a very large computational time will be required for finding the frequent fragment patterns among large molecules and a large number of molecules. To decrease the computational time for the case, a reduced representation of molecular structure was investigated here. In this representation method, particular substructures such as functional groups and rings are regarded as pseudo-atoms (or super-atoms). These reduced representations can be defined by the user, arbitrarily. Furthermore, the user can specify particular atoms (or atomic groups) to suppress trivial atom(s) in molecular structures. Using this utility, for example, the user can treat a molecular structure as a reduced graph that consists of only pseudo-atoms and hetero atoms without trivial carbons and hydrogen atoms.

The super-atom transformation of particular atomic group(s) for a molecule was implemented on the basis of a substructure search technique in the ordinary chemical graph representation. The user can specify a set of atoms to be reduced and the atomic coordinate of the super-atom. For example, in the case of benzene-ring (Fig. 3), the reduced atoms are six carbons and a representative coordinate of these atoms is set at the center of the ring. On the other hand, for the case of carboxyl group, the atoms can be reduced into a single super-atom and the representative coordinate of the super-atom is approximated to that of the carbon atom of the -COOH. Although these preprocesses possibly cause some problems in the following procedures because of the user's bias to the data set, we believe that these representations allow us not only to reduce the searching space, but also to obtain more significant search results in chemical sense.

The algorithm mentioned above has been implemented into the program, FragSearch, by Java language, and all of the computational works in the following were carried out on a PC (CPU: Pentium4 2.4GHz, main memory: 512MB) with Sun Java2 SDK, SE v1.4.0_01 on Windows XP.

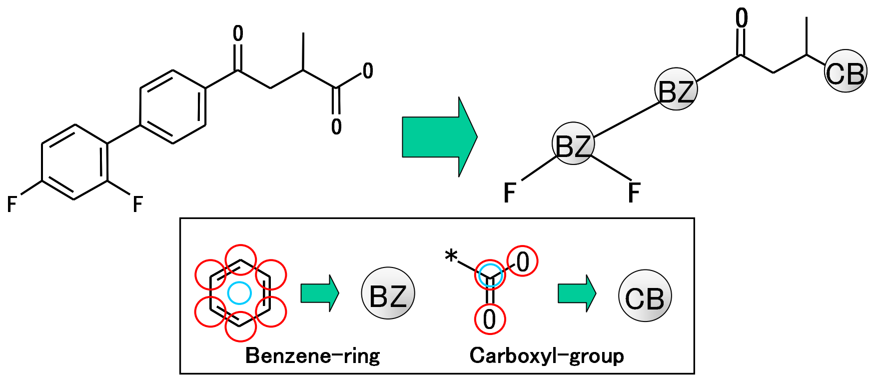


Fig. 3. Illustrative example of the reduced representation of a molecular structure

3 Results and Discussion

3.1 3D Fragment Searching on Dopamine D1 Receptor Agonists

We have prepared a target data set that contains 66 molecular structures taken from MDDR-3D (MDL Drug Report) Ver.2001.1 [9]. These molecules have the same biological activity, the dopamine D1-receptor agonist activity. The hydrogen atoms were omitted, and the above-mentioned benzene-ring reduced representation was employed. We assumed that every 3D structure of molecules used in the present analysis is rigid. Search trials were carried out under the conditions that the distance allowance is 0.5Å, the different types of elements are distinguished, and the value of frequency is 100%, 50%, and 30%. The results are summarized in Table 1. Some graphical views of the extracted fragments are shown in Fig. 4 with their parent molecular structures. Obviously, the FragSearch with the lower value of frequency mined the larger and more interesting fragments. The benzene-ring reduced representation allows us the efficient mining, and it gives us various structural features for this case.

Table 1. The result of the 3D fragment search for 66 drug molecules

Frequency [%] (Comput. Time)	Extracted Fragments	
	Size	Number
100 (13 sec.)	3	1
	2	6
	--	--
50 (8 min.)	7	1
	6	39
	5	219
30 (24 min.)	10	1
	9	2
	8	20

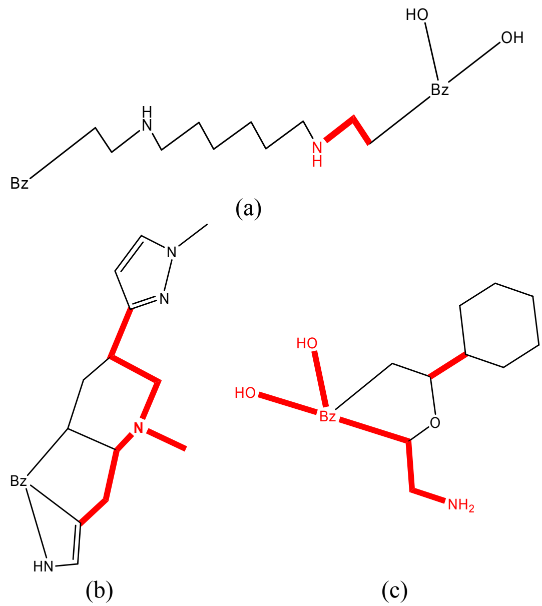


Fig. 4. Graphical views of three different 3D fragments superimposed on their parent molecular structures. The label "Bz" means the benzene-ring in the reduced representation. The fragment (a) was one of the fragment patterns extracted with the value of frequency of 100%, both (b) and (c) were obtained with the frequency of 30%.

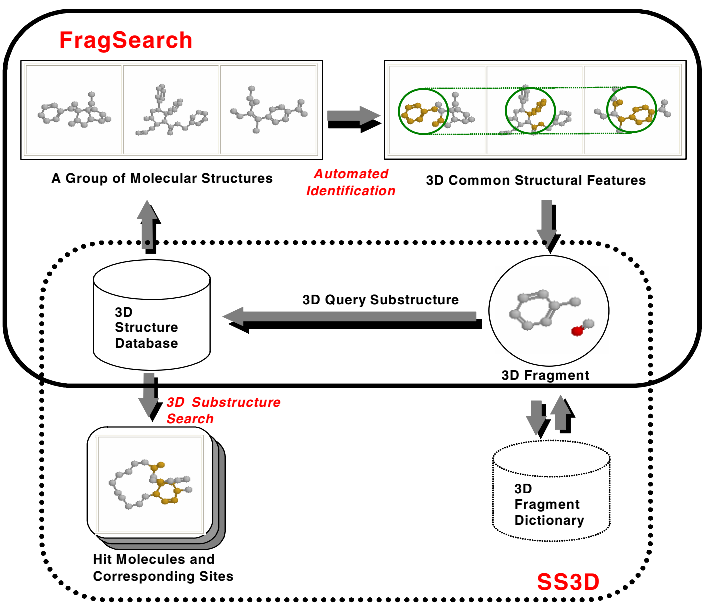


Fig. 5. 3D structural feature mining for drug molecules using the present approaches

3.2 3D Structural Feature Mining Using the Fragment Dictionary

Alternatively, in the previous work, the authors reported a computer program, SS3D, for 3D substructure searching, which allows us to identify all occurrences of a user-defined 3D query pattern [6]. More extensive analyses of 3D structural features of drug molecules can also be executed by using our program with the fragment patterns registered in the 3D fragment dictionary (Fig. 5). In the present work, we have prepared other six data sets that consist of dopamine agonists (for D2 receptor and auto-receptor), and antagonists (D1, D2, D3, and D4 receptors). The 3D pattern searching was carried out for the representative fragment shown in Fig.4 using these data sets. The number of molecules for each data set and the percentage of hit molecules are summarized in Table 2. Here, the fragment (a) is very small and very common. On the other hand, it is shown that the fragment (b) is specific to the dopamine agonists, and (c) is closely related to D1 agonists, respectively. These results suggest that the present approach is quite useful for 3D structural data mining for drug molecules.

Table 2. The results of 3D substructure searching for the 3D fragment patterns obtained in dopamine D1 receptor agonists. The hit molecules for each data set are indicated in percentage. The labels (a), (b), and (c) are corresponded to the fragments shown in Fig. 4. The number in parentheses shows the number of molecules in the data set. The total computational time was about 20 seconds.

Fragment	Dopamine Agonists			Dopamine Antagonists			
	D1	D2	Auto	D1	D2	D3	D4
	(66)	(143)	(191)	(173)	(430)	(254)	(574)
(a)	100	100	100	100	99	98	100
(b)	33	45	10	1	4	9	0
(c)	30	0	0	0	0	0	0

4 Conclusion

A computer program used for 3D structural fragments search, FragSearch, has been developed. This program can identify all the fragment patterns that appear with the frequency of occurrences specified for the given set of molecules. The computational experiment was carried out using a data set of the dopamine D1 receptor agonists of 66 compounds. The FragSearch successfully found several fragments that seemed to be characteristic for the drug activity class.

The comprehensive analysis of dopamine agonists and antagonists are now under progress using the tools described here. We believe that the 3D fragment information helps us understand structure-activity relationships of drug molecules. The graphical user interface for these tools will be also required in future works.

Acknowledgement

The authors wish to thank Prof. Takashi Okada and Dr. Masumi Yamakawa of Kwansei Gakuin University for their valuable comments. This work was supported by a Grant-in-Aid for Scientific Research on Priority Areas '*Active Mining*', from the Japanese Ministry of Education, Culture, Sports, Science and Technology.

References

1. Willett, P.: Chemical similarity searching, *J. Chem. Inf. Comput. Sci.*, **38** (1998) 983-996
2. De Raedt, L., Kramer, S.: The level-wise version space algorithm and its application to molecular fragment finding, *Proc. 17th International Joint Conference on Artificial Intelligence*, (2001) 853-862
3. Inokuchi, A. et al.: Applying the apriori-based graph mining method to mutagenesis data analysis, *J. Comput-Aided Chem.*, **2** (2001) 87-92
4. Nishimura, K. et al.: Fast apriori-based graph mining algorithm and application to 3-dimensional structure analysis, *Trans. Jpn. Soc. Artificial Intelligence*, **18** (2003) 257-268
5. Takahashi, Y. et al.: Automated recognition of common geometrical patterns among a variety of three-dimensional molecular structures, *Anal. Chim. Acta*, **200** (1987) 363-377
6. Kato, H., Takahashi, Y.: Development of a three-dimensional substructure search program for organic molecules, *Bull. Chem. Soc. Jpn.*, **70** (1997) 123-127
7. Kato, H. et al.,: Data mining based on 3D structural similarity of drug molecules, *IEICE Tech. Rep.*, **102** (2003) AI2002-88
8. Koshika, T. et al.: Development of a three-dimensional fragment search system for SAR, *26th Symp. Chem. Inf. Comput. Sci.*, (2003) 109-110
9. MDL Information Systems, Inc., MDL Drug Data Report -3D, Ver. 2001.1 (2001)

Author Index