Active Mining Project

Field number 759

Grant-in-Aid for Scientific Research on Priority Area

Interim Research Summary

September 2002

Principal Investigator

Hiroshi Motoda

(I.S.I.R., Osaka University)

Preface

The active mining project started in September, 2001. This is a four year project which is funded by the Japanese Ministry of Education, Culture, Sports, Science and Technology as Scientific Research of Priority Areas.

In this era of information overload where everybody can access easily to millions of records of information through the internet, the need is ever increasing for 1) identifying and gathering relevant data from a huge information search space, 2) mining useful knowledge efficiently and effectively from different forms of data, and 3) promptly reacting to situation changes and giving necessary feedback to both data collection and mining steps.

The goal of this project is, in response to the above needs, to explore mechanisms of 1) active information gathering in which necessary information is effectively searched and pre-processed, 2) user-centered active mining in which various forms of information sources are effectively mined, and 3) active user reaction in which the mined knowledge is easily assessed and prompt feedback is made possible.

Active mining is a collection of activities each solving a part of the above need, but collectively achieves the above various mining need. By "collectively achieving" it is meant that the total effect outperforms the simple add-sum effect that each individual effort can bring. In other words, a spiral effect of these interleaving three steps is the target to be pursued.

The research covers many activities of datamining and comprises ten independent subprojects which are grouped into the above three scopes. This interim report is a collection of research summary of each individual activity that is available at the time of printing. We hope that this helps disseminate the progress of active mining project and are looking forward to receiving any comments and suggestions.

We are grateful for your support and encouragement.

September, 2002

Hiroshi Motoda
Principal Investigator

I. A01 Active Information Gathering

A01-02 Acquiring Meta Information Resources in the WWW

A01-03 Active Information Gathering from Distributed Dynamic Information Sources

A01-04 Automation of Data Gathering and Preprocessing by using Multistage Learning

II. A02 User-Centered Active Mining

A02-05 Active Mining for Structured Data

(A01-02-1) Active Information Gathering with Interactive Document Retrieval

Principal Investigator Seiji Yamada (National Institute of Informatics)
Collaborator Masayuki Okabe (Japan Science and Technology Corporation)

Background and Aim

Web search engines are indispensable tools to access useful information which might exist somewhere on the Internet. While they have been getting higher capability to meet various information needs and large amounts of transactions, they are still insufficient in the ability to support the users who need to collect a certain number of Web pages relevant to their purpose. Based on a query (usually composed of a few words) inputted by a user, search engines return a "hit list" in which so many Web pages are presented in a certain order. However it does not often reflect the user's search intent and includes a large number of non-relevant Web pages, thus the user would waste much time and energy on judging many Web pages.

To resolve this problem and provide efficient retrieval process, we propose a system which mediates between a user and a search engine in order to select only relevant Web pages out of a hit list through the interactive process called *relevance feedback*. Given some Web pages marked with their relevancy(relevant or non-relevant) by a user, this system generates a set of rules for filtering relevant ones from many Web pages in a hit list, each of which is a logical rule to decide whether the user should look a Web page or not.

The system constructs such rules from the combination of keywords, relational operators and tags with a relational learning algorithm which is superior to learn structural patterns. We have developed this basic framework in document retrieval(Fig. 1) and found our approach was promising. In this report, we applied this method to an interactive filtering system which coordinates hit lists of search engines in order for an individual user to find their useful information easily.

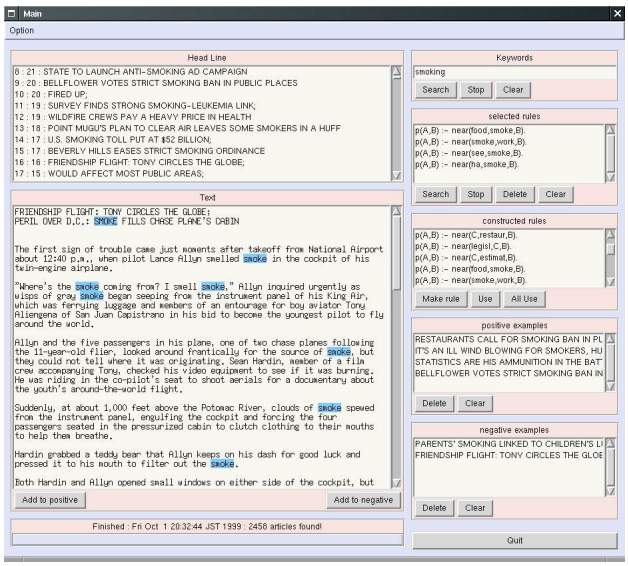


Figure 1 Intreface of interactive document retrieval

Research Plan and Approach

Fig. 2 shows an interactive Web page filtering system we propose. This process consists of six steps, each of which corresponds to the number in Fig. 2.

In step 1, a user starts to search by inputting a query to the system, and then receives the hit list. In step 2, the user evaluates and marks the relevancy(relevant or non-relevant) of its upper (more or less)10 pages in order to teach the system what kind of pages are needed. In step 3, the system makes an analysis of the marked pages by extracting extended keywords and literals which are used to construct filtering rules.

Based on the literals and a learning algorithm, in step 4, the system generates filtering rules to distinguish relevant Web pages and non-relevant ones. Fig. 3 shows that some keywords extracted from Web pages are added to a body of a rule. Adequate keywords are selected with expected information gain.

Step 5 is prepared for the case that the user noticed the initially or previously inputted query was not adequate or sufficient and thinks it's better to search again. This step is not always done. Step 6 is the revision procedure in which the system selects(re-selects) relevant pages based on the newly constructed filtering rules to provide the user with the better results.

These procedures follow the general relevance feedback process, and the steps 2 to 6 repeat until the user would collect enough relevant pages.

Main Results

Fig. 4 shows relation between judged pages and relevant pages found in the judged pages when our Web page filtering was employed and not. A system without the Web page filtering is identical to just a Web search engines. The number of relevant pages is average value per 20 topics.

About first ten pages, there is no difference because two methods judge the same ten pages. The difference of found relevant pages increases after the first feed-

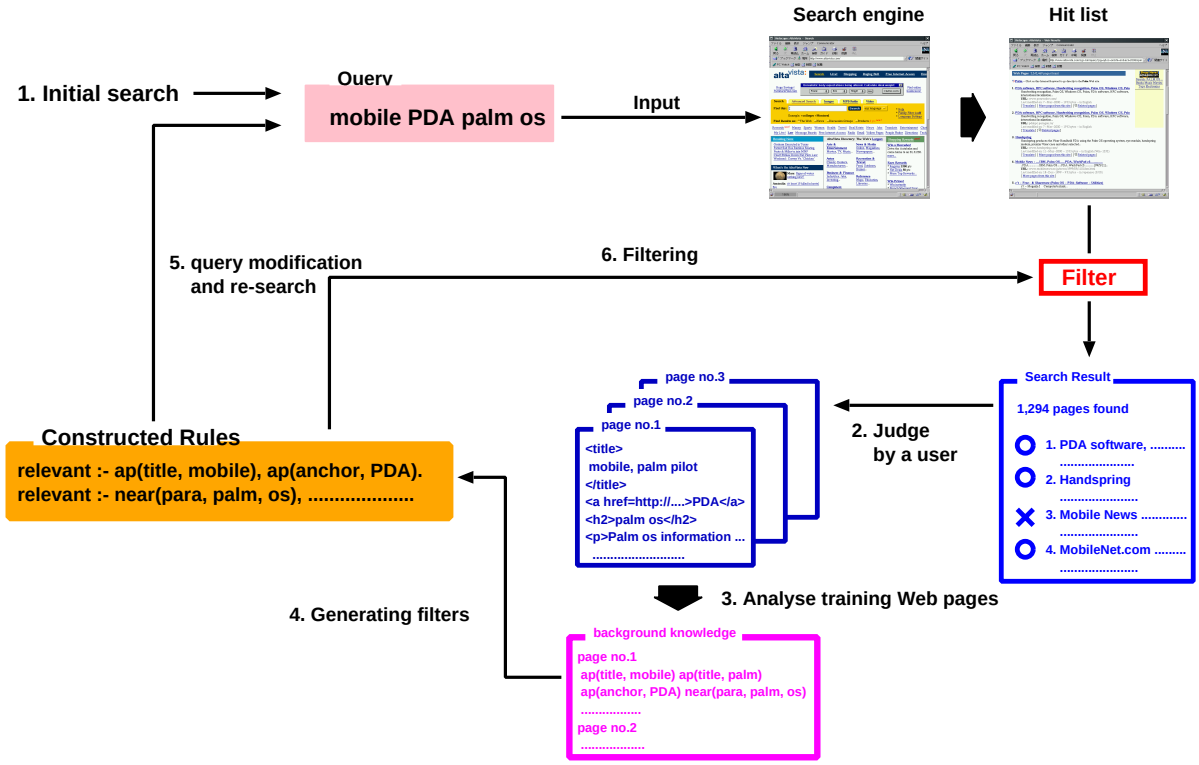


Figure 2 Web page filtering with relational learning

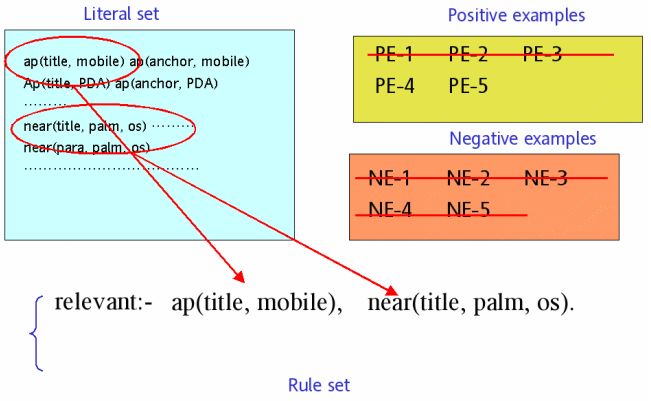


Figure 3 Learning of filter rules.

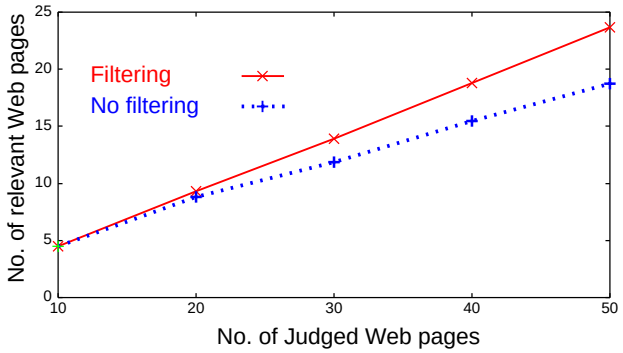


Figure 4 Average of relevant Web pages.

back at ten pages. As a result, a system with our Web page filtering got about 5 relevant pages more than retrieval1 after the fourth feedback was done.

Thus we conclude our Web filtering system enables a user to more efficiently many useful(relevant) Web pages than a search engine.

Future Plan and Expected Results

To reduce cognitive load which a user needs to judge pages, we have a plan to utilize a clustering method for grouping Web pages in a hit list. The grouped Web pages are indicated to a user instead of individual Web pages, and he/she can evaluate only representative Web pages of their groups. Since judge of such a single representative Web page makes judge of all the Web pages in the group, we can reduce the cost of user's judge. However defining similarity between Web pages is critical to correct clustering.

Contact:

Seiji YAMADA
National Institute of Informatics
2-1-2 Hitotsubashi, Chiyoda, Tokyo 101-8430, JAPAN
Tel&Fax: +81-3(4212)2562 E-mail: seiji@nii.ac.jp

(A01-02-2) Partial Update Monitoring in Web pages for WWW Information Management

Principal Investigator Seiji Yamada (National Institute of Informatics)
Collaborator Yuki Nakai (Tokyo Institute of Technology)

Background and Aim

We currently obtain various information from the WWW and utilize them for many purposes like business, education, personal use and so on. Since we can easily make, delete and modify Web pages, the WWW is growing as a huge and dynamic information resource. While one of the most important advantages of the WWW is its frequent updates of Web pages, we needs to constantly check them for acquiring the latest information and this task obviously forces much cognitive load to us. Thus a number of applications and services to automatically check and notify updates of Web pages have been developed. Unfortunately almost all of them notify updates to a user whenever any part of a Web page is updated, and most of such updates may not useful to him/her.

We developed an automatic monitoring system PUM(Partial Updates Monitoring) that constantly checks partial updates in Web pages and notifies them to a user. A user can give PUM regions in which he/she wants to know the updates in a Web page as training examples, and it is able to learn rules to identify the partial updates by classification learning. By this learning, a user do not need to directly describe the rules. Since describing such rules is significantly hard to a naive user, this learning of PUM releases a user from much cognitive load.

Research Plan and Approach

Fig. 1 shows overview of PUM. PUM is a system that identifies a region indicated by a user in a Web page, checks updates in the region and notifies a user the updates which he/she wants to know. A broken line indicates interaction between a user and PUM.

Fig. 2 shows the interface of PUM consisting of three sub-windows for Web browser, rule indication and training example indication. Through the interface, a user indicates the region in a Web pages by mouse drag and the region is given to PUM as a (positive) training example. Negative examples, which are important to inductive learning, are given by a user explicitly, or are generated automatically with heuristics. When PUM obtained training examples through interactions with a user, then a classification learning system acquires two kinds of rules for region identification and update check. Region identification rules are used to identify and extract a region in which a user wants to know its updates. Update check rules are utilized to determine whether the update is one which a user wants to know or not. After such rules were generated, PUM becomes able to identify partial updates and determine whether it is one which a user wants to know or not by using two kinds of rules.

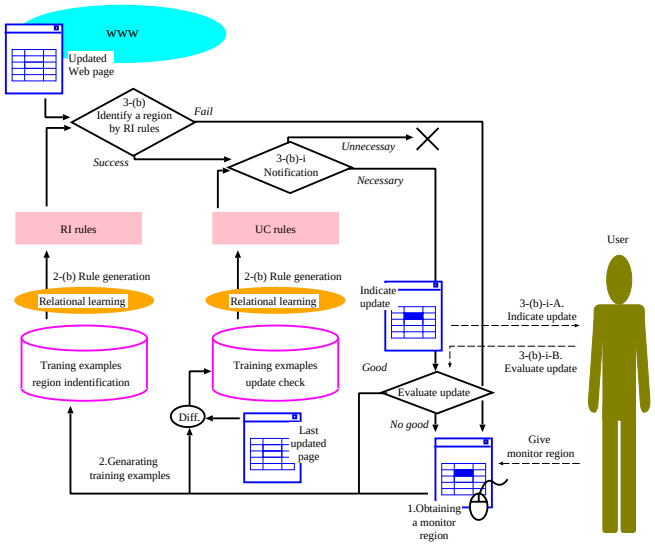


Figure 1 PUM System overview.

If PUM decides an update is useful to a user, it notifies the update to a user. Otherwise PUM indicates the updated Web page to a user and obtains his/her evaluation. PUM was implemented using Visual C++ and Ruby on Windows2000.

Main Results

We experimentally found that PUM was able to correctly detect partial updates in various Web pages. For example, available Web pages for PUM are a stock market Web pages (Fig. 3), a weather report Web page (Fig. 4) and so on.

PUM dealt with a cell in a table as a region and we consider PUM is applicable to any table in Web pages of any field. Since a table has been an important target of information extraction studies in WWW thus far, PUM sufficiently covers significant partial updates of Web pages. More strictly, PUM can deal with regions identified by the number of brother nodes in a HTML tree and cells in a table with row/column header. However PUM has limitation on its coverage.

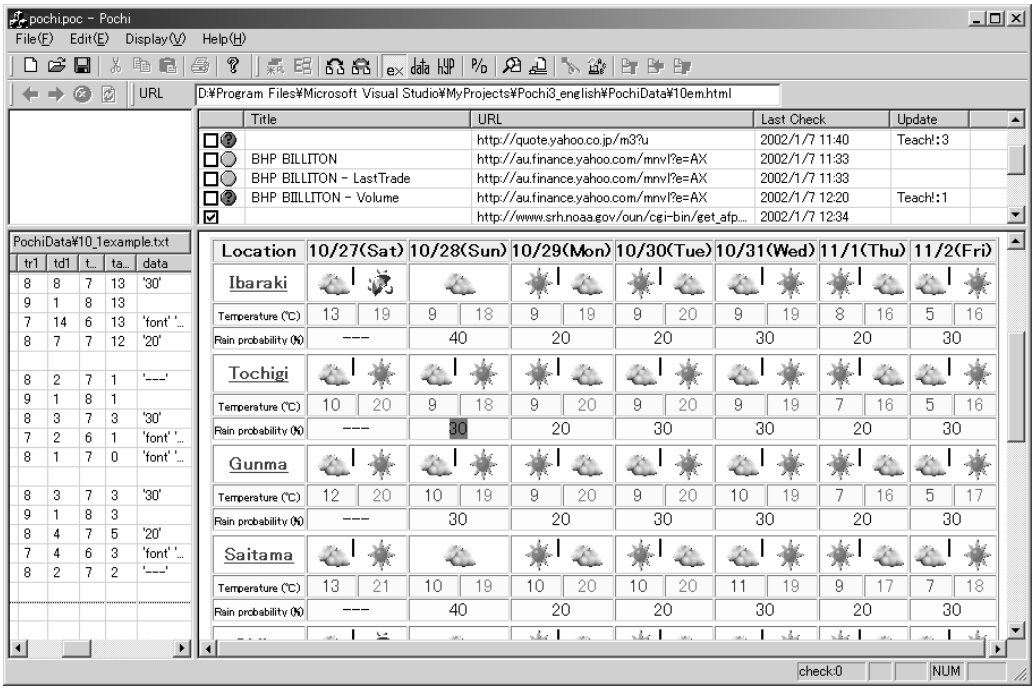


Figure 2 PUM interface.

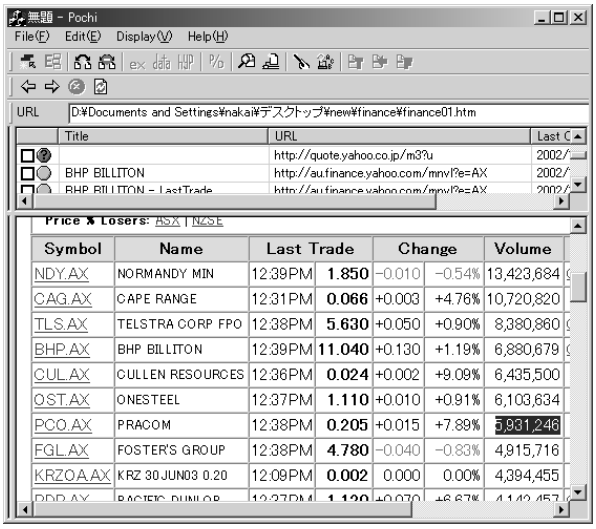
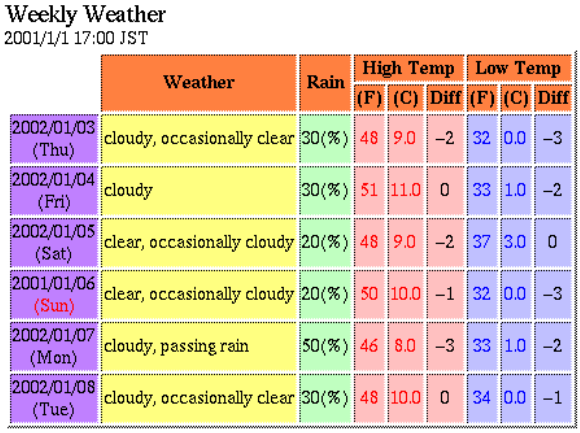


Figure 3 Stock market Web page.



Weekly Weather
2001/1/1 17:00 JST

	Weather	Rain	High Temp			Low Temp		
			(F)	(C)	Diff	(F)	(C)	Diff
2002/01/03 (Thu)	cloudy, occasionally clear	30(%)	48	9.0	−2	32	0.0	−3
2002/01/04 (Fri)	cloudy	30(%)	51	11.0	0	33	1.0	−2
2002/01/05 (Sat)	clear, occasionally cloudy	20(%)	48	9.0	−2	37	3.0	0
2001/01/06 (Sun)	clear, occasionally cloudy	20(%)	50	10.0	−1	32	0.0	−3
2002/01/07 (Mon)	cloudy, passing rain	50(%)	46	8.0	−3	33	1.0	−2
2002/01/08 (Tue)	cloudy, occasionally clear	30(%)	48	10.0	0	34	0.0	−1

Figure 4 Weather report Web page.

However there is no guarantee that a user always indicates the correct regions in a Web page in his/her evaluation. Thus PUM needs to deal with noisy training examples.

Future Plan and Expected Results

Though classification learning is rather fast, it is not sufficiently rapid for an interactive learning system like PUM. A way to improve performance is that a user directly modifies learned rules.

Fortunately symbolic rule representation is far more suitable for a user to modify learned knowledge directly than weight distribution learning like neural network learning, regression and so on. Thus we are developing a human-computer interaction framework to deal with such user's help.

Contact:

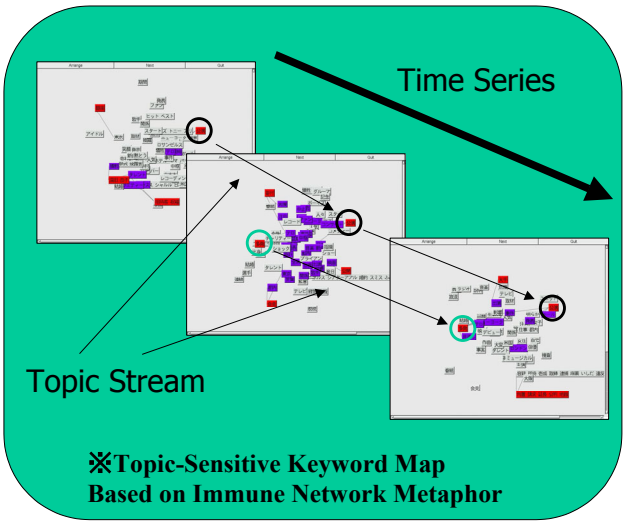
Seiji YAMADA
National Institute of Informatics
2-1-2 Hitotsubashi, Chiyoda, Tokyo 101-8430, JAPAN
Tel&Fax: +81-3(4212)2562 E-mail: seiji@nii.ac.jp

(A01-02-3) Immune network-based Clustering for WWW Information Gathering/Visualization

Investigator Yasufumi TAKAMA (Tokyo Metropolitan Institute of Technology)
Collaborator Kaoru HIROTA (Tokyo Institute of Technology)

Background and Aim

Various kinds of information issued by a lot of companies as well as individuals are available on the Web. As the information overload has been a serious problem for the users, the systems that can effectively assist the human in gathering and organizing the information on the Web should be developed. Our research aims to develop a system that can handle the dynamic nature as well as the variety of information on the Web.



Research Plan and Approaches

1) Topic-sensitive keyword map generation based on immune network metaphor

A method is proposed to visualize the topic distribution among the document set such as retrieval results and online news articles of a certain category. The proposed method employs the immune network model (Eq. (1)-(5)) to find the set of landmark keywords so that their corresponding document clusters (documents with same landmark form a cluster) cannot overlap each other. Furthermore, by utilizing the relationship among antibodies and antigens as a metaphor for the keyword map generation, topic-sensitive keyword arrangement can be obtained constantly.

$$\frac{dX_i}{dt} = s + X_i\left(f\left(h_i^b\right) - k_b\right), \tag{1}$$

$$h_i^b = \sum_j J_{ij}^b X_j + \sum_j J_{ij}^g A_j, \tag{2}$$

$$\frac{dA_i}{dt} = \left(r - k_g h_i^g\right) X_i, \tag{3}$$

$$h_i^g = \sum_j J_{ji}^g X_j, \tag{4}$$

$$f(h) = p\frac{h}{(h+\theta_1)}\frac{\theta_2}{(h+\theta_1)}, \tag{5}$$

†X_i , A_i …the concentration (activation) values of B-Cell i and antigen i, s…a source term modeling a constant cell flux from the bone marrow ,
r… a reproduction rate of the antigen,
k_b, k_g …the decay terms of the antibody and antigen,
J^b_{ij} , J^g_{ij} …the connectivity between the antibodies i and j, and that between antibody i and antigen j.

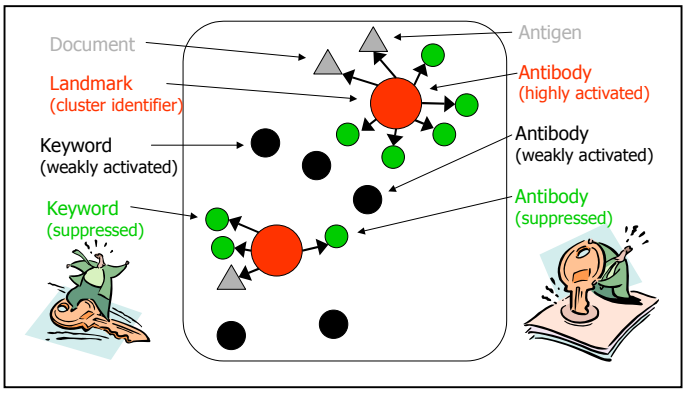


Fig. Immune Network Metaphor (Relation between Keyword Map (Left) and Immune Network Model (Right)).

2) Finding topic stream from a sequence of document sets

A sequence of document sets are frequently found on the Web, such as online news article sets or the result of a series of information retrieval processes. Based on this

notion, the method proposed above is applied to extract the major topic stream from such sequences of document sets. The property of memory cell is introduced in the above-mentioned visualization method so that the topic that has been once found can have the activation priority against other topic candidates.

Main Results

1) Topic-sensitive keyword map generation based on immune network metaphor

The proposed method are compared with the K-means clustering method based on the questionnaires, and gets better evaluations than K-means for 2 of 3 document sets. From the viewpoint of keyword map generation, it is confirmed through several experiments that the keyword arrangement that emphasizes the topic distribution found by the proposed method can be constantly obtained.

Table: Comparison between proposed method and K-means clustering based on questionnaires.

Data	Eval. Item	Proposed	K-means
Set1	# of Cllusters	5	4
	Dist. of Clusters.	0.48	3.6
	Eval. Avg.	4.33	3.90
Set2	# of Cllusters	5	4
	Dist. of Clusters.	0.32	4.625
	Eval. Avg.	3.82	3.13
Set3	# of Cllusters	5	5
	Dist. of Clusters.	0.48	4.25
	Eval. Avg.	2.3	4.00

2) Finding topic stream from a sequence of document sets

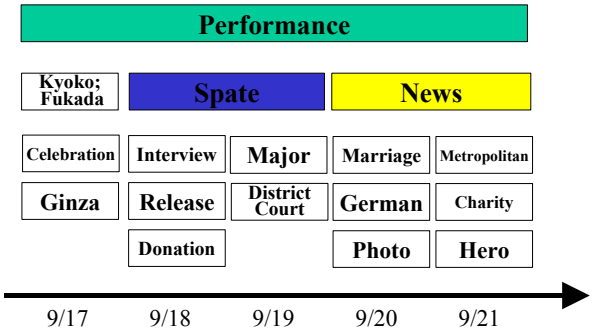
The proposed method equipped with the property of memory cells is applied to two sequences (5-day stream and 2-week stream) of online new articles, and experimental results show that the proposed method with memory cells can find the topic stream 2 – 3 times as many as the method without memory cells.

Future Plan and Expected Results

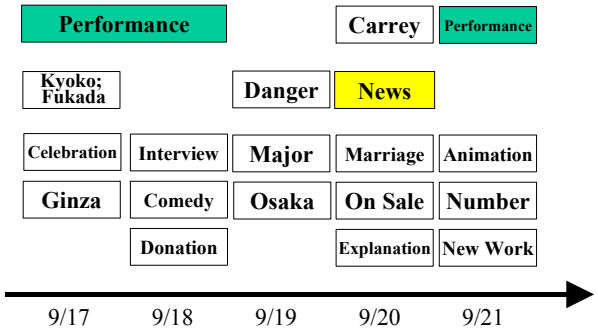
The developed system will be enhanced so that the information resource on the Web can be gathered and visualized from more various viewpoints.

1) Extraction of meta information, such as subjectivity / objectivity, as the supplemental information of the topic information represented by noun phrases.
2) Visualization of various relations among topics.
3) Considering dynamic nature of topics in time series.

Extracted Stream with Using Memory Cell



Extracted Stream without Memory Cell



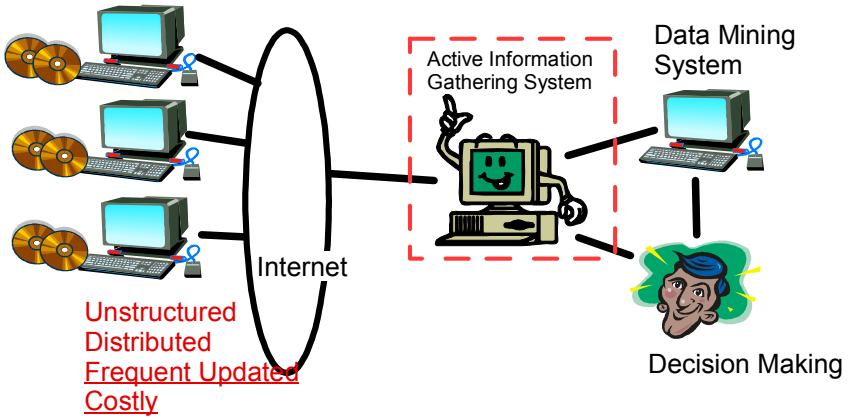
Contact:

Yasufumi TAKAMA (Investigator)
Tokyo Metropolitan Institute of Technology
6-6 Asahigaoka, Hino, Tokyo 191-0065, JAPAN
Email: ytakama@cc.tmit.ac.jp
Tel/Fax: +81-42-585-8629

(A01-03-1) Intelligent Information Gathering Technique from Distributed Dynamic Information Sources and Its Application to Discovered Knowledge Filtering

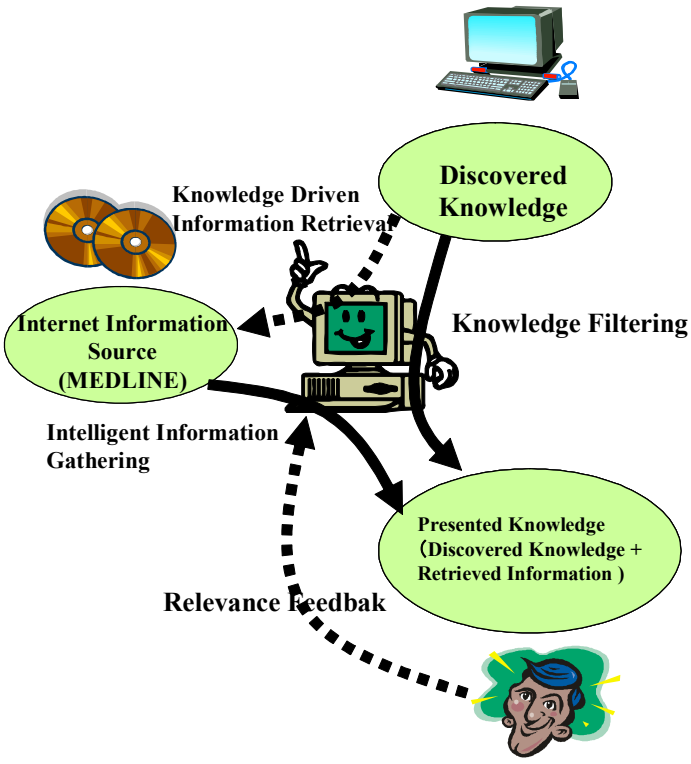
Principal Investigator Yasuhiko Kitamura (Graduate School of Engineering, Osaka City University)
Collaborators Park Keunsik (Graduate School of Medicine, Osaka City University)
Akira Iida and Takuya Murao (Graduate School of Engineering, Osaka City University)

Background and Aim



To realize active mining that exploits information sources on the Internet such as World Wide Web, we need to gather information considering the characteristics of the sources, which are summarized as follows. (1) There is a huge amount of various and unstructured information on the Internet. (2) Pieces of related information are distributed among a large number of information sites. (3) The information is updated frequently. (4) Gathering information takes time and cost. Conventional researches on information gathering mainly deal with problems raised by (1) and (2), and Semantic Web and Mediator technologies have been developed as remedies. In our project, we mainly deal with problems raised by (3) and (4), and develop essential technologies required for developing active information gathering systems which efficiently gather and integrate dynamic information from distributed sources on the Internet and apply them to the fields of data mining and decision making.

Research Plan and Approach



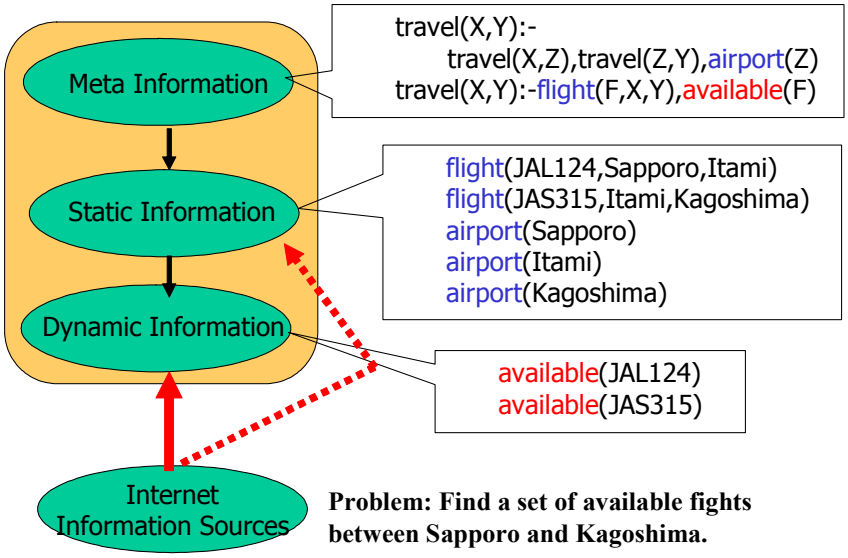
Developing intelligent information gathering method: We develop a method to gather information from distributed dynamic information sources. To this end, firstly, we develop a model to represent distributed dynamic information sources, in which the location, importance, and update frequency of information are specified. Secondly, we develop an intelligent information gathering algorithm that efficiently gather information based on the model mentioned above. Lastly, we develop a mechanism to monitor the changes of distributed dynamic information sources that are frequently updated.

Applying the method to discovered knowledge filtering: We apply the intelligent information gathering method to a data mining system. A data mining system discovers a number of pieces of knowledge by using machine learning technique, but all the knowledge discovered is not useful to the user. To improve the quality of knowledge discovery, we filter the knowledge by using information retrieved from the Internet sources. The task proceeds as follows. (1) [Knowledge driven information retrieval] we automatically generate queries to retrieve information that relates to discovered knowledge. (2) [Intelligent information gathering] we efficiently gather information specified by the queries. (3) [Knowledge filtering] we filter discovered knowledge by using gathered information. (4) [Relevance feedback] we improve the quality of filtering by using relevance feedback technique.

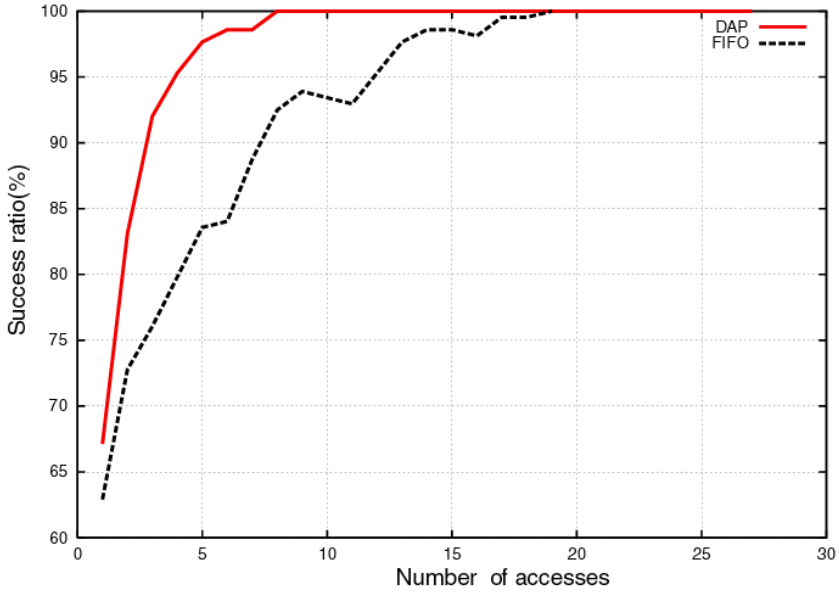
Main Results

Development of intelligent information gathering algorithm based on dynamic and static information: We developed an information gathering algorithm for flight information service, which utilizes three types of information; meta, static, and dynamic information. Dynamic information (ex. flight availability) is frequently updated and is target information to be gathered. To gather dynamic information efficiently, we can utilize static information (ex. flight schedule), which is relatively stable information that can be used to evaluate the value of the dynamic information.

Meta information (ex. flight routing) is used to integrate dynamic information and static one. Our proposed method (DAP) produces better solutions than conventional method (FIFO), which considers only the freshness of information, especially when a small number of information accesses are allowed.



Problem: Find a set of available fights between Sapporo and Kagoshima.



Future Plan and Expected Results

Generalizing intelligent information gathering algorithm: We has developed an intelligent information gathering algorithm that is applicable to a flight information service and needs to generalize it to be applicable to a wide range of application domains. In the current version, we categorize static and dynamic information formally, but actually even static information is updated in a long run. We need to make the system adaptive to the change of static information and that makes the system maintenance easy. We also need a mechanism to monitor the update of information sources distributed on the Internet.

Applying the information gathering method to discovered knowledge filtering: We apply the intelligent information gathering algorithm to filter knowledge discovered in hepatitis data mining which is a common domain of active data mining project. In gathering information that relates to discovered knowledge, it is required to efficiently gather latest information about hepatitis using a number of combinations of keywords. We keep a log of information retrieval and use it to improve the performance of information gathering considering the number of hits and frequency of information source updates. To improve the performance of our system, the medical knowledge is prerequisite and we ask an expert of medical informatics in our research group for advising.

Contact

Yasuhiko Kitamura, Graduate School of Engineering, Osaka City University
3-3—138 Sugimoto, Sumiyoshi-ku, Osaka 558-8585, JAPAN
E-mail: kitamura@info.eng.osaka-cu.ac.jp
Phone & FAX: +81-6-6605-3081

（A01-03-2）Information Gathering, Watching, and Integration by Multi-Agents

Investigator	Katsutoshi Hirayama	(Kobe University of Mercantile Marine)
Principal Investigator	Yasuhiko Kitamura	(Osaka City University)

Background and Aim

The Web is now widely used as a basic tool for broadcasting information through the world. Although information sources on the Web are built up by individual participant, they are tightly connected each other and form a world-wide database which has various kinds of information and does not maintained by administrators. It is, however, difficult for a user to obtain useful information on the Web efficiently not only because it has huge amount of information but also because it is frequently updated. In this work, we present a simple model for information gathering, watching, and integration on the Web by multi-agents and aim at developing a multi-agent system and task allocation protocols among agents in the system.

Research Plan and Approach

We present a model shown in Fig 1. This model is a typical multi-agent system which consists of information gathering agents and information integration agents. In this model, agents follow the procedure outlined below to gather, watch, and integrate information on the Web.

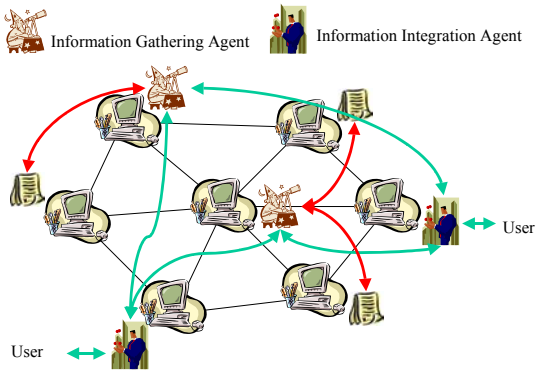


Fig 1: Model

(1) Request to watch some sources

Upon receiving a request from a user, an information integration agent generates CTAs (Conditional Task Allocations), which are pairs of a URL and an information gathering agent that watches the URL. We assume that an information integration agent can generate proper CTAs based on the abilities of information gathering agents in their solution qualities and costs. A CTA is conditional to capture a dynamic situation. Namely, when a situation is changed, an information integration agent can switch task allocation to information gathering agents by referring CTAs. With these CTAs, an information integration agent requests information gathering agents to watch some specified URLs.

(2) Watch the sources

Upon receiving a request from an information integration agent, an information gathering agent adds the URL to its schedule to watch the URL from now on.

(3) Notify of the update on the sources

If the information in the URL is updated, an information gathering agent notifies the information integration agent that requested to watch it.

(4) Execute in response to the update

An information integration agent executes some procedure (ex. Notify users, switch another URL, and so on) in response to the update notification.

Main Results

In this model, if we assume that each information integration agent asks information gathering agents to watch URLs independently, there may be a situation where watching tasks are concentrated on some agents. To avoid such a situation, we studied the following two methods.

Greedy method: upon arrival of task, allocate the task using dispatching rules while considering the current load of agents.

Dynamic reallocation method: upon arrival of task or situation changes, reallocate the overall tasks dynamically.

The key to success for the former method is obviously dispatching rules. On the other hand, for the latter method, we need to devise a mechanism for dynamic task reallocation. We currently investigate the latter method because we expect it is effective and it tackles with a problem with enough complexity. Our current achievement is as follows.

Problem formalization

We formalize task allocation problems as dynamic distributed constraint satisfaction problems (DyDisCSP) as shown in Fig 2. A DyDisCSP is a constraint satisfaction problem where variables and constraints are distributed among multiple agents. In this case, variables correspond to information gathering agents and variable domains correspond to a set of URLs. Constraints are CTAs along with the number of acceptable tasks for each information gathering agent. Those variables and constraints are distributed among information integration agents. Each information integration agent has to determine variable values (assign tasks to agents) so that all these constraints are satisfied. Note that in the task allocation problem some constraints are conditional, that means being labeled as under what condition they are valid, and the problems are thus modeled as DyDisCSP.

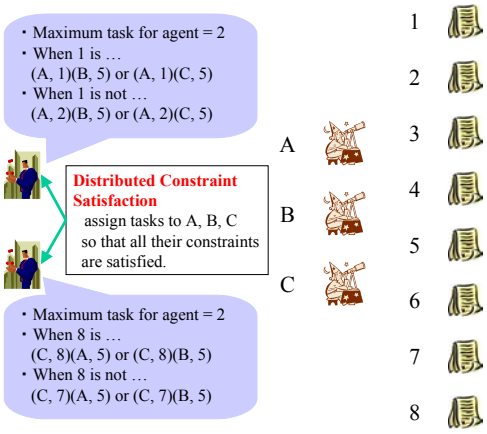


Fig 2: Formalization as DyDisCSP

Algorithm design

The above formalization enables us to apply a general distributed constraint satisfaction algorithm to the task allocation problem. However, since all the previous distributed constraint satisfaction algorithms are designed for static problems, they cannot handle problem dynamism (i.e., conditional constraints) successfully. We therefore re-formalized a DyDisCSP as a plain DisCSP including environmental variables (variables whose values cannot be changed by agents) by restating a conditional constraint as a set of environmental variables and a plain constraint. Also, we introduced a heuristic strategy where agents prefer stable variable values that would not violate constraints with environmental variables.

Future Plan and Expected Results

We are going to implement the dynamic reallocation method with the designed algorithm and make experimental evaluation on a simulator. For the greedy method, we are going to devise possible dispatching rules and make experimental evaluation on a simulator as well. Through these experimental evaluations, we will identify which method is the most appropriate for a real system and begin to develop a prototype information gathering, watching, and integration system by using the method.

Contact

Katsutoshi Hirayama（Investigator）

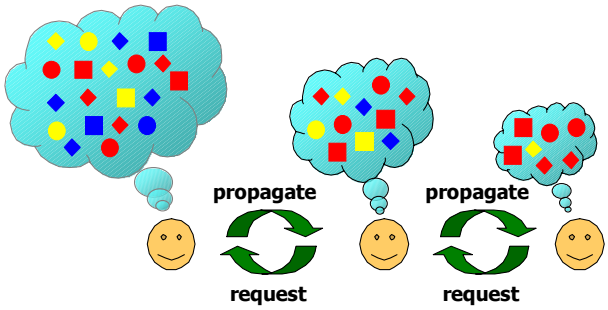
Kobe University of Mercantile Marine, 5-1-1 Fukaeminami-machi, Higashinada-ku, Kobe 658-0022, JAPAN

Email: hirayama@ti.kshosen.ac.jp　　Tel: +81-78-431-6262　　FAX：+81-78-431-6362

(A01-04-1)Chinese Whispers Approach for Information Gathering and Data Preprocessing

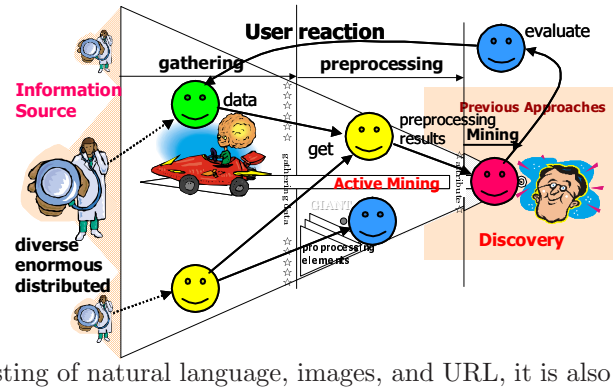
Principal Investigator	Masayuki Numao	(Tokyo Institute of Technology)
Investigator	Ryutaro Ichise	(National Institute of Informatics)
Collaborator	Yusuke Ito	(Tokyo Institute of Technology)
Collaborator	Cholwich Nattee	(Tokyo Institute of Technology)

Background and Aim



WWW and e-mail are very useful tools for communication. However, we sometimes feel uncomfortable because of flaming or mental barriers to participate in Computer-Mediated Communication (CMC). There are some important differences between CMC and direct communication.

Another problem is that computer networks deliver too many pieces of information, by which it is too hard to select useful pieces. Although search engines, such as *Yahoo*, *Goo* and *Google*, are very useful to find web pages, we need another type of tool without requiring a keyword for search. Good candidates are a mailing list and a network news system, where we need a filtering system to select only useful messages. Although content-based filtering and collaborative filtering are good solutions, the current methods have not achieved high precision and recall. This paper presents another approach by relaying a message like Chinese whispers to gather useful information, to alleviate mental barriers and to block flames.



The figures shows spread of information by word of mouth, where each person relays a message like Chinese whispers. Although a message is distorted by being passed around in the game, in a computer-assisted environment we expect that a delivered message is the same as its original. In such a process, we even have a merit that, as a result of evaluation and selection by each person, this process delivers only useful information. Each person knows whom (s)he should ask on a current topic, and retrieve a small amount that can be handled, where only interesting information survives. Even though this system is firstly designed for exchanging messages consisting of natural language, images, and URL, it is also useful for supporting data mining in exchanging data, preprocessing procedure and results, mining results, and evaluation.

Research Plan and Approach

We propose a system WAVE (Word-of-mouth-Assisting Virtual Environment) for supporting efficent information gathering and smooth communication on an electronic community. WAVE reproduces word-of-mouth communication on an electronic community in the same way as a human community. With a function as a distributed information gathering system, it forms a global information exchange network. Moreover, the user interface of WAVE is designed so that a user can seamlessly post, open to the public, browse, evaluate and retrieve information. With these characteristics, WAVE supports efficient information gathering and smooth communication better than the previous approaches.

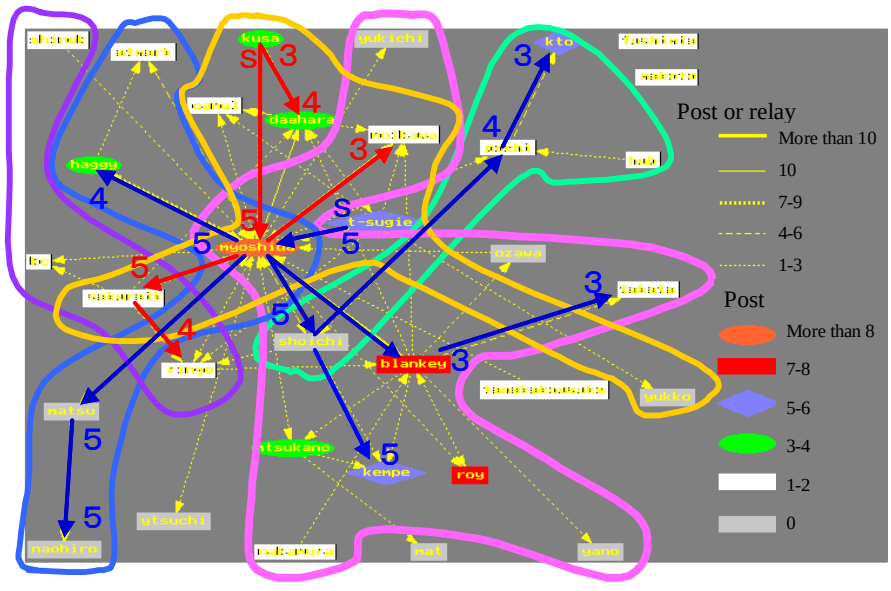
For example, WAVE supports a cooperated work for mining shared data by gathering around 10 research groups, each consists of a few laboratories from all over the country or all over the world. Each laboratory sets up a system separately. Then, local members discuss together, and other groups participate. Since it is annoying if a user from outside has to check all messages, (s)he may read only from the gatekeeper. As such, discussion in the

whole project and discussion among local members are separated. As the gatekeeper naturally appears in a human commmunity, WAVE supports his/her behavior.

As a scenario, data is firstly post on WAVE by the data provider, data then gradually spread in the community. Mining experts obtains the data and then try to preprocess them. Some additional information may be needed, then they may ask for the information or provide useful procedure for preprocessing by adding useful comments for other users. When proprocessing is complete, users discuss and evalute mining results, and better processing spreads. This approach for data mining is an important way of "Active Mining" as shown in the figure above.

Main Results

33 users test the system for 20 days. The result is visualized as shown in the map. This map is based on one by KrackPlot, which is a program for network visualization designed for social network analysts.



Each node denotes a user, whose shape denotes the number of messages (s)he posts. Here, myoshida, blankey, roy and t-sugie are opinion leaders that post many messages. A directed arc denotes that messages are retrieved and reviewed in that direction. Its thickness denotes the number of messages retrieved. In the network, we can see many triangles, each of which forms *triad* strongly connecting each other.

Two example flows of a message are shown in the figure. One flow is in red arrow. The other is in blue arrow. *S* denotes their origin. Each attached number denotes evaluation by each person. In most cases, the evaluation degrades as people relay a message.

Each island circled in the figure shows a community the authors observed, where people know each other in their real life. A message moves mainly in a community. Some people appear in multiple communities, and play a role of *gatekeeper*, who bridges information between communities.

Future Plan and Expected Results

In the current system, we also create the automatic recommendation by asking each user to give an evaluation to each message. The system then recommends a message by combining previous evaluations and user's preference. When applying this system to the data mining, other values can be useful for measuring the recommendation, such as precision and recall of the mining results. We plan to update the system to handle this kind of evaluation.

We also plan to attach a user-defined category to each message. Then, each message will firstly be classified by a user. Even if each user define his/her own category structure differently, some categories are linked to classy messages easily.

Contact

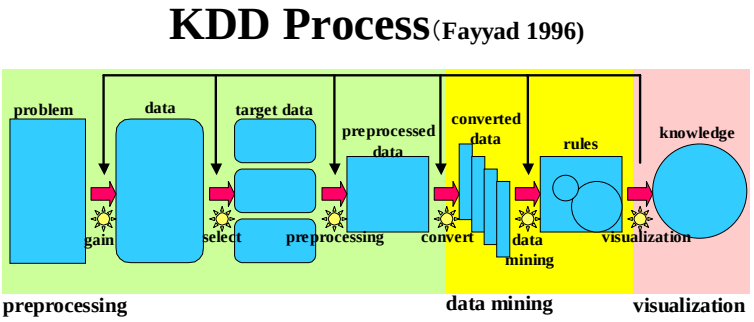
Masayuki Numao (Principal Investigator)
Graduate School of Information Science and Engineering, Tokyo Institute of Technology
2-12-1 Ookyama, Meguro, Tokyo 152-8552
E-mail: numao@cs.titech.ac.jp ; Tel: 03-5734-2684 ; Fax: 03-5734-2689

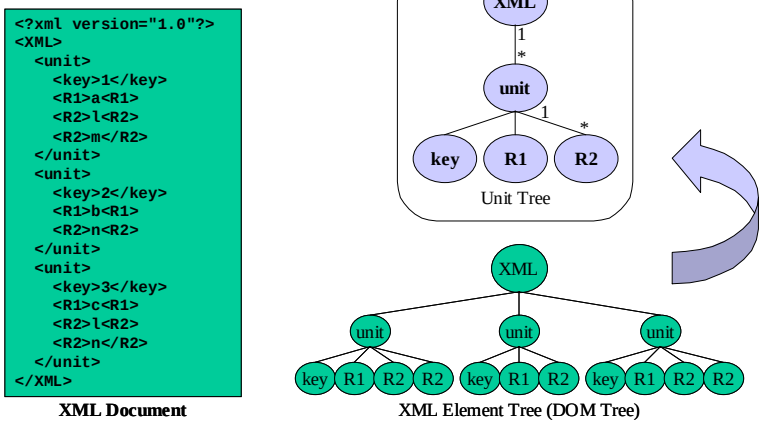
(A01-04-2) An XML-Based Tool for Data Preprocessing

Principal Investigator	Masayuki Numao	(Tokyo Institute of Technology)
Investigator	Ryutaro Ichise	(National Institute of Informatics)
Collaborator	Yukichi Yamada	(Tokyo Institute of Technology)

Background and Aim

There are many algorithms for data mining, such as association rules, decision trees, clustering, neural networks, and genetic algorithms. In order to apply these algorithms to large data sets, preprocessing is needed. The preprocessing consists of changing the structure of data or normalizing values. Therefore, due to the characteristics of data, we have to employ different processes, most of which are complicated and require enough experience. As a result, the cost for preprocessing amounts to 60% of the overall data mining process.

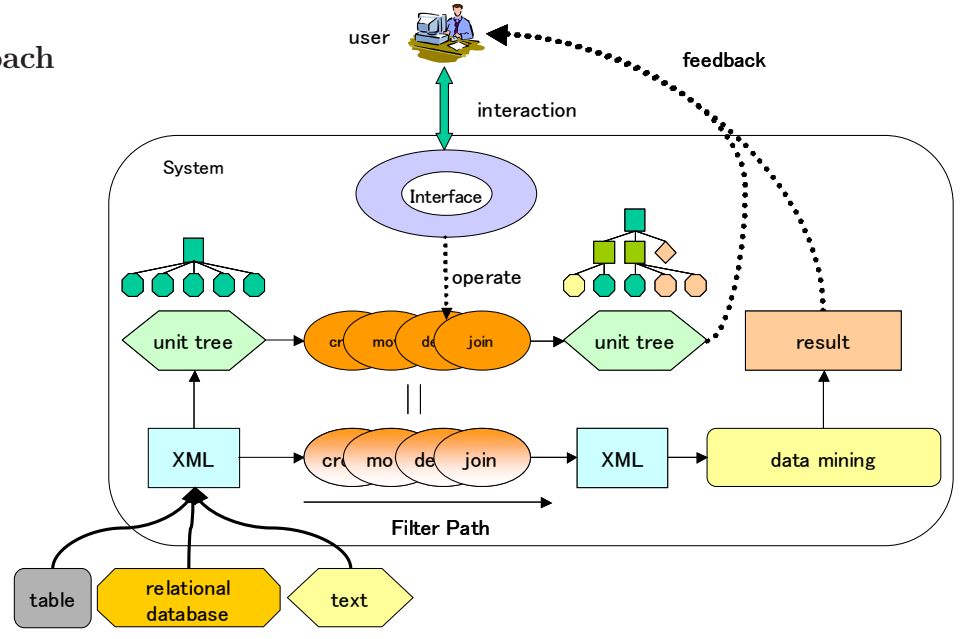




The input data of the analytic algorithms are normally flat table format but in rare cases other structures, such as a graph, are necessary. Therefore, nowaday tools for preprocessing usually support only data in table format or a relational database. With the relational database, large amount of data can be handled efficiently. However, when it is applied for preprocessing, some processes, such as creation and modification of relations between tables or generation and modification of a new column (attribute), require high costs. If a structure of data is not decided at the first step, it increases processing cost in the following steps. As such, the preprocessing frequently needs backtracking. The precise goal of preprocessing is not decided previously. It is usually found after we observe and analyze processed data. Since it is difficult to evaluate results, we have to process data in some different ways. Different analysis methods or objectives in preprocessing produce different results. Overall, we need a more powerful data structure and a backtracking mechanism in the preprocessing.

Research Plan and Approach

In this research, we precisely define and provide an approach to transform the accumulated data for the datamining algorithms. We propose an efficient data structure, an algorithm for automatic preprocessing and implement a practical system. All data in preprocessing are stored in XML format. The system visualizes the data structure when a user transforms data. It also automatically suggests a filter to be applied.



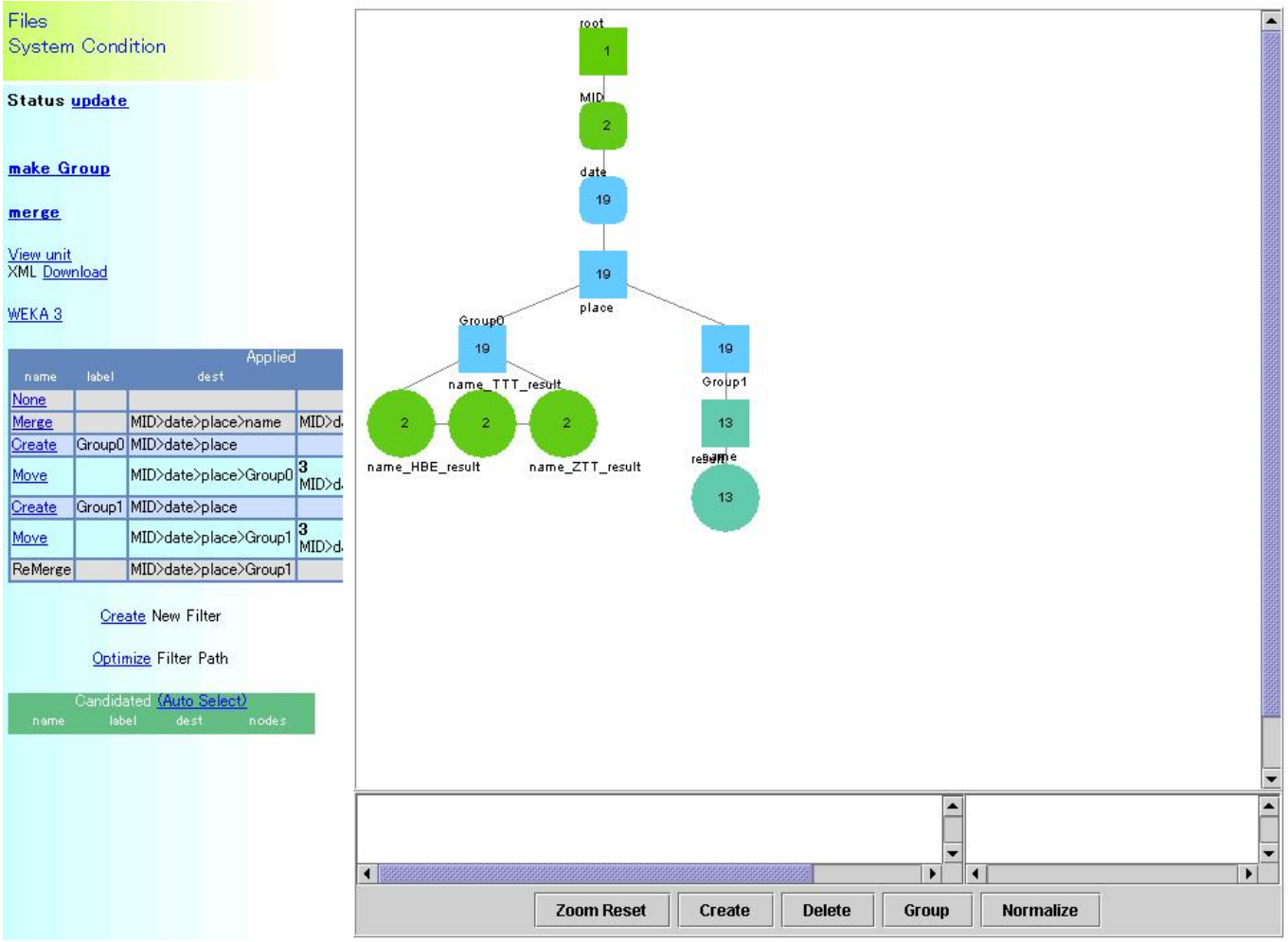
The proposed system processes only an XML format since XML has ability to represent most of general data formats. XML is recently a standard for data exchange on the Web. In order to cover a wide range of preprocessing processes, the

system has to support various levels of data, such as characters, attributes, or tuples. XML is flexible and powerful to manage data in the context structure, although it requires a long computation time.

From a distinguishing characteristic that XML represents data in the overall data mining processes, the proposed system transforms all the data into XML format. By using structure transformation, it preprocesses data more efficiently.

Main Results

With an experiment in manually preprocessing clinical data and constructing a decision tree, we found that a user works more efficiently on our XML-Based system called *TransX* than on database applications while the preprocessing procedures and the structure transformation. The following is a screen shot of TransX.



Future Plan and Expected Results

The preprocessing in data mining by using XML transformation is useful and possibly turned to practical use. In order to gain more practicality, we need an XML tool that is robust on a large amount of data, and conforms to the complete XML definition. For the future works, we plan to study an approach that utilizes more features of XML, an application of XML query language for transformation, and integration with the datamining algorithms. We expect to complete a *constructive adaptive user inteface* for data preprocessing and mining at the end of the project.

Contact:

Masayuki Numao (Principal Investigator)
Graduate School of Information Science and Engineering, Tokyo Institute of Technology
2-12-1 Ookyama, Meguro, Tokyo 152-8552
E-mail: numao@cs.titech.ac.jp ; Tel: 03-5734-2684 ; Fax: 03-5734-2689

(A01-04-3) Temporal Spatial Data Structures for Person Trip Data Analysis

Investigator Hiroyuki Kawano (Kyoto University)

Motivation

Technologies of GIS (Geographic Information System) and location service are becoming popular, and huge volume of spatial and geographic data are stored into the various data warehouses. We focus on the data from positioning systems and sensoring systems of ITS, which affect the quality and quantity of traffic planning/management systems. Then, new advanced temporal spatial queries and data structures are required, in order to analyze and mine the actual data in a traffic data warehouse effectively.

In this report, from the view point of traffic engineering, we discuss characteristics of temporal spatial queries. Next, we consider some basic problems in order to execute data mining algorithms in huge volume of actual temporal spatial database. We proposed a temporal spatial data structure, $\sum$-tree, which is based on the techniques of advanced spatial indices and the data cube.

Trip Data and Temporal Spatial Queries

In GIS and spatial database systems, many database researchers have developed various data processing technologies in order to execute typical spatial queries. Based on those previous researches, we can search spatial objects in a very large scale of spatial databases effectively. However, in order to analyze characteristics of traffic flows from the view point of traffic management and analysis, we have to pay attention to the object behavior along time axis, and we need much more advanced and complex spatial temporal queries.

In this section, we introduce typical queries for person trip data analysis. Here, we use definitions of time series, t_1, $t_2 (t_1 < t_2)$, and regions, R_1, R_2, and we discuss three following temporal spatial queries.

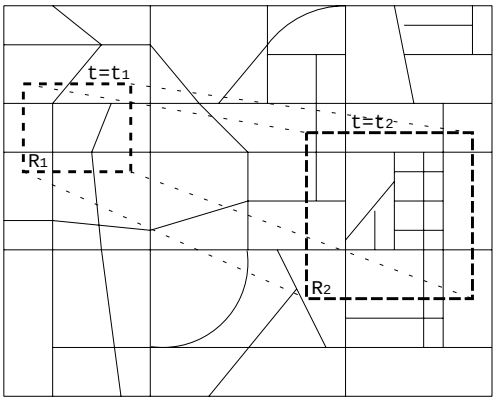


Figure 1: Example of moving query

1. **timeslice query** $Q_{ts} = (R_1, t_1)$: At time point t, objects are searched for in a region R_1.

   (ex.) Based on the results of a query, we can calculate typical traffic flow parameters, such as traffic density, average traffic velocity and others.

2. **window query** $Q_{win} = (R_1, t_1, t_2)$: shows that Moving objects are searched for in the region R_1 from t_1 to t_2.

   (ex.) By using the results of window queries, we can calculate time average velocity which has rather stable property in traffic analysis.

3. **moving query** $Q_{mov} = (R_1, R_2, t_1, t_2)$: In Fig. 1, we search the objects in a moving region, which is covered by a connecting trapezoid of (R_1, t_1) and (R_2, t_2).

   (ex.) Traffic density of moving objects on an expressways is calculated, then the traffic congestion is forecasted by using the density and other specific values.

High Dimensional Data Structures for Trip Data Analysis

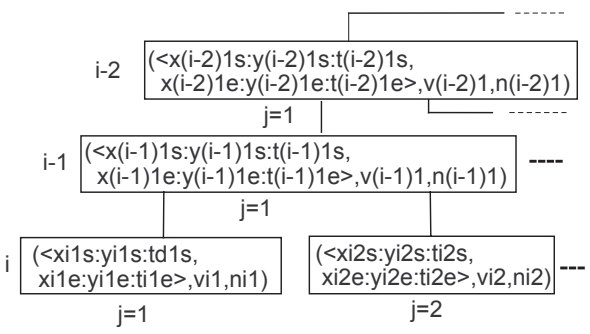


Figure 2: Preprocessed attribute values in $\sum$-tree

In the research fields of data mining, a lot of algorithms to discover knowledge in the huge volume of databases are proposed. Many spatial data mining algorithms have been also proposed, these algorithms can derive useful and meaningful patterns, trends, rules and knowledge from spatial and geographical data. For example, in order to make clusters effectively, clustering algorithms make full use of the spatial characteristics, such as density, continuity and so on. We also focused on effective clustering algorithms based on the spatial index technologies, such as R-Tree, R*-Tree, PR-Quadtree and others. Furthermore, we try to reduce computing cost of range queries based on the technologies of data cube.

In this work, we proposed the temporal spatial data structures, $\sum$-tree, which is based on the advance techniques of temporal spatial indices and data cube.

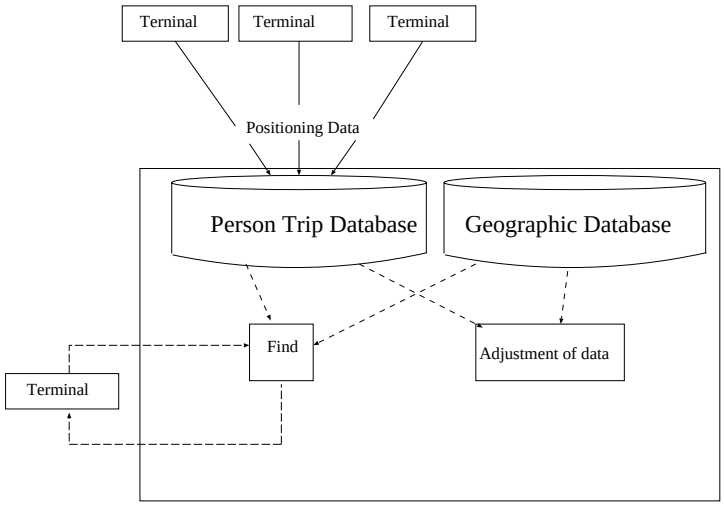


Figure 3: Architecture of person trip data collecting system

There exist n objects (x, y, t) with temporal spatial attributes, these data (x_1, y_1, t_1), $(x_2, y_2, t_2), \cdots, (x_n, y_n, t_n)$ are stored in $\sum$-tree.

In order to construct temporal spatial data structures for person trip analysis, we focus on some restrictions of spatial objects with velocity and other spatial properties. For example, we can divide objects into clusters based on the moving direction and the traffic lane, the following parent nodes $L_1, L_2, \cdots, L_N$ are constructed.

$$\begin{cases} (x_1, y_1, t_1), (x_2, y_2, t_2), \cdots, (x_i, y_i, t_i) \subset L_1 \\ (x_{i+1}, y_{i+1}, t_{i+1}), \cdots, (x_j, y_j, t_j) \subset L_2 \\ \\ (x_{k+1}, y_{k+1}, t_{k+1}), \cdots, (x_n, y_n, t_n) \subset L_N \end{cases}$$

In addition to the fundamental attributes, we can store other temporal spatial values, such as objects velocities, (x_1, y_1, t_1, v_1), $(x_2, y_2, t_2, v_2), \cdots, (x_n, y_n, t_n, v_n)$ in leaf nodes and sum of these values, $V_{L_1}, V_{L_2}, \cdots, V_{L_N}$ in the parent node.

$$V_{L_1} = \sum_{l=1}^{i} v_l,\ V_{L_2} = \sum_{l=i+1}^{j} v_l,\ \cdots, V_{L_N} = \sum_{l=k+1}^{n} v_l$$

Next, we focus on the computing cost of OLAP (On-Line Analytical Processing) in our traffic data warehouse. In order to analyze various traffic parameters of person trip data, we can store several attribute values including the number of objects n_{ij}, sum of velocities V_{ij} and so on.

We define the node space $(< x_{ijs} : y_{ijs} : t_{ijs}, x_{ije} : y_{ije} : t_{ije} >)$ using a start point $(x_{ijs}, y_{ijs}, t_{ijs})$, and an end point $(x_{ije}, y_{ije}, t_{ije})$). We store temporal spatial attribute values $(x, y, t) \subset L_u$, sum of various traffic parameters (sum of speed V_{L_u}, total number of cars N_{L_u}, and so on) in a node L_u. The sum of attribute values of child nodes are stored into the parent node recursively.

Using $\sum$-tree data structures shown in Fig. 2, we can reduce the range of a search region and the number of objects effectively in order to execute temporal spatial queries.

Future Work

In our previous works, in order to convert from raw positioning data to person trip data (in Fig. 4), we proposed the system architecture for person trip data analysis in Fig. 3 In the next stage, we discuss how to derive and mine useful patterns of person trip routes with small computing cost. In near future, we try to apply our proposed analytical and spatial mining techniques in actual traffic data warehouse in order to discover complex rules and characteristics.

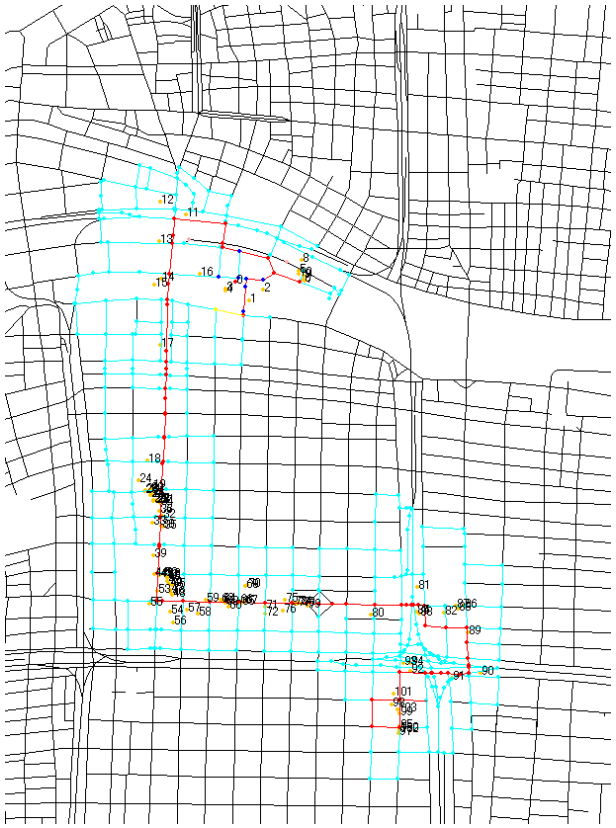


Figure 4: From location data to trip data

Contact:

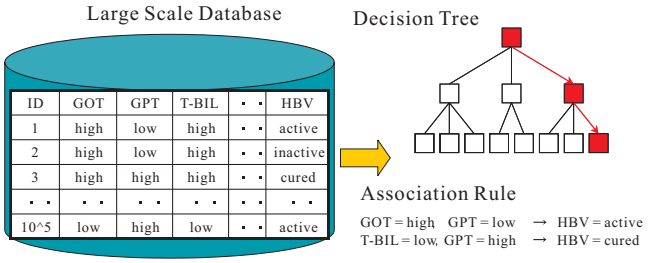
Hiroyuki Kawano (Investigator)
Graduate School of Informatics,
Department of Systems Science, Kyoto University,
Yoshida Hommachi, Sakyo, Kyoto 606-8501, JAPAN
Email: kawano@i.kyoto-u.ac.jp; Tel: 075-753-5493 ; Fax: 075-753-3358

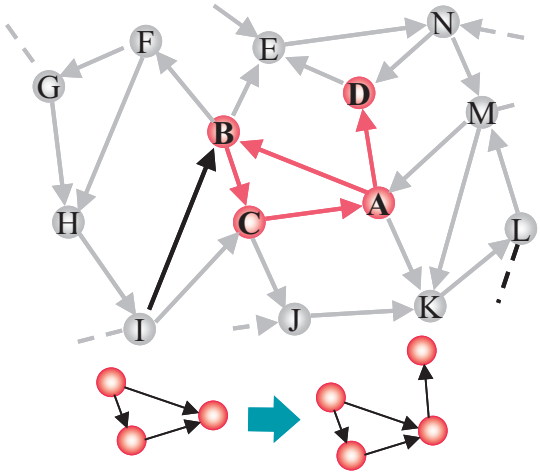
(A02-05-1) Discovery of Typical Patterns from Structured Data by Graph-based Induction

Principal Investigator	Hiroshi Motoda	(I.S.I.R., Osaka University)
Investigator	Takashi Washio	(I.S.I.R., Osaka University)
	Tetsuya Yoshida	(I.S.I.R., Osaka University)
Collaborator	Takashi Matsuda	(I.S.I.R., Osaka University)

Background and Aim

The main stream research on data mining often assumes that data is represented by a set of attribute-value pairs and focuses on developing mining algorithms based on the current database technologies. The relashionship between the target class and the values of attributes is often represented as decision trees, classification rules, or association rules (right figure). Various ideas to minimize disk scanning and precalculate various statistics have been proposed to scale up mining algorithms for large scale databases.





Extracted Rule from Graph Structured data

On the other hand, many real world datasets e.g., WWW network data, chemical data, DNA sequences, patient history data, etc., have relationships between the items in data. Such relationships form the overall structure in the dataset. Mining structured data is a new challenge since it is difficult to apply any standard mining algorithms to structured data without significant modifications. Structure in data is represented as graphs in our approach (left figure) and we aim at:

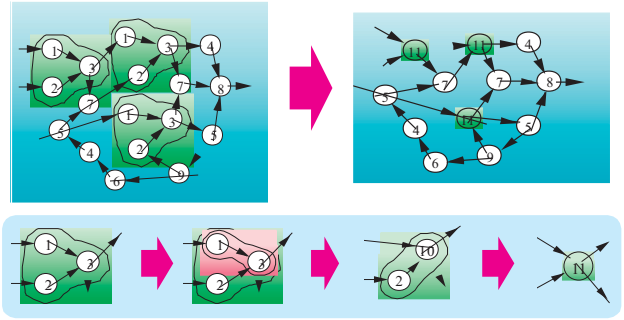
- Developing a fast algorithm for mining typical patterns from a structured data,
- Discovering task-dependent structures (patterns) by applying the developed algorithm to various real-world data by closely working with domain experts, especially, to the hepatitis data chosen as a common test bed.

Research Plan and Approach

Since the problem of finding all isomorphic subgraphs is known to be NP-complete (upper right figure), it is difficult to discover all the subgraphs (patterns) efficiently. We need to devise a way to discover typical patterns from large scale graphs in an admissible computation time. For the first two years we plan to extend the Graph-based Induction (GBI) method, which can approximately extract typical pattens from graph structured data in $O(N^2)$ (N: the number of nodes in a graph) based on frequency measure. GBI employs the stepwise pair expansion (pairwise chunking) principle. Although its greedy search strategy incurs the incompleteness of search, complex patterns can still be discovered by pairwise chunking (lower right figure).

Example of Isomorphic Graphs

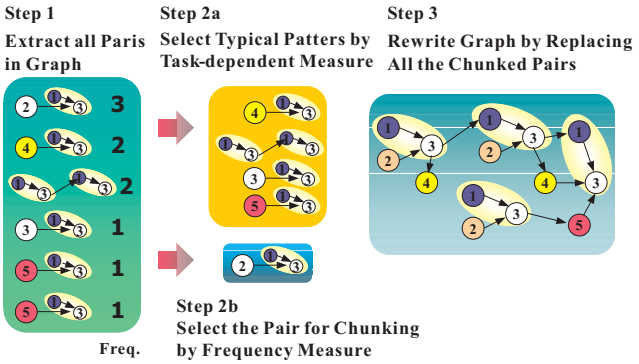
Basic Operations in GBI

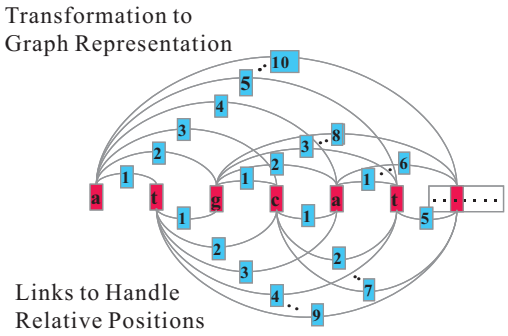


GBI is a general mining method since it is applicable to general graph structured data. However, always utilizing frequency measure for pairwise chunking can hinder effective discovery of typical patterns since "typicality" depends on tasks and domains. We plan to clarify this issue of GBI by applying it to real world data and propose a solution. Furthermore, going a step further beyond developing a mining method for graph data, we plan to work very closely with domain experts to get their feedback on how to transforme data into graph most effectively and how to show the discovered pattern in most interpretable way. The feedback from domain experts will facilitate the development of mining algorithm for task-dependent typical patterns from graph structured data.

Main Results

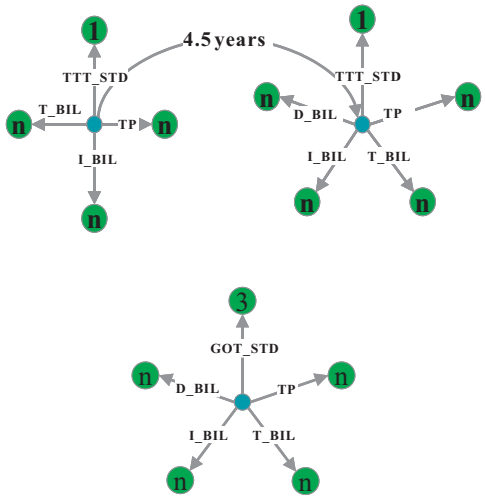
We clarified three problems in GBI, namely, 1) vagueness of the meaning of discovered patterns, 2) unknowability of undiscovered good patterns due to its greedy search strategy, 3) miscounting of isomorphic patterns. The proposed solutions for these are: 1) discremination of frequency measure for chunking from the measure for discovering typical patterns, 2) incorporation of beam search to expand search space, 3) adoption of canonical labeling for accurate enumeration of identical pattens. The new algorithm, now called Beam-wise GBI, B-GBI for short, is proposed.





B-GBI is applied to the promoter dataset in UCI Machine Learning Repository to extract patterns suitable for classifying DNA nucleotides sequences. Standard classification rule learning methods require the sequence to be alignied at a reference point (nucleotide). Graph representation solves this problem since the relative position between two nucleotides can be naturally represented as a link (left figure). The extracted patterns are treated as binary attributes of each sequence to build a classifier by C4.5Rule. The best prediction error rate is 2.8% (with LVO), which is equivalent to the result 3.8% (with 10-CV) obtained by KBANN using M-of-N expression. (Note: Fair comparison is difficult since the experimental conditions are different.) It is worth mentioning that C4.5Rule combined with B-GBI that does not use any domain knowledge induced good prediction rules (KBANN used domain knowledge).

Intensive analysis of the hepatitis dataset, which is provided by Chiba University Hospital, was carried out by collaborating with a domain expert (medical doctor). The expert defined a new index for the activity of hepatitis B virus based on our preprocessing of the dataset. By focusing on GOT, GPT, TTT, ZTT, which are short-term indices for the status of liver, their standard deviations over 6 months are added to the dataset as new attributes to indicate short-term (6 month) changes. B-GBI is then applied to the dataset to extract typical patterns for the activity of hepatitis B virus. Some of the extracted patterns are confirmed reasonable by the expert, e.g. one example: a pattern is consistent with the known medical knowledge that the change of TTT is small when the hepatitis B virus is inactive (upper right figure). Some others are evaluated as unexpected or surprising, e.g., one example: a pattern in which the change of GOT is large when the hepatitis B virus is inactive (lower right figure). Currently we have not been able to fully incorporate the feedback on the extracted patterns from the expert into the analysis. However, through the discussion on the extracted patterns with the expert, we gained knowledge to do better analysis based on medically grounded processing for the transformation of data into graph representation.



Future Plan and Expected Results

We improve B-GBI based on its quantitative evaluation. We continue to evaluate B-GBI with respect to the number of extracted patterns and computation time. We investigate more effective measure for discovery of typical patterns by evaluating the quality of extracted patterns in terms of classification accuracy. By focusing on hepatits data, we challenge the discovery of surprising new knowledge (patterns) through the close collaboration with domain experts. As an immediate future work we continue the analysis on hepatitis B by limiting the time span of links up to 2 years on both sides when transforming the data into graph expression and by discretizing the medical examination values in a more proper way. Discovery of typical patterns for the effect of interferon therapy on hepatitis C is also tackled.

Contact:

Hiroshi Motoda（Principal Investigator）
I.S.I.R., Osaka University, 8-1 Mihogaoka, Ibaraki, Osaka 567-0047, JAPAN
Email：motoda@ar.sanken.osaka-u.ac.jp ; Tel：06-6879-8540 ; Fax：06-6879-8544

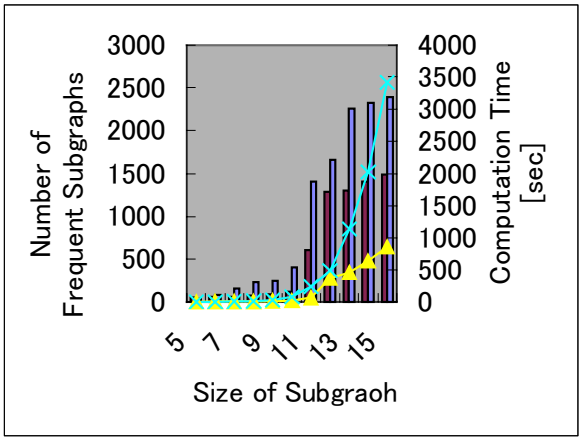
(A02-05-2) A Fast Algorithm of Frequent Subgraph Extraction Method: AGM

Investigator	Takashi Washio	(I.S.I.R.,OsakaUniversity)
	Tetsuya Yoshida	(I.S.I.R.,OsakaUniversity)
Principal Investigator	Hiroshi Motoda	(I.S.I.R.,OsakaUniversity)
Collaborator	Yoshio Nishimura	(I.S.I.R.,OsakaUniversity)

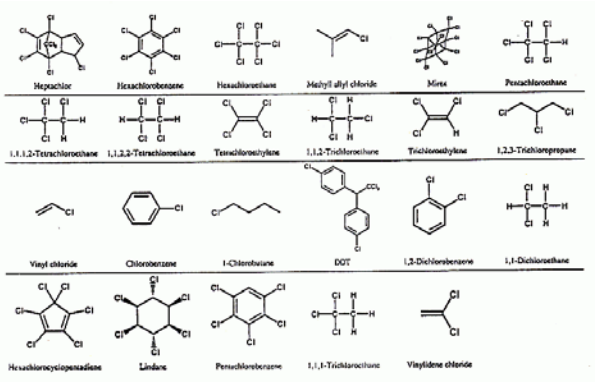
Background and Aim

The number of needs to discover important knowledge from massive graph structured data is recently increasing in various domains, e.g., analysis of relation between molecule structure of a chemical compound and its features and analysis of web link structure. Upon this background, we have been working on the development of AGM algorithm which extracts frequent subgraph patterns embedded in graph structured. This searches all subgraphs appearing in a set of massive graph data in complete fashion. For instance, the chemical descriptors consisting of molecule substructures to characterize the mutagenicity have been proposed by the knowledge and the experience of chemists. On the other hand, our algorithm enables to discover novel descriptors by applying complete search of the characteristic subpatterns included in the mutagenic chemical moledules(see the right figure). This research will provide new and effective means in the fields of medicine development, nanotechnology and new IT network development.

Relations between the size of subgraph and computation tim/number of patterns



Examples of mutagenic molecules



However, massive number of isomorphism subgraph problems must be handled in the complete search of frequent subgraph structures. This problem is mathematically known to be NP-complete where the number of computational operation required increases more than the polynomial order in terms of the size of the subgrahs to be searched. The former algorithm has a difficulty on the computational speed where it takes several weeks to derive larger size of descriptors (subgraphs) (see the left figure). Acoordingly, we will work on

- Proposal of a new constraints to accelerate the computational speed of AGM,
- While maintaining the completeness of its search.

In addition, under the review of the results obtained through our approach by some chemists, the appropriateness and the novelty of the derived descriptors is evaluated. Though out this evaluation, we confirm the practicality of our proposing approach.

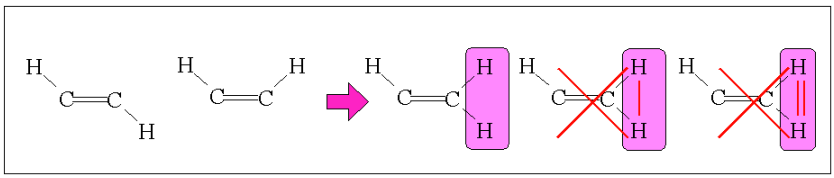
Research Plan and Approach

AGM algorithm searches a frequently appearing node in the graph data set at first. Starting from the stage, larger size of frequent subgraphs are discovered through the combination the smaller size of the frequent subgraphs which have been already discovered. For instance, any subgraph containing Nitrogen atom (N) which frequency is more than a certain level does not exist, if the frequency of the appearance of N is less than the level. Similarly, any frequent subgraph containing C-Cl does not exists, if C-Cl is not frequent. Based on this principle, AGM efficiently generates the candidate frequent subgraphs of size k+1 by combining two frequent subgraphs of size k under a certain threshold of

frequency. The number of the candidates is determined by the number of types of bonds between two nodes which are not shared by the original two frequent subgraphs. Once the candidates are generated, each frequency is counted though the check in the graph data set, and only the frequent ones are retained. Therefore, the key issue to accelerate the computation of the frequent subgraph discovery is the efficient reduction of the number of the candidates generated in the above procedure. In this work, we propose a novel and efficient reduction approach of the candidates, and apply to the discovery of important descriptors for mutagenic chemical molecule structures.

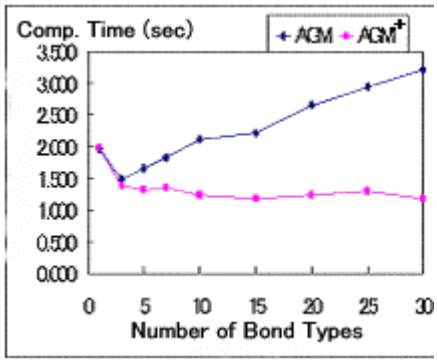
Main Results

Candidates having Size 5 are generated from two frequent subgraphs of size 4.

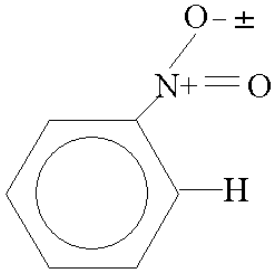


The right figure represents the scheme of the candidate generation of AGM. It combines frequent subgraphs sharing structures except one node for each. The variety of candidates caused by the bond types connecting the unshared nodes in the two original frequent subgraphs. In addition, a graph where the unshared nodes are not connected by any bond is also a candidate. IN our former method, all candidates generated in this method are applied to the frequency check in the graph data set. In contrast, our new proposing method checks if the newly generated pair consisting of the two unshared nodes and the bond has been discovered as a frequent subgraph. For example, if H-H and H=H are known to be unfrequent in the past search, the generated candidates including these pairs are removed. In this manner, the frequent subgraph mining is accelerated by checking if the newly generated node pair is a frequent subgraph, since this reduces the required frequency counts of the candidates (see the right figure). We applied this new approach to the mutagenicity data set. Because larger size of candidate descriptors became to be derived, some descriptors which have not known by the chemists were discovered. The right figure is a such example. The minus electric charge is kept at the base NO_2, whereas the plus charge is kept by the neighbor hydrogen. Thus, they are attracted one another, and the conformation of NO_2 base remains within the plain of benzene ring. This flat formation makes this molecule easily slide into the double helix of RNA, and give some damage on it. On the other hand, when the neighbor is the other atom or base which is charged by minus electron, they mutually repel, and NO_2 base does not remain within the benzene plain. This prevents the sliding of the molecule into the RNA helix, and reduces its mutagenicity.

The new AGM+ runs faster.



A generated descriptor of mutanenicity selected by chemits.



Future Plan and Expected Results

As the frequent subgraph patterns interested in many fields are connected subgraphs, we are going to develop fast algorithm to mine frequent connected subgraphs from massive graph data. This development is expected to provide more efficient and practical means and results in real world applications.

Contact:

Takashi Washio（Investigator）
I.S.I.R.,OsakaUniversity,8-1Mihogaoka,Ibaraki,Osaka567-0047,JAPAN
Email：washio@ar.sanken.osaka-u.ac.jp;Tel：06-6879-8541;Fax：06-6879-8544

(A02-05-3) A User-Centered Approach to Data Mining

Investigator	Tu Bao Ho	(Japan Advanced Institute of Science and Technology)
Collaborators	Trong Dung Nguyen	(Japan Advanced Institute of Science and Technology)
	Duc Dung Nguyen	(Japan Advanced Institute of Science and Technology)
	Saori Kawasaki	(Japan Advanced Institute of Science and Technology)

Background and Aim

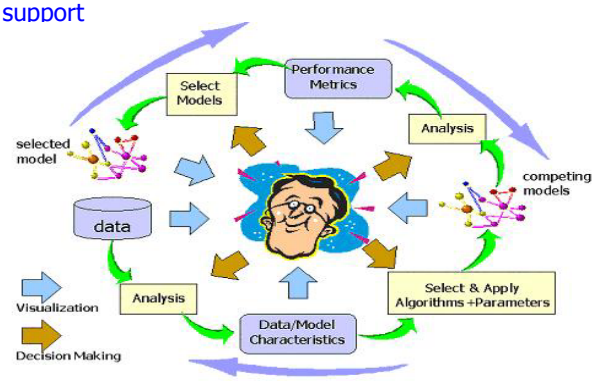
The research is motivated by the complex nature of the knowledge discovery process and the crucial role of the user in this process.

– The KDD process requires going through several steps, each step concerns with various tasks, and each task can be done by different algorithms. There are too many choices that the user often does not know which one is appropriate.
– Interesting models are hard to obtain as interestingness depends on multiple measures that are not only model validity or simplicity but also novelty or usefulness that can only be judged by the user.

This work aims to develop data mining methods and tools that support an active participation of the user in the KDD process in order to obtain expected results.

Research Plan and Approach

The user plays a central role in data mining with system support

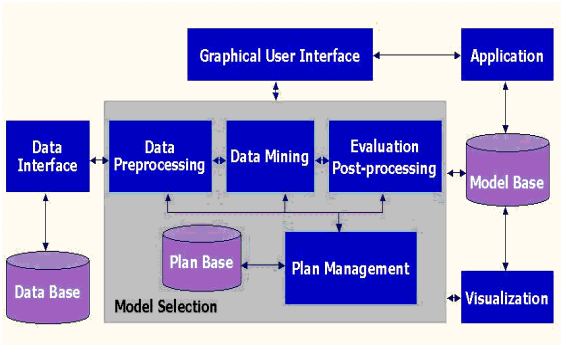


The essence of this research is model selection in KDD. Our work focuses on a *user-centered approach* to model selection. The key idea is to emphasize and facilitate the active participation of the user in the KDD process with support for a model evaluations that is combined of

– A *quantitative evaluation*: obtained by using model characteristics and performance metrics provided by data mining programs;
– A *qualitative evaluation*: obtained by the user with the support of visualization tools.

This approach is realized by a new conceptual architecture, and methods/algorithms for visualization and model characteristics.

The agent-based architecture of our system focuses on model selection and visualization, and emphasizes the cooperative mechanisms: the modules are constructed in a way that can cooperate and communicate with other modules to achieve their goals.



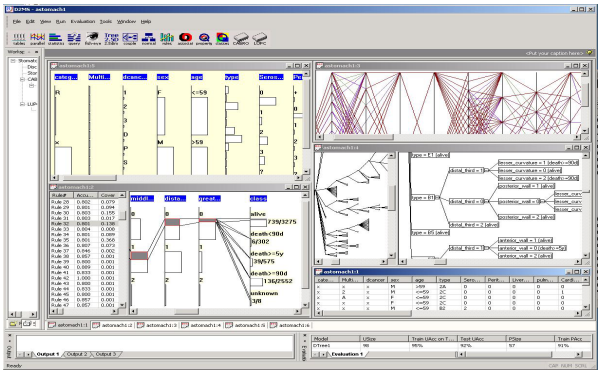
Main Results

System D2MS

The output of our research is *the system D2MS (Data Mining with Model Selection)*. D2MS supports the user participation in knowledge discovery with functions of:

– learning rules/decision trees from large datasets in accordance with the steps of the KDD process.
– selecting appropriate models with visualization support.

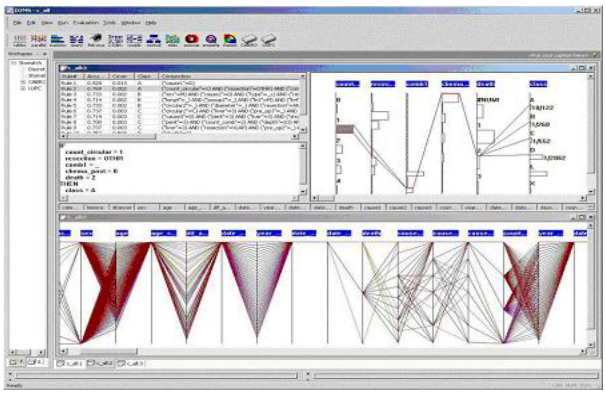
Synergistic view of data and knowledge in D2MS



Model Selection in Knowledge Discovery

– The user-centered approach provides the user with ability to easily select and combine a wide range of mining algorithms.
– By synergistically visualizing data and knowledge the user can actively select, create, and evaluate discovered models.

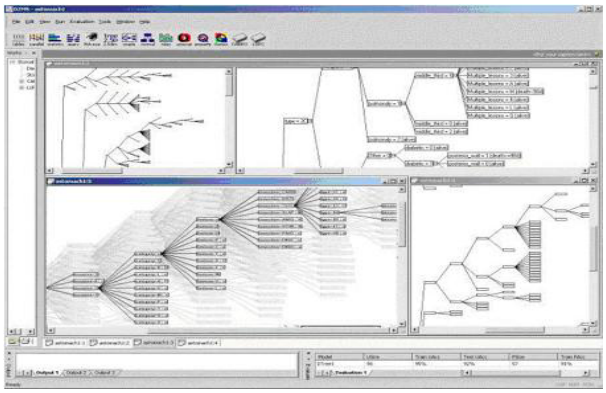
Data and rule visualization



High-Performance Data Mining Methods

– CABRO: Mining decision trees by scalable inductive algorithms;
– LUPC: Mining rules for minority classes with exclusive and inclusive constraints.

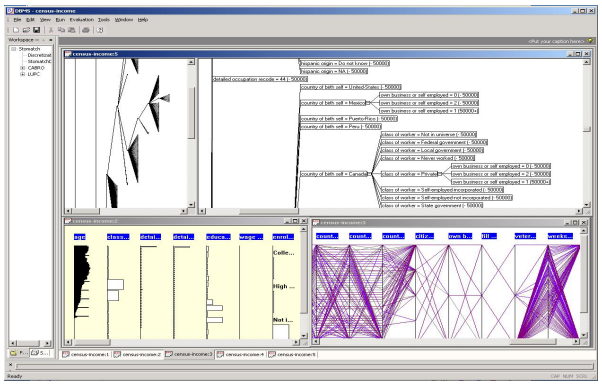
Multiple views of trees: tightly-coupled, fish-eye, T2.5D



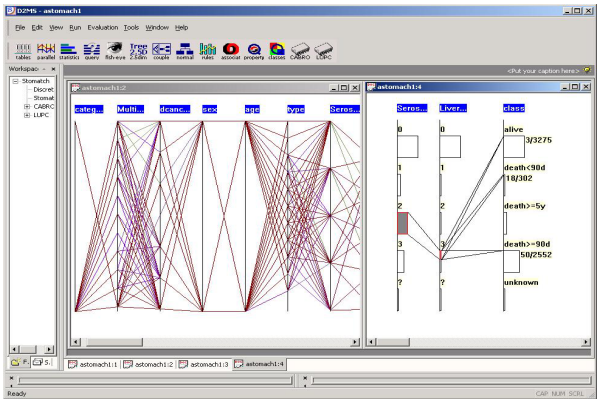
D2MS visualization tools

– Data visualizer: parallel coordinates, data summaries, and querying data.
– Rule visualizer: each rule and its covered instances, rules discovered for a class, rules related to a query.
– Tree visualizer: fish-eye view, tightly-coupled views, and T2.5D (Tree 2.5 Dimensions).

Ho, T.B., Nguyen, T.D., Nguyen, D.D., Kawasaki, S., "Visualization Support for User-Centered Model Selection in Knowledge Discovery and Data Mining", International Journal of Artificial Intelligence Tools, Vol. 10 (2001), No. 4, 691-713.



The user can detect significant hypotheses using visualizers that impose constraints in mining. For example, the detection of the symptom "metastasis" in the "alive" class of stomach cancer suggests finding rules such as "IF sex = M & type = B1 & liver_metastasis = 3 & middle_third = 1 THEN class = alive".



Future Plan and Expected Results

– To fully build and enrich D2MS with more advanced features, and with other techniques of preprocessing, post-processing, and data mining.
– To find efficient ways of integrating the user's domain knowledge into the KDD process.
– To validate and improve the effectiveness of the approach through its usage in real-life applications, in particular mining medical databases with the participation of domain experts.

Contact: Ho Tu Bao, Japan Advanced Institute of Science and Technology, Tatsunokuchi, Ishikawa, 923-1292 Japan; Phone & Fax: 81-761-51-1730, http://www.jaist.ac.jp/ks/labs/ho.

(A02-05-4) Text Mining with Tolerance Rough Set Models

Investigator	Tu Bao Ho	(Japan Advanced Institute of Science and Technology)
Collaborators	Saori Kawasaki	(Japan Advanced Institute of Science and Technology)
	Ngoc Binh Nguyen	(Hanoi University of Technology)

Background and Aim

The research is motivated by the need of suitable models for representing documents in textual data mining. It concerns with the crucial issue of improving efficiency and effectiveness in text processing and information retrieval in the following ways:

– Using inexact research with the model to improve effectiveness;

– Using cluster-based research with the model to improve efficiency.

This work aims to develop a representation model for textual documents as well clustering algorithms, cluster-based information retrieval, and other text mining methods using this model.

Research Plan and Approach

Tolerance Rough Set Model

- Tolerance rough set model TRSM, $\mathcal{R}=(\mathcal{T}, I, \nu, P)$ aims to approximate sets of index terms in spaces of **tolerance classes** determined by **tolerance relations** with reflexive and symmetric properties.
- Tolerance class of terms w.r.t. threshold $\theta > 0$

$$I_\theta(t_i)=\{t_j \mid f_{\mathcal{D}}(t_i,t_j)\geq\theta\}\cup\{t_i\}$$

$$f_{\mathcal{D}}(t_i,t_j)=\{d\in\mathcal{D}\mid t_i\in d\wedge t_j\in d\}$$

- Vague inclusion function

$$\nu(X,Y) = |X\cap Y| \;/\; |X|$$

- Tolerance lower and upper approximations of $X\subseteq\mathcal{D}$

$$\mathcal{L}(\mathcal{R},X)=\{t_i\in\mathcal{T}\mid\nu(I_\theta(t_i),X)=1\}$$

$$\mathcal{U}(\mathcal{R},X)=\{t_i\in\mathcal{T}\mid\nu(I_\theta(t_i),X)>0\}$$

We use rough set theory—introduced by Pawlak in early 1980s—as a mathematical tool to represent documents. However, instead of using the original *equivalence rough set models ERSM* based on using equivalence relations (with reflexive, symmetric, and transitive properties), we use *tolerance rough set models TRSM* using *tolerance relations* (with reflexive and symmetric properties) that we shown to be more suitable for representing textual documents.

Our research plan consists of:

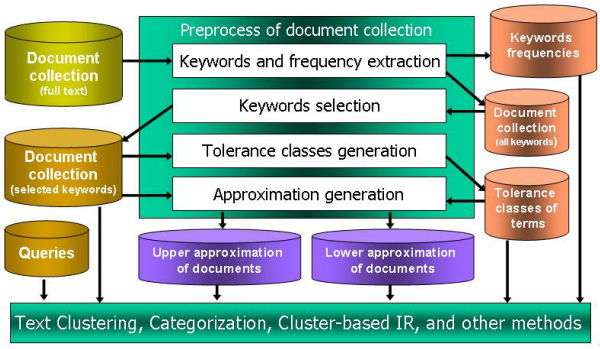
– Development of tolerance rough set models TRSM;

– Finding TRSM clustering methods;

– Development of techniques for TRSM cluster-based information retrieval, as well other tasks in text mining.

Main Results

Framework of the TRSM text mining system



Tolerance Rough Set Models

The above figure presents the framework of TRSM and TRSM-based text mining methods. A TRSM basically requires us to construct:

– A tolerance class for each word in the document database;

– The upper and lower approximations of each document in the document database.

Hierarchical and Non-Hierarchical Clustering

The research concerns with

– The TRSM representatives of clusters;

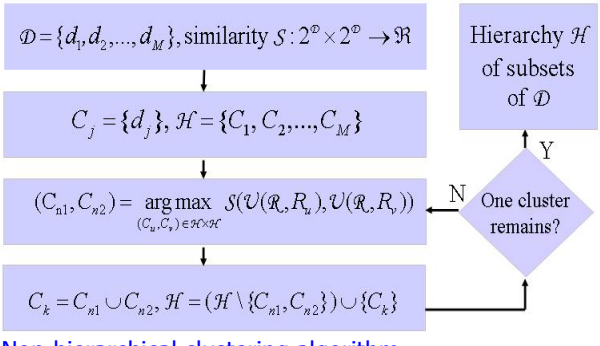
– TRSM similarities for documents and clusters;

– Hierarchical clustering algorithms;

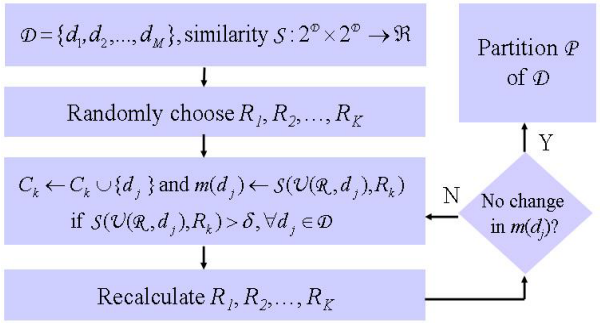
– Non-hierarchical clustering algorithms;

– Evaluations of obtained results.

Hierarchical clustering algorithm



Non-hierarchical clustering algorithm

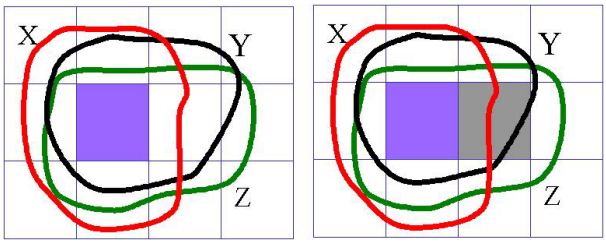


TRSM Information Retrieval

This work consist of development of

– TRSM extension of the inclusion notion;
– Tolerance rough matching algorithms using rough equality and rough inclusions;
– Ranking function and secondary ranking.

TRSM equality and inclusion



TRSM cluster-based information retrieval

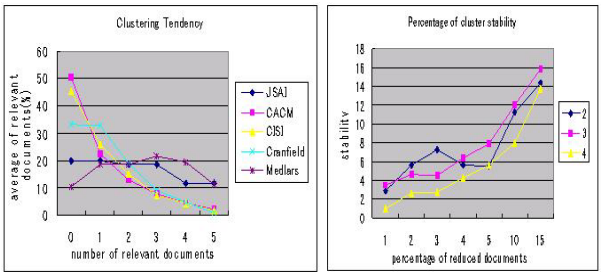
The basic idea is to retrieve relevant documents for given queries in relevant clusters created by TRSM clustering.

– Information retrieval in relation with clustering tendency;
– Information retrieval in relation with clustering stability;
– Evaluation of effectiveness (precision and recall of retrieval) and efficiency (computation time).

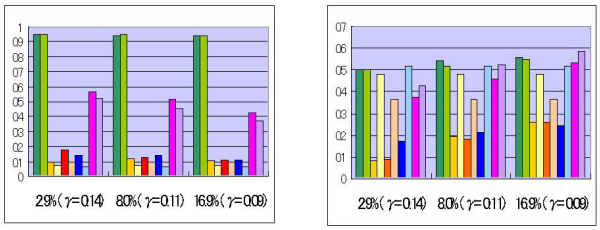
Ho, T.B., Kawasaki, S., Nguyen, N.B., "Cluster-based Information Retrieval with a Tolerance Rough Set Model", International Journal of Fuzzy Logic and Intelligent Systems, Vol. 2 (2002), No. 1, 26-32.

Ho, T.B., Nguyen, N.B., "Nonhierarchical Document Clustering by a Tolerance Rough Set Model", International Journal of Intelligent Systems, Vol. 17 (2002), No. 2, 199-212.

Stability and tendency of clustering



Evaluation of effectiveness



Evaluation of efficiency

Collection	Size (MB)	Number of queries	TRSM Time	Clustering Time	Full Search (sec)	Cluster Search (sec)	Memory (MB)
JSAI	0.1	20	2.4s	14.9s	0.8	0.1	8
CACM	2.2	64	22m2.8s	26m46.8s	13.3	1.2	201
CISI	2.2	76	13m16.8s	4m49.8s	40.1	3.4	84
CRAN	2.6	225	23m9.9s	3m6.9s	20.5	1.8	71
MED	1.1	30	40.1s	1m30.8s	2.5	0.3	25
Reuters	5	n/a	16m42.3s	173m	n/a	n/a	820

Future Plan and Expected Results

– To investigate the effect of TRSM's parameters in TRSM text mining algorithms;
– To incrementally update tolerance classes of terms and document clusters when new documents are added to the text collection;
– To extend the TRSM by considering (1) models without the symmetric property; (2) tolerance classes based on co-occurrence of more than two terms;
– To combine TRSM-based nonhierarchical and hierarchical clustering algorithms for very large text collections;
– To develop other TRSM text mining methods.

Contact: Ho Tu Bao, Japan Advanced Institute of Science and Technology, Tatsunokuchi, Ishikawa, 923-1292 Japan; Phone & Fax: 81-761-51-1730, http://www.jaist.ac.jp/ks/labs/ho.

(A02-05-5) Mining Hepatitis Data with Temporal Abstraction

Investigator	Tu Bao Ho	(Japan Advanced Institute of Science and Technology)
Collaborators	Duc Dung Nguyen	(Japan Advanced Institute of Science and Technology)
	Saori Kawasaki	(Japan Advanced Institute of Science and Technology)
	Trong Dung Nguyen	(Japan Advanced Institute of Science and Technology)

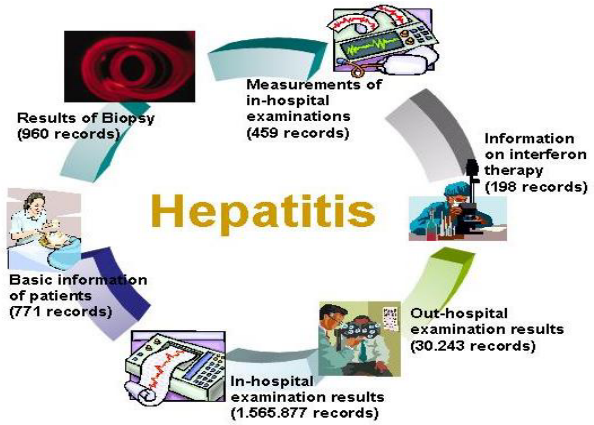
Background and Aim

The hepatitis database collected during 1982-2001 at the Chiba university hospital is an important resource to study hepatitis:

– This relational database contains irregular temporal data measured on nearly one thousand examinations.

– The medical doctors posed six problems to be investigated by data mining methods.

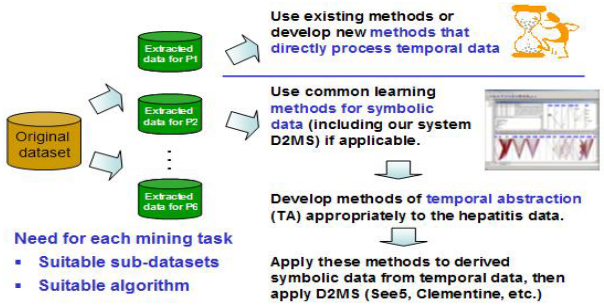
The relational database of temporal hepatitis data



This work aims to find efficient methods of temporal abstraction that generalize states and trends of hepatitis patients in suitable periods from their temporal data, then to investigate the hepatitis problems using symbolic learning methods with input as generated data by temporal abstraction.

Research Plan and Approach

Instead of directly mining multi-relational and temporal data, we extract an integrated temporal data table and apply symbolic mining methods to abstracted data by temporal abstraction methods.



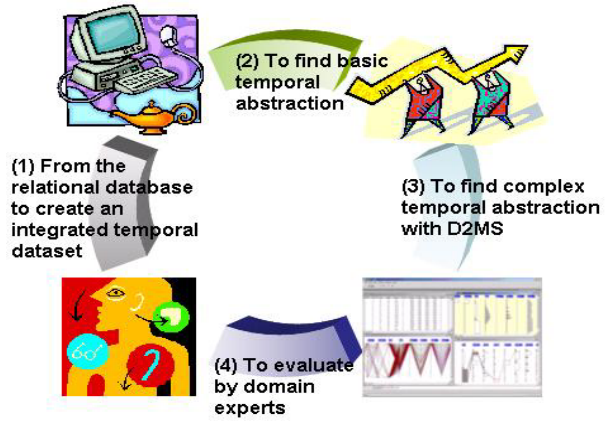
The essence of this research is *temporal abstraction* methods for finding generalizations of temporal data to be used by symbolic learning methods, in particular our visual data mining system D2MS, in order to extract novel and interesting patterns on hepatitis.

The temporal abstraction task can be defined as follows. The input includes a set of time-stamped data points (events) and abstraction goals. The output includes a set of interval-based, context-specific unified values or patterns (usually qualitative) at a higher level of abstraction. The temporal abstraction task can be decomposed into two subtasks of:

– *Basic temporal abstraction* for abstracting-time-stamped data into episodes.

– *Complex temporal abstraction* for investigating specific temporal relationships between episodes that can be generated from a basic temporal abstraction or from other complex temporal abstraction.

The process of mining hepatitis data with temporal abstraction goes through four steps



Basic temporal abstraction typically extracts *states* (e.g., low, normal, high), and/or *trends* (e.g., increase, stable, decrease) from a uni-dimensional time- series. We define a two-component structure of abstractions as

<episode, state & trend>

Main Results

Our main results so far are a methodology that could be a promising alternative way to approach the temporal, complex hepatitis data, and the output of primitive attempts that could be starting points for further investigation.

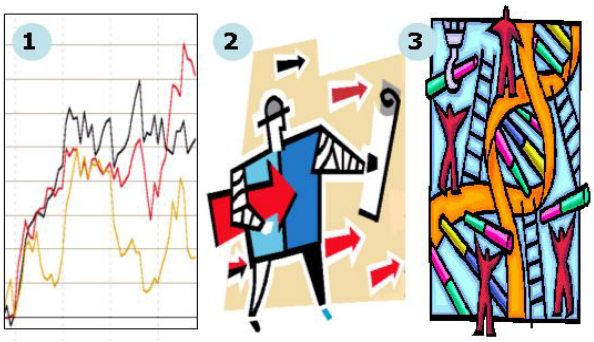
Procedures for determining context-sensitive episodes appropriate for each problem

– *Input*: Integrated dataset with irregular temporal data on examinations
– *Output*: Context-sensitive episodes with length Δ on examinations.

Procedures for determining states in episodes

– *Input*: Temporal data in each episode of patient's examination.
– *Output*: Abstracted states of examinations with values: "extremely high", "very high", "high", "normal", "low", "very low", "extremely low".

Qualitative state detection: to abstract the patient's states according to their temporal data on examinations



Procedures for detecting trends in episodes

– *Input*: Temporal data in each episode of patient's examination.
– *Output*: Qualitative trends of examinations as combination of values (obtained by piecewise linear regressions): "extremely fast increasing", "fast increasing", "increasing", "stable", "decreasing", "fast decreasing", "extremely fast decreasing".

Primitive results in finding differences in temporal patterns between hepatitis B and C (problem P1), as well the relationships between the stages of liver fibrosis and laboratory examinations (problem P2).

Qualitative trend detection: to detect the general direction that reflects the movement of the patient's temporal data over a long interval of time.

From abstracted data, our induction program LUPC yielded various rules for the above two problems, for example

Rule 2:	acc=0.833(5/6); cover=0.044(6/136)
IF	T-BIL = normal & decreasing-decreasing AND
	T-CHO = normal & decreasing-increasing AND
	U-GLU = normal & stable-stable
THEN	Stage F2

By using 10-fold stratified cross evaluation, we obtained an estimation of error rate of 17.820% ± 4.933% for findings on hepatitis B and C. Also, by using the exclusive and inclusive constraints in LUPC, we detected various rules that are highly consistent, and some could be inconsistent, with the common knowledge on short-term and long-term changes in important examinations.

Future Plan and Expected Results

- To understand better the hepatitis database, the problem and domain knowledge, and improve the pre-processing results.
- To establish the mathematical models of temporal abstraction appropriate to problems under investigation.
- To find suitable episodes for each mining task using the domain knowledge, particularly episodes with varying lengths.
- To qualitatively find and characterize trends for different length episodes in different problems.
- To do better post-processing and evaluation of results.
- We expect to find new and significant patterns or models from hepatitis data that enrich the human knowledge in this field.

Contact: Ho Tu Bao, Japan Advanced Institute of Science and Technology, Tatsunokuchi, Ishikawa, 923-1292 Japan; Phone & Fax: 81-761-51-1730, http://www.jaist.ac.jp/ks/labs/ho.

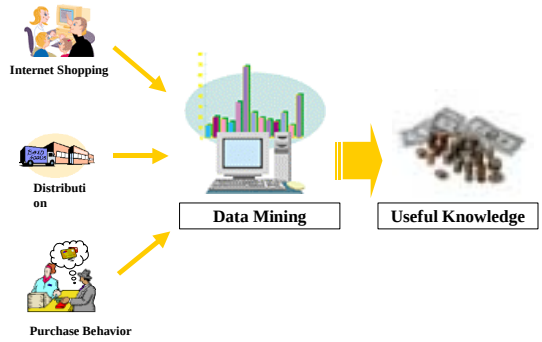
(A02-05-6) A Study on Application of Data Mining Technique to Various Types of Management Data

Investigator　Katsutoshi Yada (Kansai University, Faculty of Commerce)

Background and Aim

In recent years, under the conditions that information systems and devices are now available at lower prices and that performance characteristics of hardware have been considerably improved, enormous amount of data has been accumulated in business field. Data relating to distribution, for instance, include various types of data such as order receiving and issuance data, sales POS data, Internet access log, mail order sales data, etc. However, many business firms are not sufficiently utilizing these accumulated data for their strategic planning, production control, and marketing strategy for future. If useful information and knowledge can be found from these enormous amounts of data, it would be possible to establish more effective marketing strategy.

On the other hand, attention is now focused on the study of data mining as interdisciplinary study such as database, artificial intelligence, statistics, etc. Data mining is a technique to efficiently extract useful rules or patterns from enormous amount of business data accumulated as described above. Attempts on its application have been already made in the business field such as financing or Internet-related business although not to a great extent.



The purpose of the present study is to perform diversified fundamental studies to apply the newest technique of data mining to various types of management data and to discover the knowledge useful for business purpose. Our attempts cover extensive regions - not only the modification of the existing algorithm, but also fundamental technique such as system architecture or data structure, and further, the subjects for application such as expression of knowledge, evaluation criteria, and strategy of knowledge discovery process. In the present study, we intend to focus our attention on the data relating to distribution such as POS data with ID, WEB log in EC, etc., and we will perform comprehensive study as to how the data mining technique should be applied in the business field.

Research Plan and Approach

With the purpose of performing comprehensive study on the application of the data mining technique in the business field, it is necessary to carry out the study from the three viewpoints as given below and to unify the results of the study.

< System Architecture and Data Structure >

The current typical database systems are mostly used for the purpose of achieving simplified retrieval and sum and of attaining higher efficiency. These are not designed to generate patterns or rules between the data. In order to freely process the data and excavate crude gemstones hidden in the data, it is necessary to have system architecture and data structure

suitable for the purpose.

We are proposing original system architecture "MUSASHI", which has data mining-oriented nature. MUSASHI is a design concept of a system, which aims to attain flexibility in supporting the efficient processing of large amount of data and continuous knowledge discovery process. As implementation technique, a concept of history base is introduced, and characteristic data structure is adopted.

< Knowledge Expression and Evaluation Criteria >

In order to discover useful knowledge, which may be helpful to lead to efficient business action, we must develop expression of knowledge and evaluation criteria to cope with the problems in business field.

For instance, the basket analysis already utilized in the business field such as marketing deals with the transition of purchase based on the product brands at the time of buying. Manufacturers might be interested not only in the change of purchase at a certain time but also on the relevant purchase behavior between product brands for long run including a multiple of purchases for longer span. We introduced "Association Strength" and propose a method to extract the association between the product brands from the purchase data arranged in time series such as POS data with ID.

< Knowledge Discovery Process and Business Process >

In most cases, useful knowledge cannot be discovered by simply introducing the data into data mining software. Deep-rooted specialized knowledge and skills of high-grade data operation are needed. It is this process which is helpful in constructing the ability as a professional and this will be an important step in terms of competition.

Most of the conventional type knowledge discovery processes merely aim expression of typical patterns of data processing. No adequate model is proposed, which may give more concrete suggestion as to what kind of knowledge is actually sought and in which type of process. We propose one of the solutions of these problems by expressing the process on 2-dimensional matrix, i.e. the extent of the attention on the changes over time and analytical level.

Future Plan and Expected Results

In particular, these concepts will be introduced in actual business firms this year to elucidate the problems in practical business activities and to clarify fundamental studies required. Then, cost performance in line of business will be determined, and its usefulness will be verified. Finally, propagation into distribution industry will be promoted. Manufacturers, wholesalers, retailers as well as universities and colleges will be connected with each other and unified through network, and we are planning to propose how the cooperation should be established between industry and academic circles.

Contact:

Katsutoshi Yada (Investigator)
Faculty of Commerce, Kansai University, 3-3-35, Yamate-cho, Suita, Osaka 564-8680, JAPAN
Email : yada@ipcku.kansai-u.ac.jp ; Tel : 06-6368-1121 ; Fax : 06-6330-3106

(A02-06-1) Automatic Composition of Data Mining Applications Based on Meta-Learning

Principal Investigator	Takahira Yamaguchi	(Faculty of Information, Shizuoka University)
Investigator	Naoki Fukuta	(Faculty of Information, Shizuoka University)
	Yoshiaki Tachibana	(Faculty of Law and Letters, Ehime University)
	Noriaki Izumi	(Cyber Assist Research Center, AIST)
Collaborator	Hidenao Abe	(Graduate School, Shizuoka University)

Background and Aim

In recent years, end-users of data mining applications are faced with many problems: how to represent a target concept, model selection and so on. Conventionally, these problems are solved by trial-and-error or heuristics. This solution takes many costs. So we propose the automatic composition of data mining applications with meta-learning in order to support all the processes of developing data mining applications, constructing repositories for data pre-processing, data mining and metrics of users , as shown in Figure 1.

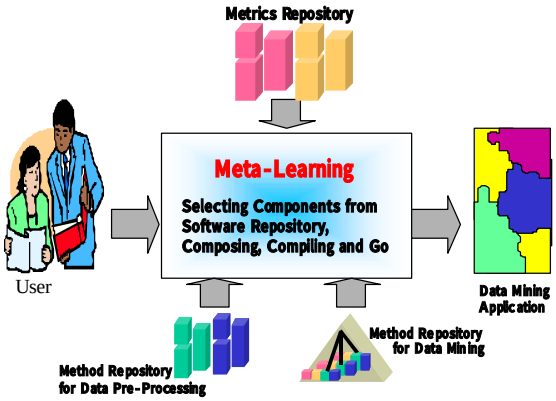


Figure 1: Automatic Composition of Data Mining Applications Based on Meta-Learning

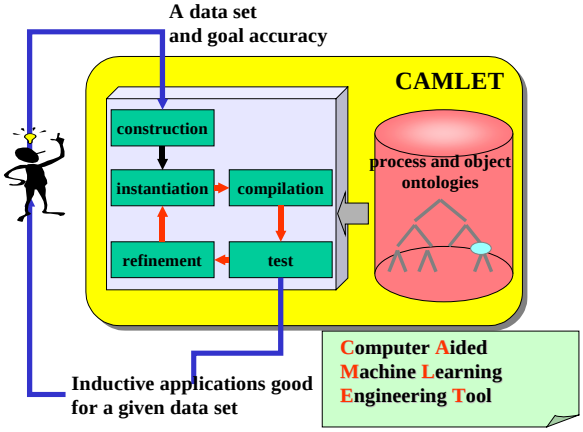


Figure 2: Overview of CAMLET

Research Plan and Approach

We have developed the tool to compose inductive applications automatically based on meta-learning, using inductive method repository, as shown in Figure 2. This tool is called CAMLET that stands for a computer aided machine learning engineering tool.

Given a data set and goal accuracy from a user, CAMLET selects just one control structure from eight different control structures, and selects the components from inductive method repository. Thus CAMLET composes the specification of inductive application. This specification is compiled and executed. While the result of execution does not go beyond the given goal accuracy, CAMLET repeats the above composition process.

We need a refined search method to compose the specifications effectively and finish the implementations of parallel CAMLET as shown in Figure 3.

Figure 3 Parallel CAMLET in progress

Main Results

We have done the case study to compare all the accuracies of inductive applications composed by CAMLET with those of popular inductive and statistical algorithms, using eight different data sets from the StatLog project. The inductive applications composed by CAMLET take the first highest accuracy on the average, taking the first to tenth highest accuracy, as shown in Table 1.

In addition, we have also implemented meta-rules (called a spec refinement rule) based on association rules as a way to achieve efficient specification refinements. While we have achieved more stable specification changes than could be achieved by random search or GA search with this approach, we need to extend the representation of the spec refinement rules in order to get to better refinement.

Table 1 Accuracy Comparison Using Common Data Sets of the StatLog Project

Algorithm	australian	diabetes	dna	letter	satimage	segment	shuttle	vehicle	Average
CAMLET	**87.3%(1)**	**76.4%(5)**	**95.0%(3)**	**82.0%(10)**	**88.3%(4)**	**96.1%(5)**	**99.8%(1)**	**79.2%(6)**	**88.0%**
BayTree	82.9%	72.9%	90.5%	87.6%	85.3%	96.7%	98.0%	72.9%	85.9%
C4.5	84.5%	73.0%	92.4%	86.8%	85.0%	96.0%	90.0%	73.4%	85.1%
NewId	81.9%	71.1%	90.0%	87.2%	85.0%	96.6%	99.0%	70.2%	85.1%
IndCart	84.8%	72.9%	92.7%	87.0%	86.2%	95.5%	91.0%	70.2%	85.0%
Cn2	79.6%	71.1%	90.5%	88.5%	85.0%	95.7%	97.0%	68.6%	84.5%
Cal5	86.9%	75.0%	86.9%	74.7%	84.9%	93.8%	97.0%	72.1%	83.9%
Dipol92	85.9%	77.6%	95.2%	82.4%	88.9%	96.1%	52.0%	84.9%	82.9%
KNN	81.9%	67.6%	85.4%	93.2%	90.6%	92.3%	56.0%	72.5%	79.9%
Ac2	81.9%	72.4%	90.0%	75.5%	84.3%	96.9%	68.0%	70.4%	79.9%
BackProp	84.6%	75.2%	91.2%	67.3%	86.1%	94.6%	57.0%	79.3%	79.4%
Alloc80	79.9%	69.9%	94.3%	93.6%	86.8%	97.0%	17.0%	82.7%	77.7%
Smart	84.2%	76.8%	88.5%	70.5%	84.1%	94.8%	41.0%	78.3%	77.3%
Cart	85.5%	74.5%	91.5%	0.0%	86.2%	96.0%	92.0%	76.5%	75.3%
QuaDisc	79.3%	73.8%	94.1%	88.7%	84.5%	84.3%	0.0%	85.0%	73.7%
LogDisc	85.9%	77.7%	93.9%	76.6%	83.7%	89.1%	0.0%	80.8%	73.5%
Radial	85.5%	75.7%	95.9%	76.7%	87.9%	93.1%	0.0%	69.3%	73.0%
Discrim	85.9%	77.5%	94.1%	69.8%	82.9%	88.4%	0.0%	78.4%	72.1%
LVQ	80.3%	72.8%	0.0%	92.1%	89.5%	95.4%	56.0%	71.3%	69.7%
Castle	85.2%	74.2%	92.8%	75.5%	80.6%	88.8%	0.0%	49.5%	68.3%
Bayes	84.9%	73.8%	93.2%	47.1%	71.3%	73.5%	0.0%	44.2%	61.0%
Itrule	86.3%	75.5%	86.5%	40.6%	0.0%	54.5%	59.0%	67.6%	58.8%
Kohonen	0.0%	72.7%	66.1%	74.8%	82.1%	93.3%	0.0%	66.0%	56.9%
Default	56.0%	65.0%	50.8%	4.0%	23.1%	24.0%	0.0%	25.0%	31.0%
Cascade	0.0%	0.0%	0.0%	0.0%	83.7%	0.0%	0.0%	72.0%	19.5%

Future Plan and Expected Results

First, we will improve how to represent spec refinement rules and establish the hybrid search with global search and local search with the adjustment of parameters with mining methods. Second, after we will extend three pu (processing units) into sixteen pu, we will compare CAMLET with other methods of meta learning. Furthermore, we will extend the single repository for data mining into three kinds of repository for data pre-processing, data mining and metrics to evaluate mining results. Finally we will apply CAMLET to chronic hepatitis dataset in order to evaluate the practice.

Contact:

Takahira Yamagiuchi (Principal Investigator)
Faculty of Information, Shizuoka University, 3-5-1, Johoku, Hamamatsu, Shizuoka 432-8011, JAPAN
Email: yamaguti@cs.inf.shizuoka.ac.jp; Tel: 053-478-1473, Fax: 053-473-6421

(A02-06-2) Rule Discovery Support Based on Clustering of Chronic Hepatitis Datasets

Principal Investigator	Takahira Yamaguchi	(Faculty of Information, Shizuoka University)
Investigator	Miho Ohsaki	(Faculty of Information, Shizuoka University)
Collaborator	Mao Komori	(Graduate School, Shizuoka University)
	Naomi Nakaya	(Graduate School, Shizuoka University)
	Yoshinori Sato	(Graduate School, Shizuoka University)

Background and Aim

This research aims to develop a data mining system for actual medical datasets such as chronic hepatitis datasets and to discover rules interesting to medical experts. It also aims to relate CAMLET from (A02-06-1) with chronic hepatitis dataset.

Research Plan and Approach

We take chronic hepatitis datasets as common datasets in the Active Mining project. Since the chronic hepatitis datasets are so ill-defined to submit into data mining systems, we need many kinds of data pre-processing in advance. They are too hard to automate the processes at present and so we will do them by hands and take a mining method based on the descretization of time series data to raise the understandability for medical experts

Main Results

We focused on the attribute of GPT that reflects the progress degree of disease, and tried to obtain the future trend of GPT resulted from the current trends of GPT and the other attributes. We prudently designed the pre-processing process, since it is an important phase especially for a clinical dataset. We conducted the pre-processing as follows.

Attribute Selection: The volume of the dataset was considerably large; the dataset included about 54000 records, and the number of attributes was 957. We extracted 42 attributes with the highest frequency of appearance among all the attributes as shown in Figure 1.

Frequency Unification: Although the cycle of medical examination for a patient was various depending on the symptom of the patient etc., a mining scheme needs its constant cycle. We examined the frequency distribution of the cycle for each attribute and unified all cycles into 28 days, which had the highest frequency of appearance. Figure 2 shows us the results.

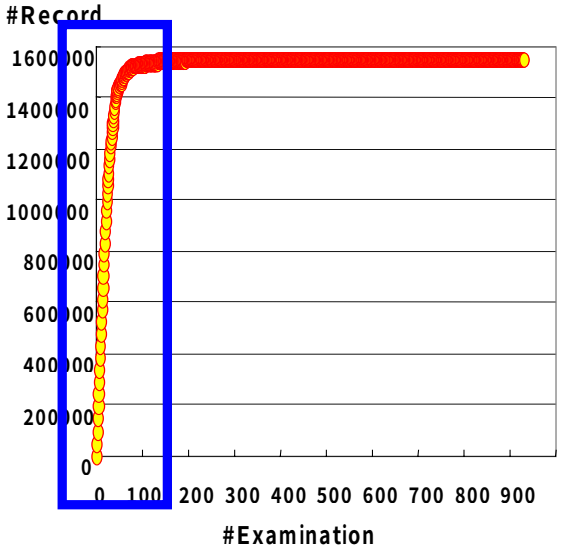


Figure 1 The number of records included in one examination

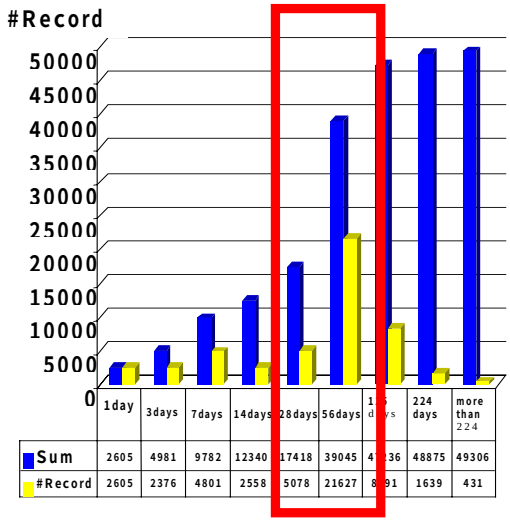


Figure 2 The cycle of medical examination

Descretization of Time Series Data: We applied the framework of Das to discretize the time series data and EM algorithm to cluster the sequential patterns of the data. Figure 3shows us the results.

Thus pre-processed datasets were given to a decision tree system such as C4.5. Although most learned rules were similar to common medical knowledge, there were some rules that medical experts were interested in. Figure 4 shows us some interesting rule. A medical expert interpreted the rule as follows; This rule explains how the state of GPT changes from rise to fall and fall to rise, and it inspires the medical expert with an idea that the activity of virus may have a cycle, since GPT reflects the activity of virus at certain level.

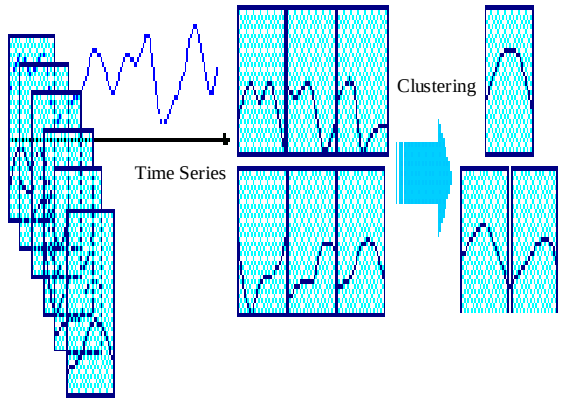


Figure 3 Clustering of sub-sequential patters

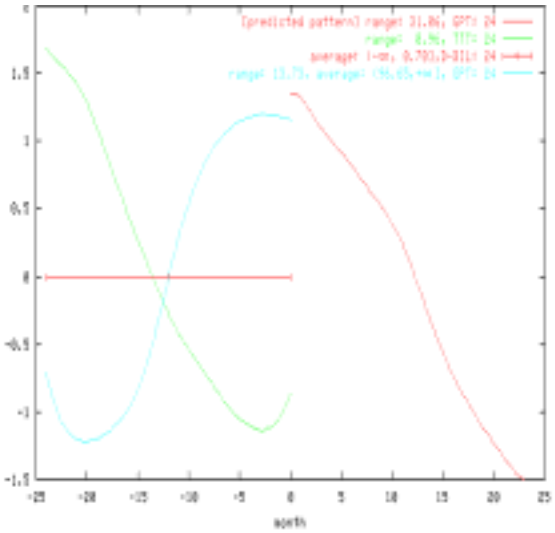


Figure 4 A rule interesting to a medial expert

Future Plans and Expected Results

This case study showed us that the post-processing of mined results are so important. Since most learned rules have no surprise with common sense and experienced of medical experts, it is important to automate the processes with hepatitis ontologies. Moreover, medical experts want to do another after evaluating learned rules as follows; in the case of that conditional parts come, how are these attributes going. So it is important to facilitate the extension processes from medical experts while doing post-processing.

Now we do the following rule discovering support: to obtain the future trend of GPT resulted from the current trends of GPT and the other attributes for mid/long terms. . At the last stage, the experience will be given back to refine CAMLET in the future.

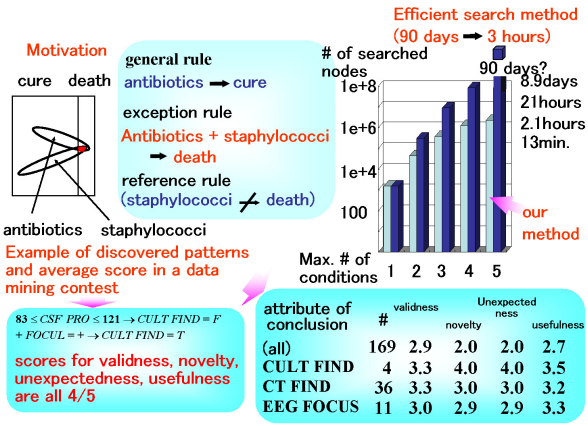
Contact:

Takahira Yamagiuchi (Principal Investigator)
Faculty of Information, Shizuoka University, 3-5-1, Johoku, Hamamatsu, Shizuoka 432-8011, JAPAN
Email: yamaguti@cs.inf.shizuoka.ac.jp; Tel: 053-478-1473, Fax: 053-473-6421

(A02-07-1) Spiral Exception Discovery

Principal Investigator	Einoshin Suzuki	(Yokohama National University)
Collaborator	Yuu Yamada, Takeshi Watanabe, Fumio Takechi,	(Yokohama National University)
	Naoki Yamaguchi, Mitsutoshi Nagahama, Yuta Choki,	(Yokohama National University)
	Kazuki Nakamoto, and Masafumi Gotoh	(Yokohama National University)

Background and Aim



An exception, which represents a different tendency from most of the rest, has been regarded as an important candidate of a useful piece of knowledge in data mining since it is often interesting and provides an opportunity of novel discovery. Our hypothesis-driven exception rule discovery obtains a rule which holds for a large number of examples with high probability and another rule which shows an exception to the rule simultaneously. As we can see from the figure, the method has been successful in terms of both interestingness of discovered pieces of knowledge and effectiveness of a discovery process.
The objective of this research is to extend this methodology to active mining methods in order to realize a spiral exception-discovery method which successively discovers exceptions based on data, knowledge, and environment; and to confirm its effectiveness by applying it to data sets such as those in medicine and commerce.

Research Plan and Approach

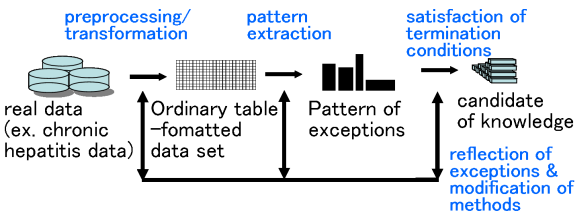
Firstly, we build the base system in order to fulfill the objective in the previous chapter. Important techniques which should be achieved are restriction of candidate patterns based on various conditions and selection of appropriate learning/discovery method based on the discovery process.

Parallel to the first process, we develop effective learning/discovery methods for important issues in data mining. These are necessary since, in spiral discovery, various methods are employed according to the discovery process.

Lastly, we apply our implemented system based on our approach to data sets such as those in medicine and commerce, and evaluate its effectiveness based on criteria such as those supplied by domain experts. In analyzing real data, the goal of the analysis cannot be determined at an early stage. Therefore, we typically consult domain experts in order to work on from important and realistic goals.

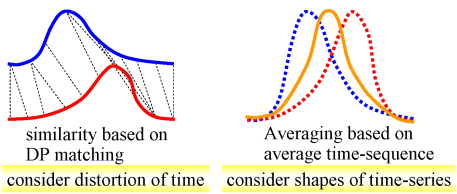
Main Results

Construction of Base System

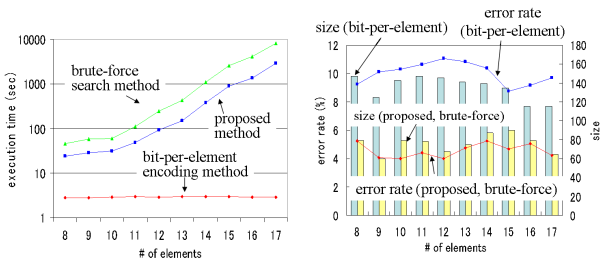


We have constructed a base system which we show in the right figure, and have realized a spiral discovery system for exception rules. This system investigates discovered exception rules from the viewpoint of data, knowledge, and environment; and obtains novel exception rules. Data, knowledge, and environment are represented as given data; previously discovered knowledge; and evaluation scores of a user for validness, usefulness, novelty, and unexpectedness. The method searches for similar pieces of knowledge for previously discovered pieces of knowledge by alternating discretization methods and probabilistic evaluation criteria. The method has been applied to the meningitis data set which has been supplied as a benchmark data set in data mining.

Development of Learning/Discovery Method



Time-series data which are regarded as highly important and from which discovery of useful pieces of knowledge are highly demanded exist in various domains. We have developed a novel data squashing method for time-series data, and have confirmed its effectiveness for clustering. This method is based on construction of an average sequence, which is shown in the right figure, and a data structure which effectively compresses time-series data.

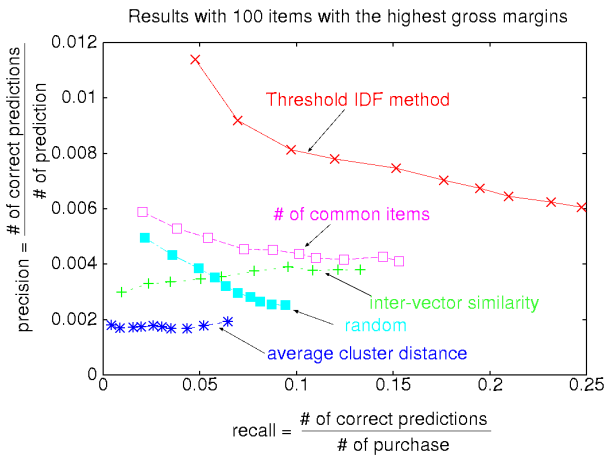


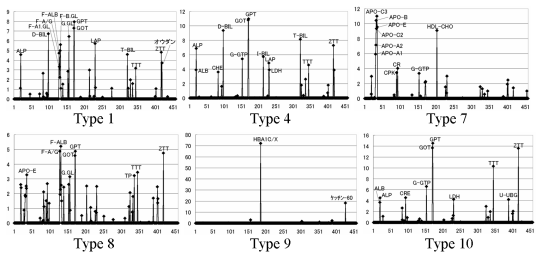
For transactional data which is common in data in medicine and commerce, a set-valued attribute which takes a set as its value is important since set inclusions typically represent essential information. We have defined a subset-split test for a set-valued attribute for decision tree induction and have developed a time-efficient method which obtains an optimal subset-split test. We have confirmed that this method outperforms other methods in terms of error rates and computational time as shown in the left figure.

We have also proposed other methods such as blooming decision tree which represents an effective classifier for a multi-dimensional class, and an outlier discovery method which demonstrates higher comprehensibility due to integration with a decision tree learner, and confirmed their effectiveness through experiments.

Analysis of Real Data

We have conducted various analyses for sales data of drug stores in commerce. Especially, we have been interested in rare items each of which brings a large gross margin and sells few, and developed a novel method for predicting exceptional customers who are likely to purchase these items. Experimental results are promising as shown in the right figure, and we consider that we have constructed an effective method for a difficult prediction problem. This method is related to cost-sensitive classification which is gaining increasing interests, and can be considered as worth further development.





We have applied statistical pre-processing methods for chronic hepatitis data, which is situated as common data of this project, and have discovered typical patterns, which we show in the left figure, for abnormal test values. These patterns are considered to be useful in transforming a huge and sparse transactional data set into an ordinary table-formatted data set.

Moreover, we are steadily contributing to determination of discovery problems through collaboration with physicians. Concerning effectiveness of interferon for C-type patients, our analysis have led to an almost complete determination of learning targets. Concerning change of conditions for B-type patients, we have conducted cluster analyses of changes of the patterns for each group of patients in terms of virus tests. Concerning prediction of progress of disease from blood tests, we have detected an exceptional group of patients each of whom heavily suffers from sickness, and their analyses are highly demanded.

Future Plan and Expected Results

We are planning to pay our primary efforts on analysis of the chronic hepatitis data, and evaluation of our spiral exception discovery method. This evaluation not only contribute to analysis of the common data, but also cooperation with other research groups of this project. The process would necessitate us to build in the learning/discovery methods that we developed in the base system. This completes a prototype system which supports exception discovery for a wide rage of tasks.

Contact:

Einoshin Suzuki (Principal Investigator)
Electrical and Computer Engineering, Yokohama National University,
79-5, Tokiwadai, Hodogaya, Yokohama, 240-8501, Japan
E-mail: suzuki@ynu.ac.jp; Tel & Fax: +81-45-339-4148

(A02-07-2) Detection of Situation Changes

Principal Investigator Einoshin Suzuki (Yokohama National University)

Background and Aim

In most of the current data mining techniques, the target of discovery and the given data are assumed to be static. However, changes of the target of discovery according to data updates pose serious challenge in a real application.

The objective of this research is to develop a detection method for situation changes based on exceptions. This method can be situated as an important technique of active mining which actively discovers useful knowledge.

Research Plan and Approach

Firstly, we build sampling theories in order to fulfill the objective in the previous chapter. These theories give required numbers of examples for a rule, which represents the most fundamental discovered pattern, and are necessary for judging a change of situations.

Secondly, we build a detection method for situation changes based on the sampling theories. This method discovers rules from data which are segmented into periods, and detects a change of situations based on exceptions among rules.

Main Results

Worst-Case Analysis of Rule Discovery

We have performed a worst-case analysis of rule discovery in order to establish a uniform criteria for detection of situation changes. As shown in the right figure, this analysis, unlike others, gives a required number of examples so that the probability of the case in which the values of multiple important criteria are not smaller than user-specified values is not smaller than a user-specified value. As we can see from the figure, our PAGA (Probably Approximately General and Accurate) discovery can be regarded as a natural extension of PAC (Probably Approximately Correct) learning, which is a standard method in classification.

describe required # of examples in terms of quality $(1-\varepsilon, 1-\varsigma)$ and luck $1-\delta$

PAC learning [Russel 95]: worst-case analysis of classification

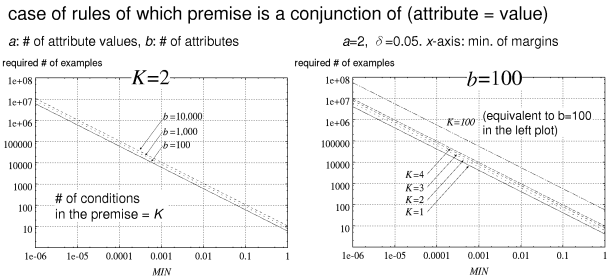
$\Pr[(\text{accuracy} \geq 1-\varepsilon)] \geq 1-\delta \Rightarrow (\#\text{ of examples}) \geq \frac{1}{\varepsilon}\ln\left(\frac{\#\text{ of classifier s}}{\delta}\right)$

PAGA discovery: worst-case analysis of rule $y \rightarrow x$ discovery

Objective: $\Pr[\Pr(x|y) \geq 1-\varepsilon, \Pr(y) \geq 1-\varsigma] \geq 1-\delta$ Method: $\hat{\Pr}(x|y) \geq \theta_y, \hat{\Pr}(y) \geq \theta_s$

$\Pr[(\text{accuracy} \geq 1-\varepsilon), (\text{generality} \geq 1-\varsigma)] \geq 1-\delta$

$m \geq \frac{\ln|\mathrm{R}| - \ln\delta}{2\mathrm{MIN}[(\theta_s-1+\varsigma)^2, \theta_s(\theta_r-1+\varepsilon)^2]}$



PAGA discovery, as shown in the left figure, gives a maximum number of required examples for rule discovery. In the figure, MIN represents the value of MIN in the inequality which gives the required number of examples m in the figure above.

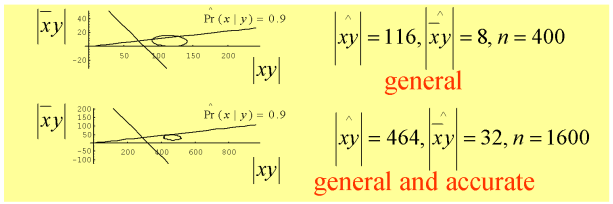
As we show in the right figure, We have also derived analytical solutions when we assume a multi-dimensional normal distribution for probabilistic variables for average-case analyses of rule discovery and exception rule discovery. These solutions describe conditions which is equivalent to the objective of PAGA discovery in terms of estimated generality and accuracy.

1. Assume $\left(|xy|, |\bar{x}y|\right)$ follows a multinomial distribution, then approximate this distribution by a 2D normal distribution
2. Since this problem is equivalent to check whether two lines are contingent to an ellipse, apply Lagrange's multiplier method

analytical solution

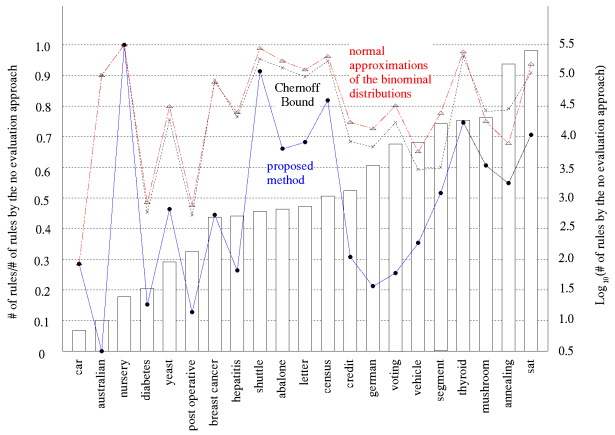
$$\left(1-\beta(\delta)\sqrt{\frac{1-\hat{\Pr}(y)}{n\hat{\Pr}(y)}}\right)\hat{\Pr}(y) \geq 1-\varsigma$$

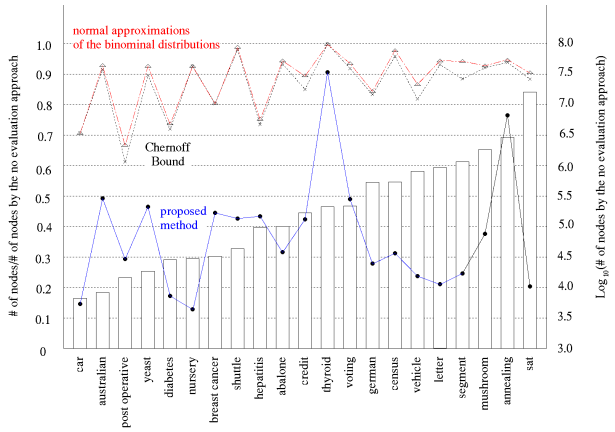
$$\left(1-\beta(\delta)\sqrt{\frac{\hat{\Pr}(\bar{x}y)}{\hat{\Pr}(xy)[(n+\beta(\delta)^2)\hat{\Pr}(y)-\beta(\delta)^2]}}\right)\hat{\Pr}(x|y) \geq 1-\varepsilon$$



These analytical solutions enable us to evaluate reliability of a discovered rule as shown in the right figure. In the figure, the case above assures only the reliability for generality, while the case below assures also the reliability for accuracy.

Moreover, the figures below show that this method is also shown to be effective for real data sets by experiments. The left figure shows that the proposed method can reduce the number of unreliable rules more effectively than other methods. The right figure shows that the proposed method can reduce the number of searched nodes more effectively than other methods. Due to these methods, we can build detection methods of situation changes for various cases which fit our objectives such as the worst case and the average case.





Detection of Situation Changes

We have developed a detection method for change of situations based on the sampling theory in the previous section. This method estimates a situation for each period by discovering multiple rules from data which are segmented into periods. A change of situations is detected based on exceptions among these rules. The proposed method has shown good results in preliminary experiments with artificial data.

Future Plan and Expected Results

PAGA discovery can be considered as promising as a fundamental theory of active mining which requests data based on necessity. We can develop discovery methods such as a method which samples/extends a data set based on the class of discovered rules, and a method which extends/limits the class of discovered rules based on the available data set.

Future work includes more realistic analyses such as an analysis for multiple-rule discovery and rule-discovery with unlimited conclusions. We would also like to consider applying our methods for real data such as WWW, which are expected to demonstrate clearer changes of situations.

Contact:

Einoshin Suzuki (Principal Investigator)
Electrical and Computer Engineering, Yokohama National University,
79-5, Tokiwadai, Hodogaya, Yokohama, 240-8501, Japan
E-mail: suzuki@ynu.ac.jp; Tel & Fax: +81-45-339-4148

(A02-07-3) Auto-Adjustment Method for Spiral Discovery

Principal Investigator	Einoshin Suzuki	(Yokohama National University)
Collaborator	Yuta Choki	(Yokohama National University)
	Shutaro Inatani	(Yokohama National University)

Background and Aim

Most of learning/discovery method necessitates specification of a set parameters which determines conditions of learning/discovery. Especially for spiral discovery, in which discovery processes occur successively, it is desirable for these parameters to be adjusted automatically according to the intermediate results. We have proposed a method which automatically adjusts discovery conditions for a hypothesis-driven exception rule discovery, and have succeeded in facilitating the use of an effective but a relatively hard-to-use method with five parameters to specify. This method can be situated as an auto-adjustment technique for pattern extraction in a data mining process.

It has been argued that data pre-processing represents a most laborious task in a data mining process, and 80 % of the total efforts are spent for the task. The objective of this research is to develop a method which automatically adjusts parameters in data pre-processing by giving feed-back from pattern extraction. Boosting, which represents a classification approach, is widely recognized as the most excellent achievement in machine learning in the 1990s since it is guaranteed to exhibit high accuracy from the viewpoints of both theories and experiments. Data squashing represents a data transformation method which enables the use of conventional learning/discovery method by appropriately compressing huge data. We have chosen boosting for pattern extraction and data squashing for data pre-processing.

Research Plan and Approach

Firstly, we construct a system which continuously refines learning results by iterating data squashing and boosting. Since boosting outputs, for each example, a weight which represents the degree of difficulty for predicting the example, we have considered to reflect this information to the degree of compression in data squashing.

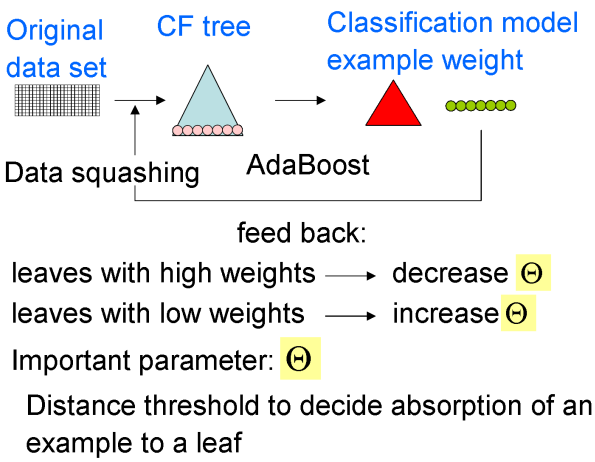
Secondly, we apply our method to data. In the experiments, we employ artificial data with which we evaluate the effectiveness of the approach, and real data with which we demonstrate its usefulness.

Parallel to these processes, we develop an outlier detection method based on data squashing and boosting. Although boosting has been proposed as a classification method, it has been shown to be able to detect hard-to-be-classified instances.

Main Results

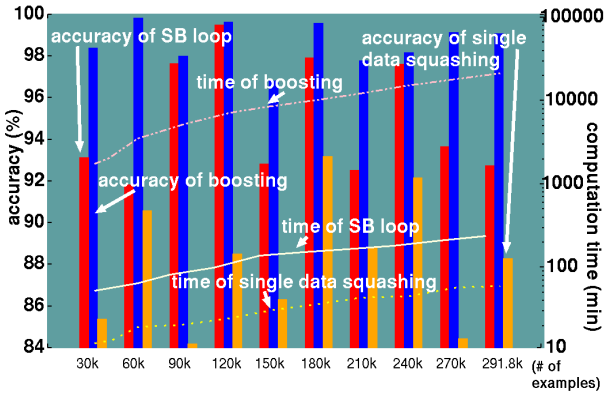
Fast Boosting Method SB Loop

As shown in the right figure, we have developed a fast boosting method SB loop. This method achieves both time-efficiency and accuracy by strongly and weakly compressing data of which classification are easy and difficult respectively. This method can be regarded as performing intelligent scheduling of discovery since it adjusts a set of parameters by iteratively applying data squashing and boosting.

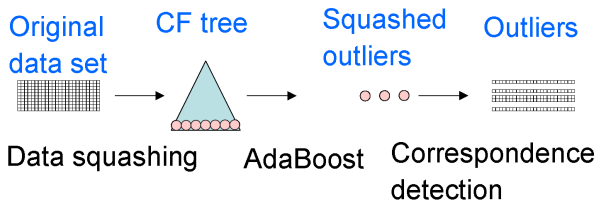


Experimental Evaluation of SB Loop

We have confirmed the effectiveness of our approach by applying it to both artificial and real data. We show the results with network intrusion data, which represents a benchmark data set of data mining, in the right figure. The figure shows that our method far outperforms a boosting method with single data squashing in terms of accuracy. Moreover, we see that our method achieves similar high accuracies as boosting without data squashing, and far outperforms it in terms of computational time.



Outlier Detection Method



Our outlier detection method based on data squashing and boosting, which is shown in the left figure, efficiently discovers abnormal instances which are extremely hard to be classified unlike most of others. We considered this method as more appropriate for spiral discovery of exceptions due to the nature of outlier detection. Therefore, we regarded application of real data as important and have not realized auto-adjustment of a set of parameters for this method.
We have applied our method to a POS (point of sales) data of drug stores and compared it with the case without data squashing. The results showed that our method is effective for POS data of drug stores when an instance represents a receipt.

Future Plan and Expected Results

Our automatic-adjustment technique for a set of parameters in pre-processing based on feed-back of pattern extraction can be applied to other problems in various ways. We are planning to develop a method which is devoted to the chronic hepatitis data, a common data of the project, and evaluate its effectiveness.

Contact:

Einoshin Suzuki (Principal Investigator)
Electrical and Computer Engineering, Yokohama National University,
79-5, Tokiwadai, Hodogaya, Yokohama, 240-8501, Japan
E-mail: suzuki@ynu.ac.jp; Tel & Fax: +81-45-339-4148

(A02-07-4) Peculiarity Oriented Mining

Investigator	Ning Zhong	(Maebashi Institute of Technology)
Collaborator	Muneaki Ohshima	(Maebashi Institute of Technology)

Background and Aim

The purpose of data mining is to discover new, surprising and interesting knowledge hidden in databases. Hence, the evaluation of interestingness (including peculiarity, surprisingness, unexpectedness, usefulness and novelty) should be done in pre-processing and/or post-processing of the knowledge discovery process.

We propose *peculiarity rules* as a new class of rules and develop new methods for mining such rule from various real-world databases. A peculiarity rule is discovered from peculiar data by evaluating their peculiarity using unified knowledge-based statistical criteria. We analyse data from a new view that is different from traditional statistical methods.

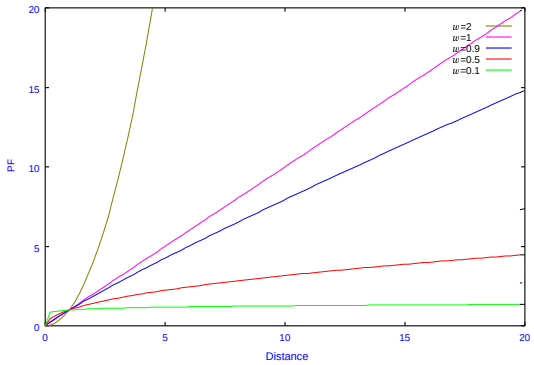
Research Plan and Approach

This study on peculiarity oriented mining can be divided into three phases: (1) developing the method of peculiarity oriented mining; (2) extending the peculiarity oriented approach for mining in multiple data sources; (3) enabling peculiarity oriented mining in a distributed and cooperative mode. Furthermore, the peculiarity oriented mining system will be integrated with the exception rule mining system proposed by Suzuki.

Main Results

Peculiarity Oriented Mining

We have developed a method of peculiarity oriented mining with the following main features of the proposed approach.



$$PF(x_i) = \sum_{j=1}^{n} N(x_i, x_j)^w$$

$$Threshold = mean\ of\ PF(x_i) + \alpha \times standard\ deviation\ of\ PF(x_i)$$

- **Peculiarity Factor Considering User Interestingness**. Peculiar data are a subset of objects in the database and are characterized by two features: (1) very different from other objects in a dataset, and (2) consisting of a small number of objects. We have developed the *Peculiarity Factor* as a measure for evaluating such peculiarity (right equations and figure). There are two merits of the measure: (1) it can handle both continuous and symbolic attributes based on a unified semantic interpretation, and (2) background knowledge represented by binary neighborhoods can be used to evaluate the peculiarity.

- **Semantic based attribute-oriented clustering**.

  In order to discover interesting knowledge, conceptual abstraction and generalization are necessary. Therefore, *attribute oriented clustering* is a useful technique as a step of the peculiarity oriented mining process. In our approach various methods of semantic based attribute-oriented clustering are provided in the mining process so that data from different sources can be handled effectively.

- **Changing the granularity of the peculiar data**. The granularity of peculiar data can be dynamically changed by using background knowledge on information granularity for mining more interesting knowledge.

- **Attribute-oriented finding**. That is, peculiar data can be searched in an attribute independent mode. Thus, the algorithm can be executed in a parallel-distributed mode for multiple attributes, relations and databases.

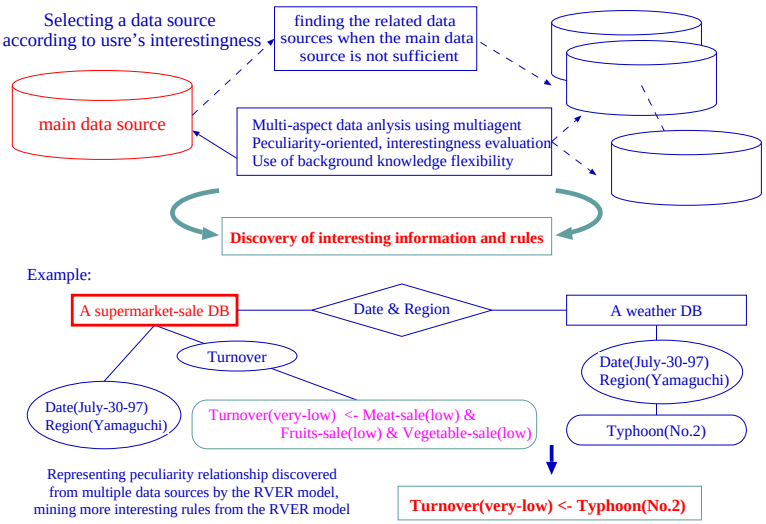
So far, a number of databases such as Japan-survey, amino-acid data, weather, supermarket, web-log, hepatitis have been tested for our approach. In particular, an initial result of peculiarity oriented mining in hepatitis data has being evaluated by a medical doctor. We are also doing a deeper, multi-aspect analysis in the hepatitis data.

Peculiarity Oriented Mining in Multiple Data Sources

We have extended our peculiarity oriented approach for mining in multiple data sources (i.e. multi-database mining). Multi-database mining involves many related topics including interestingness checking, relevance, database reverse engineering, granular computing, and distributed data mining.

By taking a unified view of multi-database mining and database reverse engineering, we have proposed the RVER (Reverse Variant Entity-Relationship) model to represent the result of multi-database mining. The RVER model can be regarded as a variant of semantic networks that are a kind of well-known method for knowledge representation. From this point of view, multi-database mining can be regarded as a kind of database reverse engineering.

A challenge in multi-database mining is semantic heterogeneity among multiple databases since no explicit foreign key relationships exist among them usually. Hence, the key issue is how to find/create the relevance among different databases. We used granular computing techniques to find/create the relevance and association among different databases by changing information granularity.



The right figure shows the framework of peculiarity oriented mining in multiple data sources. First, focus on a relation as the *main table* and find the peculiar data from this table. Then elicit the peculiarity rules from the peculiar data. If the data in the main table/database is not sufficient for finding interesting rules, search the peculiar data in the related databases. Thus, more interesting peculiarity rules can be discovered from peculiar data hidden in multiple relations/databases by searching the relevance among the peculiar data.

Future Plan and Expected Results

We have proposed *peculiarity rules* as a new class of rules. A peculiarity rule is discovered from peculiar data by searching the relevance among the peculiar data and using unified knowledge-based statistical criteria. We analyse data from a new view that is different from traditional statistical methods. So far, a number of databases such as Japan-survey, amino-acid data, weather, supermarket, web-log, hepatitis have been tested for our approach. Currently, we are working on mining from multiple mixed-media databases such as fMRI data and human brain waves.

We are also extending our approach for enabling peculiarity oriented mining in a distributed, cooperatively, multi-agent mode for performing multi-aspect data analysis as well as multi-level conceptual abstraction and learning.

Since our data mining system is very large and complex, however, we have only finished several components of the system and have undertaken to extend it for creating an organized society of autonomous knowledge discovery agents. That is, our current work takes but one step toward a multi-strategy and multi-agent KDD system.

Contact

Ning Zhong (Investigator)
Department of Information Engineering, Maebashi Institute of Technology
460-1 Kamisadori-Cho, Maebashi-City 371-0816, Japan
E-mail: zhong@maebashi-it.ac.jp; Tel&Fax: +81-27-265-7366

(A02-08-1) Part-of-speech Guessing of Unknown Word in Technical Papers

Principal Investigator Yuji Matsumoto (Nara Institute of Science and Technology)
Investigator Masashi Shimbo (Nara Institute of Science and Technology)
Collaborator Tetsuji Nakagawa (Nara Institute of Science and Technology)

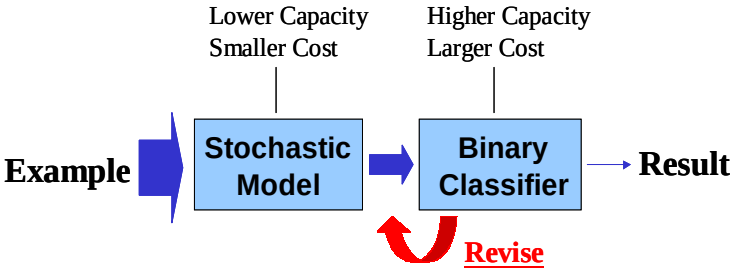
Background and Aim

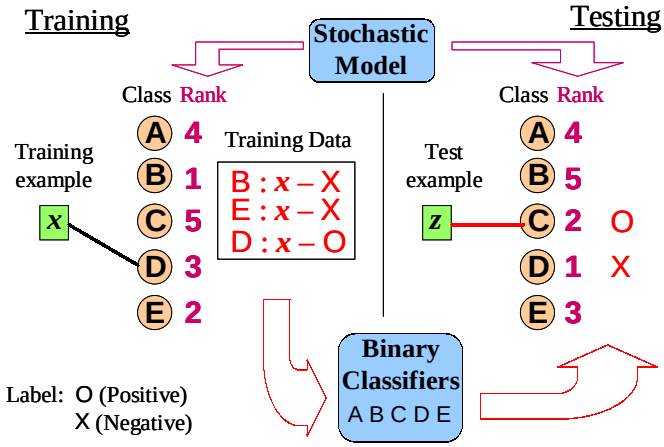
The aim of this research is to extract technical terms form medical abstracts through text mining on the Medline data. At least some linguistic analysis such as part-of-speech (POS) tagging is necessary because a number of English words have the same form both in the cases where they are used as a noun or as a verb. Moreover, since technical papers include full of words that are not registered in normal dictionaries, the unknown word problem is naturally unavoidable. Though some dictionaries for technical terms are available, they cannot cover a large part of terminologies, and what is worse technical terms are constantly growing, which makes it unrealistic to assume a full scale dictionary. The main aim of this research is, therefore, to develop a robust and accurate method for guessing the POS tags for unknown words.

Research Plan and Approach

We investigated a method to solve the unknown word problem by developing a robust POS tagger that is facilitated with POS-tag guessing for unknown words. When we perform POS tagging, all possible POS tags for each word are expanded. For unknown words, we expand all possible POS tags for those words according to their prefixes and suffixes and give them estimated probability values. Though probabilistic models like Markov models are frequently used for POS tagging, they suffer from data sparseness problem when we deal with a large scale feature space that takes into account fine grained ones like lexical forms and afixes. We once used SVMs to handle a large feature space and proved that an accurate POS tagging is possible. However, applying SVMs to such a problem requires tremendous amount of time.

In this research, we proposed a method named a **revision learning model** (Figure below), which is a cascade learning algorithm with a stocastic model and an SVM based classifier. In this model, a stocastic model is first applied with a moderate number of features and tested with the original training data. Then an SVM classifier learns where the first model successfully learns or fails to learn. The SVM model works as a revisor to identify the erroneous parts





produced by the first model. This model had two advantages over previous models: First, the size of training examples are far smaller since only the ones incorrectly learned by the first learner become the negative example. Second, the SVM is learned as a binary classifier while the previous SVM model has to learn multi-class classifiers. The production of training data and application of the learned model to test data are shown in the figure above. The performance of the system is shown in the table that compares trigram model (T3), our revision models (rows in red) and our previous full-SVM model (OVR).

Main Results

System	Number of Training Data	Training Time	Testing Time	Accuracy
T3 / Original	-	0.004	0.0076	96.59%
T3 / RL (polynomial)	1,027,840	16	0.18	96.98%
T3 / RL (linear)	1,027,840	2	0.011	96.94%
OVR	49,999,200	625	4.7	97.11%

We proposed high-performance POS tagging and unknown word guessing, proposing a new algorithm named revision learning. Our proposals are summarized as follows:

- A method for handling unknown word guessing problem within POS tagging.
- A revision learning method that combine probabilistic models and high performance binary classifiers to achieve efficient and accurate POS tagging.

Future Plan and Expected Results

Although we achieved a very good performance of POS tagging along with unknown word guessing, we have to investigate the reason that previous full-SVM model gives slightly better performance than the new one. We like to apply the revision learning model to other application area in natural language processing such as parsing and chunk identification. Besides, the performance is greatly affected by the quality of training data. We are going to use our learning framework to find out erroneous parts in the training data.

Contact

Yuji Matsumoto (Principal Investigator)
〒630-0101 8916-5 Takayama, Ikoma, Nara, 630-0101 Graduate School of Information Science, Nara Institute of Science and Technology
Email: matsu@is.aist-nara.ac.jp Tel: 0743-71-5240 FAX: 0743-72-5249

(A02-08-2) Large-scale Text Processing and Knowledge Extraction on Users' Demand

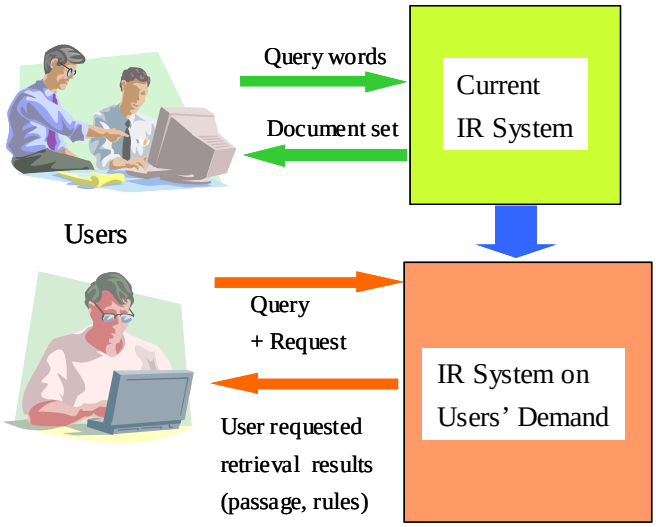
Principal Investigator Yuji Matsumoto (Nara Institute of Science and Technology)

Investigator Masashi Shimbo (Nara Institute of Science and Technology)

Hiroyasu Yamada (Japan Advanced Institute of Science and Technology)

Collaborator Taku Kudo (Nara Institute of Science and Technology)

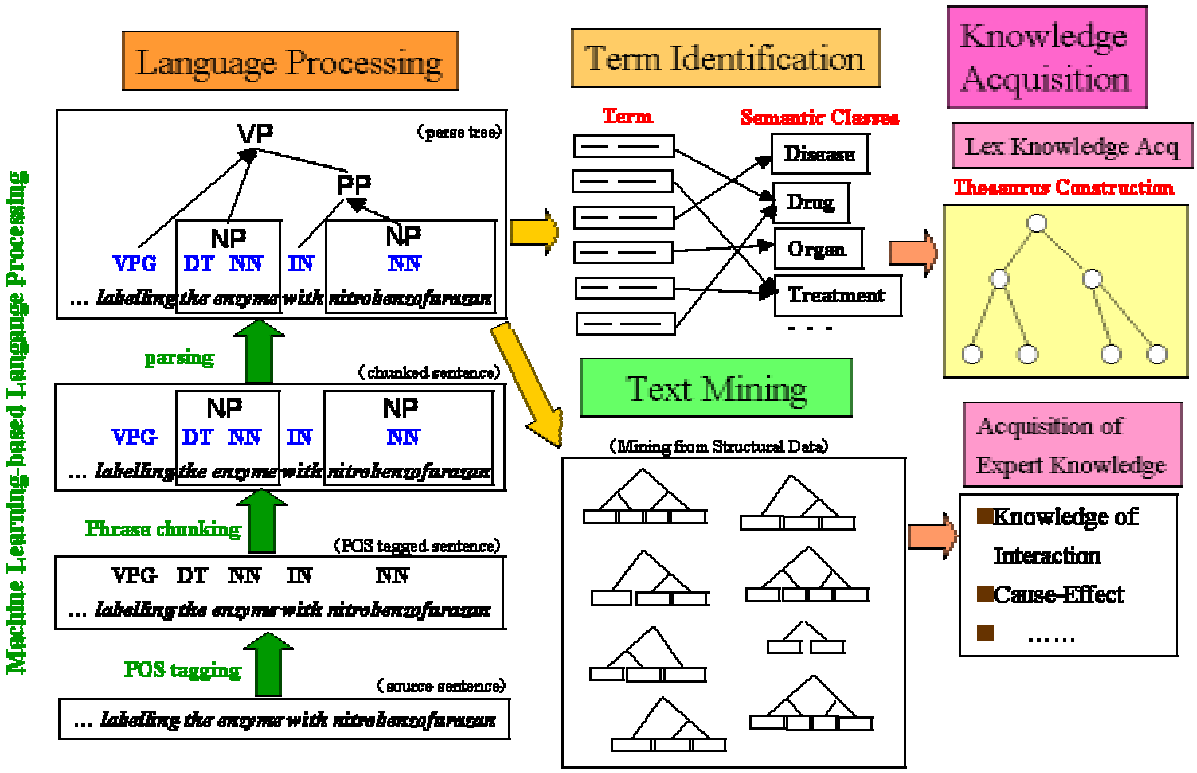
Background and Aim



Electronic text data has been proliferating in accordance with the recent extension of WWW. Simultaneously, on-line technical documents are easily obtainable. This leads to the users' demands of retrieving those documents in more effective ways. While current IR systems base on a set of query words, more sophisticated way of taking into account users' requests and intentions are called for. Since users' demands vary along with their intention, retrieval results of more flexible forms should be provided by retrieval systems (Figure). For robust analysis of large-scale text data, we aim to develop extendable language processing facilities based on statistical machine learning methodologies.

Research Plan and Approach

We plan to study flexible and active information extraction in accordance with users' demand, aiming mainly at analyzing medical paper abstracts of Medline dat. For achieving robust document processing, we investigate statistical learning models of **language processing** especially of part-of-speech tagging, syntactic dependency analysis. Those systems provide base linguistic information for other modules like **term identification/classification** and **lexical knowledge acquisition** modules for extracting semantic word similarity. Using those base language processing systems, we study information extraction methods that reflects users' intention. Another important text processing mechanism is the **text mining** that discovers frequent and collocational expressions appearing in the documents. Those interesting patterns may be specified by the user, or may be listed by the system for users' selection. They are used to extract higher level knowledge or rules such as interaction between chemical drugs, and cause-effect relationship of symptoms of diseases. (Figure on next page)



Main Results

We proposed high-performance language analysis methods making use of Support Vector Machines, and achieved technical term classification based on machine learning.

- We proposed a chunking cascade dependency paper by applying SVM learners.
- Using syntactic dependency relation, we achieved semantic classification of medical terms.
- We developed a mining system for structural text data annotated with various linguistic constraints.

Future Plan and Expected Results

The latter half of this fiscal year will be spent to incorporate top-down information into the English dependency parser to improve the accuracy, and to study text mining methods to identify useful expressions or structures for finding information on the users' demand. Extraction of knowledge of interaction and cause-effect relation between medical concepts will be investigated from the next year.

Contact

Yuji Matsumoto (Principal Investigator)
〒630-0101 8916-5 Takayama, Ikoma, Nara, 630-0101 Graduate School of Information Science, Nara Institute of Science and Technology
Email: matsu@is.aist-nara.ac.jp Tel: 0743-71-5240 FAX: 0743-72-5249

(A03-09-1) Development of the Active Mining System in Medicine Based on Rough Sets

Principal Investigator	Shusaku Tsumoto	(Shimane Medical University)
Investigator	Katsuhiko Takabayashi	(Chiba University Hospital)
	Masami Nagira	(Shimane Medical University)
	Shoji Hirano	(Shimane Medical University)

Background and Aim

Recent advances in medical examination equipment and networking technology enable us to automatically collect huge amount of temporal data on laboratory examinations. Analysis of such temporal databases has attracted much interests because it might reveal underlying relationships between temporal patterns of examination results and onset time of diseases. However, despite of its importance, large-scale analysis of time-series medical databases has rarely been performed. This is primarily due to the following problems: (1) in order to capture all of the events that have different durations, sequence should be compared in multiple observation scales. (2) such comparison scheme imposes similarity of sequences to be relative, in which triangular inequality may not hold; this property limits selection of clustering methods.

This research aims at establishing a new scheme of time-series data analysis that overcomes the above problems and enables us to discover interesting knowledge from time-series medical databases, for example common patterns appeared before/after applying some drags or treatments. It also aims at implementing the concept of active user reaction on medical data analysis, in which the feedback from users will be further used to determine strategies for the subsequent phases of data analysis and data collection.

Research Plan and Approach

The first two years of this project have been devoted to the development of the method for time-series data analysis. It includes (1) resampling method of the sequences that have different intervals and durations of data acquisition. (2) multi-scale comparison method of the sequences. (3) clustering method of the compared sequences. (4) visualization technique of the feature of the clustered sequences.

We have proposed a new methodology of data analysis that is a hybridization of multiscale matching and rough clustering. Multiscale matching is a method that compares similarity of sequences by partly changing observation scales (fig.1) and thus has an ability to capture both long- and short-term events. Since matching is performed on

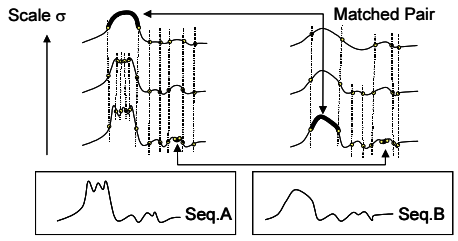


Figure 1: Multiscale Matching

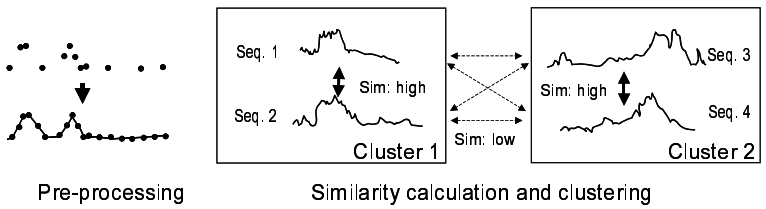


Figure 2: Procedure of Data Analysis

the basis of similarity of segments, which are subsequences between adjacent inflection points, it can be used to represent common increase/decrease patterns of the measurements. It also has a very important advantage that the connectivity of subsequences is preserved in the resultant sequence because it checks hierarchy of inflection points. While, rough clustering is a clustering method that groups up sequences based on their indiscernibility. It does not use any distance-related features, therefore, is able to produce interpretable clusters even when similarity of sequences is defined as relative ones. In our method, multiscale matching is used to calculate similarity of the sequences, and rough clustering is used to cluster the sequences according to the derived similarity (fig.2). The common patterns in the clustered sequences are then visualized using the result of multiscale matching.

Main Results

Usefulness of the method was validated on the hepatitis dataset, which is the common dataset in the active mining project. The dataset originally contained 771 cases, but 268 of them were removed from analysis because their virus types were not clearly described. For the remaining 503 cases, we clustered sequences of glutamic-pyruvic transminase (GPT), zinc sulfate turbidity test (ZTT), thymol turbidity test (TTT), albumin (ALB, F-ALB) and total cholesterol (T-CHO). The resultant clusters were then stratified according to the virus type and administration of the interferon (IFN) treatment. For clusters that had interesting compositions, we visually inspected common patterns in those

clusters. As a result, following findings were obtained. Note that each of the following figures represents matching result of two sequences grouped into the same cluster. The matched subsequences are painted in the same color.

1. Clusters on GPT well reflected effectiveness of the interferon treatment. Table 1 shows a part of the clustering result. Two types of interesting patterns were found in the clustered sequences. The first pattern was found in cluster 4, which contained remarkably many cases of type C with IFN (B/C/C(IFN) = 6/3/25). In this cluster, GPT decreased after administration of IFN and then kept flattened at low level (fig.3). This pattern represented cases where interferon successfully suppressed activity of the type C hepatitis virus. The second pattern was found in clusters 1, 5 and 7. In these clusters, GPT had continuous vibrations (fig.4). Since this pattern was commonly observed regardless of virus type and administration of IFN, it implied ineffective cases of IFN treatment.

Table 1: Clusters of GPT sequences

cluster	IFN=N		IFN=Y	total
	B	C	C	
1	24	13	42	79
2	9		7	16
3	44	25	24	93
4	6	3	25	34
5	5	4	6	15
6	1		2	3
7	42	19	31	92
⋮				
44		1		1
total	206	100	197	503

2. Clusters on ALB, F-ALB and T-CHO did not contain interesting patterns. Some common patterns were observed, but were not related to the virus types and IFN administration.
3. Sequences of TTT had continuous vibration in many cases regardless of virus types and IFN admission. Some of them had increasing values.
4. A common pattern was found on the clusters of ZTT (fig.5), although it was independent of virus type and IFN. The peak position did not correspond to the duration of IFN treatment.

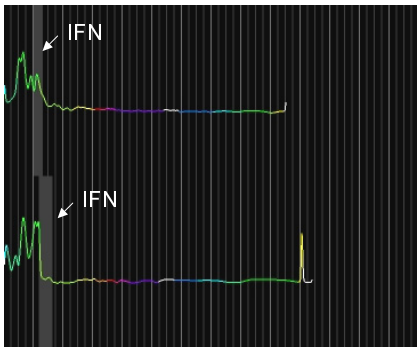


Figure 3: GPT cluster 4: #19(type C; IFN)、#158(type C; IFN)

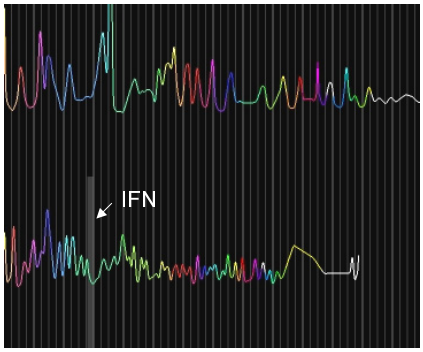


Figure 4: GPT cluster 1: #72(type C; IFN)、#892(type B)

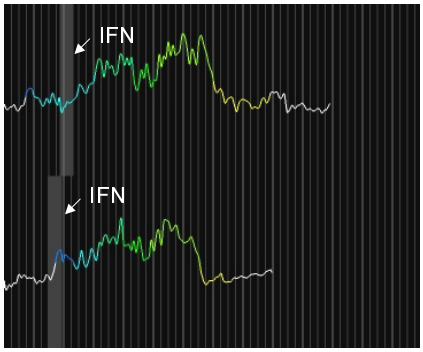


Figure 5: ZTT cluster 1: #642(Type C; IFN)、#645(Type C; IFN)

The findings on GPT sequences were very interesting and were the subject of validation by domain experts. Although ALB etc. did not show interesting patterns in this experiment, they might show some interesting patterns when stratified by the activity of virus and stage of fibrosis because they are associated with degrade of the function of the liver.

Future Plan and Expected Results

The results of the two years research enable us to automatically cluster the time-series sequences and visually recognize their features. They will be the foundation of building an active user reaction system, where the user assigns weights to the attributes for further analysis and data collection. In the next two years, we will construct the tools for manipulating feedback from the users, and establish the user reaction system in medical data mining.

Contact

Shusaku Tsumoto (Principal Investigator)
Department of Medical Informatics, Shimane Medical University
89-1 Enya-cho, Izumo, Shimane 693-8501, Japan.
E-mail: tsumoto@computer.org : Phone：0853-20-2172 : FAX：0853-20-2170

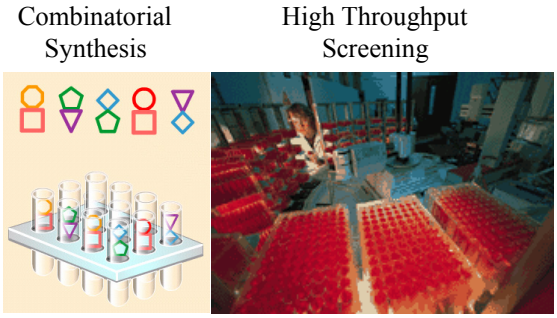
(A03–10–1) Risk Alerts for Chemical Compounds by Active Mining

Principal Investigator　Takashi Okada (Kwansei Gakuin Univ.)
Investigators　Yoshimasa Takahashi, Hiroaki Kato（Toyohashi Univ. of Technology）

Background and Aim

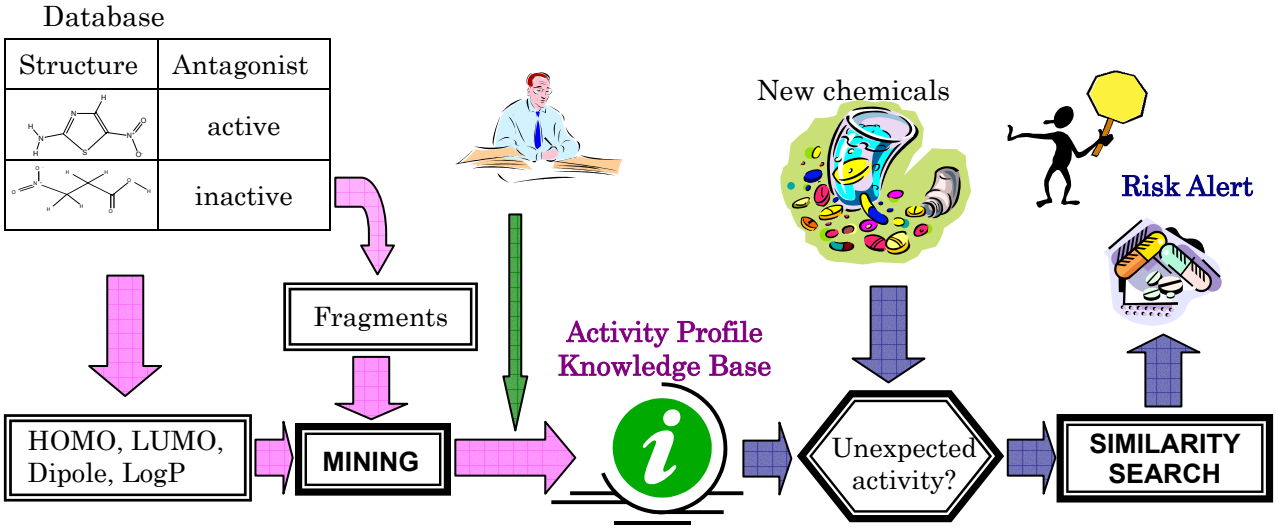
Developments in combinatorial chemistry and high throughput screening enabled the bioassay that handles a million chemicals within a week. Now, researchers of drug discovery are in the flood of data.

Aims of this research are (1) Providing a knowledge base of structure activity relationships by mining this vast amount of data, (2) Detecting a new chemical with an unexpected activity and publishing risk alerts for chemicals with a similar structure.



Research Plan and Approach

The following Figure illustrates steps for the construction of activity profile knowledge base (left) and those for the risk alerts by the similarity search of an unexpected activity (right).



Outline of the plan

- First two years. Consolidation of software infrastructure and the analysis of typical bioactivities are done at this stage. They include mutagenicity, carcinogenicity and dopamine activities.
- Latter two years. A variety of activities cited in MDDR database are analyzed one after another. GPCR (G-protein coupled receptor) related activities are treated at first. Polishing up various structural features including 3D features is also in plan.

Construction of activity profile knowledge base

Past experiences suggested the incorporation of attributes: HOMO, LUMO, dipole and LogP in addition to structural formula itself for the mining task. MM-AM1-geo procedure is employed to estimate the electronic properties, while CLogP is used to predict LogP values.

Methodologies for mining are the **Cascade Model** developed by the principal investigator and **Apriori-based Graph Mining** (developed by Research Plan A02-05). These methods can provide rules that express correlations between biological activities and the combination of various structural features. Linear fragments in a graph are used as a fingerprint representing structural characteristics of the molecule.

Survey of the datascape, viewing characteristics of data from multiple aspects, is necessary in order to invoke active users' reactions. The development of the cascade model is in progress to incorporate this facility. Open environments are planned for the compilation and publishing of knowledge base with expert's comments.

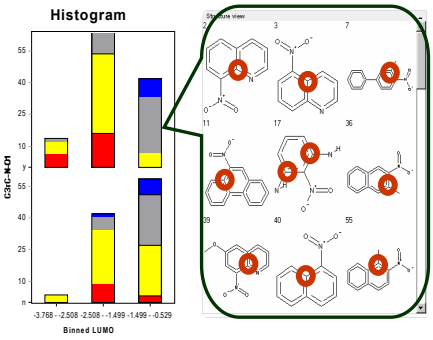
Risk alerts from exceptional chemicals

Attributes employed in this step are the same as those in the previous step. However, structural

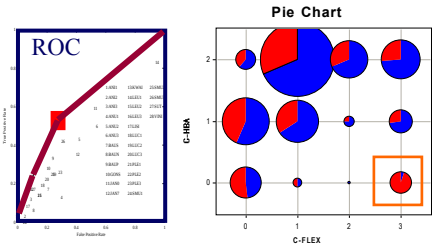
features are expressed by **Topological Fragment Spectra** (TFS) that was proposed by one of the investigators. Similarity search in the space of TFS utilizing mass and LogP is effective in order to retrieve similar chemicals from a few exceptional molecules. Also implemented is a facility in order to inspect the details of structural characteristics contributing the similarity.

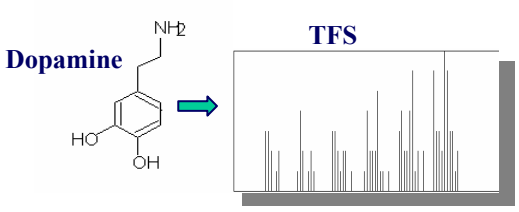
Main Results

Analysis of mutagenicity by aromatic nitro compounds has shown that a higher LUMO (right end in the bar graph) leads to lower activities (blue & gray), and its effects are very sharp in the compounds with *ortho*-substituents as shown in the upper bar chart. Visual inspection of the structures shown at the right has indicated that **the steric hindrance destroying the coplanarity of nitro substituents is responsible for the lower activity.**



Predictive toxicology challenge workshop in 2001 run a **competition to predict the rodent carcinogenicity** for 185 chemicals examined by FDA using the data of 430 chemicals provided by NTP. **Our prediction was ranked first** in its accuracy and in its comprehensibility among 14 research groups. The results of ROC analysis for female rats are depicted in the right figure, where our prediction ■ is close to the upper left corner showing its superiority. The pie charts depict the percentages of active●/inactive● compounds, where the flexibility and the number of hydrogen bond acceptors are used as x-, y-axes, respectively. We can see that flexible molecules without hydrogen bond acceptors (lower right) are mostly active.





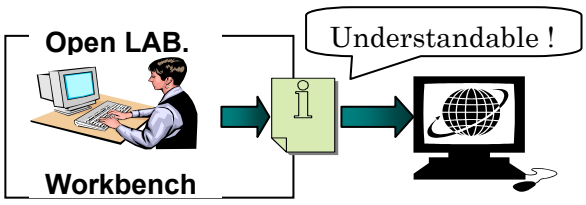
Similalrity search based on TFS was applied to test data set with 3600 chemicals. TFS was constructed from the mass of all fragments (size<6) as shown in the left Fig. Similarity search using Euclidean distance between TFS resulted in **3 active compounds among top 18 retrieved chemicals** when applied to dopamine molecule. It shows the effectiveness of the method in providing risk alerts for the unexpected activities of a drug.

Mining software for the cascade model is now equipped with facilities to **organize rules** into principal and relative rules, and to **detect ridges** in BSS measure. Preliminary application of these facilities has shown that they are **effective in surveying the datascape** by giving comprehensive rule expressions.

Future Plan and Expected Results

The following points are expected to appear.

1. **Construction of workbench.** Efforts to build up a workbench for the efficient mining task are now in progress. An experienced expert in drug discovery contributes to the establishment of a workbench effective in the real pharmaceuticals development process.
2. **Mining from chemicals with a variety of biological activities.** First, compounds related to GPCR proteins are analyzed. They include dopamine, serotonin and opioid. Other activities will also be analyzed in series.
3. **Setting of open laboratory and publication of activity profile knowledge base on the web.** Researchers in the drug development field can utilize all resources in the open laboratory to create knowledge in their interested activity. Knowledge obtained will be open on WWW, the design of which is now in progress.
4. **Improvements on fragment descriptions.** Addition of 3D geometries and branching information to the fragment is in our plan. It will be useful to establish pharmacophores.



Contact:

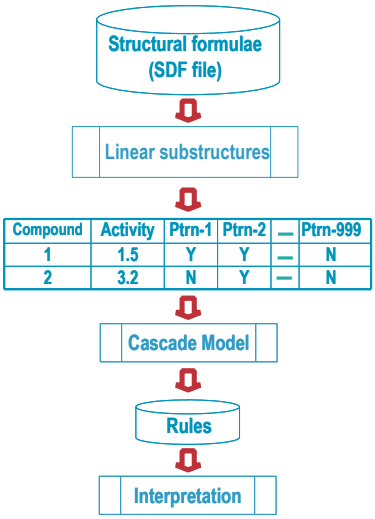
Takashi Okada (Principal investigator)
Center for Information & Media Studies, Kwansei Gakuin University
1-1-155 Uegahara, Nishinomiya, Hyogo, Japan 662-8501
Email: okada@kwansei.ac.jp ; Phone: 0798-54-6042 ; FAX: 0798-51-0913

(A03–10–2) Mining Structural Characteristics of Bioactive Molecules

Principal Investigator Takashi Okada (Kwansei Gakuin Univ.)

Background and Aim

Structure activity relationship (SAR) study is the core problem in the development of new drugs and the prediction of hazardous chemicals. The aim of this research project is to clarify SAR for a lot of biological activities using the modern data mining technology, and to construct the comprehensible knowledge base of SAR's.



Research Plan and Approach

Activities examined in the trial stage are mutagenicity and carcinogenicity. The right figure shows the flowchart of mining using the cascade model, starting from the database of structures and their activities. Descriptors employed are,

- Typical properties such as HOMO, LUMO, LogP
- Linear fragments brought out of structural formulae

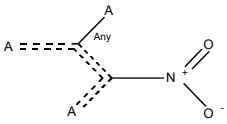
MDDR database contains 120,000 medicines covering 800 activities. Successive examinations of these activities are in our research plan, followed by the compilation and publication of the knowledge base.

Main Results

Mutagenicity of aromatic and heteroaromatic nitro compounds

Mutagenicity data of 230 compounds compiled by Debnath was examined by the cascade model. 38 rules expressed in 3 rule sets have made possible detailed analysis of SAR, aided by the inspection of the supporting chemical structures.

For example, one rule has indicated that a higher LUMO tends to give lower activities, as anticipated. The most powerful rule used higher LUMO as its main condition, but it contained the presence of the following substructure as the precondition.



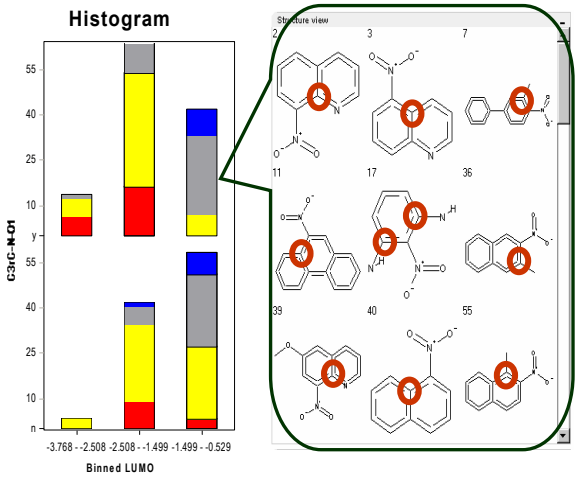
V II

Figure at the right shows the distribution of the activity concerned with the strongest rule. Here, the upper and lower bar charts correspond to the compounds with and without the substructure above, respectively. A bar chart illustrates the distribution of activity levels (■ high– ■ low) using LUMO categories as the x-axis. We can see that the effect of a high LUMO level is stronger in the upper chart. Part of the structures contained in the right most bar of the upper chart are shown at the right. Substitutions at the *ortho* position to the nitro group are observed in all compounds. And we can say that an important factor affecting the mutagenicity is the distortion of the nitro group from the coplanar position by the steric hindrance of the *ortho* substituents.

Other rules has led to valuable pieces of knowledge as shown below,

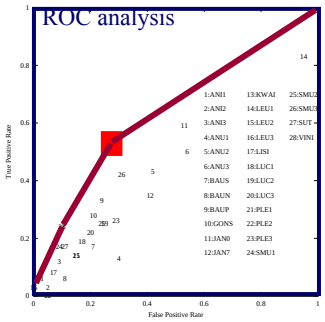
- *p*-Nitrobiphenyls show higher activities, unless they are contained in pericyclic rings.
- Single aromatic ring system attached by NO_2 group shows a lower activity, when the LogP value is low.

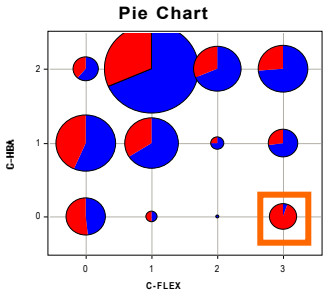
These results were already published as an article in *JCAC*.



Rodent carcinogenicity

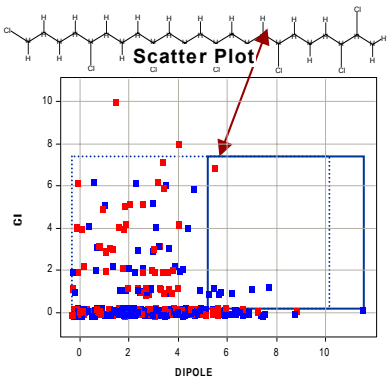
Predictive toxicology challenge workshop was held in 2001, where the problem was to predict rodent carcinogenicity of 185 compounds examined by FDA utilizing the NTP data collected from 430 compounds. Comparison of the predictive accuracies and the comprehensibility of the models by the committee placed our results at the 1st rank among 14 research groups attended. The result of ROC analysis for female rats is shown in the right Figure. In this figure, the point at the upper left corner correspnds to the perfect discrimination, while points on the diagonal line have no capacity of classification. We can see that our result (■) is the closest to the corner showing its superiority.





The left figure shows the characteristic activity distribution indicated by the strongest rule from the male mice data. Here, x- and y-axes show the flexibility and the number of hydrogen bond acceptors respectively, while each pie chart illustrates the ratio of active(●) and inactive(●) molecules. The lower right pie chart illustrates that flexible molecules without hydrogen bond acceptors show very high active ratio. Correlation with other variables and inspection of the structures have indicated that this observation is applicable to halogenated alkanes and alkenes.

Another rule states that the carcinogenicity by chlorinated compounds is restrained if their dipole moments are high. X- and y-axes of the scattergram at the right show the dipole moment and the number of Cl atoms. The solid blue rechtangle is the region indicated by the rule. In fact, almost all molecules are inactive (■) in this region, but there exists an exceptional active molecule at the upper left corner of this region. The compound is a chlorinated linear alkane as shown above the graph, suggesting that its dipole moment is not high at the time of biological interaction. The result was accepted for publication in *Bioinformatics*.



Future Plan and Expected Results

The results so far obtained have proved the capacity of the mining method in the construction of comprehensible SAR knowledge base. Research is in progress in the following directions.

1. **Construction of workbench**. Efforts to build up a workbench for the efficient mining task are now proceeding. Experts in medicinal chemistry can carry out the task by themselves, including the computation of physical properties. An expert in drug discovery contributes to the establishment of a workbench that is effective in the real pharmaceuticals development process.
2. **Application to many biological activities**. First, compounds related to GPCR (G-protein coupled receptor) proteins are analyzed. They include dopamine, serotonin and opioid activities. Others will also be analyzed in series for those cited in MDDR database.
3. **Open laboratory for mining and publication of the knowledge base**. A laboratory will be open to researchers, where they can utilize all resources to create knowledge in their expertise area. Knowledge obtained will be accessible via WWW, the design of which is now in progress. The mining software is already open to public.
4. **Improvements on the capacity of descriptors**. Branching information will be added to the fragment description. Characterization of 3D geometry is also in our plan. These capabilities are expected to lead to clearer mining results leading to precise images of pharmacophores.

Contact:

Takashi Okada (Principal investigator)
Center for Information & Media Studies, Kwansei Gakuin University
1-1-155 Uegahara, Nishinomiya, Hyogo, Japan 662-8501
Email: okada@kwansei.ac.jp ; Phone: 0798-54-6042 ; FAX: 0798-51-0913

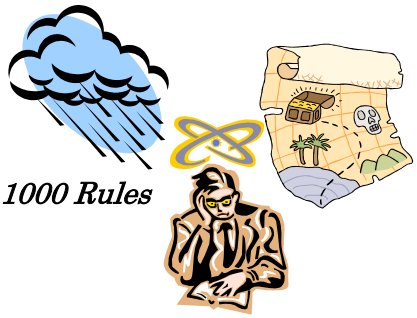
(A03–10–3) Surveying Datascape and New Expression of Rules

Principal Investigator Takashi Okada (Kwansei Gakuin Univ.)

Background and Aim

Mining characteristic rules from data is useful in data analysis. However, there appear thousands of rules by the association rules, a typical characteristic rule miner, and it is difficult to inspect them by an analyst. The methodology used in this project is the cascade model, in which the strength of each rule is described by a *BSS* value (between-groups sum of squares). Ordering rules using *BSS* helps the selection of important rules, but the number of resulting rules is still large.

'Datascape' is a new word proposed in this project. It refers to the image of a scenic view of a data set from the perspective of the analyst. We think that a datascape survey is essential for invoking an active user reaction, and the aim of this research is to provide rules that help users in exploring the datascape.



Research Plan and Approach

Datascape can be obtained by the visualization of data if its distribution is represented in 2 or 3 dimensional space. However, high dimensional data is usually too complex to be illustrated on a single map. Therefore we utilize rules to explore characteristic patterns hidden in data, and to set viewpoints for visualization. The following two functions are introduced for the datascape survey using rules.

- Organization of rules into principal and relative rules
- Detection of ridge area where the *BSS* value drops sharply

We propose the theoretical framework for these aims, and implement the functions in DISCAS software. Application to a medical diagnosis has proved the usefulness of datascape survey guided by the rules.

Main Results

Rule strength in the cascade model is defined by the *BSS* value in the following formulae,

$$TSS = \frac{n}{2}\left(1-\sum_{\alpha} p(\alpha)^2\right)$$

$$WSS^g = \frac{n^g}{2}\left(1-\sum_{\alpha} p^g(\alpha)^2\right)$$

$$BSS^g = \frac{n^g}{2}\sum_{\alpha}\left(p^g(\alpha)-p(\alpha)\right)^2$$

Here, n is the number of cases, and $p(\alpha)$ shows the probability of value α for the objective attribute. Absence (presence) of a superscript g indicates those before (after) the application of the main condition of a rule.

The first step is greedy optimization of main and preconditions of a rule so that its *BSS* takes the maximum value. Many rules are expected to converge to the same expression, causing the decrease in number of rules.

Organization of rules

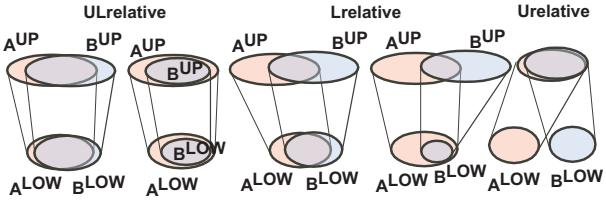
There arises a question whether two rules A➔C and B➔C are independent or not, when most of their supporting cases overlap. We think that these are two different aspects of a single phenomenon, and organize them into a principal and a relative rule. Consequently, we have smaller number of principal rules, and we inspect its relatives when we need to explore the surrounding datascape in detail.

Rlv value defined by the following equation is used to judge the relevance between rules. If it is more than *min-rlv*, two rules are relevant.

$$rlv^{UL}(A,B) = \max\left(\frac{cnt\left(A^{UL}\cap B^{UL}\right)}{cnt\left(A^{UL}\right)}, \frac{cnt\left(A^{UL}\cap B^{UL}\right)}{cnt\left(B^{UL}\right)}\right)$$

where A, B show 2 rules, and *cnt* gives the number of cases of the rule. Values of superscript *UL* are *UP* (*LOW*), indicating the relevance before (after) the application of the main condition.

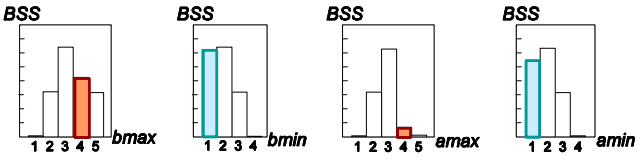
Relative rules can be classified to 3 kinds: *ULrelative*, *Lrelative* and *Urelative* by the values of rlv^{UP} and rlv^{LOW}. Figure at the right shows typical overlaps of cases for these relatives. In these situations, two rules, A and B, can give alternative explanations for the overlapping data.



Sample data (Y: pos/neg)

A	B				
	1	2	3	4	5
1	10/10	14/6	14/6	10/10	10/10
2	14/6	18/2	18/2	10/10	10/10
3	14/6	18/2	18/2	10/10	10/10
4	10/10	6/14	6/14	10/10	10/10
5	10/10	6/14	6/14	10/10	10/10

IF [A: 2-3] added on [B: 2-3]
THEN Y: (0.62 0.38) => (0.9 0.1)



Detection of ridges

Suppose that the problem is the discrimination of *Y* values (*pos*/*neg*) using *A* and *B*. The left figure shows the summary table and the derived rule. Moving the upper boundary, *bmax*, of the precondition [*B*: 2-3], *BSS* value of the rule changes as depicted in (1). The large decrease of *BSS* at *bmax*=4 makes sense as the precedence of [*Y*: *pos*] diminishes at [*B*: 4] in the summary table. A steep drop of *BSS* also appears when the upper boundary of the main condition, *amax*, reaches 4. These regions are called **ridges** and are useful to identify characteristic patterns in datascape survey.

A region, ΔX, in the following formula is defined as a ridge, if $\Delta BSSrate$ surpasses a threshold value. This formula expresses the normalized change of *BSS* when the rule region shifts from X to $X+\Delta X$. n gives the number of cases, and the normalization is incorporated for the change in the number of cases.

$$\Delta BSSrate = \frac{BSS(X+\Delta X) - \left(n(X+\Delta X)/n(X)\right)\cdot BSS(X)}{\left|n(\Delta X)\right|} \Big/ \frac{BSS(X_0)}{n(X_0)}$$

Application to meningoencephalitis diagnosis

Differential diagnosis is important to decide whether the disease is bacterial or viral meningitis. 140 records in a publicly available dataset of this disease (http://wwwada.ar.sanken.osaka-u.ac.jp/pub/washio/jkdd/jkddcfp.html) was analyzed using the new expression of rules.

Starting from 250 rules obtained, optimization procedure decreased the number of rules to 19. After automatic omission of 2 useless rules, 17 rules were organized to principal rules and their relatives. When we set minimum relevance value to 0.7, there appeared 8 principal rules as well as 9 relatives.

The strongest principal rule and two of its relatives shown in the right table indicates high probability of bacterial meningitis. The principal rule, supported by 30 patients, indicates the importance of the count of polynuclear cells. Two preconditions added in the first *ULrelative* rule give us more detailed interpretation of the rule,

Rule	Main condition	Preconditions
Principal	[Cell_Poly > 300]	No
ULrelative	[Cell_Poly > 50]	[AGE > 20] [STIFF > 0]
ULrelative	[CRP > 3]	[SEIZURE = 0] [FOCAL = 0]

applicable to 23 overlapping records with the principal rule. The overlap in the second *ULrelative* rule is only 12 cases, but it gives a completely different interpretation to its segment.

The steepest ridge has appeared in the rule shown below suggesting the bacterial meningitis.

Main condition : [CT_FIND: abnormal]
Predonditions : [FEVER=<6] [LOC=<1] [BT=<39]

The main condition has high correlations with attributes: LOC_DAT, SEX, FOCAL, CSF_PRO too. Here, body temperature [BT≥39] is recognized as a ridge. [CT_FIND: abnormal] is not important in this ridge, as bacterial probability is already high before the application of the main condition.

Future Plan and Expected Results

Facilities of rules organization and ridge detection have enabled an easy extraction of knowledge from rules. In order to help easier survey of datascape, faster computation and the interface agent to the visualization software are in plan.

Contact:

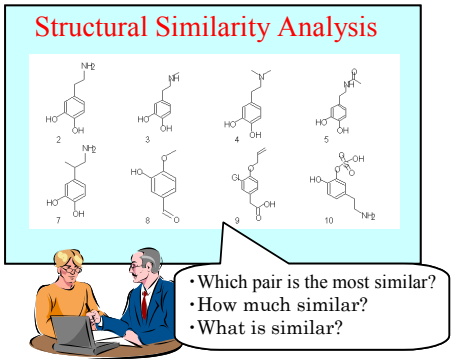
Takashi Okada (Principal investigator)
Center for Information & Media Studies, Kwansei Gakuin University
1-1-155 Uegahara, Nishinomiya, Hyogo, Japan 662-8501
Email: okada@kwansei.ac.jp ; Phone: 0798-54-6042 ; FAX: 0798-51-0913

(A03-10-4) Chemical Data Mining Based on Structural Similarity

Investigator Yoshimasa Takahashi (Toyohashi University of Technology)
Collaborator Satoshi Fujishima (Toyohashi University of Technology)
Kyoko Yokoe (Toyohashi University of Technology)

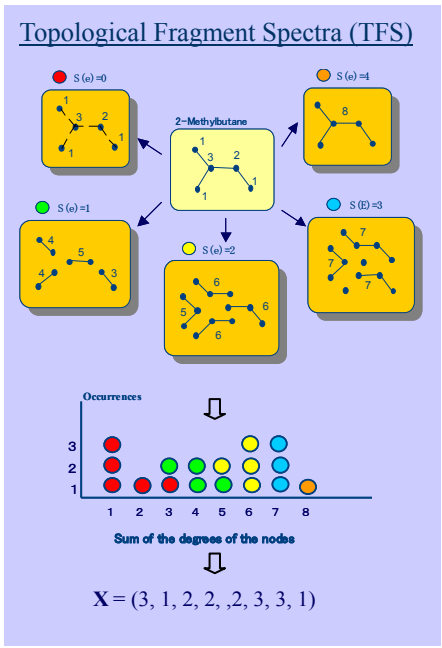
Background and Aim

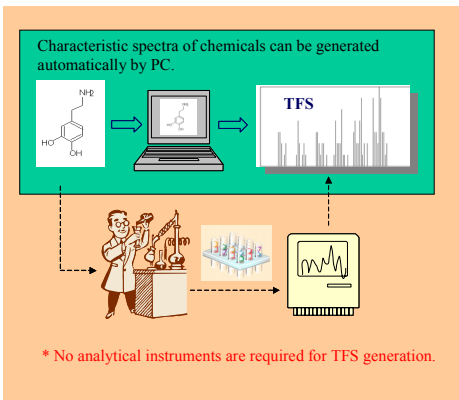
We often say that "A is similar to B" or "C is similar to D in terms of xyz". "Similarity" is very important concept in solving problems in science. This is true in chemistry. The use of molecular similarity methods, especially structural similarity, is under active development in the area of drug design, for the selection of candidate analogs as new chemicals and for the estimation of molecular properties. Nevertheless, the concept of structural similarity is quite important for the further intelligent use of computers in the chemical field. The basic idea behind it is that structurally similar compounds are likely to possess similar molecular properties and similar biological activities. Most of the approaches for the evaluation are based on finding particular functional atoms or atomic groups defined in advance. However, the result of such a structural similarity analysis depends on the chosen set of substructures defined as the descriptors. In that case, an approach is required to process structural information in a more flexible way in order to allow somehow the automatic evaluation of the more ambiguous structural similarity; in other words, a method to examine the similarity of structures when they are regarded as whole entities. The aim of this research project is in establishing a basis of chemical data mining based on structural similarity without any set of substructures defined in advance.



Research Plan and Approach

In the previous work, the authors proposed Topological Fragment Spectral (TFS) method as a tool for the description of the topological structure profile of a molecule. Here we investigate a more flexible way of structure handling based on TFS method and its application to chemical data mining based on structural similarity. The TFS is based on enumeration of all possible substructures from a chemical structure and numerical characterization of them. For a given structure represented as a chemical graph (hydrogen suppressed graph), all the possible subgraphs embedded in it are enumerated. Subsequently, every subgraph is characterized with a specific numerical quantity. To perform the characterization we have used two methods in the present study as follows: (i) the overall sum of degrees of the nodes composing each subgraph. (ii) The overall sum of the mass numbers of the atoms (atomic groups) corresponding to the nodes of the subgraph. With the first method the chemical structure is represented by a simple graph thus the characterization of the structure depends only on the topology of the structural skeleton. An illustrative scheme of the procedure is shown in the figure. For the second method, attached hydrogen atoms are taken into account as augmented atoms and are represented by weighting correspondingly their respective nodes in the graph. This is similar representation of mass spectra of chemicals. It is considered that the TFS is a function of chemical structure. In this project, the applicability of the TFS method will be validated for similar structure-based risk reporting. In addition to this, discrimination of pharmacological activity classes of chemicals would be investigated using artificial neural network with the input signals of TFS descriptors.

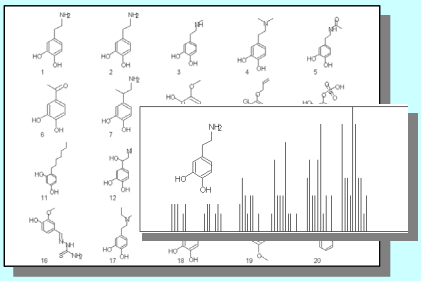




Main Results

Risk report based on structural similarity: To validate an instance -based chemical risk report approach based on structural similarity, TFS-based similar structure searching was employed for identification of active molecular analogues. The TFS database that consists of 3,600 drugs taken from World Drug Index (WDI) was prepared and used for the trial. It was shown that the TFS-based similarity searching gave us successful result for the purpose. For the search trial with a query of dopamine resulted that three of first 20 similar compounds have dopamine activity. The result shows that the TFS is powerful tool for similar structure-based risk report of chemicals.

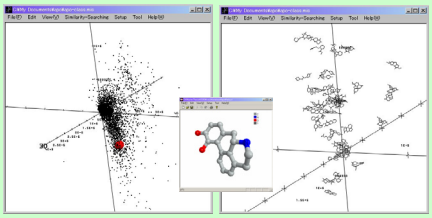
Structural similarity searching based on TFS method



Result of structural similarity searching for a query of dopamine by TFS method. (First twenty similar compounds by Euclidean distance)

Visualization of TFS similarity space: A desktop software tool, MolSpace, has been developed for visualizing massive molecular data space. MolSpace can project a set of massive multivariate data (e.g. TFS data) onto a visual space (2D or 3D space) by means of principal component analysis. MolSpace allows users not only to draw a scatter diagram of the data but also to display their 2D or 3D molecular structures as the objects in the space. With a probe (a molecular object) the user can navigate vast data spaces, thus facilitating understanding of the data structure. In addition, partial space searching is also available that is based on similarity searching techniques. It is possible to interrogate a 3D structure of a chemical compound that corresponds to each object on the space in real time.

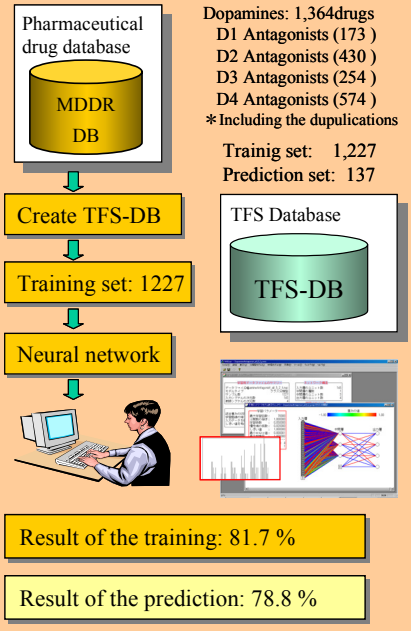
Visualization of massive chemical data space by MolSpace



Three-dimensional Virtual data space of TFS and the partial space near to a probe molecule (red colored point in left side figure)

Classification of pharmacological activity using TFS/ANN: The applicability of the TFS was validated in discriminating active classes of pharmaceutical drugs. Dopamine antagonists of 1,227 that interact with different type of receptors (D1, D2, D3 and D4) are used for training an artificial neural network(ANN) with their TFS to classify the type of action. The ANN classified 87% of the drugs into their own classes correctly. Then, 79% were correctly predicted for a prediction set of 137 prepared in advance. The result shows that TFS is very powerful tool to describe structural information of chemicals and should be suitable as input signal to artificial neural network for the classification of pharmaceutical drug activity.

Classification of pharmacological activity of chemicals by artificial neural networks



Future Plan and Expected Results

From the present work, it is resulted that TFS characterized by the sum of atomic mass numbers works successfully for similar structure searching. Because many instances are required for predictive risk assessment and risk report, more large set of real data should be used in further work. For the purpose, a large size of TFS database of 120,000 pharmaceutical drugs is under preparation, and it would be used to improve the classification model and to find the similar molecules using similar structure searching. Additional system that can be used for identification and interpretation of TFS peaks will be also developed.

Contact:

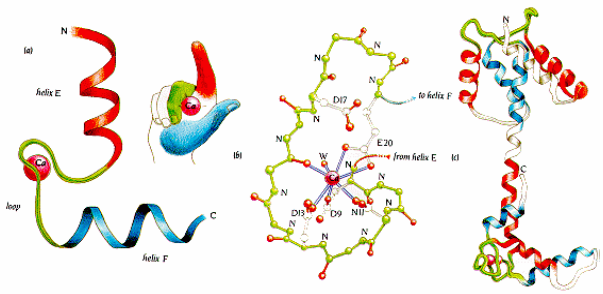
Yoshimasa Takahashi (Investigator)
Department of Knowledge-based Information Engineering, Toyohashi University of Technology,
1-1 Hibarigaoka,Tempaku-cho, Toyohashi 441-8580 JAPAN
Email:taka@mis.tutkie.tut.ac.jp; Tel: 0532-48-6878 ; Fax: 0532-48-6873

(A03-10-5) Construction of a Three-Dimensional Motif Dictionary for Protein Structural Data Mining

Investigator	Hiroaki Kato	(Toyohashi University of Technology)
	Yoshimasa Takahashi	(Toyohashi University of Technology)
Collaborator	Hiroyuki Miyata	(Toyohashi University of Technology)
	Shin-ichi Chikamatsu	(Toyohashi University of Technology)

Background and Aim

With the rapidly increasing number of proteins of which three-dimensional structures are known, the protein structure database is one of the key elements in many attempts being made to derive the knowledge of structure-function relationships of proteins. However, it is almost impossible to manually search 3D local structural features called motifs within proteins (e.g. the figure in right side shows an illustrative example of EF-hand motif observed in calcium binding proteins) because of increasing number and their complex structures. For the reason, computational methods are required for a systematic search for the 3D features of proteins in such a database. The purpose of our research project is knowledge discovery based on three-dimensional structural feature analysis of proteins. The project consists of two major steps: (1) construction of 3D motif dictionary that is corresponded to the sequence motif of PROSITE, (2) systematic extensive analysis of 3D protein structures based on the 3D motif dictionary established.



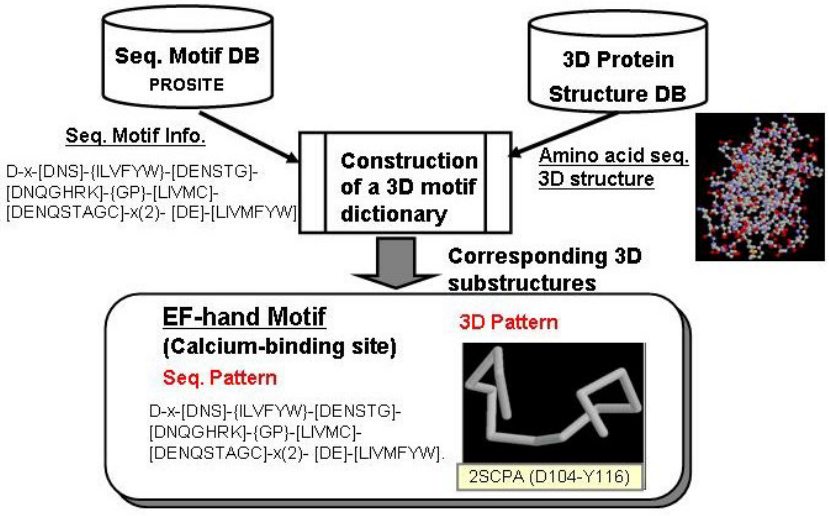
An illustrative example of 3D motif of protein (EF-hand motif)

Example of sequence motif representation in PROSITE

Motif	Pattern
Kringle	[FY]-C-R-N-P-[DNR].
Zinc finger C2H2	C-x(2,4)-C-x(3)-[LIVMFYWC]-x(8)-H-x(3,5)-H.
EF-hand	D-x-[DNS]-{ILVFYW}-[DENSTG]-[DNQGHRK]-{GP}-[LIVMC]-[DENQSTAGC]-x(2)-[DE]-[LIVMFYW].

Researcch Plan and Approach

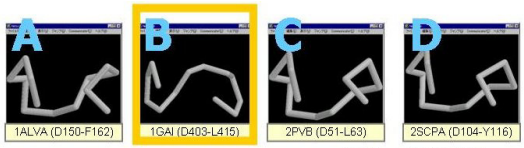
As shown in the Table above, each sequence motif pattern in PROSITE database is described with regular expression. In this research project, using the sequence motif patterns, the corresponding sites of them are extensively explored on the three-dimensional structures of proteins taken from the PDB database. The segments found by the searching are collected for constructing a 3D motif segment database. The 3D segments found with a particular sequence pattern will be clustered on the basis of similarity or dissimilarity of their 3D geometrical patterns. Then, a representative pattern for each cluster obtained will be also identified. Alternatively, in the previous work, the authors reported a computer program for 3D structural feature searching, which allows us to identify all occurrences of a user-defined 3D query pattern or a 3D motif consisting not only of chain-based peptide segments but also of a set of disconnected amino acid residues. More extensive analyses of 3D structural features of proteins will be also done by using our program with the representative patterns determined here. On the basis of the results, the 3D motif dictionary will be refined. The dictionary compiled in this work should be open-to-public for the academic researchers who have interests in the use.



Basic concept of 3D motif dictionary of proteins

Main Results

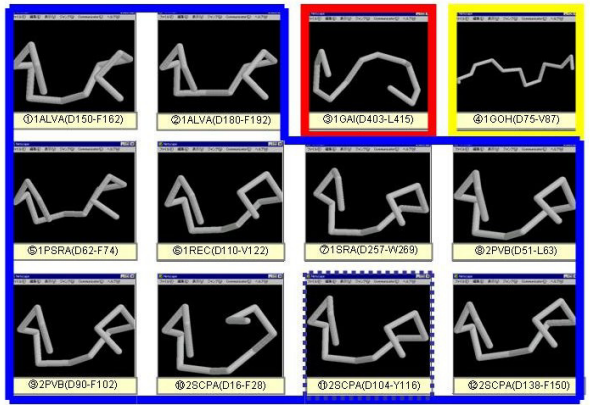
Classification of 3D structural features of proteins: In the present work, we defined the average error of Euclidean distances between the corresponding a-carbons of 3D structure segments to be compared as a measurement of dissimilarity. The dissimilarity matrix was calculated for the 3D substructures that have a particular common structural feature. Using the dissimilarity matrix, the 3D substructures are clustered into several groups. Then, for each group, a representative 3D feature pattern was determined on the basis of minimum variance of the distances between the representative and others. In the figure (right), two clusters are obtained, and B forms a single cluster and the remainder is grouped into another cluster. In the latter case, the segment B is chosen as the representative pattern for the three-membered cluster.



(※ threshold : 100)

	A	B	C	D
A	0	811	64	62
B	811	0	792	786
C	64	792	0	53
D	62	786	53	0

Three-dimensional pattern clustering based on the dissimilarity matrix of the segments.



Three-dimensionsegment corresponded to the EF-hand motif.

Construction of 3D motif dictionary and feature analysis: We have prepared a trial database that involves 902 proteins taken from Protein Data Bank (PDB). For 1,299 sequence patterns that are available on the PROSITE, the 3D features are extensively explored to the trial database. As the result, 464 of the patterns were found in the trial database. For the EF-hand motif described above, twelve segments (including multiple segments on a single protein) corresponded the sequence pattern were found in the database. They are grouped into three clusters (a ten-membered cluster and two single clusters). (See figure in left) The representative pattern of the ten-membered cluster was identified with 2SCPA (D104-Y116). Other two are 1GAI (D403-L415) and 1GOH (D75-V87). But it was realized that the latter two are not true for EF-hand motif site because of missing the typical feature characterized with two helices on the preceding and the following part.

3D Pattern Searching: Three-dimensional pattern searching was carried out for the representative pattern obtained in the above analysis. Several sites that are different from the previous ones were identified. The pattern of 1B47A (D229-F241) shows one of them. (see figure in right) It was realized that the site is true for the EF-hand but that has a different residue in the sequence pattern reported in the PRISITE. The result shuggests that the present approach is quite useful for 3D structural feature analysis of proteins.

EF-hand motif site detected by our program (Sequence pattern: DL**T**CNDYISVFEF)

Future Plan and Expected Results

All the data set of PDB should be used to construct the 3D motif dictionary for practical use in further work. Then, some sort of refinement technique to get more precise information of the related motifs is definitely required. The graphical user-interface for using the dictionary system will be also required too. The authors now are doing investigation on the development of filtering tool to get alternative features of protein too. We believe that the 3D motif dictionary described here will be more and more important in post genomic research to understand structure-function relationships of proteins.

Contact:

Hiroaki Kato (Investigator)
Department of Knowledge-based Information Engineering, Toyohashi University of Technology.
1-1 Hibarigaoka, Tempaku-cho, Toyohashi, 441-8580, JAPAN
Email: hiro@cilab.tutkie.tut.ac.jp ; Tel: 0532-44-6876; Fax: 0532-44-6873

(A03 Planned Research)

Knowledge Evaluation and Selection based on Human-System Interaction

Principal Investigator Yukio Ohsawa (University of Tsukuba)

Investigators Takao Terano and Kenichi Yoshida (University of Tsukuba)

Background and Aim

We study scientific methods for acquiring useful knowledge, i.e., easy to comprehend and smooth to use for achieving user's goal. For having such knowledge established, we *must* (more than *should*) stimulate and have the user, not machine, discover for him/herself. So this purpose, we introduce the effects of the subjective awareness of human(s) and go beyond the frames of KDD (knowledge discovery and data mining) which are confined to a misleading dungeon of objectivity.

Research Plan and Approach

A feature of living bodies is that it has homeostasis and complex flickering, leading to ambiguity in behaviors. For this reason, the genuine process of life is extremely hard to understand, even if the observation is under a perfect condition. In the case of disease, the causal factors are hidden and relevant to each other – the data beyond the understanding of medical doctors become useless.

The text books of medical science are confined to a reductive descriptions on observable causal factors, and it is desired to understand the cases as complex systems. The data-mining approach is well directed to this demand. We aim at two-fold framework of active mining:

(a) Reevaluate existing knowledge of medical doctors and construct a sound and certain structure of knowledge

(b) Discover unnoticed knowledge contributing to jumping evolutions of medial science

As stated later, we aim at the former approach from the aspect of Evidence-Based Medicine (EBM) and the latter approach on the recent methods of Chance Discovery which have been studied and realized for understanding novel events.

In the evaluation and selection of knowledge, the subjective desires and the utilities should be considered. In fact, the bound of "natural" and "meaningless" knowledge contain the most useful knowledge.

(a) Meta-{analysis & learning} for certain knowledge

Since 1995, it has been pointed out that studies rising and negated in the area of immunology come from the problems in the biases of each researcher in the area. For avoiding this, it has been proposed to evaluate 30 or more relevant studies. Hence appeared meta-analysis, an analysis of multiple relevant and conflicting studies.

In meta-analysis, multiple studies for validating the same proposition are unified into a new overall direction by integrating the statistical values in the results. However, the complex structure of knowledge acquired by KDD methods cannot be dealt in exiting meta-analysis methods. For coping with this obstacle, we introduce meta-learning to be resolved with meta-analysis. In meta-learning, different sets of training data are used to lean multiple knowledge and their integration realizes globally meaningful knowledge across various institutes and researchers. It is inevitable that the variety of sampling causes the bias in each study, but meta-learning is giving a light to the way to progress. Our approach hear is to integrate meta learning and meta-analysis for acquiring certain and sound knowledge medical doctors can use.

(b) Chance Discovery for new awareness on rare events

Studies on Chance Discovery (CD) have been contributing to aiding human awareness of such chances, i.e., events significant for decision-making. The symptom in the initial step of disease is rarely observed but significant for the diagnosis. The process of chance discovery has the state transitions of human mind i.e. the

context-shifting cycles carrying human from *concern*, *understanding* of the chance, and *decision/actions*, returning to a new concern of another chance. Data mining is positioned as a stimulating tool of this spiral process.

The latest model of chance discovery, called *double helix*, has two helical sub-processes as in Fig.1. One is the spiral process of chance discovery by human, which substantially is of the cycles above into deeper awareness. The other helix is the process of computer(s), which receives and mines the data ("DM" in Figure 1). The simultaneous run of these pair of helixes is for monitoring "the subject(s)-data," words in the thought of the subject.

Main Results For approach (a), we integrated meta-learning and structural equation modeling enabling us to describe flexible models from background knowledge. Through the empirical study, we have got the conclusion that the framework of meta-learning is achieves a comprehensible and have good predictive performance.

For approach (b), the double helix model, for the process in which human discovers what we call chances, has been exemplified for cases of social survey. We also ran the double-helical processes for other areas as biology, seismology, and marketing. For the time being, we are applying the CD method to medical data. Figure 2 is the result on the way, showing the relations between items to be checked in inspections of liver. Here we find CHE, T-CHO, NA etc. as signs of the increase in GPT. This trend could not be seen for patients without GPT-increase.

Future Plan and Expected Results We are applying the method of chance discovery above, for various domains. In the case of medical data, we should finally go to a objectively correct knowledge as well as useful. On the other hand, the market data should create more *active* knowledge which can change reactively for fitting customers demands. Looking at such variety, we are developing a generalized framework of active mining.

Contact: Yukio Ohsawa, Associate Professor,
University of Tsukuba Tokyo 112-0012 Japan,
Fax: +81-3-3942-6829

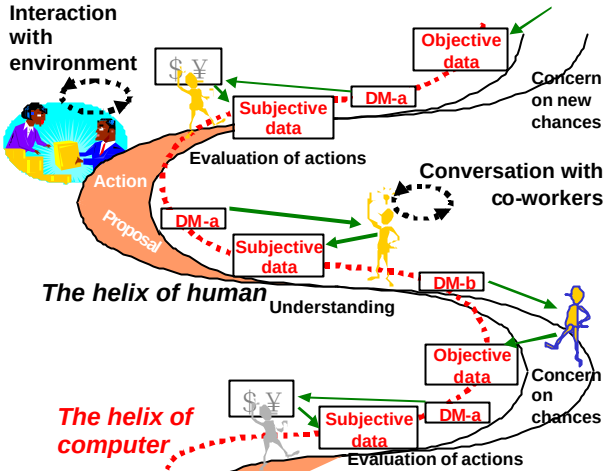
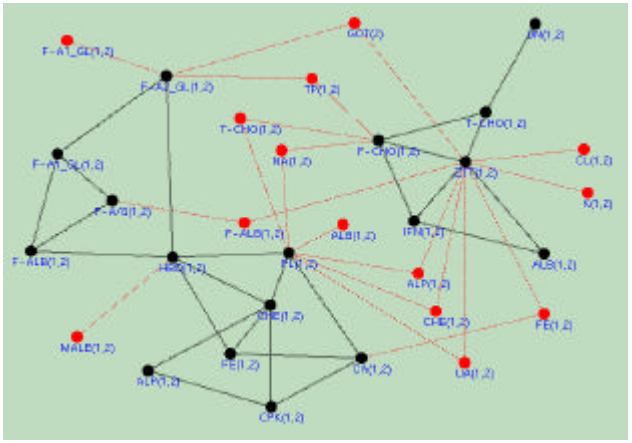
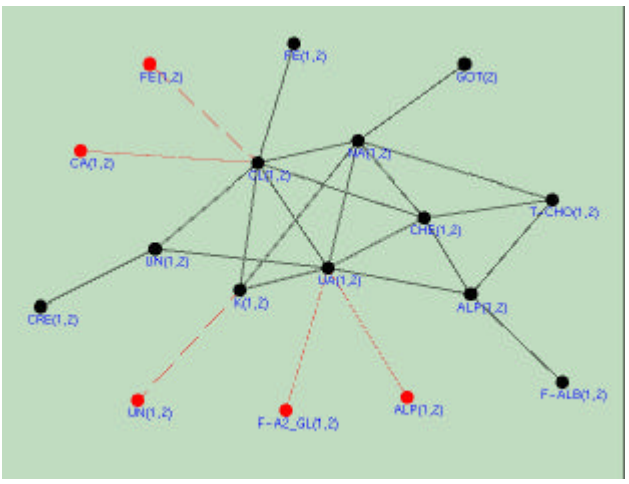


Fig.1 The double-helix process of chance discovery



a. Before the increase in GPT



b. After the increase in GPT

Fig.2 χ-KeyGraph applied to patient data before and after the increase of GTP values.

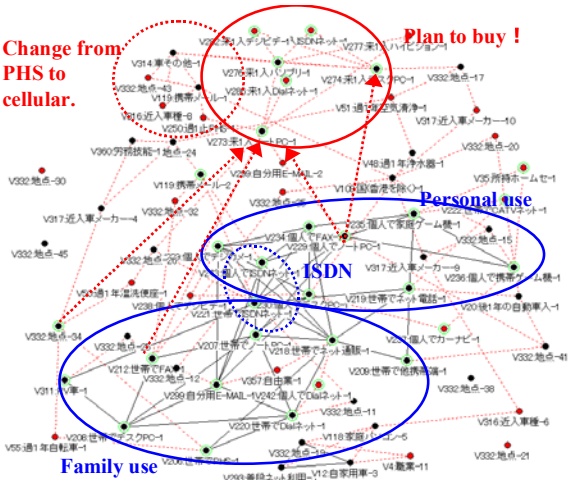
(A03-11-1) Data Visualizer for Chance Discovery

Principal Investigator Yukio Ohsawa (University of Tsukuba)

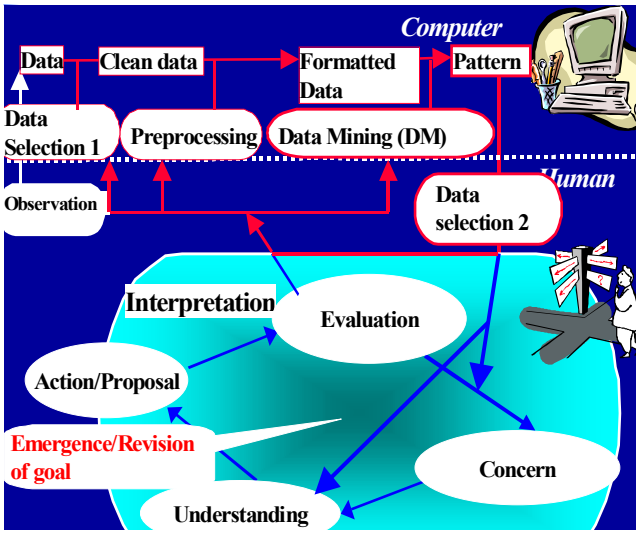
Background and Aim

Similar to meritorious books tending to be difficult to understand, a useful piece of information is not always easy to understand. Rather, a more important aspect is to develop a process for understanding the significance of information on sudden events in face of its difficulty, as well as we do in the case of catching the underlying meaning of important books fully using his/her own ability of imagination. As a result, we can obtain comprehensible translation of the source information, typically expressed in a simple manner.

On the other hand, the misguided belief that the goal of user must be explicitly defined before introducing data-mining has been disturbing user from taking the full advantage of the data. That is, data has a power to help user be aware of the user's own goal as far as the data was acquired following his/her own implicit goal. This study is dedicated to developing a new mental process of human from the concern on, acquisition of, the understanding of, and the action on data, a cyclic process calling deepened cycles.



Data Visualizer: An application implementing some tools for visualizing the structure among events, on their co-occurrence. This figure shows a case where user and machine co-operatively discovered a chance of marketing, i.e., if old people changes from PHS to cellular, he/she is inclined to buy new IT tools.



Chances come to a prepared mind: In the process of chance discovery, as well as the discovery of significant signs of events, the user's concern on the chances is more essential than other factors. The Data Visualizer does not create anything for themselves, but only supports the cycles.

Research Plan and Approach

Based on user's own potential concern, he/she gathers corresponding data and understands the interesting part of the data from the result of data mining. The data visualier developed in this study aids in human creativity to discover chances and make actions on the discovery. Data visualizer is a visualization tool for showing the structure of event-relations underlying the data, and we are also developing an HCI framework for accelerating the cyclic model of chance discovery by human, using the data visualizer. Among others, we are dealing with the chance discovery of a group of people.

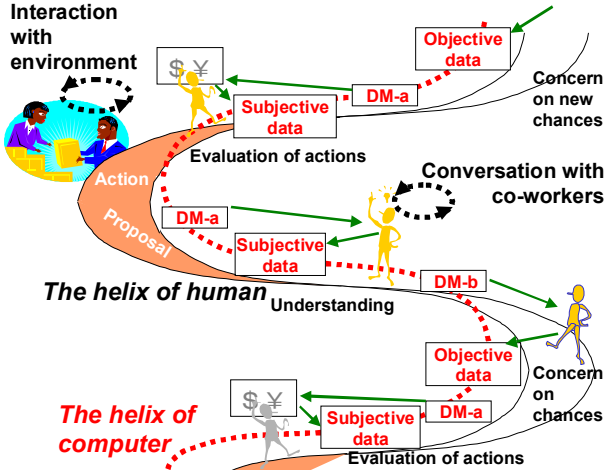
The most significant point for realize this is to pay enough research attention to *subjective awareness* of user, rather than objective generalization of knowledge, i.e., knowledge *useless* in any particular situation.

In medical science, it is desired to understand human bodies complex systems. We aim at two-fold framework of active mining: (a) Reevaluate existing knowledge of medical doctors and construct a sound and certain structure of knowledge, and (b) Discover unnoticed knowledge contributing to jumping evolutions of medial science. Here in particular, I aim at the goal (b).

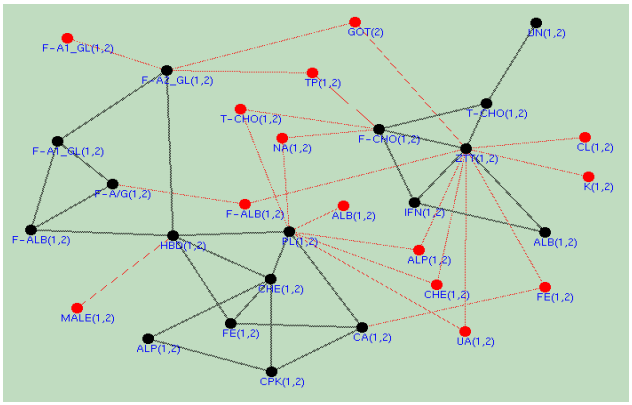
[Chance Discovery for new awareness on rare events]

Studies on Chance Discovery (CD) have been contributing to aiding human awareness of such chances, i.e., events significant for decision-making. The symptom in the initial step of disease is rarely observed but significant for the diagnosis.

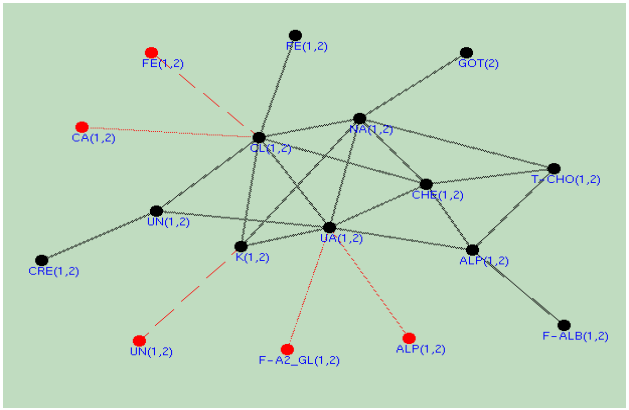
The latest model of chance discovery, called *double helix*, has two helical sub-processes. One is the spiral process of chance discovery by human, which substantially is of the cycles shown above, i.e., from concerns into actions. The other helix is the process of computer(s), which receives and mines the data ("DM"). The parallel running of these pair of helixes is for monitoring "the subject(s)-data," words in the thought of the subject.



The double-helix process of chance discovery



a. Before the increase of GPT



b. After the increase GPT (NA, CHE, T-CHO increased)

Main Results Following the double helix model, for the process in which human discovers what we call chances, has been exemplified for cases of social survey. We also ran the double-helical processes for other areas as biology, seismology, and marketing. For the time being, we are applying the CD method to medical data. The left-hand figure is the result on the way, showing the relations between items to be checked in inspections of liver., by χ-KeyGraph applied to patient data before and after the increase of GTP values. Here we find CHE, T-CHO, NA etc. as signs of the increase in GPT. This trend could not be seen for patients without GPT-increase.

Future Plan and Expected Results The hurdle currently most hard to overcome is the interaction speed between the busy medical doctors and I, another busy in to be involved in the communication for running the double-helix. The next most important step for us is the design of a new interface for accelerating the human-human interactions on the process of chance discovery. The visualization effect of Data Visualizer is the main key to open the way to this significant step.

Contact: Yukio Ohsawa, University of Tsukuba
Tel: 03-3942-7141 Fax: 03-3942-6829,
e-mail: osawa@gssm.otsuka.tsukuba.ac.jp

(A03-11-2) Developing Human-in-a-Loop knowledge Validation Methodology

Investigator Takao TERANO (University of Tsukuba)

Background and Aim

In this research, we will develop a novel framework: Knowledge Validation with Human-in-a-Loop, by introducing users' interaction with mining systems so as 1) to improve the performance of total active mining processes and 2) to evaluate the validity of the acquired knowledge. Figure 1 shows the basic idea of the framework.

To achieve the aim, we will implement the methodology and the corresponding support tools, then apply these methodology and tools to practical application domains. The domains include continuous bio-chemical plants, where we must have the behaviors engineering systems and medical data analysis, where deep heuristic knowledge is necessary to understand the results.

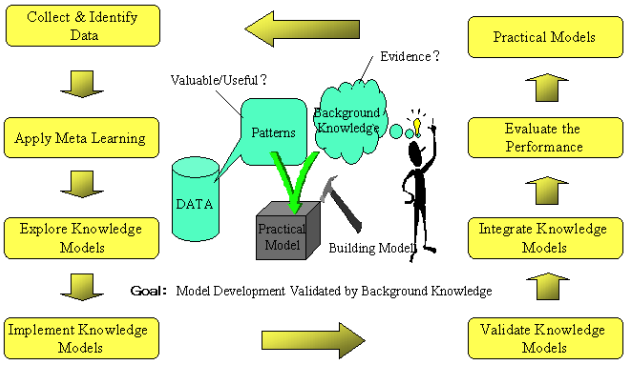


Figure 1 Framework for the Knowledge Evaluation Methodology

Research Plan and Approach

From theoretical and technological viewpoints, we will research machine learning methods with evolutionary computation and the integration of statistical techniques for hypotheses testing and exploratory data analysis.

About the machine learning, we will focus upon Learning Classifier Systems and Interactive Evolutionary Computation, both of which are promising ones reported in the literature. A Learning Classifier System is a integrated problem solving architecture with 1) production systems for rule ore classifier execution, 2) reinforcement learning for rule tuning, and 3) genetic algorithms for rule generation, Interactive Evolutionary Computation is a framework to introduce users' interaction to the selection process of Genetic Algorithms. Both concepts will be good theoretical and technical candidates to achieve the active mining.

About statistical contexts, we will focus on the methodology to develop causal models from plural data sources, human experts, and background knowledge. We divide the modeling process to meta-learning phase and model implementation phase.

In the meta-learning phase, we explore the predictability of implicit target concepts in given data sets using linear regression, artificial neural networks, inductive learning and so on as mining techniques. In the model implementation phase, we will develop corresponding causal models based on the results of the previous phases and background knowledge by utilizing hypothesis testing type co-variance structural analysis methods.

We will target the following two practical domains.

First, we will implement a novel process response model from a large amount of time series data obtained from a continuous chemical plant. Based on this, then, we will develop process prediction models to predict future status of a give plant from currently observed data. At the same time we will develop heuristic methods to explore plant control rules to give operation guidance to a plant operator.. The basic ideas of the developments are maximization of correlation coefficients among time series data, time series prediction by artificial neural networks, and association rule classier learning by a Learning Classifier System with Minima Description Length and Rule Improvement Ratio evaluation functions.

Second, using common clinical data sets on the whole project that reflects the liver function for a hepatitis data base, we will focus on the result of the ICG inspection (indocyanine green test). This

domain requires medical statements to interpret the meanings for the predicting model of the inspection data. To answer the requirement, users must positively interact the model development processes.

Main Results

1) Application to Plant Time Series Data

The proposed technique has been applied to the time series data obtained a continuous bio-chemical plant shown in Figure 2.

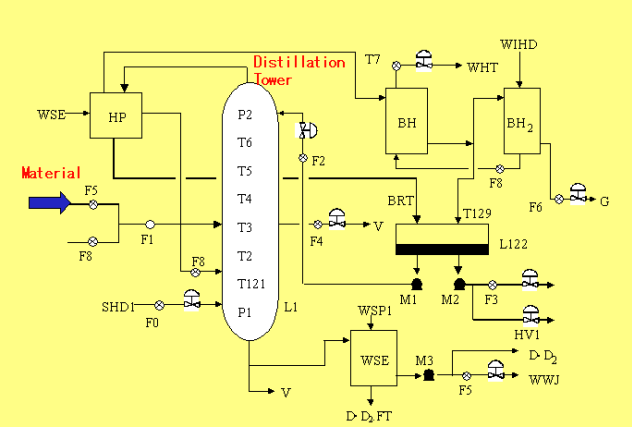


Figure 2 Plant Configuration of the Domain

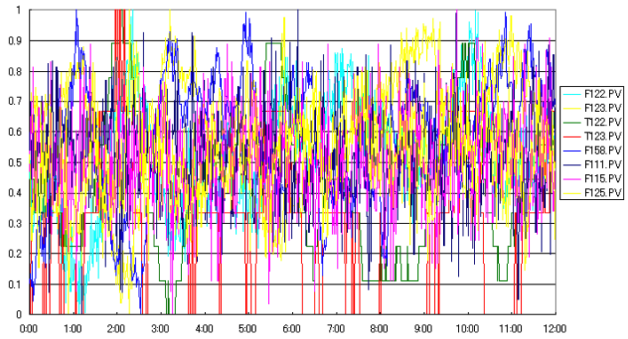


Figure 3 Plant Time Series Data

The time series data is so complex shown in Figure 3 that conventional prediction methods are impossible to apply without fine physical plant models. The application has derived the following simple and comprehensive knowledge:

If 25% < F3 <= 50% and 75% < F4 and F3 is down Then 75% < T2 (If F3 flow is between 25% and 50% and F4 flow is over 75% and F3 flow is down , then T2 temperature becomes 75% or more.)

2) Application to Liver Function Data

We have applied the proposed method to develop a causal model from clinical examination data and corresponding background knowledge related to the capacity of the liver functions. Here, we have assumed that we will develop the causal model to practical use for predicting the result of the ICG inspection (indocyanine green test) from the other inspection data. The experiment has suggested that the explanation power is enough high using the simple model in 4 developed from the proposed method.

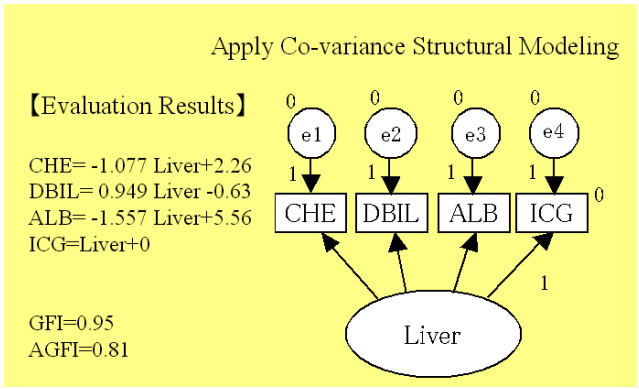


Figure 4 Acquired River Function Prediction Model

Future Plan and Expected Results

Future plan of the research include 1) theoretical one for Learning Classifier Systems and its further application to the other domains, 2) development of interactive evolutionary computation methods for plural users, and 3) development of modeling methodology with a meta-learning architecture and plural domain experts. We believe the results are promising.

Contact:

Takao Terano, University of Tsukuba

Tel: +81-3-3942-7141 Fax: +81-3-3942-6829,

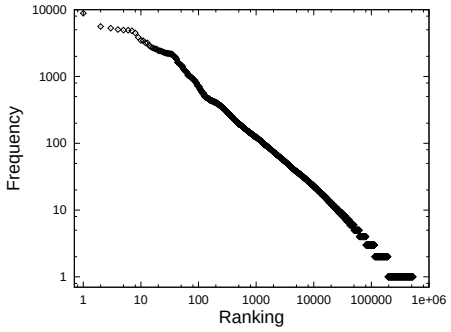
e-mail: terano@gssm.otsuka.tsukuba.ac.jp

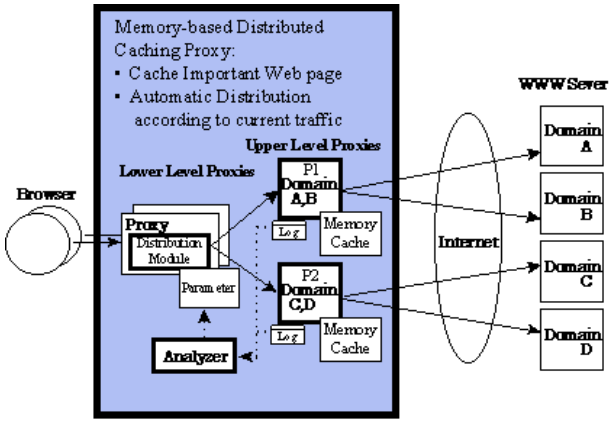
(A03-11-3) Decision Supports of Internet Operations

Principal Investigator	Yukio Ohsawa	(G.S.B.S., University of Tsukuba)
Investigator	Kenichi Yoshida	(G.S.B.S., University of Tsukuba)

Background and Aim

Internet operations require careful observation on daily network traffics. Operators have to find abnormal events based on the knowledge they acquire from daily observations. They also tune network equipments to handle traffics well. For example, if they find a fact: "WWW access patterns follow Zipf's raw and have strong lean toward famous URLs.", skilled operators setup distributed caching proxys so that proxys cache famous URLs well and give high hit rate of the cache.



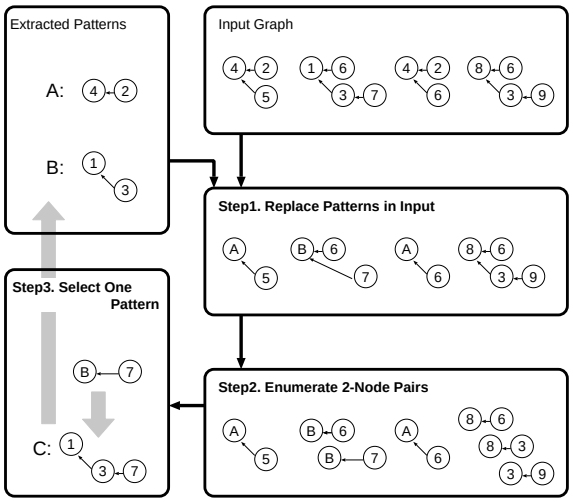


Although the internet popularization increases demands for such skilled operators, limited human resources sometimes keep the network conditions apart from ideal status. However, keeping internet, i.e., an important social infrastructure, in good condition is crucial to increase the productivity of the information-oriented society. In this research project, we try to establish decision support method for the internet operation to improve this productivity of the information-oriented society.

Research Plan and Approach

To realize decision support system of internet operation, we focus on the following two technologies:

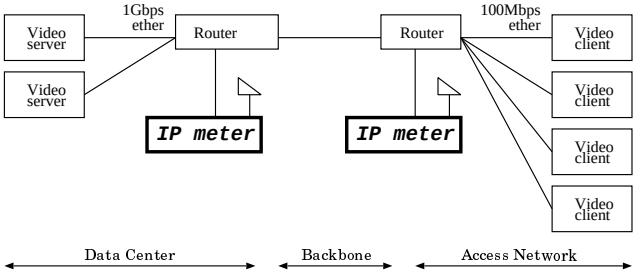
- The analysis of traffic data is important for the internet operation. The important characteristics of traffic data is its graph structure. To analyze traffic data, we have to investigate the method to extract characteristics from data with graph structure.
- Unfortuately, the measurement technology of the internet traffic data has a room of improvement. We also have to investigate the internet measurement technology.

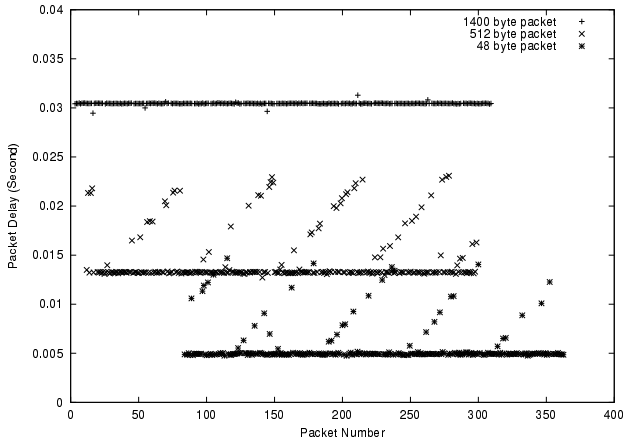


From the view point of the internet operation, the multimedia traffic increases its volume and the analysis of the time delay of multimedia traffic becomes important. We choose this new field, i.e., the decision support for multimedai application operation, as the target of our research. Moreover emerging new application of internet requires a feedback loop of observation, analysis, and operation. We will investigate a decision support system which support such feedback process.

Main Results

As the first step towards the operation support system, we have investigated a technology for packet delay measurement. We showed that we can acquire one path delay with micro second unit. Since the traditional methods could only measure round trip delay with milli second unit, we can acquire accurate data with this new measurement system.



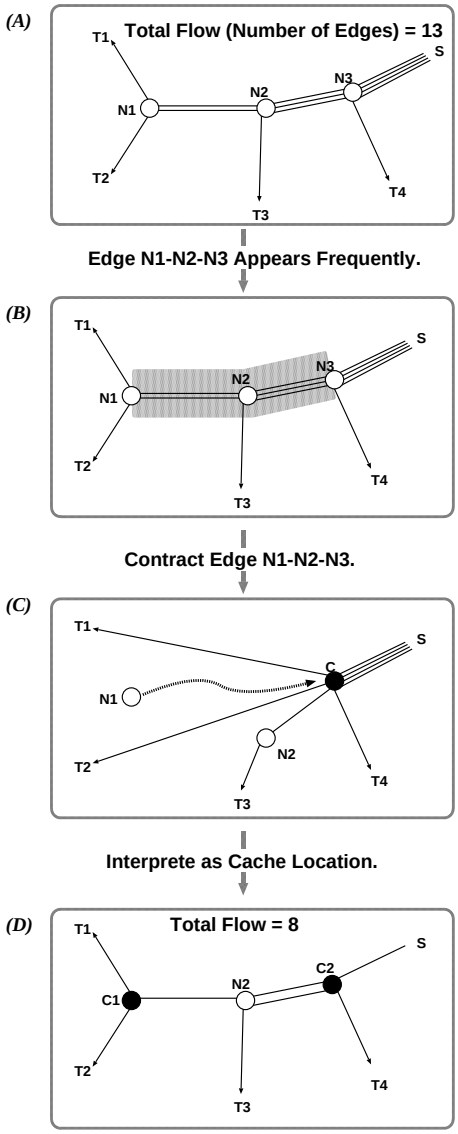


Left figure shows a typical example of measured data. To make the left figure, we have measured the packet delays from site A to another site B which is about 50 Km away from site A. Although the delay of reverse direction is roughly constant, the delay of shown direction has strange jitter. Since this type of jitter has strong effects on multimedia applications, we have to analyze the phenomena further to find the best operation under this network conditions.

Future Plan and Expected Results

Future plan of this research project includes the investigation on the feedback method of the measured delay data toward the tuning of network equipments. We have developed a method to setup WWW distributed proxies based on the WWW traffic. GBI is used to layout proxies based on the WWW traffic.

Based on the traffic analysis methods we developed, we will combine the delay analysis. The measurement technologies we have investigated in the first stage of this project give the input for this analysis phase. GBI will be used to extract important multimedia traffic patterns. Measured delay will be used to select the patterns which result abnormal events. Finally will investigate methods to tune-up network equipments to compensate the found abnormal events, We also improve GBI method so that it can find important traffic patterns of multimedia traffic.



Contact:

Kenichi Yoshida (Investigator)
Graduate School of Business Science, University of Tsukuba,
Otsuka 3-29-1, Bunkyo, Tokyo 112-0012, Japan
Email: yoshida@gssm.otsuka.tsukuba.ac.jp ;
Tel: 03-3942-6982; Fax: 03-3942-6829