Implementation of Discovered-Rule-Filtering System

Yasuhiko Kitamura† and Yatsuho Shibata†

Data mining systems semi-automatically discover knowledge by examining large volumes of data, but the knowledge discovered is not always novel to users. We proposed a discovered-rule-filtering approach that uses information retrieval results from the Internet to assess rules discovered by data mining and find those that are novel to the user. To realize this approach, we developed 2 methods: the micro view method and the macro view method. In the micro view method, we extract keywords from each discovered rule and rank the rule referring to the number of hits returned when the keywords are submitted to an appropriate database. In the macro view method, we refer to the number of hits by submitting every pair of extracted keywords and form keyword clusters according to the results. We developed a discovered-rule-filtering system called DRFS which employs the two methods. DRFS can rank discovered rules in the field of medical domain according the results of information retrieval from the MEDLINE database.

1. Introduction

Active mining[1)] is a new approach to data mining; it tries to discover "high quality" knowledge that meets the user's demand in an efficient manner by integrating information gathering, data mining, and user reaction technologies. This paper introduces a discovered-rule-filtering method[2)~5)] that extracts from the large number of rules output by a data mining system a small number of novel rules by using information retrieved from the Internet.

Data mining is an automated method to discover useful knowledge by analyzing large volumes of data mechanically[6)]. Generally speaking, conventional data mining methods try to discover statistically significant patterns from a large volume of raw data contained in a given database. Unfortunately, considering only statistically significant features yields a large number of rules, most of which are already known to the user. To cope with this problem, our discovered-rule-filtering approach winnows the large number of rules returned by a data mining system to find the small number of rules that are novel to the user. To judge whether a rule is novel or not, we utilize an information source on the Internet and judge its novelty according to the number of retrieved documents that relate to the rule.

This paper shows the concept of discovered-rule-filtering and two methods for discovered-rule-filtering: the micro view method and the macro view method in Section 2. We then show an implementation of discovered-rule-filtering system in which two methods are implemented in Section 3.

2. Discovered Rule Filtering

The target of our active mining project is a clinical examination database of hepatitis patients[7)] and an example of discovered rule is like "If, for the past 24 months, D-BIL stays unchanged, TTT decreases, and GPT increases, then GPT will decrease for the following 24 months," which shows a relation among GPT(Glutamic Pyruvic Transaminase), TTT (Thymol Turbidity Test), and D-BIL (Direct Bilirubin)[8)].

When a set of discovered rules is output by a data mining system, the discovered-rule-filtering system first retrieves information related to the rules from the Internet and then filters the rules based on the information retrieval results. In our study, we are interested in discovering knowledge in a hepatitis clinical database, and it is not easy to gather information related to hepatitis from the Internet by using a naive search engine because Internet information sources generally contain a huge amount of various and noisy information. We use the MEDLINE (MEDlars on LINE) database as the information source since it is the largest bibliographical database in the medical and biological domain. PubMed (http://www.ncbi.nlm.nih.gov/entrez/query.fcgi)

† Kwansei Gakuin University

is a free MEDLINE search service on the Internet run by NCBI (National Center for Biotechnology Information).

The discovered-rule-filtering process consists of two steps.

2.1 Extracting keywords from discovered rules

We need to develop a set of proper keywords to retrieve MEDLINE documents that relate to a discovered rule. Such keywords are extracted from the rule and the domain of data mining as follows.

2.1.1 Keywords directly extracted from discovered rules

These keywords represent attributes in discovered rules. For example, keywords that can be extracted directly from the discovered rule shown before are GPT, TTT, and D-BIL because they explicitly appear in the rule. If any abbreviation is not acceptable to Pubmed, it is translated into its full normal name. For example, TTT and GPT are translated into "thymol turbidity test" and "glutamic pyruvic transaminase", respectively.

2.1.2 Keywords related to the mining domain

These keywords represent the purpose or the domain of the data mining task. Together with keywords extracted from the rule, they are submitted to the Pubmed as the common keywords to improve the quality of retrieved documents. For hepatitis data mining, "hepatitis" is a domain keyword. The domain keywords are implicit keywords, and we do not directly refer to such keywords hereafter.

2.2 Filtering discovered rules

How to filter discovered rules according to the search result of MEDLINE document retrieval is a most important issue of this work. We have two methods; the micro view method and the macro view method, to filter discovered rules[4].

2.2.1 Micro view method

The micro view method retrieves documents directly related to a discovered rule. It utilizes the document retrieval results not only to filter the discovered rules, and shows the rules and the documents to the user. This allows the user to expand her insights on the rule and the data mining task[2]. The micro view method is quite simple and is based on the following hypotheses.

[Hypotheses] (Micro View Method)

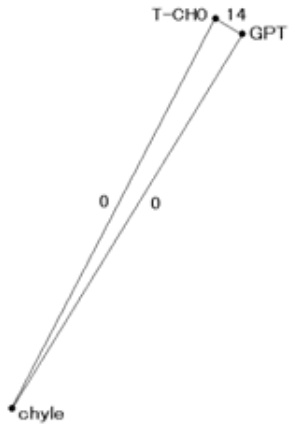

Fig. 1 Keyword co-occurrence graph.

(1) The number of documents related to a known rule is large.
(2) The number of documents related to an unknown rule is small.
(3) The number of documents related to a garbage rule is zero

We hypothesize that known rules have been the subject of many papers. On the other hand, unknown rules have been the subject of only a few papers. Nobody has any interest in garbage or nonsense rules, and so no papers are related to them.

2.2.2 Macro view method

The macro view method tries to roughly observe the research trends implicit in each discovered rule. Given a rule, it submits every pair of keywords extracted from the rule, not the whole sequence of the keywords, to the Pubmed system, and integrates the results in the form of a keyword co-occurrence graph to judge the novelty of the rule.

Fig. 1 shows a keyword co-occurrence graph. In a graph, a node represents a keyword and edge length represents the inverse of the frequency of co-occurrences of the keywords connected by the edge. The score attached to the edge represents the frequency of co-occurrence. Hence, the more documents related to a pair keywords are retrieved from Pubmed, the closer the keywords are located in the graph.

For example, Fig. 1 shows that the relation between T-CHO and GPT is strong, but that between chyle and either of T-CHO or GPT is rather weak.

We then form clusters of keywords by us-

ing the Hierarchical Clustering Scheme[9]. As a strategy to form clusters, we adopt the complete linkage clustering method (CLINK). In the method, the distance between clusters A and B is defined as the longest among the distances of every pair of keywords in cluster A and a keyword in cluster B. The method initially forms a cluster for each keyword. It then repeatedly merges clusters within a threshold length into one or more clusters.

We consider that the number of clusters is strongly related to research activity as follows.

[Hypothesis] (Macro View Method)

(1) The number of clusters concerning a known rule is 1.

(2) The number of clusters concerning an unknown rule is 2.

(3) The number of clusters concerning a garbage rule is more than 3.

A rule with only one cluster is regarded as a known rule because a large number of papers that use all pairs of keywords in the rule have been published. A rule with two clusters is regarded as an unknown rule. This is because each cluster represents a lot research activity, but little cross-cluster research has been done. A rule with more than two clusters is regarded as a garbage rule. Such a rule is too complex to understand because the keywords are partitioned into many clusters and the rule consists of many unknown factors.

For example, if we set the threshold of CLINK to 1 (the frequency of co-occurrences is 1), the rule in Fig. 1 are merged into two clusters; one cluster consists of GPT and T-CHO and another consists of chyle only. Hence, the rule is judged to be unknown.

A preliminary evaluation of the micro view method and the macro view method is discussed in[5].

3. DRFS: An Implementation of Discovered-Rule-Filtering System

We implemented two methods, the micro view method and the macro view method, for discovered-rule-filtering into a Web-based system called DRFS which is accessible through the Internet.

The discovered-rule-filtering process in DRFS is summarized as follows.

(1) We first register rules discovered by data mining systems into the DRFS.

(2) DRFS then retrieves bibliographical information concerning registered discovered rules.

(3) DRFS ranks rules depending on the results of information retrieval.

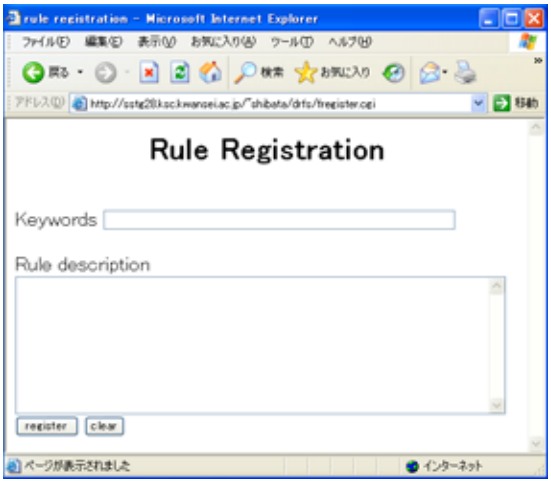

Fig. 2 Rule registration page.

We discuss the detail of each process in the following subsections.

3.1 Registering discovered rules

In the first stage, we register rules to filter or to rank which are discovered by data mining systems and we have two ways to do so. In the first way, which is suitable for connecting between a data mining system and DRFS, rules are registered into the DRFS through a file. Each rule is represented as a couple of the keywords list and the rule description.

Currently a keyword list to be submitted to the Pubmed is made manually by extracting attribute keywords from a rule description. The current version of DRFS handles only the conjunctive form of attribute keywords. Actually a rule represents some relations among the attribute keywords, but it is difficult to represent the relations as a query to the Pubmed at present. The issue is left as future study.

In the second way, rules can be registered through the registration page as shown in Figure 2. Users can register rules into DRFS by inputting keywords and rule description through this form.

3.2 Retrieving bibliographic information from the Pubmed

In the second stage, the DRFS retrieves bibli-

ographical information from the Pubmed. The DRFS submits the whole set of the keywords to the Pubmed system for the micro view method and every pairs of the keywords for the macro view method, and receives the number of hit documents.

Some attribute keywords have aliases to submit to the Pubmed, so we can register aliases for each attribute keyword and the DRFS automatically submits the aliases one by one.

Recently the Pubmed became available as a Web service. For example, ESearch accepts various type of parameters, not only list of keywords, but also search field (title, author, language, journal, and others), date range of publication, retrieval mode (which specifies the data format of output), and others. The search result can be specified in the XML format, so it is easy to extract the data required for the DRFS.

3.3 Ranking rules according to the results of document retrieval

In the third stage, discovered rules are shown with the results of information retrieval as shown in Figure 3.

Each row is composed of ID, rule description, keywords, the result of the micro view method and that of the macro view methods. For example, the first rule's ID is 00001. its rule description is "If ALB decreases, then ALB will become fixed to be around 4.3 after increases rapidly." The keyword extracted from this rule is only "ALB." The result of the micro view method is 1800 which is the number of hits when "ALB" and the domain keyword "hepatitis" are submitted to the Pubmed. The result of the macro view method is 1 because there is only 1 keyword to cluster.

Initially rules are sorted according to their ID code, but they can be re-sorted according to the ID, the result of the micro view method, or that of the macro view method by clicking one of the three items on the top row.

By clicking the number in the "Micro View" column, the result of the micro view method appears as shown in Fig. 4(above).

It consists of rule description, keywords, the number of hits, and the references that appear as a group of 20 references at one time. The next group is shown if the "next 20" link is clicked. When "prev 20" link appears, the previous group is shown if it is clicked. If the title of reference is clicked, the reference information appears in the lower frame. The information consists of PMID (PubMed ID), the title of journal, the title of paper, the authors' names and affiliation, and the abstract of the paper. Each keyword in the abstract is highlighted using an individual colour.

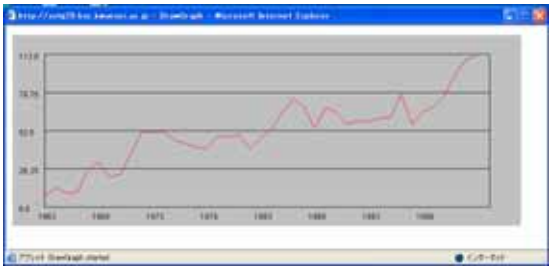

Fig. 5 Yearly trend of hits.

By clicking the "yearly trend" link in the "Total" of hits row, a graph which shows the yearly trend of publication appears as shown in Fig. 5. Referring to this graph, users are able to know whether the rule is a hot topic in the field or not.

By clicking the number in the "Macro View" column of the main page, the result of the macro view method appears as shown in Fig. 4(below).

It consists of rule description, keywords and the number of clusters. In addition, the number of hits of every pair of keywords is shown with the keyword co-occurrence graph which shows how frequent the keywords are co-occurred in the MEDLINE database as discussed in the previous section.

4. Summary and future work

We developed a discovered-rule-filtering system called DRFS for a clinical data mining domain. It can rank discovered rules according to the results of information retrieval from the MEDLINE database.

At present, DRFS just submits a sequence of keywords extracted from each discovered rule and documents retrieved from the MEDLINE are not always very related to the rule. Out future study includes to improve the performance of document retrieval. To this end, we are planning to employ techniques of natural language processing[10].

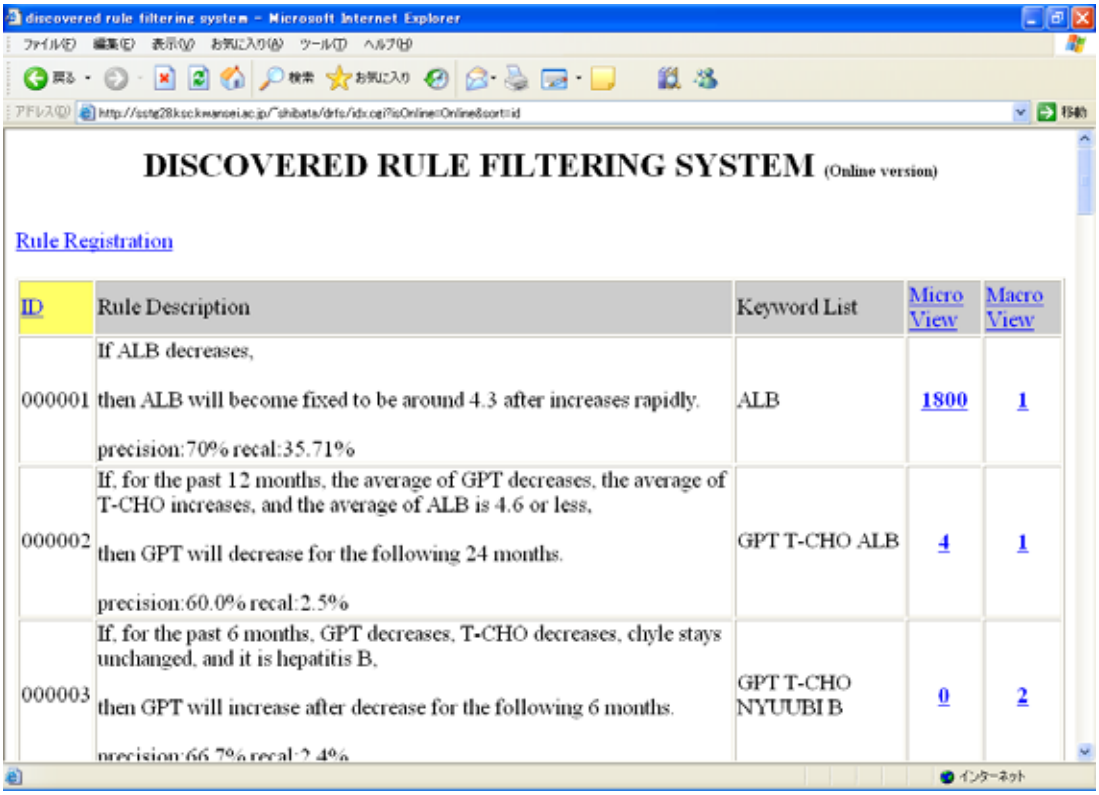

Fig. 3 Main page.

Acknowledgement

This work is supported by a grant-in-aid for scientific research on priority area by the Japanese Ministry of Education, Science, Culture, Sports and Technology.

References

1) H. Motoda (Ed.), Active Mining: New Directions of Data Mining, IOS Press, Amsterdam, 2002.
2) Y. Kitamura, K. Park, A. Iida, and S. Tatsumi. Discovered Rule Filtering Using Information Retrieval Technique. Proceedings of International Workshop on Active Mining, pp. 80-84, 2002.
3) Y. Kitamura, A. Iida, K. Park, and S. Tatsumi, Discovered Rule Filtering System Using MEDLINE Information Retrieval, JSAI Technical Report, SIG-A2-KBS60/FAI52-J11, 2003.
4) Y. Kitamura, A. Iida, K. Park, and S. Tatsumi, Micro View and Macro View Approaches to Discovered Rule Filtering. Proceedings of 2nd International Workshop on Active Mining, pp.14-21, 2003.
5) Y. Kitamura, A. Iida, and K. Park. Preliminary Evaluations of Discovered Rule Filtering Methods. Proceedings of 3rd International Workshop on Active Mining, pp.53-62, 2004.
6) U. M. Fayyad, G. P. Shapiro, P. Smyth, and R. Uthurusamy (Eds.), Advances in Knowledge Discovery and Data Mining, AAAI Press, 1996.
7) H. Yokoi, S. Hirano, K. Takabayashi, S. Tsumoto, Y. Satomura, Active Mining in Medicine: A Chronic Hepatitis Case - Towards Knowledge Discovery in Hospital Information Systems -, Journal of the Japanese Society for Artificial Intelligence, Vol.17, No.5, pp.622-628, 2002. (in Japanese)
8) M. Ohsaki, Y. Sato, H. Yokoi, and T. Yamaguchi, A Rule Discovery Support System for Sequential Medical Data - In the Case Study of a Chronic Hepatitis Dataset -, Proceedings of International Workshop on Active Mining, pp. 97-102, 2002.
9) S. C. Johnson, Hierarchical Clustering Schemes, Psychometrika, Vol.32, pp.241-254, 1967.
10) T. Yamasaki, M. Shimbo, and Y. Matsumoto: A MEDLINE document search system using section information, JSAI, SIG-KBS-A301-05, 2003.

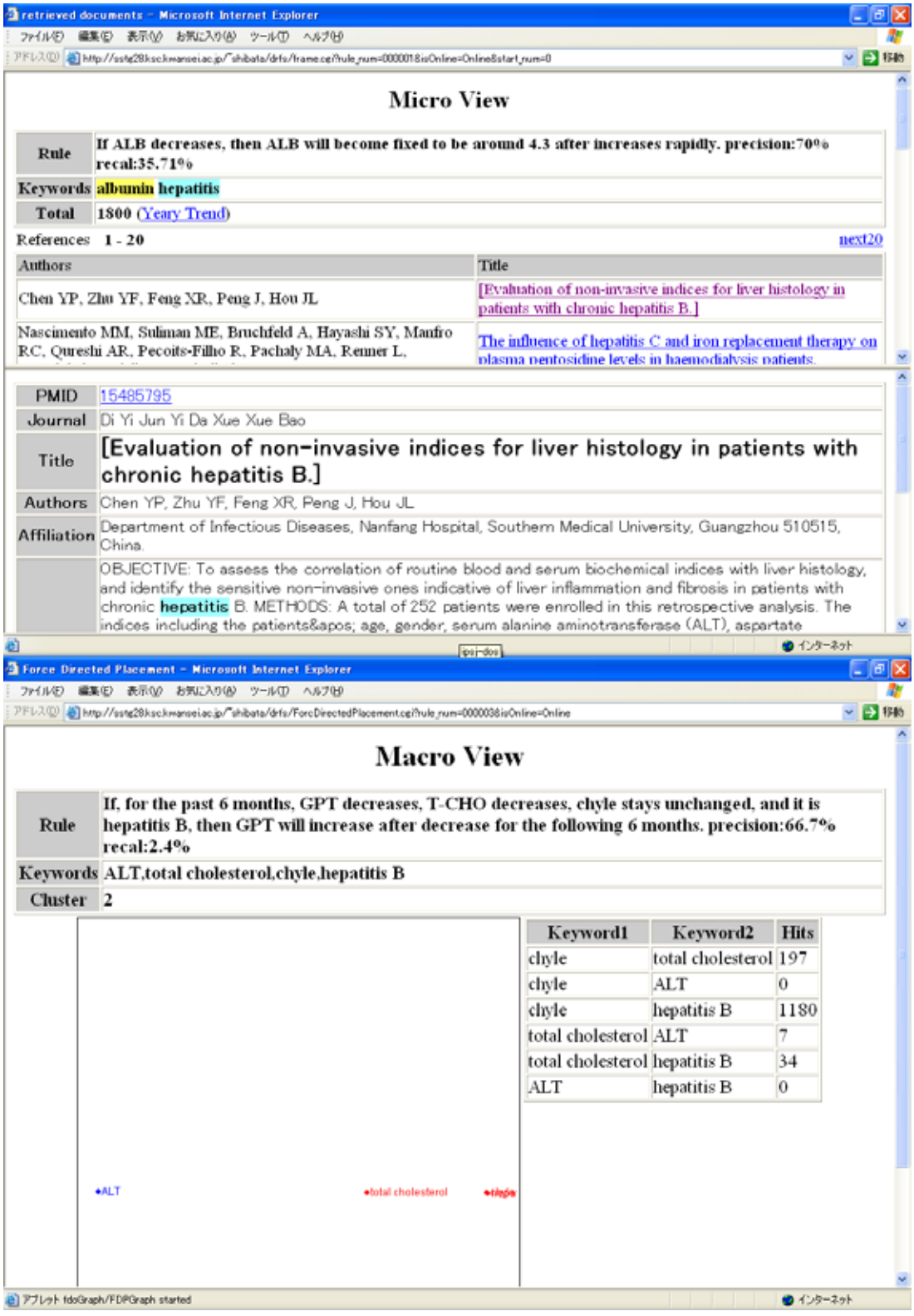

Fig. 4 Results of the micro and macro view methods.