Preliminary Evaluations of Discovered Rule Filtering Methods

Yasuhiko Kitamura[1], Akira Iida[2], and Keunsik Park[3]

[1] School of Science and Technology, Kwansei Gakuin University,
2-1 Gakuen, Sanda, Hyogo 669-1337, Japan
ykitamura@ksc.kwansei.ac.jp
http://ist.ksc.kwansei.ac.jp/~kitamura/
[2] Graduate School of Engineering, Osaka City University,
3-3-138, Sugimoto, Sumiyoshi-ku, Osaka, 558-8585
{iida, tatsumi}@kdel.info.eng.osaka-cu.ac.jp
[3] Graduate School of Medicine, Osaka City University,
1-4-3, Asahi-Machi, Abeno-ku, Osaka, 545-8585
kspark@msic.med.osaka-cu.ac.jp

Abstract. Data mining systems semi-automatically discover knowledge by mining a large volume of data, but discovered knowledge is not always novel to users. We discuss a discovered rule filtering method to filter rules discovered by a data mining system into ones that are novel to the user by using information retrieval results from the Internet. We have two methods; the micro view and the macro view methods, to achieve discovered rule filtering. In the micro view method, we extract keywords from a discovered rule and rank the rule referring to the number of hits when the keywords are submitted to the MEDLINE database. In the macro view method, we first retrieve documents by submitting every pair of the extracted keywords and then form keyword clusters according to the results. We evaluated the methods by sending out a questionnaire to medical students. The evaluation indicates that the macro view method is promising as a discovered rule filtering method.

1 Introduction

Active mining [1] is a new approach to data mining, which tries to discover "high quality" knowledge that meets users' demand in an efficient manner by integrating information gathering, data mining, and user reaction technologies. This paper argues a discovered rule filtering method [3,4,5] that filters a large number of rules obtained by a data mining system to be a small number of novel rules by using an information retrieval technique from the Internet.

Data mining is an automated method to discover useful knowledge by analyzing a large volume of data mechanically [6]. Generally speaking, conventional data mining methods try to discover significant patterns in the statistical sense from a large volume of raw data contained in a given database, but if a system pays attention to only statistically significant features, it may produce a large number of rules that have been known to users. To cope with this problem, we are developing a discovered rule fil-

tering method that filters a large number of rules discovered by a data mining system to be a small number of rules that is novel to the user. To judge whether a rule is novel or not, we utilize an information source on the Internet and judge the novelty of rule according to the number of retrieved documents that relate to the rule.

In this paper, we show two discovered rule filtering methods called the micro view method and the macro view method and evaluate the methods by conducting a questionnairing to medical students. We show the concept and the process of discovered rule filtering with an application example to clinical data mining in Section 2. We then show two discovered rule filtering methods; the micro view and the macro view methods in Section 3 and evaluate them in Section 4 by a questionnaire method. Finally we conclude this paper with our future work in Section 5.

2 Discovered Rule Filtering

The target of our active mining project is a clinical examination database of hepatitis patients, which is offered by the Medical School of Chiba University, on which 10 research groups cooperatively work as a common data source [7]. Several groups have already discovered some sets of rules. For example, Yamaguchi et al. in Shizuoka University analyzed sequential trends between GPT (Glutamic Pyruvic Transaminase), which represents the progress of hepatitis, and other blood test data, and has discovered a number of rules, as one of them is shown in Fig. 1 [8].

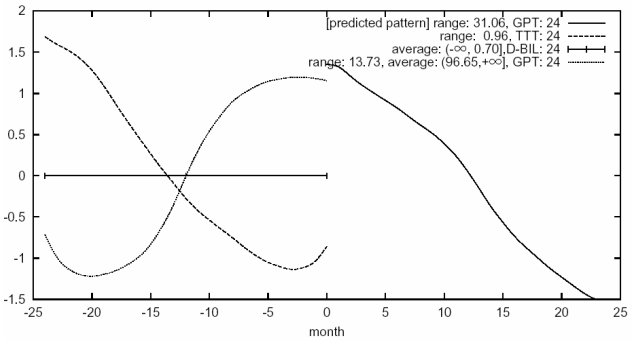

Fig. 1**.** An example of discovered rule [8].

This rule shows a relation among GPT, TTT (Thymol Turbidity Test), and D-BIL (Direct Bilirubin) and means "If, for the past 24 months, D-BIL stays unchanged, TTT decreases, and GPT increases, then GPT will decrease for the following 24 months." A data mining system can semi-automatically discover a large number of rules by analyzing a set of given data, but the discovered rules may include ones that have been known to users. Showing all the discovered rules to a user just results in putting a burden on her. We need to develop a method to filter the discovered rules into a small set of rules that is novel to her. To judge whether a rule is novel or not, we utilize an information source on the Internet and judge it according to the number of retrieved documents relate to the discovered rule.

When a set of discovered rules are given from 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 result of information retrieval. In our project, we aim at discovering knowledge from a hepatitis clinical database, and it is not easy to gather information related to hepatitis from the Internet by using a naïve search engine because the Internet information sources generally contain a huge amount of various and noisy information. We instead use the MEDLINE (MEDlars on LINE) database as the source of retrieving information, which is the largest bibliographical database in the medical and biological domain. PubMed (http://www.ncbi.nlm.nih.gov/entrez/query.fcgi) is a free MEDLINE search service on the Internet run by NCBI (National Center for Biotechnology Information). By using the Pubmed, we can retrieve MEDLINE documents by submitting a set of keywords just like a normal search engine. In addition, we can retrieve documents according to the year of publication and/or the category of documents. These functions are not provided by normal search engines.

The discovered rule filtering process takes the following steps.

Step 1: Extracting keywords from a discovered rule

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

- **Keywords extracted from a discovered rule.** These keywords represent attributes in a discovered rule. For example, keywords that can be extracted directly from a discovered rule shown in Fig. 1 are GPT, TTT, and D-BIL because they are explicitly appeared in the rule. If any abbreviation is not acceptable for the Pubmed, it is translated into a normal name. For example, TTT and GPT are translated into “thymol turbidity test” and “glutamic pyruvic transaminase” respectively.
- **Keywords related to the mining domain.** These keywords represent the purpose or the domain of the data mining task. With keywords extracted form a rule, they should be submitted to the Pubmed as the common keywords to improve the quality of retrieved documents. For our hepatitis data mining, “hepatitis” is a domain keyword. The domain keywords are implicit keywords to be submitted to the Pubmed, and we do not explicitly indicate the keywords in the following discussion.

The rule shown in Fig.1 includes information not only about relations among attributes but also about how the attributes change, but it is difficult to represent the latter information in a sequence of keywords. This problem is left as our future work.

Step 2: Filtering Discovered Rules

We filter discovered rules by using the result of MEDLINE document retrieval. We have two methods called the micro view method and the macro view method to filter discovered rules. The details of the methods are discussed in the following section.

3. Two Methods for Discovered Rule Filtering

How to filter discovered rules according to the search result of MEDLINE document retrieval is a most important issue of this work. We have two methods; the micro view method and the macro view method, to filter discovered rules [5].

3.1 Micro View Method

The micro view method retrieves documents directly related to a discovered rule. It utilizes the result of retrieving documents not only to filter discovered rules, but also to show the documents to the user. By showing a rule and documents related to the rule together, the user may expand her insights on the rule and the data mining task [3]. Filtering rules by the micro view method is quite simple and is based on the following hypotheses.

[Hypotheses] (Micro View Method)
1. The number of documents related to a known rule is large.
2. The number of documents related to an unknown rule is small.
3. The number of documents related to a garbage rule is zero

We hypothesize that research activities on a known rule have been done a lot and that a large number of papers related to the rule have been published. On the other hand, those on an unknown rule have been done a little, and the number of papers related to the rule must be small. Nobody has interest in a garbage or nonsense rule, and the number of papers related to the rule must be zero.

As a strategy of discovered rule filtering, we filter out the garbage rules at the first stage. In the above hypotheses, the border between known rules and unknown one is vague, so we rank rules as the number of the related documents ascends.

However, the micro view method depends much on the performance of document retrieval, and it is actually difficult to retrieve appropriate documents rightly related a rule because of the low performance of keyword-based document retrieval technique. Generally speaking, when a rule is simple with a small number of attributes, the Pubmed system returns a large number of unrelated noisy documents. When a rule is complicated with a large number of attributes, it returns only few documents.

3.2 Macro View Method

The macro view method tries to roughly observe the research trend of discovered rules. 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 keyword co-occurrence graph to judge the novelty of the rule.

Fig. 2, 3, and 4 show keyword co-occurrence graphs. In each graph, a node represents a keyword and the length of edge represents the inverse of the frequency of co-occurrences of keywords connected by the edge. The score attached to the edge repre-

sents the frequency of co-occurence. Hence, the more documents related to a pair keywords are retrieved from the Pubmed, the closer the keywords are located in the graph.

For example, Fig. 2 shows that the relation between any pair from ALB, GPT, and T-CHO is strong. Fig. 3 shows that the relation between T-CHO and GPT is strong, but that between chyle and either of T-CHO and GPT is rather weak. Fig. 4 shows that the relations among GPT, female, and G-GTP are strong, but the relation between hemolysis and G-GTP and those between "blood group a" and the other keywords are weak.

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

We can regard keywords in a cluster are strongly related and research activities concerning the keywords have been done much, so we have a hypothesis to filter rules in the macro view method 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 concerning every pair of keywords in the rule have been published. A rule with two clusters is regarded as an unknown rule. This is because research activities concerning keywords in each cluster have been done much, but those crossing the clusters have not 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 be 1 (the frequency of co-occurrences is 1), the rule in Fig. 2 is regarded as a known rule because all the keywords are merged into a single cluster. Keywords in Fig. 3 are merged into two clusters; one cluster consists of GPT and T-CHO and another consists of chyle only. Hence, the rule is judged to be unknown. Keyowrds in Fig. 4 are merged into 3 clusters as GPT, G-GTP, and female form a cluster and each of hemolysis and "blood group a" forms an individual cluster.

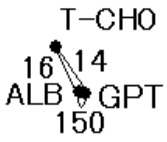

Fig. 2. The keyword co-occurrence graph of rule including GPT, ABL, and T-CHO.

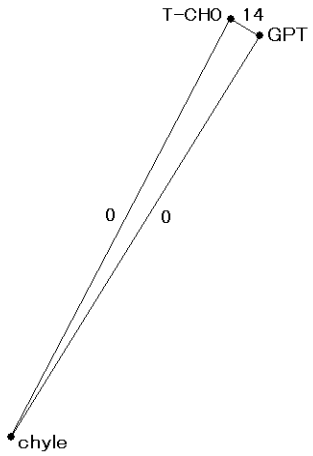

Fig. 3. The keyword co-occurrence graph of rule including GPT, T-CHO, and chyle.

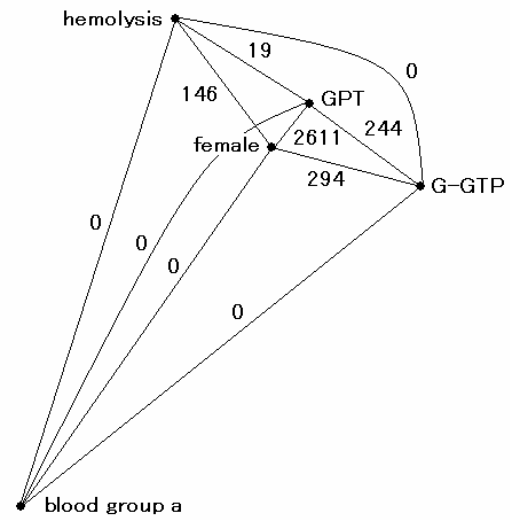

Fig. 4. The keyword co-occurrence graph of rule including GPT, G-GTP, hemolysis, female and "blood group a".

4 Evaluation of Discovered Rule Filtering Methods

4.1 Questionnaire

We performed an evaluation of discovered rule filtering methods by a questionnaire method. In the evaluation, we verified the hypotheses that are the basis of the

micro view and the macro view methods. We first made a questionnaire with two questions shown in Fig. 5. 20 items in Q1 are made from rules discovered by the data mining group in Shizuoka University [8] by extracting keywords from the rules. The reason why we do not show the discovered rules to the subjects is because we would like to evaluate the performance of our rule filtering method that submits only attribute keywords extracted from the rules. If we show the discovered rules directly to the subjects, the subjects judge them considering more than just the relation among attribute keywords, ex. how the attributes change, and that makes the evaluation not proper.

20 items in Q2 are randomly chosen from keywords in the discovered rules. The purpose of Q2 is to show that the number of retrieved documents correlates with the evaluation of medical students when we limit to the number of submitted keywords to be 2, as discussed in Section 4.2.

We sent out the questionnaire to 47 medical students in Osaka City University. The students were just before the state examination to be a medical doctor, so we suppose they are knowledgeable about the medical knowledge in text books.

Q1: How do you guess the result when you submit the following keywords to the Pubmed system? Choose one among A, B, and C.

A (Known): Documents about a fact that I know will be retrieved.
B (Unknown): Documents about a fact that I do not know will be retrieved.
C (Garbage): No document will be retrieved.

(1) [A B C] ALT and TTT
(2) [A B C] TTT, Direct-Bilirubin, and ALT
(3) [A B C] ALT, Total-Cholesterol, and Hepatitis C
(4) ….

Q2: Choose one among four choices about the relation between the following items.

A: The items have a strong relation with each other.
B: The items have a medium relation with each other.
C: The items have a weak relation with each other.
D: The items have no relation with each other.

(1)[A B C D] ALT, Total-Bilirubin
(2)[A B C D] ALT, Total-Cholesterol
(3)[A B C D] ALB, Total-Cholesterol
(4) …

Fig. 5. Questionnaire sent out to medical students.

4.2 Evaluation of the micro view method

We here evaluate the micro view method by analyzing the relation between the number of documents hit by the keywords and the ratio of choices in Q1 made by medical students. Fig. 6 shows the result. We plot the relation between the ratio of choice and the number of retrieved documents for each item in Q1. We also add regression lines to show the relation more clearly and assessed the significance by using the t-test method at the risk level of 5%, but we cannot find any significant relation.

The reason why the micro view method, in which all the keywords extracted from a rule are directly submitted to the Pubmed, does not work well is because the number of hits seems to depend much on the number of keywords submitted to the Pubmed. Generally speaking, the more the number of submitted keywords is, the less the number of retrieved documents is.

However, if we limit the number of submitted keywords to be 2, the method shows a better performance. Fig. 7 shows the relation between the number of documents and the evaluation of medical students obtained through Q2 in the questionnaire. The number of keywords used in Q2 is fixed to be 2. We put the score 3, 2, 1, and 0 to the choice A, B, C, and D respectively. The averaged score and the number of documents have a significant correlation since the correlation coefficient is 0.54. Hence, if we limit the number of submitted keywords to be 2, the result reflects the evaluation of medical students.

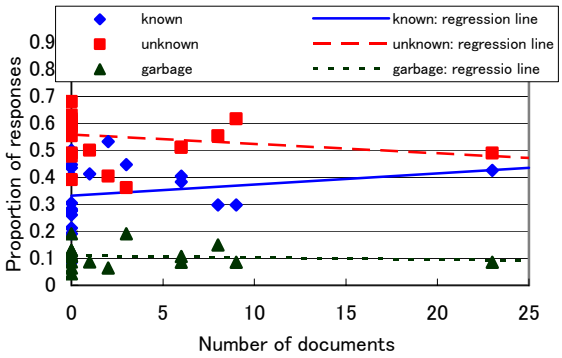

Fig. 6. The relation between the ratio of choice and the number of documents.

4.3 Evaluation of the macro view method

We verify the hypothesis of the macro view method by using the result of Q1 of the questionnaire. We show the relation between the number of clusters and the average ratio of choice in Fig. 8. The threshold of CLINK is 1. At the risk level of 5%, the graph shows two significant relations.

- As the number of clusters increases, the average ratio of “unknown” increases.
- As the number of clusters increases, the average ratio of “known” decreases.

The result does not show any significant relation about “garbage” choice because the number of students who chose “garbage” is relatively small to the other choices and does not depend on the number of clusters. We suppose the medical students hesitate to judge that a rule is just garbage.

The hypotheses of the macro view method are partly supported by this evaluation. The maximum number of clusters in this examination is 3. We still need to examine how medical students or experts judge rules with more than 4 clusters.

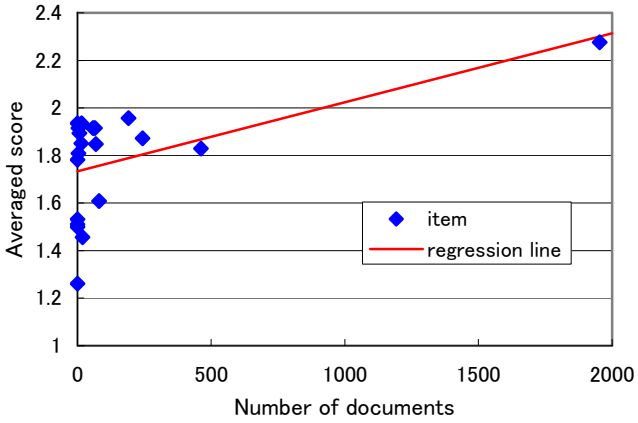

Fig. 7. The relation between the number of documents and the evaluation of medical students when the number of submitted keywords is 2.

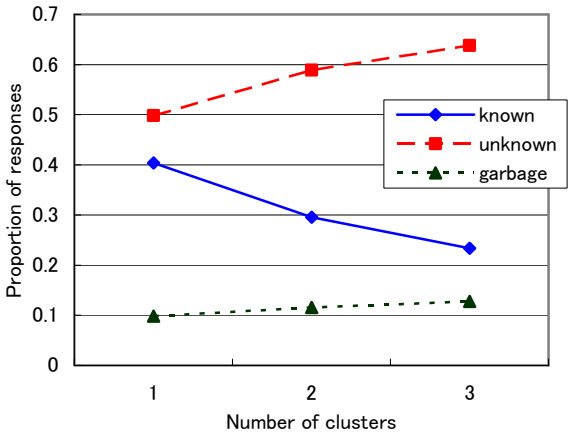

Fig. 8. The relation between the number of clusters and the evaluation of medical students.

5. Summary

We discussed two discovered rule filtering methods, the micro view and the macro view methods, which filters rules discovered by a data mining system into novel ones by using the information retrieval technique from the Internet. We evaluated the methods by using the questionnaire method. The result supports that the output of the macro view method reflects the evaluation of medical students.

Our future work is summarized as follows.

- We need to improve the performance of the information retrieval technique which is based on a naïve keyword search. We plan to improve the performance by applying natural language processing techniques [10].
- We apply the discovered rule filtering methods to a practical application domain such as hepatitis data mining, and evaluate its feasibility.

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

References

1. H. Motoda (Ed.), Active Mining: New Directions of Data Mining, IOS Press, Amsterdam, 2002.
2. R. Baeza-Yates and B. Ribeiro-Neto, Modern Information Retrieval, Addison Wesley, 1999.
3. Y. Kitamura, K. Park, A. Iida, and S. Tatsumi. Discovered Rule Filtering Using Information Retrieval Technique. Proceedings of International Workshop on Active Mining, pp. 80-84, 2002.
4. Y. Kitamura, A. Iida, K. Park, and S. Tatsumi, Discovered Rule Filtering System Using MEDLINE Information Retrieval, JSAI Technical Report, SIG-A2-KBS60/FAI52-J11, 2003.
5. Y. Kitamura, A. Iida, K. Park, and S. Tatsumi, Micro View and Macro View Approaches to Discovered Rule Filtering. Proceedings of 2nd International Workshop on Active Mining, pp.14-21, 2003.
6. U. M. Fayyad, G. P. Shapiro, P. Smyth, and R. Uthurusamy (Eds.), Advances in Knowledge Discovery and Data Mining, AAAI Press, 1996.
7. H. Yokoi, S. Hirano, K. Takabayashi, S. Tsumoto, Y. Satomura, Active Mining in Medicine: A Chronic Hepatitis Case – Towards Knowledge Discovery in Hospital Information Systems -, Journal of the Japanese Society for Artificial Intelligence, Vol.17, No.5, pp.622-628, 2002. (in Japanese)
8. M. Ohsaki, Y. Sato, H. Yokoi, and T. Yamaguchi, A Rule Discovery Support System for Sequential Medical Data – In the Case Study of a Chronic Hepatitis Dataset -, Proceedings of International Workshop on Active Mining, pp. 97-102, 2002.
9. S. C. Johnson, Hierarchical Clustering Schemes, Psychometrika, Vol.32, pp.241-254, 1967.
10. T. Yamasaki, M. Shimbo, and Y. Matsumoto: A MEDLINE document search system using section information, JSAI, SIG-KBS-A301-05, 2003.