# Micro View and Macro View Approaches to Discovered Rule Filtering

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

[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/index.htm
[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.** A data mining system can semi-automatically discover knowledge by mining a large volume of data, but the discovered knowledge is not always novel and may contain unreasonable facts. We try to develop a discovered rule filtering method to filter rules discovered by a data mining system to be novel and reasonable ones for the user by using information retrieval technique. In this method, we rank discovered rules according to the results of information retrieval from an information source on the Internet. In this paper, we show two approaches toward discovered rule filtering; micro view approach and macro view approach. The micro view approach tries to retrieve and show documents directly related to discovered rules. On the other hand, the macro view approach tries to show research activities related to discovered rules by using the results of information retrieval. We discuss advantages and disadvantages of micro view approach and possibilities of macro view approach by using an example of clinical data mining and MEDLINE document retrieval.

## 1 Introduction

The 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 the discovered rule filtering method [3,4] that filters rules obtained by a data mining system based on documents retrieved from an information source on the Internet.

Data mining is an automated method to discover useful knowledge for users by analyzing a large volume of data mechanically. Generally speaking, conventional methods try to discover significant relations among attributes in the statistics sense from a large number of attributes contained in a given database, but if we pay attention to only statistically significant features, we often discover rules that have been known by the user. To cope with this problem, we are developing a discovered rule filtering method that filters a large number of rules discovered by a data mining system to be novel to the user. To judge whether a rule is novel or not, we utilize information sources on the Internet and try to judge the novelty of rule according to the search result of document retrieval that relates to the discovered rule..

In this paper, we show the concept and the procedure of discovered rule filtering using an example of clinical data mining in Section 2. We then show two approaches toward discovered rule filtering; the micro view approach and the macro view approaches in Section 3. Finally we conclude this paper with our future work in Section 4.

## 2 Discovered Rule Filtering

As a target of data mining, we use a clinical examination database of hepatitis patients, which is offered by the Medical School of Chiba University, as a common database on which 10 research groups cooperatively work in our active mining project. Some groups have already discovered some sets of rules. For example, a group in Shizuoka University analyzed sequential trends between a set of blood test data (GPT), which represents a progress of hepatitis, and other test data and has already discovered a number of rules, as one of them is shown in Fig. 1.

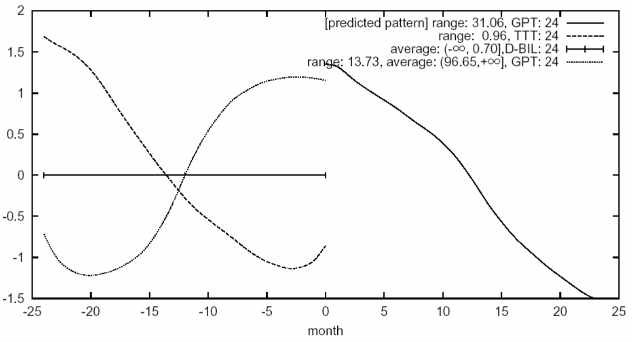

**Fig. 1.** An example of  discovered rule.

This rule shows a relation among GPT (Glutamat-Pyruvat-Transaminase), TTT (Thymol Turbidity Test), and D-BIL (Direct Bilirubin) and means “If, for 24 months, D-BIL stays unchanged, TTT decreases, and GPT increases, then GPT decreases for 24 months.” A data mining system can semi-automatically discover a large number of rules by analyzing a set of data given by the user. On the other hand, discovered rules may include ones that are known by the user. Just showing all of the discovered rules to the user may not be a good idea and may result in putting a burden on her. We need to develop a method to filter the discovered rules into a small set of unknown rules for her. To this end, in this paper, we try to utilize information retrieval technique from an information source on the Internet.

When a set of discovered rules are given from a data mining system, a discovered rule filtering system first retrieves information related to the rules from an information source on the Internet and then filter the rules based on the result of information retrieval. In our project, we aim at discovering rules from a hepatitis database, but it is not easy to gather information related to hepatitis from the Web by using a naïve search engine because the Web information sources generally contain a huge amount of various and noisy information. We instead use the MEDLINE (MEDlars on LINE) database as the target of retrieving information, which is a bibliographical database (including abstracts) that covers more than 4000 medical and biological journals that have been published in about 70 countries. It has already stored more than 11 million documents since 1966. PubMed (http://www.ncbi.nlm.nih.gov/entrez/query.fcgi) is a free MEDLINE search service on the Internet run by NCBI (National Center for Biotechnology Information). By using Pubmed, we can retrieve MEDLINE documents by submitting a set of keywords just like an ordinary search engine. In addition, we can retrieve documents according to the year of publication and/or a category of documents. These functions are not available in ordinary search engines.

A discovered rule filtering process takes the following steps.

**Step 1: Extracting keywords from a discovered rule**

At first, we 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 related to attributes of a discovered rule.** These keywords represent attributes of a discovered rule. For example, keywords that can be acquired from a discovered rule shown in Fig. 1 are GPT, TTT, and D-BIL because they are explicitly shown in the rule. When abbreviations are not acceptable for Pubmed, they need to be translated into normal names. For example, TTT and GPT should be translated into "thymol turbidity test" and "glutamic pyruvic transaminase" respectively.
- **Keywords related to the domain.** These keywords represent the purpose or the domain of the data mining task. They should be included as common keywords. For hepatitis data mining, "hepatitis" is the domain keyword.

**Step 2: Gathering MEDLINE documents efficiently**

We then perform a sequence of MEDLINE document retrievals. For each of discovered rules, we submit the keywords obtained in Step 1 to the Pubmed system. However, redundant queries may be submitted when many of discovered rules are similar, in other words common attributes constitute many rules. The Pubmed is a popular system that is publicly available to a large number of researchers over the world, so it is required to reduce the load to the system. Actually, too many requests from a user lead to a temporal rejection of service to her. To reduce the number of submissions, we try to use a method that employs a graph representation to store the history of document retrievals. By referring to the graph, we can gather documents in an efficient way by reducing the number of meaningless or redundant keyword submissions.

**Step 3: Filtering Discovered Rules**

We filter discovered rules by using the result of MEDLINE document retrieval. More precisely, based on a result of document retrieval, we rank discovered rules. How to rank discovered rules by using the result of document retrievals is a core method of discovered rule filtering.

We assume the number of MEDLINE documents hit by a set of keywords shows a trend of research activity related to the keywords, so we may say that the more the number of hits is, the more the rule that contains the keywords is commonly known in the research field. The published month or year of document may be another hint to rank rules. If many documents related to a rule are published recently, the rule may contain a hot topic in the field.

## 3 Two Approaches to 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 approaches; micro view approach and macro view approach, to realize discovered rule filtering.

### 3.1 Micro View Approach

In the micro view approach, we retrieve and show documents related to a discovered rule directly to the user.

By using the micro view approach, the can obtain not only novel rules discovered by a data mining system, but also documents related to the rules. By showing a rule and documents related to the rule at once, the user can get more insights on the rule and may have a chance to start a new data mining task. For the detail, please refer to [3].

However, it is actually difficult to retrieve appropriate documents rightly related a rule because of the low performance of information technique. Especially, when a rule is simple as it is composed of a small number of attributes, the IR system returns a noisy output, documents including a large number of unrelated ones. When a rule is complicated as it is composed of a large number of attributes, it returns few documents.

To see how a micro view approach works, we performed a preliminary experiment of discovered rule filtering. We used 20 rules obtained from the team in Shizuoka University and gathered documents related to the rules from the MEDLINE database. The result is shown in Table 1.

In this table, "ID" is the ID number of rule and "Keywords" are extracted from the rule and are submitted to the Pubmed. "No" shows the number of submitted keywords. "Hits" is the number of documents returned. "Ev" is the evaluation of rule by a medical doctor. He evaluated each rule, which was given in a form depicted in Fig. 1, and categorized into 2 classes; R (reasonable rules) and U (unreasonable rules).

**Table 1.** The preliminary experiment of discovered rule filtering.

| ID | Ev. | Hits | No. | Keywords |
|---|---|---|---|---|
| 1 | R | 6 | 4 | hepatitis, gpt, t-cho, albumin |
| 2 | U | 0 | 4 | hepatitis b, gpt, t-cho, chyle |
| 3 | U | 0 | 4 | hepatitis c, gpt, lap, hemolysis |
| 4 | R | 0 | 5 | hepatitis, gpt, got, na, lap |
| 5 | R | 0 | 6 | hepatitis, gpt, got, ttt, cl, (female) |
| 6 | U | 0 | 5 | hepatitis, gpt, ldh, hemolysis, blood group a |
| 7 | R | 7 | 4 | hepatitis, gpt, alb, jaundice |
| 8 | R | 9 | 3 | hepatitis b, gpt, creatinine |
| 10 | R | 0 | 4 | hepatitis, ttt, t-bil, gpt |
| 11 | U | 0 | 4 | hepatitis, gpt, alpha globulin, beta globulin |
| 13 | U | 8 | 4 | hepatitis, hemolysis, gpt, (female) |
| 14 | U | 0 | 4 | hepatitis, gpt, ttt, d-bil |
| 15 | U | 0 | 3 | hepatitis, gpt, chyle |
| 17 | R | 0 | 5 | hepatitis, gpt, ttt, blood group o, (female) |
| 18 | R | 2 | 3 | hepatitis c, gpt, t-cho |
| 19 | R | 0 | 6 | hepatitis, gpt, che, ttt, ztt, (male) |
| 20 | R | 0 | 5 | hepatitis, gpt, lap, alb, interferon |
| 22 | U | 0 | 7 | hepatitis, gpt, ggtp, hemolysis, blood group a, (female), (age 45-64) |
| 23 | U | 0 | 4 | hepatitis b, gpt, got, i-bil |
| 27 | U | 0 | 4 | hepatitis, gpt, hemolysis, i-bil |

As we can see, except Rule 13, rules with hits more than 0 are categorized in reasonable rules, but a number of reasonable rules hit no document. It seems that the number of submitted keywords affects the number of hits. In other words, if a rule is complex with many keywords, the number of hits tends to be few.

This result tells us that it is not easy to distinguish reasonable or known rules from unreasonable or gabage ones by using only the number of hits. It shows a limitation of macro view approach.

To cope with the problem, we need to improve the performance of micro view approach as follows.

(1) **Accurate document retrieval.** In our current implementation, we use only keywords related to attributes contained in a rule and those related to the domain, and the document retrieval is not accurate enough and often contains documents unrelated to the rule. To improve the accuracy, we need to add adequate keywords related to relations among attributes. These keywords represent relations among attributes that constitute a discovered rule. It is difficult to acquire such keywords directly from the rule because, in many cases, they are not explicitly represented in the rule. They need

to be included manually in advance. For example, in the hepatitis data mining, “periodicity” should be included when the periodicity of attribute value change is important.

(2) **Document analysis by applying natural language processing methods.** Another method is to refine the results by analyzing the documents using natural language processing technique. Generally speaking, information retrieval technique only retrieves documents that contain the given keyword(s) and does not care the context in which the keyword(s) appear. On the other hand, natural language processing technique can clarify the context and can refine the result obtained by information retrieval technique. For example, if a keyword is not found in the same sentence in which another keyword appears, we might conclude that the document does not argue a relation between the two keywords. We hence can improve the accuracy of discovered rule filtering by analyzing whether the given keywords are found in a same sentence. In addition, if we can analyze whether the sentence argues the conclusion of the document, we can further improve the accuracy of rule filtering.

### 3.2 Macro View Approach

In the macro view approach, we try to roughly observe the trend of relation among keywords. For example, the number of documents in which the keywords co-occur approximately shows the strength of relation among the keywords. We show two methods based on the macro view approach.

**(1) Showing research activities based pair-wise keyword co-occurrence graph**

We depicted a graph which shows a research activities related to a discovered rule by using the number of co-occurrences of every two keywords found in the rule. A node in the graph represents a keyword which specifies an attribute found in the rule and an edge represents the number of co-occurrences of two keywords connected by the edge.

Fig. 2 shows a graph concerning rule 1.

This graph shows that the number of co-occurrences of "albumin" and "gpt" is 150, that of "total cholesterol" and "albumin" is 16, and that of "gpt" and "total cholesterol" is 14. As shown in Table 1, the rule is judged as "reasonable" by a medical doctor and we can see each attribute is interconnected to other attributes strongly.

Fig. 3 shows research activities related to rule 2. This rule is judged as "unreasonable" and the number of hits is 0. Research activities look weak because only 14 documents related to "total cholesterol" and "gpt" are retrieved.

Fig. 4 shows research activities related to rule 4. This rule is judged as "reasonable", but the number of hits is 0. Contrasting with rule 2, research activities related to rule 4 look active because a number of documents are retrieved except documents related to "na" and "lap".

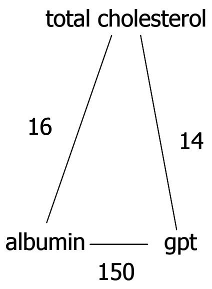

**Fig. 2.** Research activities related to rule 1.

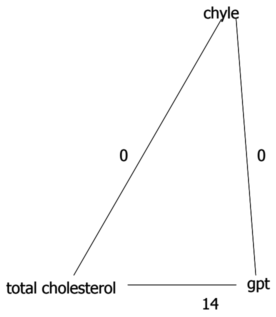

**Fig. 3.** Research activities related to rule 2.

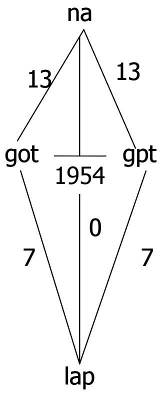

**Fig. 4.** Research activities related to rule 4.

As a conclusion, the graph shape of reasonable rules looks different from that of unreasonable rules. But, when given a graph, how to judge whether the rule is reasonable or not is our future work.

**(2) The yearly trend of research activities**

The MEDLINE database contains bibliographical information of bioscience articles, which includes the year of publication, and the Pubmed can retrieve the information according to the year of publication. By observing the yearly trend of co-occurrences, we can see the change of the research activity. For example, we can have the following interpretations as shown in Fig. 5.

(a) If the number of co-occurrences moves upward, the research topic related to the keywords is hot.

(b) If the number of co-occurrences moves downward, the research topic related to the keywords is terminating.

(c) If the number of co-occurrences keeps high, the research topic related to the keyword is commonly known.

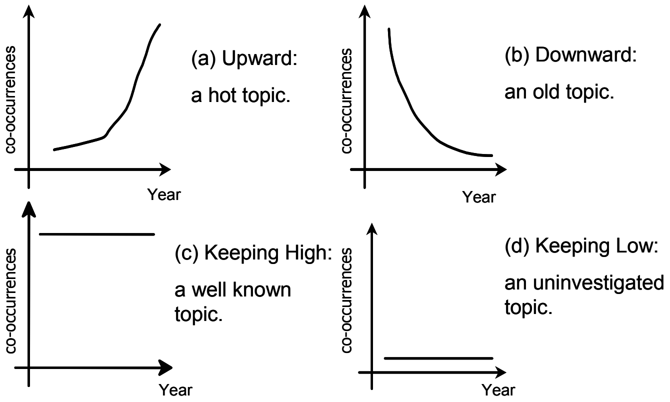

**Fig. 5.** Yearly Trends of co-occurrences.

(d) If the number of co-occurrences keeps low, the research topic related to the keyword is not known. Few researchers show interest in the topic.

To evaluate a feasibility of this method, we submitted 4 queries to the MEDLINE database and show the results in Fig. 6.

(a) "hcv, hepatitis"

The number of co-occurrences has been increasing since 1989. In 1989, we have an event of succeeding HCV cloning. HCV is a hot topic of hepatitis research.

(b) "smallpox, vaccine"

The number of co-occurrences has been decreasing. In 1980, the World Health Assembly announced that smallpox had been eradicated. Recently, we see the number turns to increasing because discussions about smallpox as a biochemical weapon arise

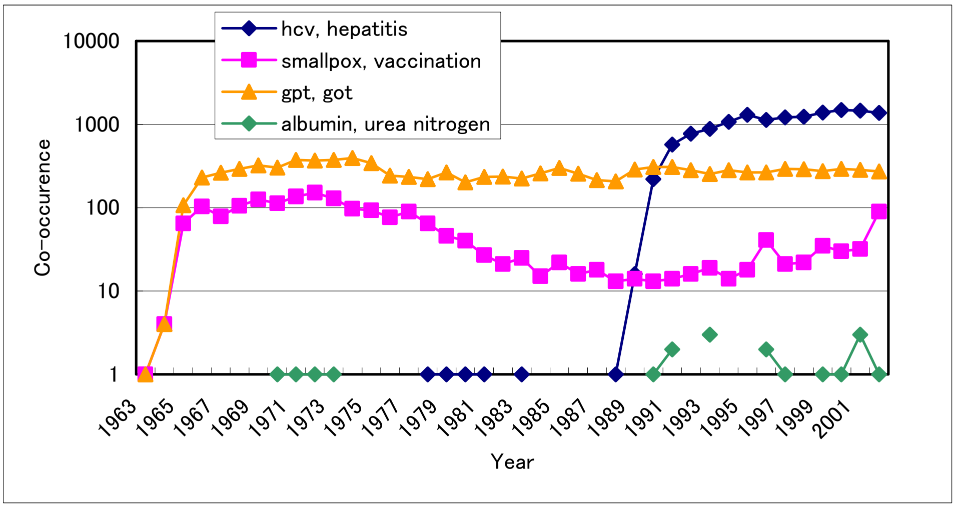

**Fig. 6.** The yearly trend of research activities.

(c) "gpt, got"

The number of co-occurrences stays high. GPT and GOT are well known blood test measure and they are used to diagnose hepatitis. The relation between GPT and GOT is well known in the medical domain.

(d) "albumin, urea nitrogen"

The number of co-occurrences stays low. The relation between albumin and urea nitrogen is seldom discussed.

From above results, the yearly trends well correspond with historical events in the medical domain, and can be a measure to know the research activities.

## 4 Summary

We discussed a discovered rule filtering method which filters rules discovered by a data mining system into novel ones by using the IR technique. We proposed two approaches toward discovered rule filtering; the micro view approach and the macro view approach and showed merits and demerits of micro view approach and possibilities of macro view approach.

Our future work is summarized as follows.

- We need to find a measure to distinguish reasonable rules from unreasonable one, which can be used in the macro view method. We also need to find a measure to know the novelty of rule.
- We need to improve the performance of micro view approach by adding keywords that represent relations among attributes and by using natural language processing techniques. The improvement of micro view approach can contribute the improvement of macro view approach.
- We need to implement the macro view method in a discovered rule filtering system and apply it to an application of hepatitis data mining.

**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.