Micro View and Macro View Approaches to Discovered Rule Filtering

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/index.htm
[2] Graduate School of Engineering, Osaka City University,
3-3-138, Sugimoto, Sumiyoshi-ku, Osaka, 558-8585
iida@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 tries to semi-automatically discover knowledge by mining a large volume of raw 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 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; the micro view approach and the 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 the trend of research activities related to discovered rules by using the results of information retrieval. We discuss advantages and disadvantages of the micro view approach and feasibility of the 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 statistic 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 ones 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 first discuss the principle of integrating data mining and information in Section 2, and we show the concept and the process of discovered rule filtering using an example of clinical data mining in Section 3. We then show two approaches toward discovered rule filtering; the micro view approach and the macro view approaches in Section 4. In Section 5, we show an evaluation of the macro view approach. Finally we conclude this paper with our future work in Section 6.

2 Integration of Data Mining and Information Retrieval

Data mining process can be defined to discover significant relations among attributes from a large volume of data set $D(\subseteq A_1 \times A_2 \times \ldots \times A_n)$ where A_i $(1 \leq i \leq n)$ is an attribute of data. For simplicity of discussion, we assume each attribute value is 0 or 1. Hence, data mining process can be viewed as a function $dm(D) \subseteq R = \{<A_{c1}, A_{c2}, \ldots, A_{cp} \rightarrow A_d>\}$ which takes a data set D as an input and produces a set of rules representing relations among attributes. As methods to produce such a set of rules, we normally use statistical methods that consider precision and/or recall measure. When we try to discover novel rules, we often sacrifice the recall measure.

On the other hand, information retrieval process can defined to count the co-occurrences of the specified keywords from a large volume of document set $D' \subseteq B_1 \times B_2 \times \ldots \times B_m$ where B_j $(1 \leq j \leq m)$ is a keyword. Hence, information retrieval process can be viewed as a function $ir(D', \{B_{k1}, B_{k2}, \ldots, B_{kq}\}) \in \mathbf{Int}$ which takes a set of keywords as an input and produces the number of co-occurrences of the keywords, where **Int** is a set of integer. Practically, the function produces a list of documents which contain the keywords rather than just the number of co-occurrences.

Now what can we get by integrating data mining and information retrieval. If we have a proper function $c(A_i)=B_j$ that associate an attribute A_i in data mining process with a keyword B_j in information retrieval process, we can associate the result of data mining with that of information retrieval. For example, let us assume that we have a rule $<A_{c1}, A_{c2}, \ldots, A_{cp} \rightarrow A_d>$ as a result of data mining. We can get the number k that is the number of co-occurrences when the keywords in the discovered rule are used for information retrieval. Formally it is given as $ir(D', \{c(A_{c1}), c(A_{c2}), \ldots, c(A_{cp})\})=k$. Then, we can rank discovered rules according to k. If k is large, it is probable that the discovered rule has been known. On the other hand, if k is small, it is probable that the discovered rule is novel.

Some information retrieval systems accept additional keywords and parameters. For example, a document retrieval system accepts a parameter that specifies range of publication as its input. By utilizing this function, we can recognize whether discovered rules deal with a latest research topic or not.

3 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 [5]. Some groups have already discovered some sets of rules. For example, a group in Shizuoka University analyzed sequential trends of the relation 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.

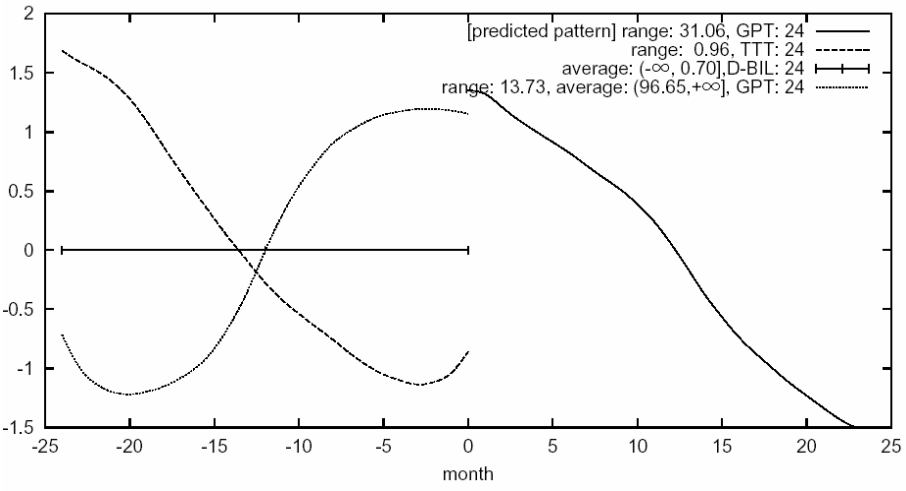

Fig. 1. An example of discovered rule.

This rule shows a relation among GPT (Glutamat-Pyruvat-Transaminase), TTT (Thymol Turbidity Test), and D-BIL (Direct Bilirubin) and means "If, for 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. On the other hand, discovered rules may include ones that are known and/or uninteresting to the user. Just showing all of the discovered rules to the user 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 and interesting rules to her. To this end, in this paper, we try to utilize information retrieval technique from 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 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 Internet by using naïve search engines 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 the Pubmed, we can retrieve MEDLINE documents by submitting a set of keywords just like an ordinary search engine.

A discovered rule filtering process takes the following steps.

Step 1: Extracting keywords from a discovered rule

At first, we need to find a set of proper keywords to retrieve MEDLINE documents that relate to a discovered rule. Such keywords can be acquired from a discovered rule, the domain of data mining, and the interest of the user. These are summarized 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 the Pubmed, they need to be converted into normal names. For example, TTT and GPT should be converted into "thymol turbidity test" and "glutamic pyruvic transaminase" respectively.
- **Keywords related to a relation 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.
- **Keywords related to the domain.** These keywords represent the purpose or the background of the data mining task. They should be included in advance as the common keywords. For hepatitis data mining, "hepatitis" is the keyword.
- **Keywords related to the user's interest.** These keywords represent the user's interest in the data mining task. They can be acquired directly by requesting the user to input the keywords.

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 [2]. 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, as shown in Fig. 2, 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. The graph in Fig. 2 shows pairs of submitted keywords and the number of hits. For example, this graph shows that a submission including keywords "hepatitis," "gpt," and "t-cho" returns nothing. It also shows that the combination of "hepatitis" and "gpt" is better than the combination of "hepatitis" and "total cholesterol" because the former is expected to have more returns than the latter.

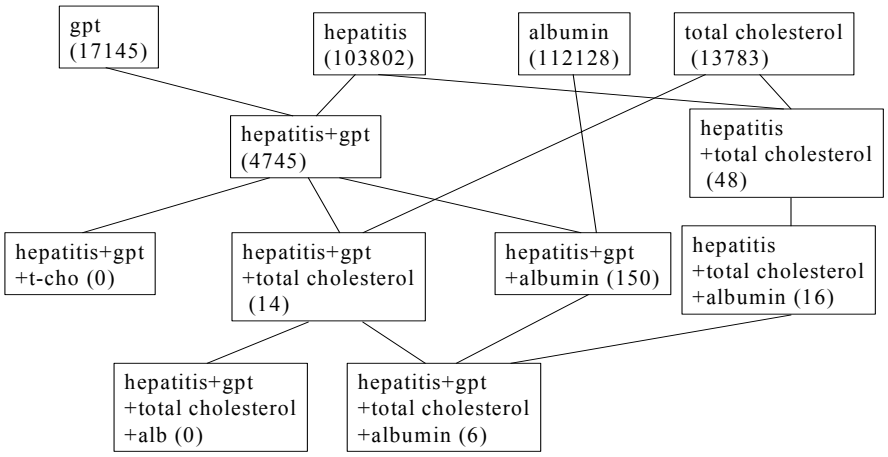

Fig. 2. A graph that represents document retrieval history.

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.

Basically the number of documents hit by a set of keywords shows the correlation of the keywords in the MEDLINE database, so we can assume that the more the number of hits is, the more the combination of attributes represented by the keywords is commonly known in the research field. We therefore use a heuristic such that "If the number of hits is small, the rule is novel." We discuss the detail in the next section.

The published month or year of document can 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.

Step 4: Estimating User's Preference

Retrieving documents by simply submitting keywords obtained in Step 1 may produce a wide variety of documents. They may relate to a discovered rule, but may not to the user's interest. To deal with this problem, we may request the user to input additional keywords that represent her interest, but this may put a burden to her.

Relevance feedback is a technique that indirectly acquires the preference of the user. In this technique, the user just feedbacks "Yes" or "No" to the system depending on whether she has interest in a document or not. The system uses the feedbacks as a clue to analyze the abstract of the document and to automatically find keywords that show the user's interest, and uses them for further document retrievals.

4 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; the micro view approach and the macro view approach, to realize discovered rule filtering.

4.1 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 user 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. In our preliminary experiment, at first we showed a discovered rule alone, shown in Figure 1, to a medical doctor and received the following comment (Comment 1). The doctor seems to take the discovered rule as a commonly known fact.

Comment 1: "TTT shows an indicator of the activity of antic body. The more active the antic bodies are, the less active the hepatitis is and therefore the amount of GPT decreases. This rule can be interpreted by using well known facts."

We then retrieved related documents by using the rule filtering technique. The search result with keywords "hepatitis" and "TTT" was 11 documents. Among them, there was a document, shown in Fig. 3, in which the doctor shows his interest as mentioned in a comment (Comment 2).

Comment 2: "This document discusses that we can compare type B virus with type C virus by measuring the TTT value of hepatitis virus carriers (who have not contracted hepatitis). It is a new paper published in 2001 that discusses a relation between TTT and hepatitis, but it reports only a small number of cases. The discovered rule suggests the same symptom appears not only in carriers but also in patients. This rule is important to support this paper from a standpoint of clinical data."

The effect shown in this preliminary examination is that the system can retrieve not only a new document related to a discovered rule but also a new viewpoint to the rule, and gives a chance to invoke a new mining process. In other words, if the rule

alone is shown to the user, it is recognized just as a common fact, but if it is shown with a related document, it can motivate the user to analyze the amount of TTT depending on the type of hepatitis by using a large volume of hepatitis data. We hope this kind of effect can be found in many other cases.

1: Hepatol Res 2001 Sep;21(1):67-75

Comparison of clinical laboratory liver tests between asymptomatic HBV and HCV carriers with persistently normal aminotransferase serum levels.

Murawaki Y, Ikuta Y, Koda M, Kawasaki H.

Second Department of Internal Medicine, Tottori University School of Medicine, 683-8504, Yonago, Japan

We examined the clinicopathological state in asymptomatic hepatitis C virus (HCV) carriers with persistently normal aminotransferase serum levels in comparison with asymptomatic hepatitis B virus (HBV) carriers. The findings showed that the thymol turbidity test (TTT) values and zinc sulfate turbidity test (ZTT) values were significantly higher in asymptomatic HCV carriers than in asymptomatic HBV carriers, whose values were within the normal limits. Multivariate analysis showed that the independent predictor of serum TTT and ZTT levels was the HCV infection. In clinical state, simple and cheap tests such as TTT and ZTT are useful for mass screening to detect HCV carriers in medical check-ups of healthy workers.

PMID: 11470629 [PubMed - as supplied by publisher]

Fig. 3. A document retrieved.

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 the 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 garbage ones by using only the number of hits. It shows a limitation of micro 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.

4.2 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

Fig. 4, 5, and 6 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 number attached to the edge represents the frequency of co-occurrence. Hence, the more documents related to a pair keywords are retrieved from the Pubmed, the closer the keywords are in the graph.

For example, Fig. 4 shows that the relation between any pair among ALB, GPT, and T-CHO is strong. Fig. 5 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. 6 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 [7]. 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 Approach)
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.

Rule with only one cluster are regarded as known rules because a large number of papers concerning every pair of keywords in the rule have been published. Rules with two clusters are regarded as unknown rules. This is because research activities concerning keywords in each cluster have been done much, but those crossing the clusters have not been done. Rule with more than two clusters are regarded as garbage rules. Such a rule is too complex to understand because 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. 4 is regarded as a known rule because all the keywords are merged into a single cluster. Keywords in Fig. 5 are merged into two clusters; one cluster consists of GPT and T-CHO and another consists of chyle only. Hence, the rule is judged to be unknown. Keywords in Fig. 6 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.

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.

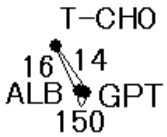

Fig. 4. The keyword co-occurrence graph of rule including GPT, ABL, and T-CHO.

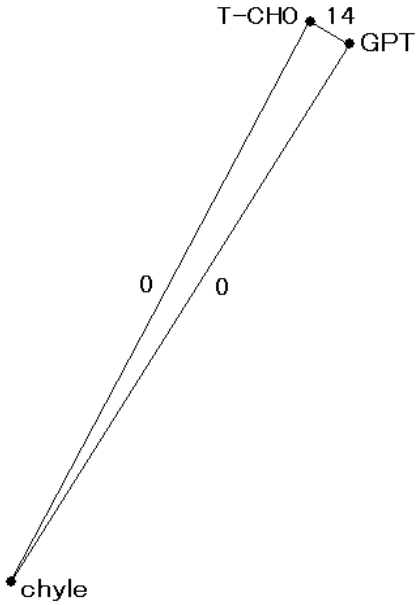

Fig. 5. The keyword co-occurrence graph of rule including GPT, T-CHO, and chyle.

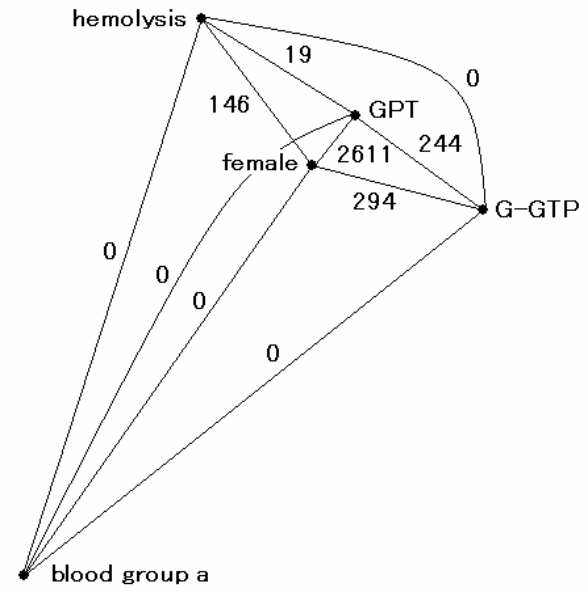

Fig. 6. The keyword co-occurrence graph of rule including GPT, G-GTP, hemolysis, female and "blood group a".

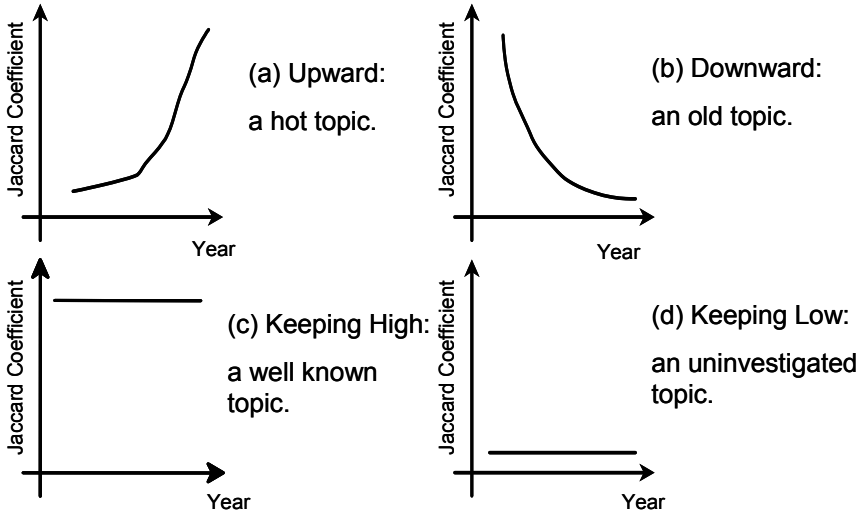

Fig. 7. Yearly trends of the Jaccard co-efficient..

(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 publication, we can see the change of research activity.

Generally speaking, the number of documents contained in the MEDLINE database increases rapidly year by year, so the number of documents hit by most sets of keywords increases. Hence, we use the Jaccard coefficient as an alternative to measure the yearly trend of publication. Given keywords $K_1, K_2, \ldots, K_n$, its Jaccard coefficient is defined as

$$\frac{h(K_1, K_2, \ldots, K_n)}{h(K_1) + h(K_2) + \cdots + h(K_n)}$$

where $h(L)$ is the number of documents hit by the set of keywords L. The Jaccard coefficient is a measure to show the strength of association among multiple keywords.

For example, we can have the following interpretations as shown in Fig. 7.

(a) If the Jaccard co-efficient moves upward, the research topic related to the keywords is hot.

(b) If the Jaccard co-efficient moves downward, the research topic related to the keywords is terminating.

(c) If the Jaccard co-efficient keeps high, the research topic related to the keyword is commonly known.

(d) If the Jaccard co-efficient 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. 8 through Fig. 11.

(a) "hcv, hepatitis" (Fig.8)

The Jaccard co-efficient 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" (Fig.9)

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

(c) "gpt, got" (Fig.10)

The Jaccard co-efficient 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" (Fig.11)

The Jaccard co-efficient stays low. The relation between albumin and urea nitrogen is seldom discussed.

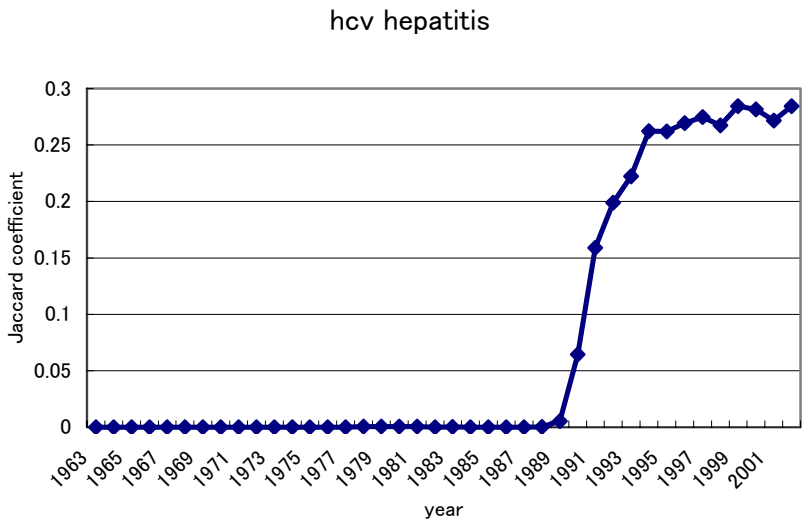

Fig. 8. The yearly trend of the Jaccard co-efficient concerning “hcv” and “hepatitis”.

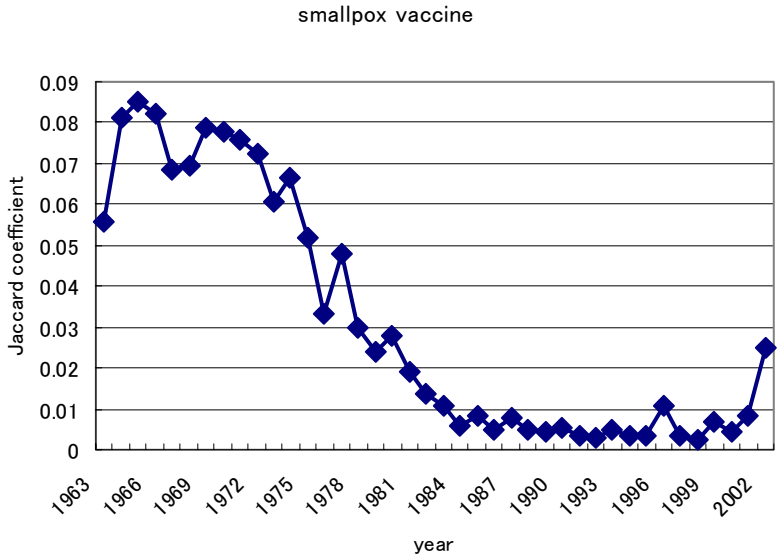

Fig. 9. The yearly trend of the Jaccard co-efficient concerning “snallpox” and “vaccine”.

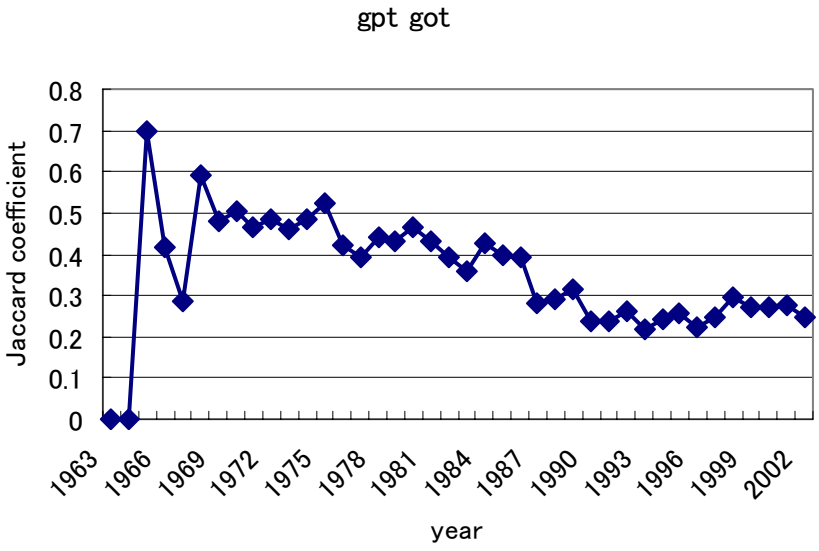

Fig. 10. The yearly trend of the Jaccard co-efficient concerning “gpt” and “got”.

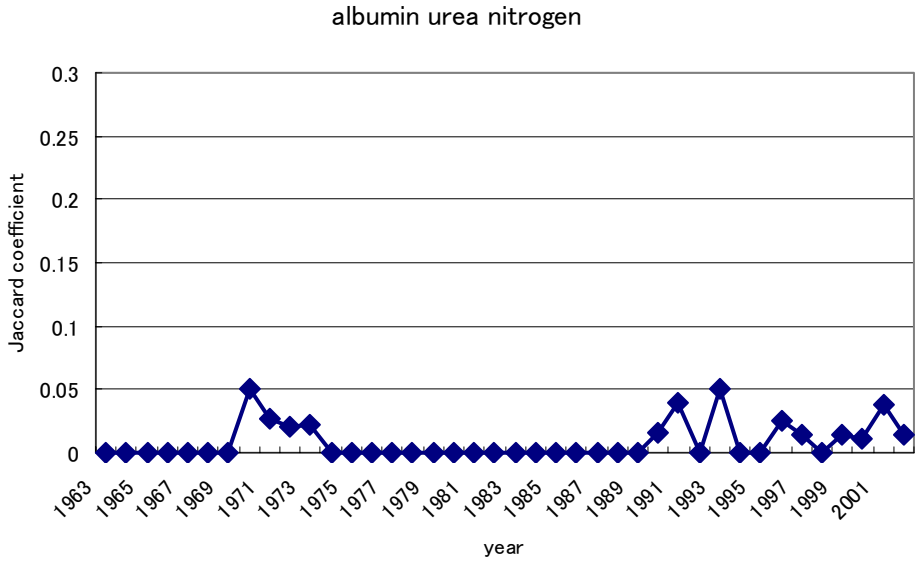

Fig. 11. The yearly trend of the Jaccard co-efficient concerning "albumin" and "urea nitrogen".

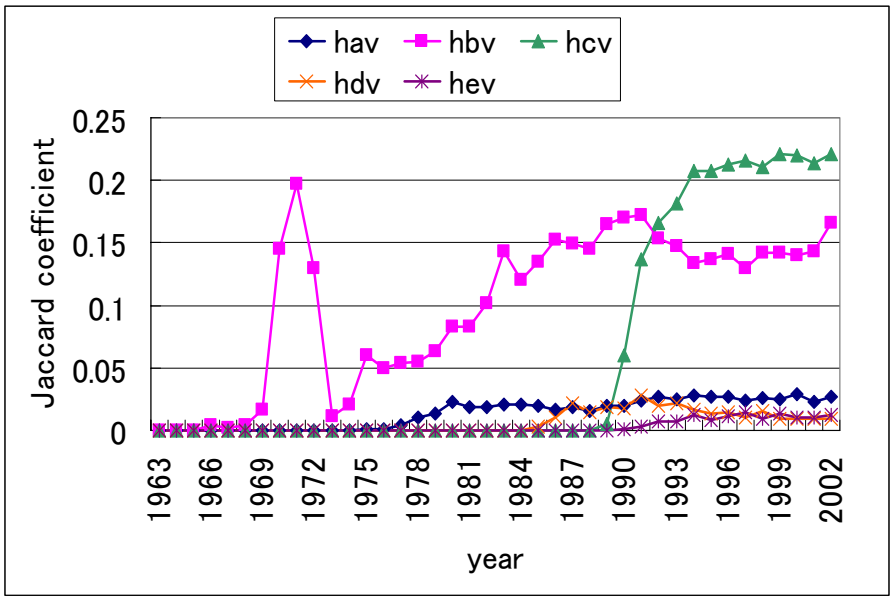

Fig. 12. The yearly trend of the Jaccard co-efficient concerning hepatitis viruses.

To show feasibility in the hepatitis data mining domain, we measured the yearly trend of Jaccard co-efficient of "hepatitis" and each of five representative hepatitis viruses

(hav, hbv, hcv, hdv, and hev) and show the results in Fig. 12. We also show the history of hepatitis virus discovery in Table 2. There is apparently a co-relation between the Jaccard co-efficient and the discovery time of hepatitis viruses. In hepatitis research activities, works on hbv and hcv are major and especially those on hcv rapidly increase after its discovery.

Table 2. History of hepatitis viruses discovery

1965	Discovery of Australian antigen. This is the first step toward hbv discovery. (B.S. Blumberg)
1973	Discovery of hav. (S.M. Feinstone)
1977	Discovery of delta antigen. This is the first step toward hdv discovery. (M. Rizzetto)
1983	Detection of hev by reverse transcription-polymerase chain reaction. (M.S. Balayan)
1989	Success of cloning hcv. (Q.L. Choo)

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.

5 Evaluation of Macro View Approach

We performed an evaluation of the macro view approach by the questionnaire method. We first made a questionnaire shown in Fig. 13. 20 items are made from rules discovered by the data mining group in Shizuoka University [6] by extracting keywords from the rules. 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.

We verify the hypothesis of the macro view method by using the result of the questionnaire. We show the relation between the number of clusters and the average ratio of choice in Fig. 14. 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 approach are partly supported by this evaluation. The maximum number of clusters in this examination is 3. We still need to examine how medical experts judge rules with more than 4 clusters.

Q: 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 are retrieved.
B (Unknown): Documents about a fact that I do not know are retrieved.
C (Garbage): No document is 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) ….

Fig. 13. Questionnaire sent out to medical students.

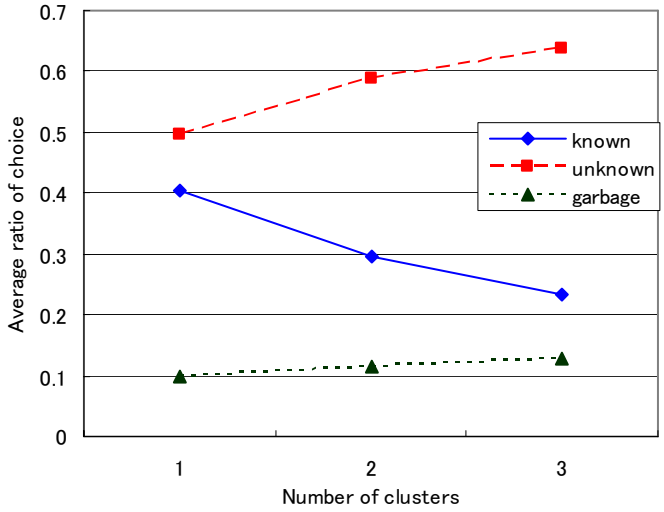

Fig. 14. The relation between the number of clusters and the evaluation of medical experts.

6 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 feasibility of macro view approach.

Our future work is summarized as follows.

- We need to find a clear 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.
5. 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)
6. 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.
7. S. C. Johnson, Hierarchical Clustering Schemes, Psychometrika, Vol.32, pp.241-254, 1967.