社団法人 電子情報通信学会
THE INSTITUTE OF ELECTRONICS,
INFORMATION AND COMMUNICATION ENGINEERS

信学技報
TECHNICAL REPORT OF IEICE

# Macro View Approach to Discovered Rule Filtering

Yasuhiko KITAMURA　Akira IIDA †　Keunsik PARK ‡　and　Shoji TATSUMI †

School of Science and Technology, Kwansei Gakuin University　2-1, Gakuen, Sanda-shi, Hyogo, 669-1337

† Graduate School of Engineering, Osaka City University　3-3-138, Sugimoto, Sumiyoshi-ku, Osaka, 558-8585

‡ Graduate School of Medicine, Osaka City University　1-4-3, Asahi-Machi, Abeno-ku, Osaka, 545-8585

E-mail: ykitamura@ksc.kwansei.ac.jp, † {iida, tatsumi}@kdel.info.eng.osaka-cu.ac.jp,
‡ 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 interesting to the user. We have proposed a discovered rule filtering method to filter rules discovered by a data mining system to be novel and interesting to the user by using information retrieval technique. In this method, we rank discovered rules according to the results of information retrieval from the Internet. In this paper, we show two approaches toward discovered rule filtering; micro view approach and macro view approach, and mainly discuss the latter. Our macro view approach utilizes the yearly trend of Jaccard coefficient and judges whether a discovered rule is a hot topic or not. By using a concrete example of clinical data mining and MEDLINE document retrieval, we show advantages of macro view approaches toward discovered rule filtering.

**Keyword** discovered rule filtering, data mining, information retrieval, MEDLINE database, macro view approach, Jaccard coefficient

# 発見ルールフィルタリングへのマクロビューアプローチ

北村　泰彦　飯田　暁 †　朴　勤植 ‡　辰巳　昭治 †

関西学院大学理工学部　〒669-1337　三田市学園 2-1
† 大阪市立大学大学院工学研究科　〒558-8585　大阪市住吉区杉本 3-3-138
‡ 大阪市立大学大学院医学研究科　〒545-8585　大阪市阿倍野区旭町 1-4-3

E-mail: ykitamura@ksc.kwansei.ac.jp, † {iida, tatsumi}@kdel.info.eng.osaka-cu.ac.jp,
‡ kspark@msic.med.osaka-cu.ac.jp

**あらまし**　データマイニングシステムは与えられた大量のデータに隠されている知識を自動的に発見してくれる．しかしながら発見された知識は新規なもので，また利用者にとって興味深いものとは限らない．そこでわれわれは情報検索技法を用いることで，発見されたルール形式の知識を新しく，利用者にとって興味のあるものにフィルタリングする手法を提案する．本手法では，インターネット上の情報源からの検索結果に応じて発見ルールをランク付けする．本論文では，発見ルールフィルタリングへのアプローチをミクロビューアプローチとマクロビューアプローチに分類し，マクロビューアプローチに関して議論を行う．マクロビューアプローチでは情報検索により得られる Jaccard 係数の経年変化を観測することにより，発見されたルールがホットトピックであるかどうかを判定する．本論文では，医療データマイニングと MEDLINE 文献検索という具体的な事例を用いて，マクロビューアプローチの有効性について議論する．

**キーワード**　発見ルールフィルタリング，データマイニング，情報検索，MEDLINE データベース，マクロビューアプローチ，Jaccard 係数

## 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 obtained by 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 may discover (1) too many rules to be handled by the user, (2) rules that have been known to the user, and (3) rules that are not interesting to the user. To cope with these problems, we are developing a discovered rule filtering method that filters a large amount of knowledge discovered by a data mining system in a rule format 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.

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. In section 4, we show a feasibility of macro view approach by retrieving documents from the MEDLINE database. Finally we conclude this paper with our future work in Section 5.

## 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 Figure 1.

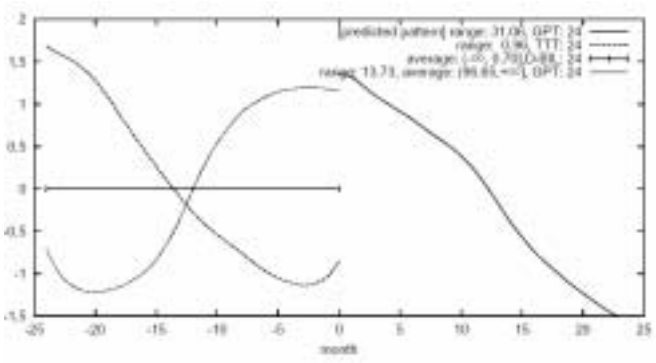

Figure 1. An example of discovered rule.

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 and/or uninteresting to 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 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 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.

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 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 Figure 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 converted into normal names. For example, TTT and GPT should be converted into "thymol turbidity test" and "glutamic pyruvic transaminase" respectively.
- **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 common keywords. For hepatitis data mining, "hepatitis" is the 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.

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. 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. In the next section, we discuss how to filter rules based on the result of information retrieval.

## 3. Micro View Approach and Macro View Approach

How to filter discovered rules according to the search result of MEDLINE document retrieval is a most important issue of this research work. We have two approaches; micro view approach and macro view approach, toward 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. It is effective when the number of related documents is small. We have two ways to show appropriate documents to the user.

(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 and those related to the user's interest as follows.

- **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 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 or indirectly by using a relevance feedback technique [2]. 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 classify documents into interesting ones and uninteresting ones.

(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.

Another measure is the Jaccard coefficient. The Jaccard coefficient between keywords K1 and K2 is defined as h({K1,K2})/(h({K1})+h({K2})) where h(X) specifies the number of hits when a set of keywords, X, is given to a search engine. It is a measure to show a relative strength of relation among keywords.

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 Jaccard coefficient, we can see the change of relation among keywords. For example, we can have the following interpretations as shown in Figure 2.

(a) If the Jaccard coefficient moves upward, the research topic related to the keywords is hot and promising.

(b) If the Jaccard coefficient moves downward, the research topic related to the keywords is closing.

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

(d) If the Jaccard coefficient keeps low, the research topic related to the keyword is not known. Few researchers show interest in the topic.

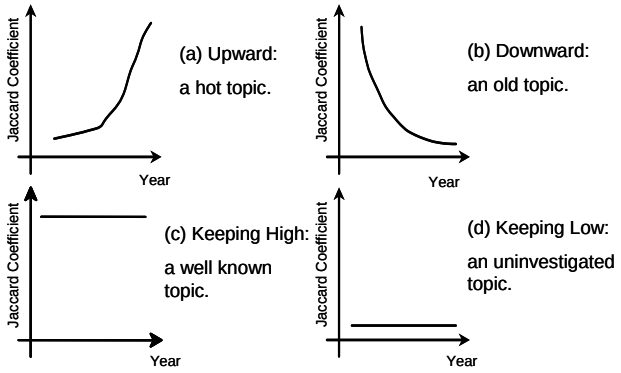


Figure 2: Yearly Trends of Jaccard Coefficient.

### 3.3. Integrating the micro view approach and the macro view approach

An ideal effect of discovered rule filtering is not only that it retrieves documents that relates to the rule, but also that it makes the number of documents small enough for the user to glance through them. As shown in [3], a user can have a meaningful insight from a combination of a discovered rule and related documents. As shown in Figure 3, this case is categorized in (A).

On the other hand, as categorized in (B), we have a rule that has few significant documents. Generally it means there is few research works related to the rule, but hopefully the unknown rule might lead to a new discovery.

As categorized in (C), we have a rule that hits a number of documents. In this case, it is difficult for the user to read through all of them and the macro view approach works well.

We have two ways to integrate the micro view and the macro view approaches. As shown in (X), even if a rule hits a number of documents, the user actually shows interest in a part of them. It is effective to categorize documents and to show the user according to the category. The relevance feedback is an applicable technique in this case.

As shown in (Y), even if a rule hits few documents, we may have a number of documents by generalizing the rule. A way of generalizing a rule is to eliminate one or more keywords specified in the rule. When a generalized rule hits a number of documents, the macro view approach is available.

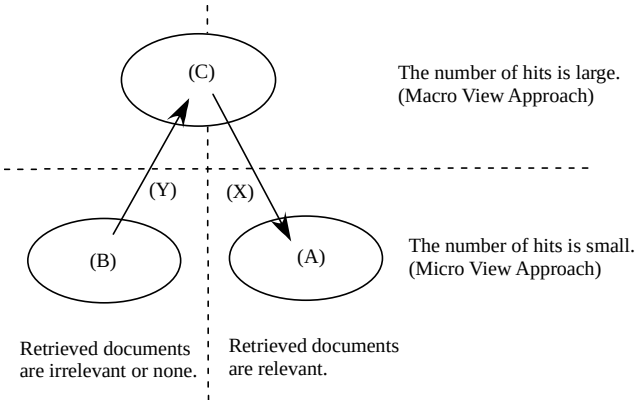


Figure 3: Integration of Micro View and Macro View Approaches.

## 4. Feasibility of Macro View Approach

To evaluate a feasibility of macro view approach, we submitted 4 queries to the MEDLINE database and examined 4 cases shown in Figure 4 through 7.

(a) "hcv, hepatitis" (Figure 4)

The Jaccard coefficient has been increasing since 1989. In 1989, we have an event of succeeding HCV cloning. HCV is a hot topic and has been researched.

(b) "smallpox, vaccine" (Figure 5)

The Jaccard coefficient has been decreasing. In 1980, the World Health Assembly announced that smallpox had been eradicated. Recently, we see the coefficient turns to increase because discussions about smallpox as a biochemical weapon arise

(c) "gpt, got" (Figure 6)

The Jaccard coefficient 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" (Figure 7)

The Jaccard coefficient stays low. The relation between albumin and urea nitrogen is seldom discussed.

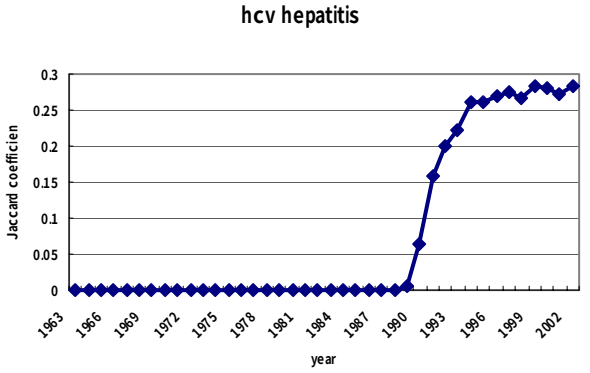


Figure 4: The yearly trend of Jaccard Coefficient between "hcv" and "hepatitis".

.

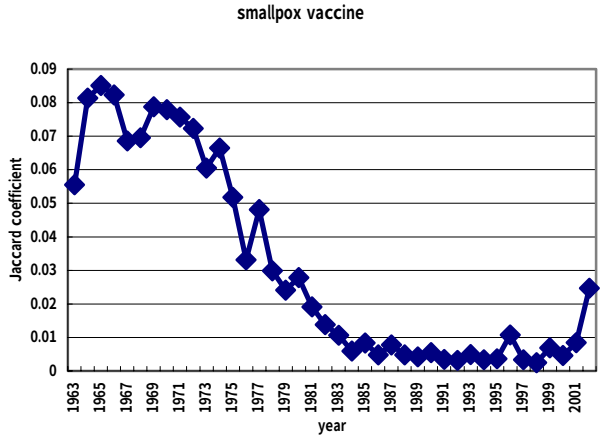


Figure 5: The yearly trend of Jaccard Coefficient between "smallpox" and "vaccine".

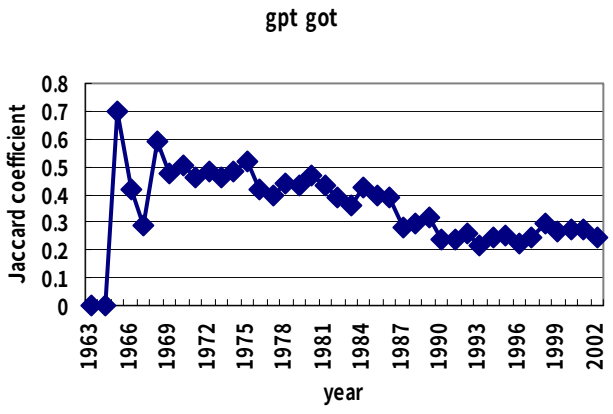


Figure 6: The yearly trend of Jaccard Coefficient between "gpt" and "got".

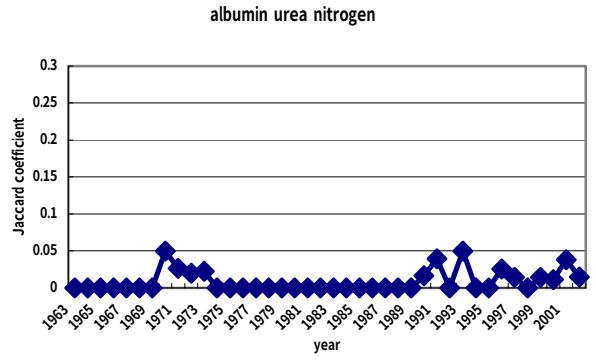


Figure 7: The yearly trend of Jaccard Coefficient between "albumin" and "urea nitrogen".

## 5. Summary

We discussed the discovered rule filtering method to filter the results from a data mining system to be novel ones for the user. We proposed two approaches; the micro view approach and the macro view approach toward the discovered rule filtering, and showed the feasibility of macro view approach. Our future work is summarized as follows.

- Evaluating the effect of discovered rule filtering. We need to examine the relation between the novelty of discovered rule and the result of information retrieval.
- Improving the performance of information retrieval. By using the relevance feedback and natural language processing techniques, we need to improve the performance of information retrieval to meet the user's interest.
- Developing a discovered rule filtering system. We need to develop a system that automatically performs the process of discovered rule filtering.
- Applying the discovered rule filtering technique to real-world research domains. We are going to apply our system to support tasks of mining hepatitis data and show the effectiveness of the system.

**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.