Discovered Rule Filtering Using Information Retrieval Technique

Yasuhiko Kitamura*, Keunsik Park**, Akira Iida*, and Shoji Tatsumi*
*Graduate School of Engineering, **Graduate School of Medicine
Osaka City University, JAPAN
kitamura@info.eng.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 propose a discovered rule filtering method to filter rules discovered by a data mining system and to produce ones that are novel and interesting to the user by using information retrieval technique. In the method, we rank discovered rules according to the result of information retrieval from the Internet. In this paper, we show the steps of discovered rule filtering by using a concrete example of clinical data mining and MEDLINE document retrieval. Preliminary results show that this method has merits in not only filtering discovered rules but also providing a new viewpoint to the rules to give a chance to invoke a new data mining process.

1. Introduction

As information technology becomes indispensable for our daily life, a huge amount of information is proliferated in the world. The speed and amount of the proliferation has been further accelerated by the advent of Internet and the available information is almost flooding us. From such a huge amount of various and noisy information, we need new tools to discover useful information or knowledge that meets demands of individual user. Active mining is a new direction of data mining and aims at discovering valuable knowledge for users in an efficient way by integrating data mining, information retrieval, and user reaction techniques [1].

As an approach to active mining, we have interest in integrating data mining and information retrieval techniques [3]. By using a data mining system, we can semi-automatically discover a number of rules hidden in a set of data, but each of the discovered rules can be classified according to the following characteristics.

(1) Does the rule express an important fact or not?
(2) Does the rule express a novel fact or a known fact?
(3) Does the rule express a fact that is interesting to the user or not?

Of course, we would like to discover rules that are important, novel, and interesting to the user. Conventional data mining systems mainly try to deal with the characteristic (1); the importance of rules, for example, by using a statistics approach. Some systems rank rules by using the precision and recall value of each rule. However, it is not easy to deal with other characteristics; the novelty of rule and the significance to the user because the novelty may change as the time goes on and the significance depends on the user's preference or interest.

To deal with characteristics (2) and (3), we try to utilize information retrieval results from the Internet. On the Internet, a huge amount of information is stored and is updated frequently. By retrieving latest information from the Internet, we may check whether a discovered rule is novel or not. Moreover, by monitoring the user's behavior of retrieving her preferred information, we may learn her preference and interest and may utilize them to check whether a discovered rule is interesting to her.

In this paper, we discuss the discovered rule filtering technique based on information retrieval from the Internet. In Section 2, we discuss the steps of discovered rule filtering. We here show an example when we apply the technique to a data mining task from a clinical examination database about hepatitis. We show preliminary results in Section 3 and conclude this paper in Section 4.

2. Discovered Rule Filtering in Hepatitis Data Mining

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.

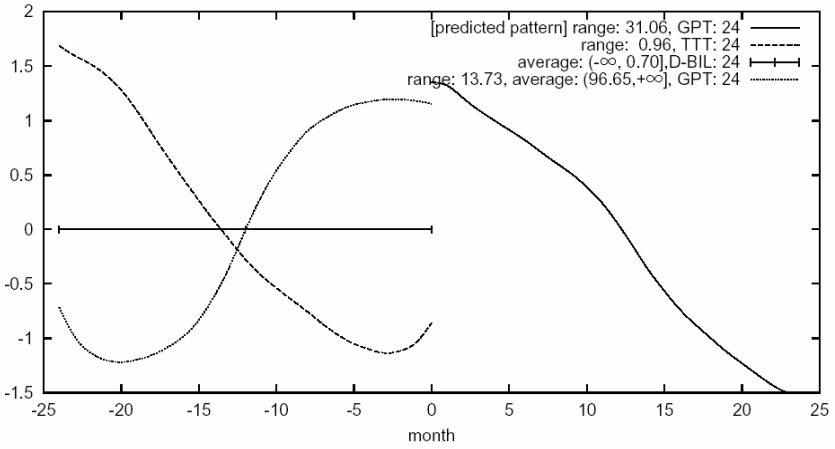

Figure 1. An example of discovered rule.

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.

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 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 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 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 or indirectly by using a relevance feedback technique as mentioned in Step 4.

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 Figure 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 Figure 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 "total cholesterol" is better than the combination of "hepatitis" and "gpt" because the former is expected to have more returns than the latter.

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."

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

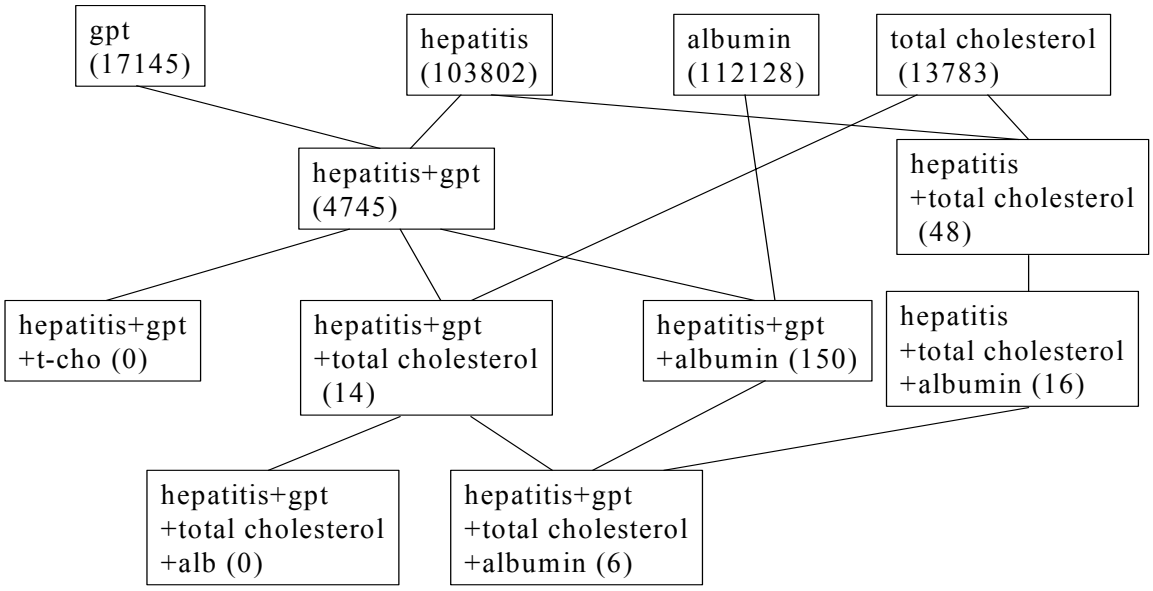

Figure 2. A graph that represents document retrieval history.

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]

Figure 3. A document retrieved.

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.

3. Preliminary Results and Discussion

To evaluate the feasibility of discovered rule filtering technique, we manually examined the number of retrieved documents for each of 30 rules discovered by Takahira Yamaguchi group at Shizuoka University. We used keywords only that are related to attributes of discovered rules and the domain. For 12 rules in 30, we could succeed to retrieve 7.3 documents in average. For 18 rules, we retrieved no documents. However, no hit does not always mean that the rule has a novel fact. Even when the rule contains no important facts, in other words it is just a garbage, it is likely that the system retrieves no documents. When the reliability of output from a data mining system is low, the discovered rule filtering does not work well.

We here discuss advantages of discovered rule filtering to deal with the characteristics (2) and (3) mentioned in Section 1. If we submit a set of proper keywords to the Pubmed system, we can roughly know how much work related to the keywords has been done in the research field. For example, if we submit "hepatitis GPT TTT," we have only 3 documents. On the other hand, if we submit "hepatitis GPT GOT," we have 1878 documents. The difference of the numbers is quite reasonable because GPT and GOT are well known blood test data to examine hepatitis. In addition, we know the relation between GPT and TTT has not been studied very much in the research field of hepatitis. Therefore, we may be able to conclude that a rule with attributes TTT and GPT looks more attractive than one with attributes GPT and GOT.

The number of retrieved documents changes depending on whether a user has a special interest. Let us assume a user has interest in the periodicity of attribute value. If we submit "hepatitis and GPT," we receive 4798 documents, but if we add "periodicity" to the keywords, we receive only 12 documents.

Of course, our information retrieval method based on keywords submission tends to produce noisy documents. We still need to improve the performance and we expect that the relevance feedback technique plays an important rule because it can narrow the space of document appropriately by using feedbacks from the user. We have not quantitatively examined how effectively the rule filtering technique works and left it as our future work.

However, we would like to report a side effect of showing discovered rules and related documents to a user (a medical doctor). In our preliminary experiment, at first we showed a discovered rule alone, shown in Figure 1, to the user and received the following comment (Comment 1). The discovered rule looks a part of common facts to the user.

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 Figure 3, in which the user 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.

4. Conclusions and Future Study

In this paper, we proposed the discovered rule filtering by integrating data mining and information retrieval techniques and showed the steps to develop a system. In a preliminary experiment, we show the technique contribute not only filtering discovered rules but also providing users a new viewpoint toward discovered rules and a motivation to invoke a new mining process. We believe this is a new approach to active data mining. Our future works are 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 other 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 task of mining hepatitis data and show the effectiveness of the system.

Acknowledgement

We would like express out thanks to Professor Takahira Yamaguchi for giving us rules discovered at his laboratory for our preliminary experiment and also helpful comments for our work. This work is partly supported by Grant-in-Aide for Scientific Research (13131209) from the Ministry of Education, Culture, Sports, Science and Technology.

References

[1] Motoda, H. (Ed.), Active Mining: New Directions of Data Mining, IOS Press, Amsterdam, 2002.

[2] Kitamura, K., Nozaki, T., and Tatsumi, S. "A Script-Based WWW Information Integration Support System and Its Application to Genome Databases," Systems and Computers in Japan, Vol.29, No.14, 1998, pp.32-40

[3] Baeza-Yates, R. and Ribeiro-Neto, B. Modern Information Retrieval, Addison Wesley, 1999