Effective Browsing of the Discovered Association Rules Using the Web

Yiming Ma, Ching Kian Wong, Bing Liu
School of Computing
National University of Singapore
3 Science Drive 2, Singapore 117543
{maym, wongck, liub}@comp.nus.edu.sg

ABSTRACT

Interpreting the discovered knowledge to gain a good understanding of the domain is one of the important phases of the KDD process. It usually requires the user to browse a set of discovered rules during the post-processing phase. Trying to interpret a set of association rules is not a trivial task as the number of rules in the complete set is usually very large. This is because association rule mining aims to discover all the rules present in the data. One way of overcoming this is to employ some form of summarization so that only the essential subset of the rules is presented to the user. In this paper, we describe a system, called DS-WEB, which assists users in interpreting a set of association rules. We first use a summarized set of rules known as direction-setting rules (DS rules) to build a hierarchical structure for easy browsing of the complete set of rules. We then propose to publish this hierarchy of rules via multiple web pages connected using hypertext links. By utilizing a web interface, we inherit much of the web's advantage like accessibility and user-friendly interface. We are also able to make use of the Internet's interconnected network as an infrastructure for communication and information sharing. This allows us to create a virtual web workspace where multi-users can share opinions on the rules. This ultimately contributes towards comprehension of the domain.

Keywords

Domain understanding, post-processing, association rules, Web.

1. INTRODUCTION

Interpreting the discovered rules to gain a good understanding of the domain is one of the important phases of the KDD process. It usually requires the user to browse a set of discovered rules or patterns. Typical techniques employ visualization, documentation or summarization methods to translate the discovered rules into a more understandable form.

Trying to interpret a set of association rules is not a trivial task as the number of rules in the complete set is usually too large. This is because association rule mining aims to discover all the patterns present in the data. Furthermore, there is no logical ordering of the rules other than a ranking based on the confidence or the support level of the rules. However, using confidence and support might not be a good way of organizing the set of rules, as they are not directly related to how interesting a rule is to the user. It will however be interesting if we can interpret the large set of rules since it contains the complete knowledge of the domain [AIS93, AS94].

One way of overcoming this is to employ some form of summarization so that only the essential subset of the rules are presented to the user at the beginning. Several research efforts have been focused on trying to find a summarized subset of the complete set of rules. In [LHM99] the authors proposed a novel method to summarize the complete set of rules to only the essential relationships in the domain. The rules in this subset are called the *direction-setting rules* (DS rules). We propose to use DS rules to organize the large number of rules into a hierarchical structure to facilitate browsing. At the top of the hierarchy are all the DS rules with bottom levels made up of following non-DS rules. This technique will be discussed in detail in section 3.

In recent years, the Internet has become the single most important medium for communication and information sharing. Among all its applications, the world-wide-web is the most popular. The web owes its popularity to its accessibility and also its intuitive interface. We can easily inherit some of the web's advantages if we can deliver our knowledge for post-processing via the web.

1) Contents published on the web can be accessed from any computer in the world. This eliminates the need for the distant party to re-execute the learner system in order to get access to the same knowledge.
2) The web offers a familiar interface environment to browse the discovered knowledge. This vastly reduces any learning curve that might otherwise be required for other formats like graphs or maps.

Rules can be easily adapted and published as web pages since rules, in its natural form, is already sentence-like. Our proposed hierarchical structure can be easily implemented by utilizing hypertext links to simulate the next level of rules. Users start from the top level DS rules and then interactively drill down to access the non-DS rules. This aids the analysis and interpretation process.

Besides offering a web interface for easy browsing of the rules, the Internet also provides a basic infrastructure for co-operative work to be performed through its interconnected network. Users can individually browse, evaluate and comprehend the rules while browsing the web pages and at the same time choose to share whatever comments, insights or ideas about them with others. This creates a virtual multi-user workspace. We believe that it is logical and realistic to have multiple users to analyze the complete set of rules because (1) many users might be interested in the rules, and (2) the size of the rule set could overwhelm a single user, and (3) by combining the expertise of different users, we could eliminate any bias a single user might have towards certain domain knowledge.

In this paper, we present our post-processing framework and the DS-WEB system that implements the framework. The paper is organized as follows: Section 2 discusses the related work. Section 3 describes the direction setting rules used in building our hierarchical structure. Section 4 reports our system, DS-WEB, which is based on the proposed framework. Section 5 gives an account of our application experience with the system. Finally, we'll conclude in section 6.

2. Related Work

Several commercial products offer the capability of performing post-processing of discovered association rules. Current popular association rule miners in, e.g. SGI MineSet [SGI], IBM Intelligent Miner [IBM], Clementine [CLE], DBMiner [DBM] etc., present discovered association rules either in graphical or text form. Typically, the association rules are ranked by their confidence and support levels. This usually led the user to focus only on the high confidence and high support rules. However, high confidence or support level may not imply the importance or interestingness of a rule. We feel that this simple strategy does not fully explore the potentials of the discovered association rules. Users should be allowed to browse and analyze the rules more comprehensively without being overwhelmed by too many rules. Our framework proposes to use direction-setting rules to summarize the discovered rules and then use it to create a logical hierarchical structure so that users can interactively drill down to an interesting topic.

There are a number of reported works on the topic of discovering interesting association rules in the KDD literature [ST95, SA96, SA95, LHM98, etc.]. The focus is to use some interestingness measures to prune the discovered associations. We can easily adapt these algorithms into our proposed framework. In fact, in the process of discovering DS rules, our system performs pruning to remove those insignificant rules before summarizing the remaining significant rules to a set of DS rules. We will give more details about this in the next section.

Many existing systems have adopted a visual approach to analyzing discovered knowledge, most notably, the MineSet system from SGI [SGI]. In MineSet, an innovative method is used to offer a visual tour of a wide variety of concept descriptions (e.g. decision trees, decision tables, regression trees, scatter graphs, etc). Users can interact with these visual models to incrementally extract interesting knowledge. Our work promotes a similar idea of allowing users to interact directly with the concept description for incremental understanding. However, we also actively help the user by doing post-processing, i.e. summarization and organization. In addition, our framework builds an infrastructure to allow for collaborative work between users via the web.

There are also a number of vendors that offer web-based analysis tools. JWAVE [VNI] from Visual Numerics is a web-based software that can be used to produce visually comprehensive models of a company's data. However, it differs from our work in that it analyses attributes but not concept descriptions discovered from a mining process. WebSemba [ASO] from ASOC also utilizes the web as an infrastructure for delivering derived knowledge. However, it focuses on utilizing the knowledge and not on performing post-processing in order for the user to understand the domain.

3. Post-processing the Discovered Associations through Summarization

In this section, we present a systematic way of summarizing the discovered associations. We summarize the discovered associations to a small set of *direction setting rules* (DS rules). The DS rules give a summary of the behavior of the discovered associations. They represent the essential relationships or structure (or skeleton) of the domain. The non-DS rules simply give additional details. Using the DS rules as a summary, the user can interactively focus on the key aspects of the domain and selectively view the relevant details (non-DS rules).

3.1 General Ideas

In this work, we focus on association rule mining from a relational table, which consists of a set of records described by a number of attributes. An item is an attribute value pair, i.e., (attribute = value) (numeric attributes are discretized). Association rule mining in such data is typically targeted at a specific attribute because the user normally wants to know how other attributes are related to this *target* attribute (which can have many values) [LHM98, BAG99]. With a target attribute, we can express an itemset as follows (instead of a set of items as in [AIS93]):

$$X \rightarrow y$$

where y is an item (or a value) of the target attribute, and X is a set of items from the rest of the attributes. For simplicity, we call this itemset a *rule* hereafter, regardless of whether it is significant or not. We also say a rule is *large* if it meets the minimum support.

The proposed technique first prunes the discovered associations to remove those insignificant ones, and then finds a special subset of the unpruned associations to form a summary of the discovered associations. In the following subsection, we use an example to briefly introduce how we summarize the set of unpruned associations. The detailed descriptions of the technique and also how pruning is performed can be found in [LHM99].

3.2 Summarizing the unpruned rules

Pruning can reduce the number of rules substantially. However, the number of rules left can still be very large. We find a subset of the rules, called direction-setting rules (or DS rules), to summarize the unpruned rules. Essentially, DS rules are significant association rules that set the directions for non-DS rules to follow. The direction of a rule is the type of correlation it has, i.e., *positive correlation* or *negative correlation* or *independence*, which is computed using χ^2 test (see [LHM99]). Let us see an example.

Example: We have the following discovered rules:

R1: Job = yes → Loan = approved [sup = 40%, conf = 70%]
R2: Own_house = yes → Loan = approved [sup = 30%, conf = 75%]

χ^2 analysis shows that having a job is positively correlated to the grant of a loan, and owning a house is also positively correlated to obtaining a loan. Then, the following association is not so surprising to us:

R3: Job = yes, Own_house = yes → Loan = approved [sup = 20%, conf =90%]

because it intuitively follows R1 and R2. We can use R1 and R2 to provide a summary of the three rules. R1 and R2 are DS rules as they set the direction (positive correlation) that is followed by R3. In real-life data sets, a large number of associations are like R3.

From the example, we see that the DS rules give the essential relationships of the domain. The non-DS rule is not surprising if we already know the DS rules. However, this, by no means, says that non-DS rules are not interesting. Non-DS rules can provide further details about the domain. For example, the non-DS rule above (R3) gives a higher confidence, which may be of interest to the user. Using DS rules to form a summary is analogous to summarization of a text article. From the summary, we know the essence of the article. If we are interested in the details of a particular aspect, the summary can point us to them in the article. In the same way, the DS rules give the essence of the domain and points the user to those related non-DS rules. Non-DS rules are basically combinations of DS rules. In the above example, R3 is a combination of R1 and R2.

Experiment results, including real-life applications, show that although the number of discovered rules can be huge, the number of DS rules is very small [LHM99]. They can be analyzed manually by the human user to obtain the essential relationships in the data. He/she can then focus his/her attention on those interesting aspects of the relationships, and to see the relevant non-DS rules. This process can be easily facilitated by an interactive user interface [LHM99]. The user can now obtain a complete picture of the domain without being overwhelmed by a huge number of rules.

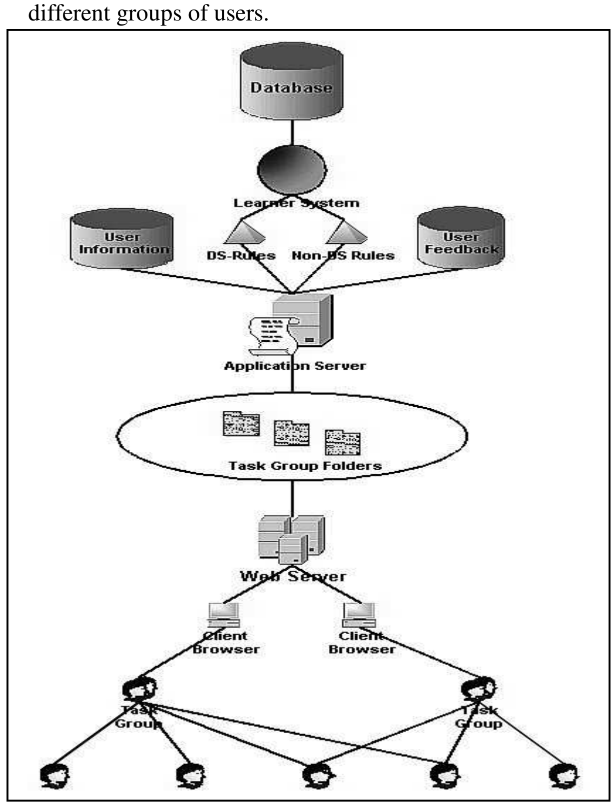
4. Post-processing system: DS-WEB

In this section, we present our system DS-WEB. Its framework exploits the web-based client/server architecture. It basically consists of a set of dynamically generated web pages that describes the discovered knowledge. These pages can be mounted, through a web server, onto either the Internet or a corporate intranet to enable sharing and collaboration between users. In the rest of this section, we introduce our server program in the sub-section 4.1, and client program in the sub-section 4.2.

4.1 Server program

We present our server-client architecture in Figure 1. There are two servers in the diagram:

1) Application Server: Responsible for organizing the mining results and the user information.
2) WEB server: Responsible for presenting the mining results to different groups of users.

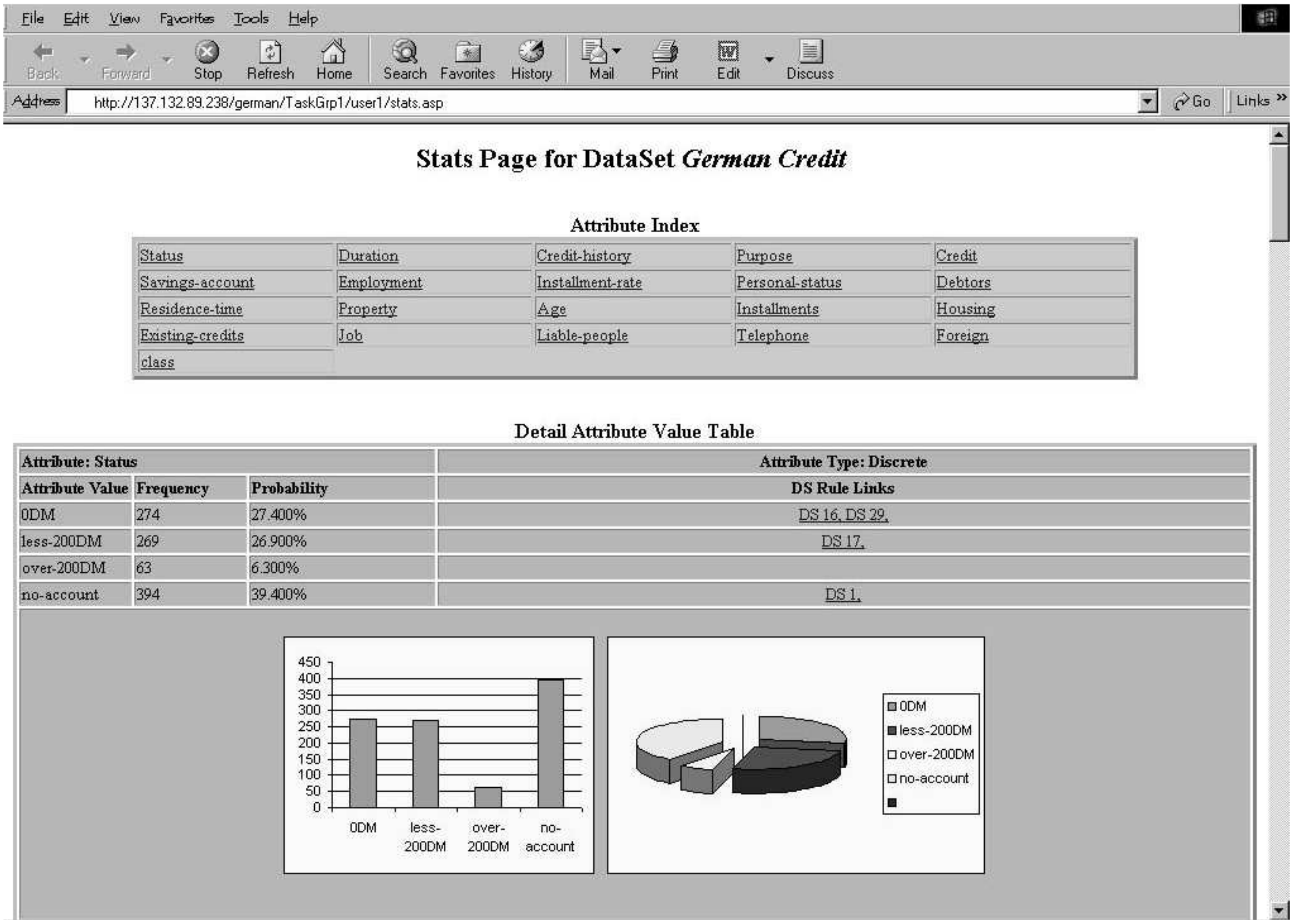


(Figure 1. High-level view of DS-WEB)

4.1.1 Application server

Application server mainly organizes the following three types of information:

1) User information: Different users have different access rights to the mining results. On the server, we define the task groups, and store them in a user information database.
2) DS rules and non-DS rules (mining results): Based on the task groups, the server further filters the rules, and set up *task group folders* (a virtual folder to dynamically provide the relevant DS, non-DS rules to different task groups).
3) User feedback information: A user in a task group can write his/her comments on the rules and also rate the rules. These are kept in a database, and organized by the server.



(Figure 2. Attribute Distribution Page)

4.1.2 WEB server

The group folder basically stores a set of rule indexes to the rule database. Based on the requests from different task groups, WEB server dynamically retrieves the relevant rules from the respective task group folder and presents them as a set of HTML pages to the users.

4.2 Client program

Once the client program connects to the server, all the accessible task group folders are shown to the users. Users can open these task group folders, and retrieve the relevant rules from each folder. Based on the DS-rules and non-DS rules on the WEB server, we present a set of web pages to users. In the following sub-sections, we introduce the process for a systematic analysis and understanding of the association rule space.

4.2.1 Browsing the attribute distribution information

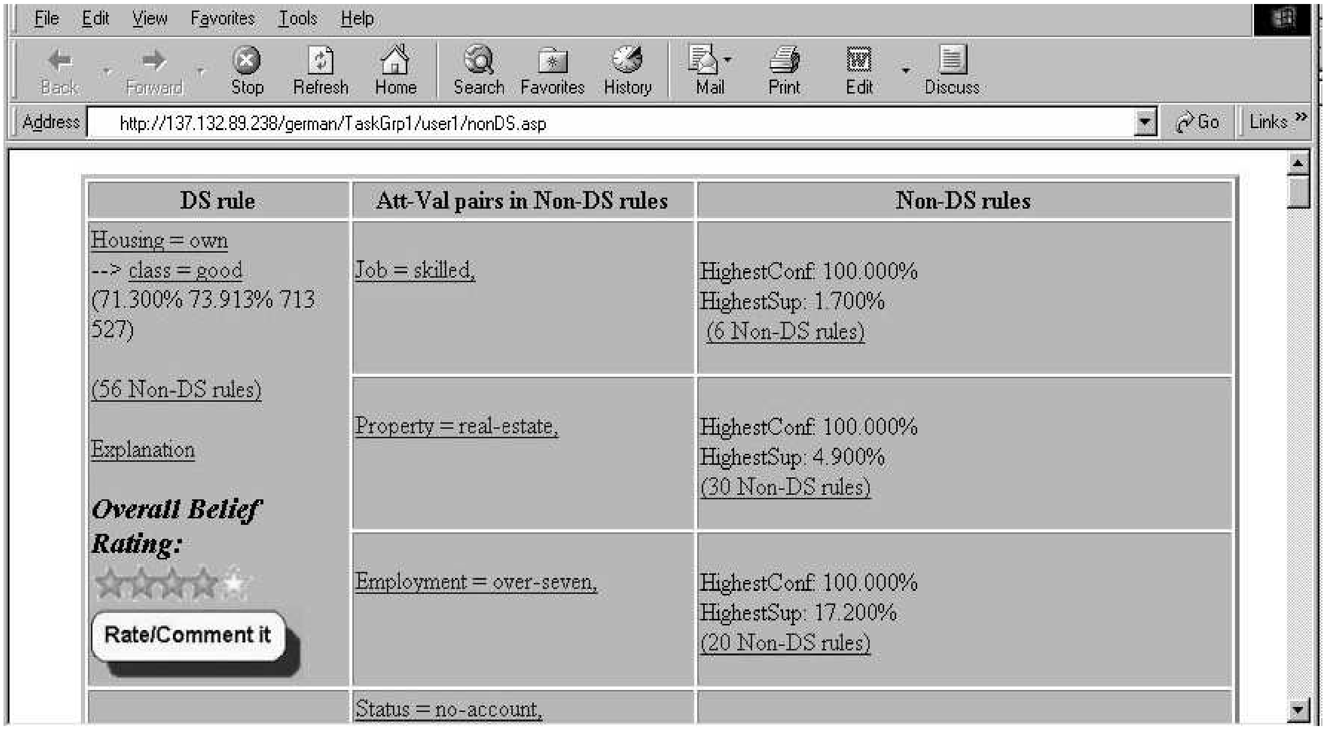
Attribute distribution information can be very useful to the users. From this information, the user is able to get a basic feel of the data. All our real-life applications have shown that users would like to understand the attribute distribution before other mining results are presented.

Therefore, the first web page presented to the user is the attribute distribution information of the databases. We also provide simple graph plots to help users understand the attribute distributions. The user also can export the information to other statistical packages.

More importantly, this page provides a starting point to the DS-rules. After analyzing the attribute distributions, the user can directly go to the DS rules, and start to explore the relevant knowledge. All the DS-rules are indexed. The set of DS-rule links are properly placed under relevant attribute-value pairs or conditions (Figure 2) (the sample used in this section is taken from *German Credit* dataset [MM96]).

4.2.2. Exploring a DS-rule and its non-DS rules

To help users understand a rule (DS or non-DS), we present our rules in two forms, one for expert users and the other for novice users. The expert rule format (Figure 3) presents a rule in a conjunctive form. The advantage of this format is that expert users can view the relevant information quickly. In Figure 3, the first



(Figure 5. Navigating non-DS rules)

number within the brackets is the support, the second is the confidence, the third is the support count of the conditions and the fourth is the support count of the conditions and the consequent. However, a novice user might encounter problems in understanding the rule in this form. To overcome this, we offer a novice format, which is presented in a more understandable form (Figure 4). By making use of the support and confidence information from the rule, we rewrite the rule, and add in some qualitative measures (i.e. Large portion of data, High confidence etc., which are suggested by domain experts).

Housing = own
Employment = over-seven-years
→ class = Loan Approved
(17.200%, 100.000%, 172, 172)

⇒

A **large portion** of data (17.200%) having value:
Housing = own

Employment = over-seven-years
leads to: class = Loan Approved with **very high confidence 100.000%** compare to the distribution of the: class = Loan Approved of 70.000%

(Figure 3. Rule Format for Expert)

(Figure 4. Rule Format for Novice)

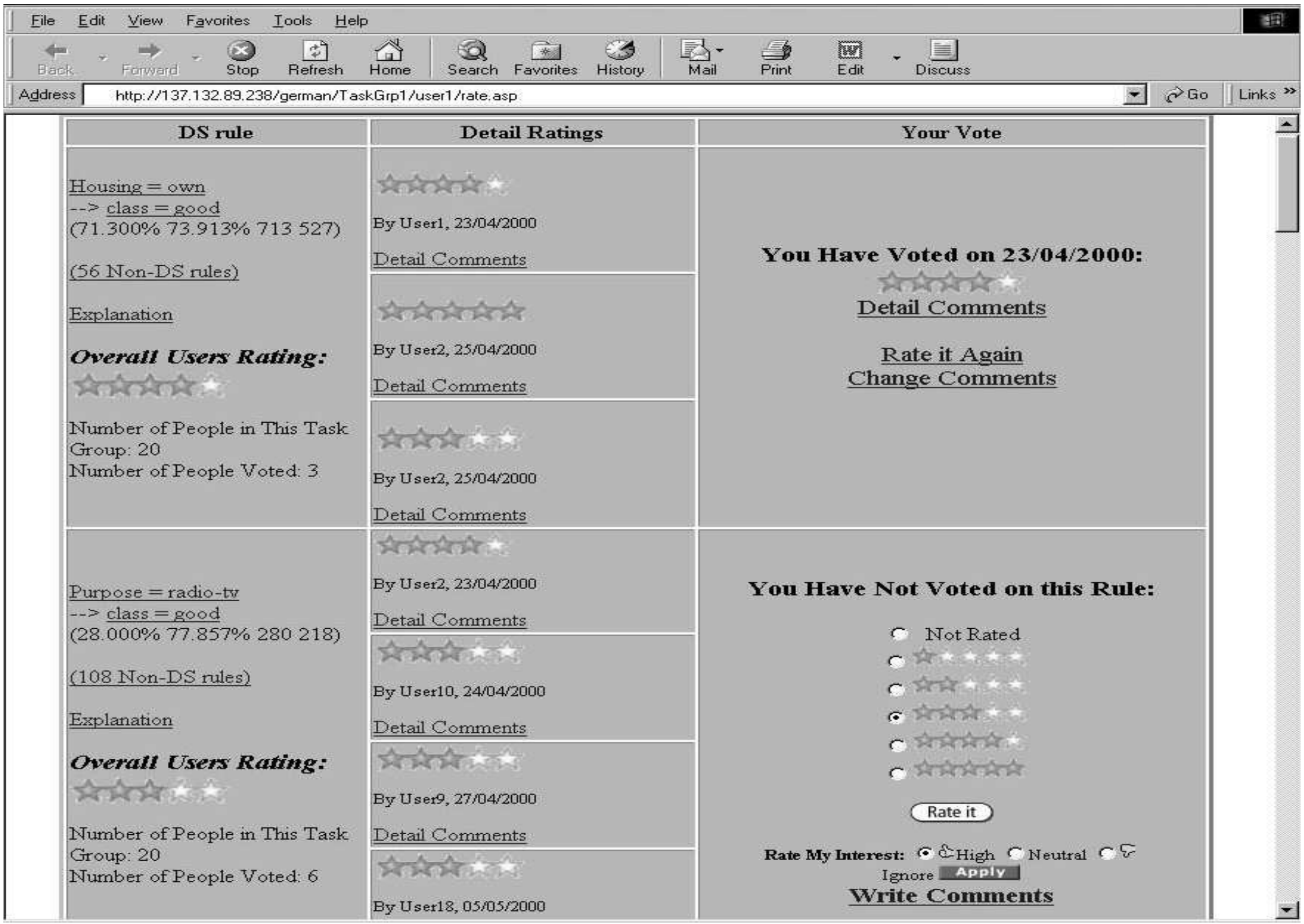
As we mentioned earlier, non-DS rules are the followers of the DS rules with more conditions and higher confidences. Some of the high confidence non-DS rules could also be important for understanding the domain. We use the following HTML page to explore the non-DS rules.

In column 1 of the above sample HTML page (Figure 5), we present the DS rules in the expert format (Figure 3). The user can directly jump to the relevant non-DS rules from the non-DS rules hyperlink. He/she can also switch to the novice rule format (Figure 4) by following the *Explanation* hyperlink. At the bottom of each DS rule, we provide the overall user *belief rating* of the rule. This rating gives the user some ideas of how believable this rule is to other users in the same task group. A user can see the detailed ratings and give his/her own views of the rule by clicking the *Rate/Comment it* button, which leads to Figure 6.

From our application experiences, the users typically would like to focus on a few important attributes in exploring the non-DS rules. Hence, in column 2, we present the additional conditions (attribute value pairs) that non-DS rules have. They give a rough knowledge of what conditions are involved in the non-DS rules. The user can easily choose the interesting conditions to start off.

Column 3 contains a hyperlink to the set of all such non-DS rules. We also let the user know how many such non-DS rules there are in total and give the highest confidence, and support level in of the non-DS rules. By viewing the highest confidence and support information, users will have some ideas of the non-DS rules involved.

It is also interesting to understand the relationship between DS rules. In our system, we group the similar DS rules together. By carefully comparing and contrasting the DS-rules, we could provide the user more useful information.
On top of these features, we also have rule search utilities to help



(Figure 6. WEB based Review Form)

the user search for the relevant rules that match certain criteria.

4.2.3. Collaboration Feature

In DS-WEB, we construct a virtual web workspace where users can share opinion on the perceived believability and the interestingness/usefulness of a particular rule. As mentioned earlier, users can rate each DS-rule or non-DS rule, and log their comments. Such features are not only useful in a multi-user environment but also useful in a single user case. Figure 6 gives a screen shot of the web-based review (rating and commenting) form.

In DS-WEB, we allow two types of rating to be given by users.

1). *Belief rating*: The user uses this rating to express his/her opinion on the believability of the rule.

2). *Interestingness rating*: The user uses this rating to indicate his/her interest in the rule. This information is used in collaborative filtering (see below).

In Figure 6, column 1 lists each individual rules and their overall consolidated belief rating together with information about the number of users that have contributed to the rating. This gives the user an idea of the perceived quality of the rule at one glance. Details of how each individual user rated this rule are listed in column 2. Users are also allowed to attach annotations or comments to each rule. This serves as an explanation of why a particular rating is given to a rule. Links to these comments also appear in column 2. Column 3 provides a means for the user to submit his/her own belief rating, interestingness rating and comments on a rule. Submitted information will then be consolidated by the application server and become visible to the other users. Note that the interestingness rating of a user is only used by the system (see below) and not revealed to others users. The system also provides an option to apply different weights to ratings from different users, based on certain discriminating criteria like seniority or user expertise in a particular area.

The user can also request for a listing of rules sorted based on either his/her own or on overall consolidated ratings (belief or interestingness ratings). A sorted list of rules based on the user's own ratings reminds the user of his/her previous works while a sorted list of rules based on overall consolidated ratings gives the user an idea of how others in the task group think. The system can compare a user's rating with the overall rating given by others in the group on the different rules to give a *Conformity Level.* This shows the user how inline his/her views are as compared to the group.

The interestingness rating (Figure 7, also in Figure 6) not only allows the user to record down those interesting rules, but also allows the system to perform collaborative filtering, i.e., the system can suggest a list of other rules that might be of interest to the user based on the records of others with the same interest.

(Figure 7. Indicating interest level to a rule)

5. Applications of DS-WEB

We have used our system in a number of real-life applications. We are still enhancing the system to suit the users' needs. In this section, we briefly describe one of these applications we have done for an education institution.

This institution needs to analyze its students cross the years. One of the tasks is to study the profile of the students. We have the complete set of student data. There are 45 attributes. It includes education, family background, sex, ages, etc. In this application, it is not sufficient to just analyze the statistical distribution of each attribute. The institution is more interested in correlations among the attributes. More importantly, they would like to have a good understanding of their students (or the domain).

When we run the normal association rule-mining engine, it generates a large number of associations (more than 20,000), which are very difficult for the users to understand. When we run our DS-WEB on this set of data, the number of DS rules generated is 138, and non-DS rules is 421. It now becomes possible for the human users to interpret them manually.

There are 4 users working on this project. Since they are all familiar with the WEB interface, they can use our system without much difficulty. They only used less than an hour to select the relevant rules, and started to write their comments on the rules. By using our review form (Figure 6), they easily gathered the opinions from one another. After working for half a day, they identified 17 DS rules and 45 non-DS rules that are important for understanding the domain. The collaboration features of our system make it easy for users to communicate with each other.

In this application, we find that the WEB interface is very useful. Users are familiar with it, and can work in their offices using a web browser without installing any software systems on their computer. Based on our experiences, we are now working on a more generic and powerful system.

6. Conclusion

This paper proposes a framework to aid the user in analyzing a set of association rules. The set of rules is first organized into a hierarchical structure using DS rules to facilitate browsing. The framework then uses the web as an infrastructure to create a virtual workspace where multi-users can co-operatively work towards evaluating rules in the rule set to filter out noisy and imprecise knowledge by sharing opinions. This contributes significantly towards the final comprehension of the discovered knowledge. Although this paper focus primarily on association rules, this framework can also be applied to other concept descriptions with little modification.

We envision that not only will the web be the primary channel for knowledge delivery in the near future but much of the KDD process will also shift towards a distributed multi-participant environment. Performing post-processing as proposed in our framework is very natural.

REFERENCES

[AIS93] Agrawal, R., Imielinski, T., Swami, A. "Mining association rules between sets of items in large databases." *SIGMOD-1993* , 1993.

[AS94] Agrawal, R. and Srikant, R. 1994. "Fast algorithms for mining association rules." *VLDB-94*, 1994.

[AS96] Agrawal, R. and Shafer J. 1996 "Parallel mining of association rules." *IEEE Knowledge & Data Engg*. 1996.

[ASO] http://www.asoc.com/ehome.html

[B98] Roberto J. Bayardo. "Efficiently mining long patterns from databases." *SIGMOD-98*, 1998.

[BAG99] Bayardo, R., Agrawal, R, and Gunopulos, D. "Constraint-based rule mining in large, dense databases." *ICDE-99*, 1999.

[CLE] http://www.spss.com/clementine/

[DBM] http://www.dbminer.com/

[FI93] Fayyad, U. M., Irani, K. B. "Multi-interval discretization of continuous-valued attributes for classification learning." *IJCAI-93*, 1022-1027.

[IBM] http://www-4.ibm.com/software/data/iminer/

[KJLM94] Kohavi, R., John, G., Long, R., Manley, D., and Pfleger, K. 1994. MLC++: a machine learning library in C++. *Tools with artificial intelligence*, 740-743.

[LHM98] Liu B., Hsu W. and Ma Y. "Integrating Classification and Association Rule Mining." *KDD-98*, 1998

[LHM99] Liu B., Hsu W. and Ma Y. "Pruning and Summarizing the Discovered Associations." *KDD-99*, 1999

[MM96] Merz, C. J, and Murphy, P. UCI repository of machine learning database.

[http://www.cs.uci.edu/~mlearn/MLRepository.html].

[PCY95] Park, J. S., Chen, M-S, and Yu, P. S. "An Effective Hash Based Algorithm for Mining Association Rules" *SIGMOD-95*, 1995.

[Q94] Quinlan, J. *C4.5: programs for machine learning* Morgan Kaufmann, 1994.

[SA95] Srikant, R. and Agrawal, R. "Mining generalized association rules." *VLDB-1995*, 1995.

[SA96] Srikant, R. and Agrawal, R. "Mining quantitative association rules in large relational tables." *SIGMOD-96*. 1996.

[SGI] http://www.sgi.com/software/mineset/

[ST95] Silberschatz, A., and Tuzhilin, A. "On Subjective Measure of Interestingness in Knowledge Discovery" *KDD-95*, 1995

[VNI] http://www.vni.com/products/wpd/jwave/index.html