Web for Data Mining Applications

Bing Liu, Yiming Ma and Ching Kian Wong

School of Computing
National University of Singapore
3 Science Drive 2, Singapore 117543
{liub, maym, wongck}@comp.nus.edu.sg

Abstract

The Web not only contains a huge amount of information, but also provides a powerful infrastructure for communication and information sharing. While mining for valuable information or resources from the Web is an active research area, in this paper, we focus on discussing the use of the Web for data mining applications. In particular, we show that the Web can be used to facilitate delivering and interpreting a set of discovered rules.

1. Introduction

It is well known that the Web provides an excellent source of information for data mining. We can mine the content of the Web to discover useful information or resources, or mine user access patterns for e-commerce applications. The Web also provides a powerful infrastructure for communication and information sharing. In this position paper, we argue that data mining can make use of the Web for mining as well as for analyzing the mining results.

We envision that in the near future many data mining applications will use the following framework. A comprehensive database is first mined. All the discovered patterns are stored in a DBMS. A Web interface is then employed to enable users to view and to analyze the patterns based on their needs and access rights. Different users can also share their opinions and views about the patterns. As an example, we will describe a Web-based system called DS-WEB that helps users to browse and to understand a set of association rules.

2. Analyzing the Discovered Rules

Analyzing and interpreting the discovered rules to gain a good understanding of the domain is an important phase of data mining. It usually requires the user to browse a large number of rules. Typical techniques that assist the user in the process include templates [2], expectations [3, 7], and summarization [4, 5].

Analyzing a set of association rules to find those interesting ones is a very difficult task [2, 7]. The main reason is that there are often too many rules. To overcome this problem, [4] proposes a novel method to summarize the complete set of rules to only a subset of essential relationships. The rules in this subset are called *direction-setting rules* (DS rules).

DS rules give the basic structure of the domain. Non-DS rules are 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. We propose to use DS rules to organize the discovered rules into a hierarchical structure to facilitate browsing. We then employ the Web to organize and publish the DS rules and non-DS rules.

3. The Web

In recent years, the Web has become the single most important medium for communication and information sharing. The Web owes its popularity to its accessibility and its intuitive interface. If we can deliver rules for analysis via the Web, we can inherit these advantages:

1) Contents published on the Web can be accessed from any computer in the world. This eliminates the need for a distant party to re-execute the mining system in order to gain access to the same rules.
2) The Web offers a familiar interface environment to browse the discovered rules. This vastly reduces any learning curve that might otherwise be required if a conventional interface is used.

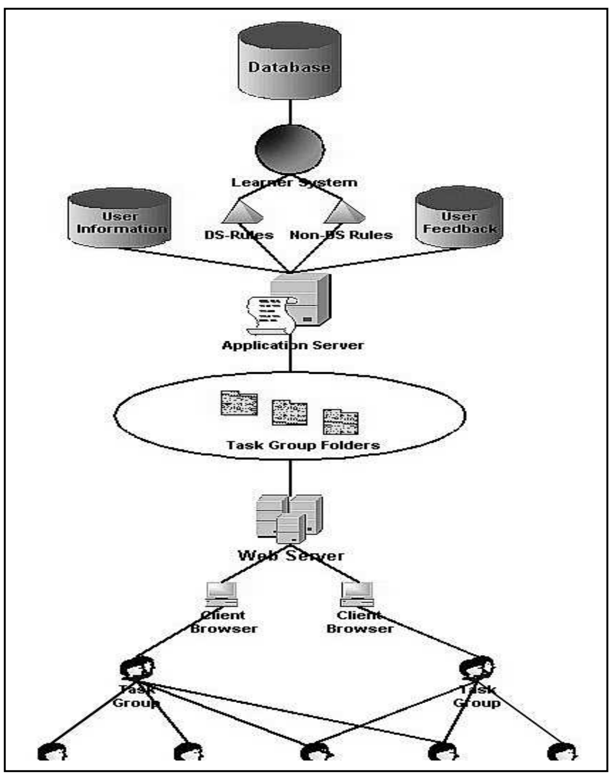
Rules can be easily published as Web pages. Our proposed hierarchical structure can be implemented by utilizing hypertext links to simulate the next level of rules. Users can start from the top level DS rules and then interactively drill down to access the non-DS rules. This aids the rule analysis and interpretation process.

Besides offering a Web interface for easy browsing of the rules, the Web 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 the rules with others. This creates a virtual multi-user workspace. It is both logical and realistic to have multiple users to analyze the rules because

(1) many users may be interested in the same set of rules (each may be interested in a different subset), and
(2) by combining the expertise of different users, we can eliminate any bias a single user might have towards certain domain knowledge.

4. The DS-WEB System

The DS-WEB system uses the Web-based client/server architecture. It basically consists of a set of dynamically generated Web pages that describes the discovered rules. These pages can be mounted, through a Web server, onto either the Internet or a corporate intranet to enable sharing and collaboration between users. DS-WEB's server-client architecture is given in Figure 1.



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

There are two servers in the diagram:

1) Application Server: Responsible for organizing the mining results and the user information:
   - User information: Different users have different access rights to the mining results. On the server, we define task groups, and store them in a user information database.
   - DS rules and non-DS rules: Based on the task groups, the server further filters the rules, and sets up *task group folders* (a virtual folder to dynamically provide the relevant DS and non-DS rules to different task groups).
   - User feedback information: A user in a task group can write his/her comments on rules and also rate rules. These are kept in a database. The rating information is used in collaborative filtering.
2) WEB server: Responsible for presenting the mining results to different groups of users.

Details on the DS-WEB system can be found in [6].

5. Conclusion

This paper shows that we can exploit the Web infrastructure to create a virtual workspace where multiple users can co-operatively work towards analyzing and understanding the discovered rules. This contributes significantly towards the final comprehension of the underlying application domain. The DS-WEB system has been used in a number of applications. We are still enhancing the system to suit our users' needs. Our initial application experiences show that DS-WEB is much more powerful than a conventional system. We envision that not only will the Web be the primary channel for knowledge delivery in the near future but much of the data mining process will also shift towards a distributed multi-participant environment.

References

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

[2] M. Klemetinen, H. Mannila, P. Ronkainen, H. Toivonen, and A. Verkamo, "Finding interesting rules from large sets of discovered association rules." *CIKM-1994*, 1994.

[3] B. Liu, and W. Hsu. "Post-analysis of learned rules." *AAAI-96*, 1996.

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

[5] B. Liu, M. Hu, and W. Hsu, "Multi-level organization and summarization of the discovered rules," *KDD-2000*, Aug, 2000, Boston, USA.

[6] Y. Ma, C-K. Wong, and B. Liu. "Effective browsing of the discovered association rules using the Web." ACM SIG-KDD-2000 Workshop on *Post-Processing in Machine Learning and Data Mining*, 2000.

[7] A. Silberschatz, and A. Tuzhilin. "On subjective measure of interestingness in knowledge discovery" *KDD-95*, 1995.