INTERACTIVE DATA MINING AND ITS IMPACT ON THE WORLD WIDE WEB

Bhavani THURAI SINGHAM
The MITRE Corporation
Burlington Road
Bedford, MA 01730
U. S. A.
Phone: 617-271-8873
Fax: 617-271-2780
thura@mitre.org

ABSTRACT

This paper first examines the integration of data management, data mining, and visualization technologies. In particular, issues on interactive data mining are discussed. Then the impact of these technologies on the world wide web (WWW) is examined. In particular, Internet database management, data mining on WWW, as well as the application of visualization techniques, are discussed.

Key Words: Data Management, Data Mining, Visualization, World Wide Web

INTRODUCTION

Data has become a critical resource in many organizations and effectively managing the data has become a major need. Various types of database management systems have been developed to manage the multimedia data types such as text, voice, and video, and the increasing number of database systems are now interoperating with one another. Extracting useful information from the data has also become an urgent requirement. Data by itself will not be of much use unless one can assign meanings to the data. Data mining techniques enable users to pose queries and extract information from the data often previously unknown. Visualization techniques have helped the users to graphically capture the data in the database as well as help the user in the mining process. Therefore, database management, visualization, and data mining are three technologies that have to be integrated to provide an interactive data mining environment that can help the users a great deal to uncover previously unknown information. The world wide web also adds another dimension to interactive data mining. During recent years the explosion of the users on the Internet and the increasing number of world wide web servers are rapidly advancing digital libraries technology. Numerous databases are now scattered across the various sites. The goal is for users to access these databases in a transparent manner. Furthermore, it is also desirable for the users to extract information from these data sources through data mining. The data and information that are extracted from the databases have to be formatted in a way that can be understood by the users. Therefore, visualization techniques play a major role for data management and mining on the Internet.

This paper focuses on two aspects. We first describe issues on interactive data mining. Then we describe the impact of the WWW on
Interactive data mining. Finally, future directions are given.

INTERACTIVE DATA MINING

Data mining is the process of extracting information and patterns from the data often previously unknown. Various tools are now available for data mining. Many of the data mining techniques need to be guided by the user. That is, the user steers the data mining tools so that useful results are obtained. This process is called interactive data mining as there is interaction between the user and the data mining tool. Figure 1 illustrates interactive data mining showing the relationship between the database management system (DBMS), the visualization tool, and the data mining tool.

Data mining on visualized results

Visualization Tool  Data Mining Tool

Direct DBMS  Data mining

DBMS responses

Figure 1. Interactive Data Mining

Visualization plays a major role in interactive data mining. For example, the visualization tool may be used directly to graphically display the results extracted from the database possibly via a DBMS. The graphical display can then be used by the data mining tool to extract patterns. For example, suppose an automobile company wants to sell automobiles. It then queries the database to obtain information about the various neighborhoods in the town. The DBMS may graphically display the map of the town and information about the people in the town. From this picture the data mining tool could determine the types of cars to sell to the people in various neighborhoods.

Visualization tools may be applied directly to the results obtained by the data mining tool. In this case the data mining tool extracts information from the database via the DBMS and this information is visualized by the visualization tool. Based on this information, the data mining tool may pose additional queries to the DBMS to get more data so that it can deduce additional patterns.

Note that while the first approach applies data mining techniques to visualization, the second approach applies visualization techniques to data mining. That is, in the first approach, the results from the DBMS are visualized and the data mining tool is applied on the graphical displays. In the second approach, visualization techniques are used on the results of the data mining tool so that the results can be graphically displayed. In both cases, visualization tools guide the data mining process. In an earlier paper we have identified a third approach where visualization techniques compliment the data mining process. That is, the data mining tool is used to uncover patterns; but visualization tools may still be used on the data to have a better understanding of the data [Grinstein-Thuraisingham95].

IMPACT ON THE WWW

Now let us examine the impact of interactive data mining on WWW. First of all, there is now a lot of research on accessing databases on the Internet. For example, languages such as SQL are being embedded into Internet programming languages to access relational database management systems. Special protocols for database access are being developed. Furthermore, special techniques for query processing, transaction management, metadata management, security, and integrity are being examined. Integrating the heterogeneous data sources on the Internet is also an important topic. For a discussion on data management for the Internet we refer to [Thuraisingham96].

The biggest challenge on the Internet is the explosion of data and information. Users have too much data to deal with and often this is very confusing. What is needed is a way to make the data manageable for the user. Visualization tools are needed to make the data easier to understand. Furthermore, data mining tools to extract information from the databases is critical. Figure 2 illustrates database management, visualization, and data mining on the Internet.
One of the major problems with applying current visualization and data mining tools on the WWW servers is that many of the existing tools have been developed for structured data such as relational databases. Many of the databases on the Internet are unstructured. That is, they contain multimedia data such as text, audio, and video. Tools for such databases are not mature. Although there is some progress on visualizing unstructured data such as images, data mining tools for unstructured databases are still in the research stage. Therefore, it is important to start developing such tools. One approach would be to extract structure from the unstructured databases and then apply the data mining tools on the structured databases. This could be a temporary solution.

![Diagram](image)

Figure 2. Interactive Data Mining and the WWW

Metadata management is a major challenge for the WWW. First of all, one has to define the metadata. Metadata may not only include information about the data servers, metadata may also include information about the various resources, security policies, and access patterns of users. In many cases, the metadata has to be visualized. Furthermore, mining metadata is also important.

Another challenge in the WWW environment is integrating multiple data mining tools. Each tool may act on a collection of data sources and produce some results or patterns. However, the results of multiple tools have to be integrated so that users can extract information at a global level. Integrating the results of various data mining tools is illustrated in figure 3.

Intelligent agents also play a role in locating the resources on the WWW as well as processing the information. Various types of agents are needed. One is the locator agent to locate the resources, another is the integrator agent to integrate the various heterogeneous information sources, a third is the filtering agents to filter out the information not needed by the user, and the fourth type is the mining agent to extract information from the various data sources. Figure 4 illustrates the various types of agents on the WWW.

![Diagram](image)

Figure 3. Integrated Data Mining

![Diagram](image)

Figure 4. Agents on the WWW

Finally, distributed object management technology plays a major role in integrating the various agents and tools on the WWW. With this approach, each module (which may an agent, a data mining tool, a database management system, or a visualization tool) is encapsulated as an object. Each object has well defined interfaces in accordance with an Interface Definition Language (IDL). The objects interact with each other through some type of Object Request Broker (ORB). An example of a distributed object management system is one based on the Object Management Group's Common Object Request Broker Architecture [Mowbray-Zahavi95]. Figure 5 illustrates object communicating with each
other on the WWW through an (ORB). An IDL is the language used for communication.

![Diagram of communication through objects]

Figure 5. Communication Through Objects

**DIRECTIONS**

This paper has described some preliminary issues in interactive data mining as well as the impact of this technology on WWW. There are several areas that need further work. First of all, tools for interactive data mining have to be developed. That is, the DBMS, visualization tool, and the data mining tools have to be integrated to extract useful information and patterns.

The challenges are much greater on the WWW environment. There is still a lot of research to be carried out on database access through the Internet. Visualization tools have to be integrated with these data management systems. Data mining tools are needed to operate on unstructured data. Various data mining tools have to be integrated so that users can extract patterns at the global level. Agent technology plays a role in locating the information sources as well as mining and visualizing the information. Finally, distributed object management technology enables the different tools and agents to interoperate on the WWW.

This paper has provided some preliminary ideas on interactive data mining and its impact on the WWW. The issues discussed here need to be investigated further so that viable solutions can be developed. With the explosion of users on the Internet, we believe that this will be an important topic for research for the next several years.

**REFERENCES**

