MineSet™: A System for High-End Data Mining and Visualization

Database Mining and Visualization Group

Silicon Graphics, Inc.
Mountain View CA 94043-1389
http://www.sgi.com

1 Introduction

MineSet™ is a highly integrated suite of client-server tools for the high-end mining and visualization of very large enterprise databases. MineSet represents the confluence of several important software and hardware technologies: data mining algorithms, fast multiprocessing database servers, novel techniques for interactive 3-D data visualization, and powerful graphics workstations. MineSet provides integrated facilities for the extraction of data from varied sources, algorithms for mining the extracted data, and tools for the 3-D visualization of results. The mining algorithms allow for the discovery of previously unknown nuggets of information, while the sophisticated visualizations allow the user to get a better understanding of both the data mining results, and also the interrelationships among the raw data ("visual data mining"). Underlying all these techniques is the ability to handle large data sets.

2 System Architecture

MineSet is built using a client-server architecture, with the server responsible for data extraction and mining, and the client supporting user interaction and data visualization. Functionally, MineSet consists of the following four components:

1. **Tool Manager:** The tool manager is the central point of user control. Using a GUI running on the client, it allows the user to direct the data extraction and mining activities on the server as well as the 3-D interactive visualization on the client.
2. **Data Mover:** The data mover is the server software that supports requests for data extraction and preparation for use by the mining algorithms. The data mover can use traditional files as well as one of several RDBMSs (currently Informix, Oracle, and Sybase).
3. **Mining Tools:** These tools support the analysis of data to find hidden patterns, with an emphasis on high performance and scalability to large data sets. Mineset currently supports tools for finding associations, classification, feature selection, and discretization.
4. **Visualization Tools:** These tools allow the user to explore mined information or raw data in an animated 3-D landscape to take advantage of the natural human ability to navigate in a three dimensional space, recognize patterns, track movement, and make comparisons between objects of different sizes. The visual tools display information in many different ways, including fly-through hierarchies, 3-D bar charts based on maps or grids, 3-D scatter plots, etc. Animation and user interaction is an important aspect of each tool.

2.1 Hardware support

MineSet currently runs on the Silicon Graphics hardware platform. The data mover and mining tools run on the CHALLENGE™ line of servers. These servers are based on the R10000 64-bit RISC architecture and are capable of supporting up to 36 processors. The data visualization tools take advantage of the advanced 3-D graphics support available on all SGI workstations.

3 Future Challenges

Future versions of the tool suite will have support for greater parallelism in data-extraction, both from relational databases and flat files. The mining tools will exploit both I/O parallelism and compute parallelism.

In addition, we are planning approaches to provide access to the visualization and mining tools through the world wide web, using Java and VRML.

Proceedings of the 22nd VLDB Conference
Mumbai(Bombay), India, 1996