Data Mining for the Corporate Masses?

Neal Leavitt

For several years, proponents have touted data mining as a powerful tool for finding patterns hidden in large databases. They promise many benefits, such as increased revenues for companies that use the technology to fine-tune their marketing by digging out customers' buying patterns from mountains of sales data.

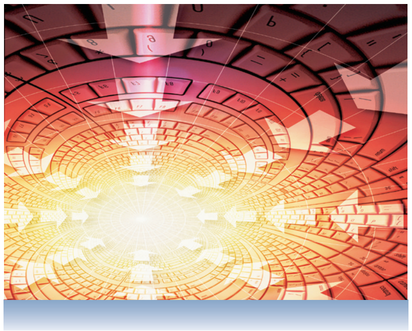
Until recently, however, data mining has been a complex, expensive, somewhat limited tool adopted primarily by large companies. This pattern may be changing, though, because of new techniques and technologies.

Insurance companies and stock exchanges, for instance, are now using data mining to detect customer-activity patterns that indicate fraudulent behavior. And doctors are using data mining to predict the effectiveness of surgical procedures, tests, or medications for various types of conditions.

DRIVING DATA MINING

Data mining, which automates the detection of complex patterns in databases, began formalizing as a technology in the mid-1990s. According to Dan Vesset, research director for market research firm IDC, the worldwide market grew from $455 million to $539 million last year, and is expected to continue increasing to $1.85 billion in 2006.

Today's leading data mining vendors include HNC Software, KXEN, Microsoft, NCR's Teradata Division, Oracle, and Quadstone. According to Aaron Zornes, research director for the Meta Group, a market research firm, the market leader is SAS Institute, followed by SPSS and IBM.



The two most significant challenges driving changes in data mining are scalability and performance.

Organizations want data mining to become more powerful so that they can analyze and compare multiple data sets, not just individual large data sets, as is traditionally the case. They also want to break up data into finer-grained categories for analysis.

Scalability is critical because databases have become very large. Terabyte-class databases have become more common today, particularly as the cost of storage has decreased and the amount of e-commerce has increased. IBM, for example, claims more than 200 customers with data warehouses larger than a terabyte, several of which are more than 30 terabytes. And industry observers predict that within a few years, some users will have 100-terabyte data warehouses.

As companies capture more data, managing and mining the information become more complex. In some cases, data miners have tried to solve this problem by analyzing small samples of large data sets.

However, said Dan Graham, IBM's director of business-intelligence solutions, this approach is a concession to the limitations of today's data mining hardware and software. Data mining yields better results if more data is analyzed, he explained.

The growth in e-commerce is driving the need for data mining approaches that work with online Web businesses. IBM, for example, has incorporated data mining into its WebSphere Commerce Analyzer, a tool for analyzing e-commerce activity.

Many organizations want improved data mining techniques they can integrate across decision-support-related systems so that they can make better decisions faster and more proactively, said Eric Apps, president of Angoss Software, a data mining vendor. Financial services organizations are already integrating data mining into their credit-related systems to reduce losses by continuously modeling and changing their operations, he explained.

NEW TRENDS

Several key data mining trends have emerged.

Data analysis

Traditionally, data mining tools have used algorithms to analyze samples of large data sets. However, performance improvements in processors, servers, and database and data mining software have made it possible to analyze large data sets in their entirety, said IBM's Graham.

Today, data mining products are using scalable tree-based classifiers to build models for analyzing very large databases, according to the Meta Group's Zornes. These trees help generate rules for classifying information in a data set. "In effect," Zornes said, "they put structure into unstructured data, even for very large 20-terabyte data warehouses."

Improvements in online analytical processing (OLAP) software, which analyzes different dimensions of data in a database, let mining tools work with more categories of information than in the past. This means users can divide data sets into more categories and thereby conduct more meaningful and finer-grained analysis.

Predictive capabilities

Zornes said models for using data mining to predict consumer behavior, such as fraud or changes in purchasing patterns, are becoming more accurate. Figure 1 shows a typical predictive data mining system.

However, improved predictive models aren't necessarily due just to better statistical techniques, said Quadstone's president, Mark Smith.

In the past, Smith noted, statistical programmers involved in data mining have been divorced from business considerations and thus produced models that, while statistically accurate, did not necessarily best address the important business issues. Now, he says, a new breed of marketing analyst with statistical skills frequently drives the process with business goals in mind.

Integration into the database

Following Microsoft's lead in 1998, several vendors are now integrating fundamental data mining capabilities into database engines. This integration, explained the Meta Group's Zornes, lets customers run complex data mining tasks in parallel inside the database, which decreases response times.

He said it's faster and less expensive to operate on information within the database than to extract it first and then analyze it within a separate and redundant storage infrastructure.

Also, IBM's Graham noted, "By delivering tight integration between the database and data mining functionality, users can drill down into very granular levels of detail." This can generate more user familiarity with the data and, ultimately, more accurate results.

Moreover, said Angoss's Apps, integration improves data mining software's performance by giving it immediate access to data and thus a better workflow.

Companies have already integrated the following products: IBM's DB2 Intelligent Miner and DB2 database, NCR's Teradata Warehouse Miner and Teradata Database, and Oracle's Darwin data mining suite and 9i database.

Easier to use, less expensive

Data mining can be expensive for organizations because it requires powerful software, servers, and storage hardware to handle large amounts of data. Data mining also requires extensive employee training and frequent intervention by analysis experts. Now, however, data mining tools are becoming easier to use by employees without extensive training or operational and statistical skills.

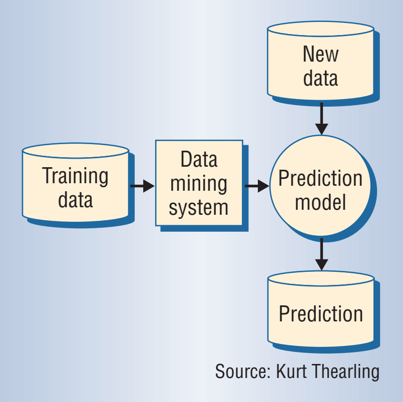
In fact, said Quadstone's Smith, "The biggest shift in data mining in recent years is not algorithms or even computing. [It] is opening data mining up to a broader user base." This makes the technology useful to more employees and reduces the time and money spent on training, operations, and data analysis.

Quadstone provides access to its data mining tools via a graphical interface that does not require the user to have programming skills. It also lets the user visualize data in multiple ways and more easily optimize the data mining model.

"All of the major data mining vendors have focused on the usability aspects of their complex products by providing such graphical environments," said the Meta Group's Zornes.

"This advance has driven the acceptance of data mining as a more widely used business tool," said Colin Shearer, SPSS's vice president of data mining.

Meanwhile, SPSS's Clementine 6.0 data mining product simplifies usage via prebuilt templates that come out of the box ready to work with specific business applications.



Source: Kurt Thearling

Figure 1. A typical predictive data mining process starts with using data to train the system how best to analyze and make predictions from a data set. The system develops a predictive model from this training data. The system then yields predictions based on new data fed into the model.

Standards

Organizations are starting to develop standards for various aspects of data mining.

PMML. With the XML-based Predictive Model Markup Language, initially developed by the National Center for Data Mining at the University of Illinois at Chicago, a set of data mining statistical models will work across platforms and on different vendors' applications. PMML lets different applications exchange the essential information required for one application to use data mining models produced by another application.

PMML, now managed by the Data Mining Group (http://www.dmg.org), an industry standards organization, defines the inputs and outputs of the data mining models, as well as the parameters that define the models themselves.

Most major data mining vendors support PMML, which has been widely deployed. The Data Mining Group plans to release version 2.1 of the standard this August.

"The challenge now being tackled is to extend PMML to cover the process of cleaning, transforming, and aggregating data—a task that is more complex and not so easily standardized," said Robert Grossman, director of the National Center for Data Mining and chair of data mining vendor Magnify Inc.

Object models. The Object Management Group has developed the Common Warehouse Metamodel (http://www.omg.org/technology/cwm/) to provide the syntax and semantics necessary for the interchange of metadata among data-warehousing tools and metadata repositories in heterogeneous environments.

Process standards. The Cross-Industry Standard Process for Data Mining Special Interest Group, an industry organization, developed CRISP-DM (http://www.crisp-dm.org/) to define a data mining process that applies across diverse industry sectors. The process is designed to make large data mining projects faster, less expensive, and more manageable.

Web standards. The data mining industry is developing standards for exploratory data analysis on Internet-based data webs. For example, the National Center for Data Mining's data space transport protocol lets users retrieve and explore all related data about a particular object from Web sites, files, and other structures, even if they are distributed across servers in a network.

DATA MINING RESEARCH

With data mining's increasing popularity, research is taking place in several areas.

Distributed mining of huge data sets

The University of Illinois at Chicago has launched the Terra Wide Data Mining Testbed (http://www.dataspaceweb.net/terra-v3.htm) for analyzing data webs over optical networks. Grossman said the project remotely analyzes, in real time, data sets of at least a terabyte. For example, researchers use the testbed to learn about global weather patterns by studying multiple data sets from the US National Center for Atmospheric Research and about the spread of diseases from United Nations World Health Organization data.

"Just as networks become more interesting as the number of nodes increases, so does the analysis of distributed data," said Grossman. Often, he said, the most interesting insights into data come from comparing one data set to another.

With this in mind, researchers from several institutions, including the University of Illinois at Chicago, have built DataSpace (http://www.dataspaceweb.net/index.htm), a testbed for mining distributed data.

Text mining determines patterns and predicts outcomes from large volumes of text-based data.

Text mining

Data mining traditionally analyzes structured data—such as statistics—that can be easily placed into defined categories. However, estimates suggest that at least 80 percent of today's data is in an unstructured textual format.

This accounts for the emergence of text mining, which determines patterns and predicts outcomes from large volumes of text-based data. SPSS and SAS recently announced plans to release text-mining technology.

SAS product manager Wayne Thompson said text mining works with large collections of documents, such as e-mail messages, survey responses, or patient health records. The technique analyzes the text, places the information into predefined categories, and converts the data into a structured format for use with traditional mining techniques.

The University of California, Berkeley, and IBM are developing tools for mining the text of Internet newsgroups, which market researchers, social scientists, and others see as potentially containing a wealth of valuable information.

Fiona Neil, Quadstone's marketing manager, predicted that data mining will continue to become more scalable, which will be important for the technology's future.

In addition, data mining will continue its integration into the database, said Alexander Linden, research director for Gartner Inc., a market research firm. "We'll also see more multimedia data mining," he noted.

Moreover, said the Meta Group's Zornes, data mining will be embedded in a growing number of business applications during the next three to five years.

Data mining still faces key challenges. Researchers must work on making predictive models easier to integrate into companies' operational systems and on making data mining even easier to use by business analysts, said SAS's Thompson. Another challenge is finding ways to effectively manage the many data mining models a large enterprise organization creates.

Regardless of what happens, data mining won't replace skilled business analysts or managers, said Herb Edelstein, president of Two Crows Corp., a data mining consultancy. "Data mining tools need to be guided by users who understand the business, the data, and the general nature of the analytical methods involved," he explained.

Or, as Angoss's Apps said, "Data mining is just a tool and has to be used by people who understand data, statistics, and the business. A fool with a tool is still a fool." ■

Neal Leavitt is president of Leavitt Communications, an international marketing communications firm with affiliate offices in Paris, France, and Hamburg, Germany. He writes frequently on technology topics and can be reached at neal@leavcom.com.

Editor: Lee Garber, Computer, 10662 Los Vaqueros Circle, PO Box 3014, Los Alamitos, CA 90720-1314; l.garber@computer.org