Hidden Messages

BY NEIL J. RUBENKING MAY 22, 2001

Data mining helps you uncover what your data is trying to tell you.

Government agencies, businesses, and research firms are churning out raw data on every subject imaginable, at an ever-increasing rate. NASA's space probes keep phoning home with data, e-businesses accumulate information about customer habits, and Web servers log every user interaction. The cost of storage keeps dropping, so there's no difficulty finding a place to warehouse these terabytes of data. Although there may be important patterns and knowledge buried in the data, the sheer amount of information has grown beyond human analytical capacity. Data mining allows computers to take over the task of finding the patterns.

Data mining is the process of automatically extracting useful information and relationships from immense quantities of data. Scientists use it to separate signals from noise in astronomical data and to find genes within DNA sequences. Your company can use it to gain valuable knowledge about customers, site visitors, and business practices. Using this knowledge, you can target advertising campaigns and evaluate their success. You can personalize Web pages and suggest related purchases. And you can predict customer behavior, making your site more effective.

Development of Data Mining

Data mining is a logical evolution in database technology. The earliest databases, which served as simple replacements for paper records, were data repositories that provided little more than the capability to summarize and report. With the development of query tools such as SQL, database managers were able to query data more flexibly. A manager could, for example, determine how many cell phones were sold in Kalamazoo during June of 1980 or which salespeople brought in the most customers. Almost any quantitative question can be answered using these tools.

OLAP (Online Analytical Processing) tools aid by making patterns visible. Given the correct view of the data, you might discover that trailer hitch sales in Texas are twice as high in February as in any other month, letting you know that you should adjust production to match or work on raising sales in the other months. These tools make patterns in data easier to see, but the manager still has to manipulate the data, look for patterns, and decide which patterns are important. A totally unexpected or hidden pattern can go unnoticed simply because nobody thought to look for it.

Data mining automates the process of locating and extracting these hidden patterns and knowledge. In its purest form, data mining doesn't involve looking for specific information. Rather than starting from a question or a hypothesis, data mining simply finds patterns that are already present in the data.

Intriguing Patterns

To wring knowledge from raw data, data mining software uses a wide variety of complex algorithms including neural networks, rule induction, decision trees, and genetic algorithms. Typically, the software performs its analysis on a portion of the data to obtain rules and patterns, then validates the results by testing them against the held-back data. In a scientific setting, this process can reveal

relationships that aren't obvious or sift out real data from mere noise. In a business setting, this knowledge can be used to set policy, exploiting favorable patterns and avoiding bad ones.

Part of the challenge for data mining software involves generating results in human-understandable terms. Among the more intelligible pattern types are associations, sequences, and clusters. Association or affinity patterns simply identify database elements that occur together in a statistically significant fashion. For example, analysis of a huge database of customer shopping carts could reveal that nine out of ten visitors who bought calendars also bought pens. Sequence patterns are similar but with a time factor thrown in. A bank's records might show that 80 percent of customers declaring bankruptcy had obtained three or more new credit cards within the past year.

To identify clusters, the software models a multidimensional space in which each of possibly thousands of dimensions represents an attribute of the data. The program then segments the data into clusters based on their proximity in this imaginary space. Further analysis, either in software or with human intervention, selects clusters that have useful characteristics. A simple example would be an analysis of a sales database that uncovers a cluster of customers who make a very large number of small purchases. The company could reduce processing costs by offering an incentive for customers to combine their small orders into fewer, larger ones.

Data mining is also used to devise predictive patterns. Given a large database of customer transactions and a specific subset that are known to be fraudulent, the software could be directed to determine what simpler characteristics distinguish the fraudulent transactions from the rest. If successful, this will yield a rule that predicts which future transactions are likely to be fraudulent, and the company can give extra scrutiny to those dealings.

Of course, most data sets are full of patterns that don't represent useful knowledge. You won't be impressed if a data mining tool reports that every customer in ZIP code 10016 has a New York address, or that every patient in the gynecology department is female. But the same technique might uncover a pattern of double billing or reveal new avenues for targeted advertising. Human intervention is required to distinguish patterns that are useful.

Using the most accurate data is essential. As a precursor to serious data mining, companies will usually establish a data warehouse—a collection of data designed to support management decision-making. The warehouse includes data from across the enterprise at a single point in time. As much as possible, the data is cleansed of errors and redundancy, and perhaps transformed into a format suitable for the mining program.

Data Mining and the Web

Every time you click on a URL, your browser requests the corresponding Web page, which a Web server supplies, logging the transaction. Further transactions may be required—to download images on the page, for example. The server's log of low-level transactions is referred to as clickstream data. A large e-commerce site can generate millions of clickstream lines every day.

Data mining software can find significant patterns in the clickstream logs alone, but that data becomes substantially more useful in combination with customer registration data. Linking clickstream entries with a specific customer lets you track that customer's travels through your site—this alone provides a vast new realm for discovery. The richest lode for data mining is a data warehouse containing clickstream data, user profile information, and all of the company's other relevant databases.

other relevant databases.

Practical Applications

One of the most cited publications in the data mining field is a 1991 doctoral dissertation by Usama Fayyad. As a graduate student, Fayyad worked with General Motors on extracting useful knowledge from an immense database of car repair data; the algorithms he devised became the basis for his dissertation. Fayyad went on to develop data-mining systems for NASA and Microsoft before founding digiMine in March 2000. In our January 16, 2001 issue, the article "Know Who I Am" explored digiMine and other data mining companies.

Data mining is not cheap. digiMine's hosted service starts at $7,500 a month; fully installed solutions cost hundreds of thousands. To see what data mining can do for you, we'll look at a few of the reports generated by digiMine.

We see a fairly simple report that reveals affinity relationships among the categories Accessories, Men's Clothing, and Women's Clothing within the general category Cycle Shop > Mountain Bike. More than half of those who browsed accessories also browsed men's clothing. Slightly more than half the visitors who looked at accessories also looked at women's clothing. Over a third of the customers who toured accessories also browsed both men's and women's clothing. The manager viewing this data will have to dig deeper to determine the reason for these patterns. The customers could be buying accessories and clothing, in which case cross-selling makes sense. Another possibility is that the customers looked in one category, didn't find what they wanted, and switched to the other. In that case, certain products could be placed in different or multiple categories.

As noted above, data mining software can plot immense quantities of data in multidimensional space to find items with similar characteristics. After the software has done the heavy lifting, the database manager studies the characteristics of each cluster and identifies those that seem useful. We also show that, digiMine has identified eight significant clusters within a group of over 300,000 visitors, and the manager has named several of the clusters based on their characteristics. Comparison of the characteristics of One Hit Wonders and Return Visitors might suggest techniques for turning more of the former into the latter.

Once an interesting cluster has been identified, you can study it further using more traditional forms of analysis, such as the funnel report has shown. A funnel report identifies how many users successfully negotiate each step of a multistep process. In the figure, the process is a purchase transaction. Just a few percent of the users drop off at each step; an unusually large drop at one step would be a red flag for the Webmaster to examine the Web page or form corresponding to that step.

Of course, you'll only get the benefit of data mining if you take action based on the knowledge it provides (and keep the data warehouse up to date). If you've identified a cluster of users based on certain simple characteristics, you can personalize your site for first-time visitors that fit those characteristics, or you can create targeted advertising campaigns. If the software shows that certain products are purchased together very frequently, link the two in your catalog. If half of the customers who begin the checkout process drop out before completing the transaction, revamp the checkout system. Your company's data is full of undiscovered gems; start digging!

Neil J. Rubenking is the contributing technical editor of PC Magazine.