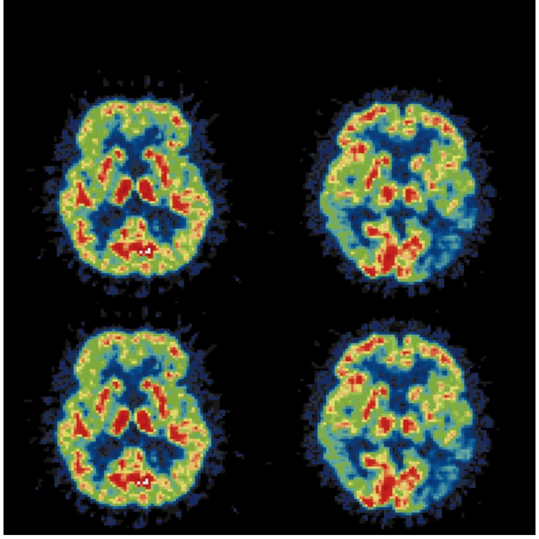


Data mining is such a hot topic that it has become an obscured buzzword. Find out what it is and what answers it can provide.

Bhavani Thuraisingham

A Primer for Understanding and Applying Data Mining

Data mining can be a powerful tool for extracting useful information from tons of data. But it can just as easily extract erroneous and useless information if it's not used correctly. Key to avoiding the pitfalls is a basic understanding of what data mining is and what things to consider in planning a data mining project.

WHAT IS DATA MINING?

Although it has only recently become a hot topic, the techniques used in data mining have been around since the 1980s. This emerging technology combines statistical analysis, machine learning, and database management to extract information from large database systems. Researchers first began using neural networks and other machine-learning techniques during the past decade, and numerous commercial data mining products and research prototypes are now available.

Vendors of most major database management systems and data analysis products now market data mining tools that manipulate the data in relational database tables. In contrast, mining unstructured databases—which combine text, image, or video data—remains a major challenge. Mining the vast quantities of Web data to extract meaningful information is also an active research area. Another challenge is mining Internet e-commerce usage patterns. For example, how can Web site owners mine e-commerce data for information to make their sites more effective and increase their business?

Inside

Data Mining Outcomes and Techniques

Resources

HOW DO PEOPLE USE DATA MINING?

Data mining collects, stores, and organizes data for use in areas such as medicine, finance, intelligence, law enforcement, defense, logistics, education, and process control. Many applications use data mining for promotions, marketing, and sales. We can also use this technology for diagnostics and anomaly detection. Examples of uses of data mining include

- a credit bureau decides on loans based on observations of people with similar buying patterns, income, and credit;
- a supermarket organizes its merchandise based on buying patterns and information about associations between products;
- a pharmaceutical company analyzes prescriptions to send promotional material to target customers;
- an intelligence agency reviews spending patterns and travel data to detect abnormal behavior by its employees;
- a physician analyzes X-ray images to detect abnormal patterns; and
- an airline reservation system uses information about travel patterns and trends to maximize seat utilization.

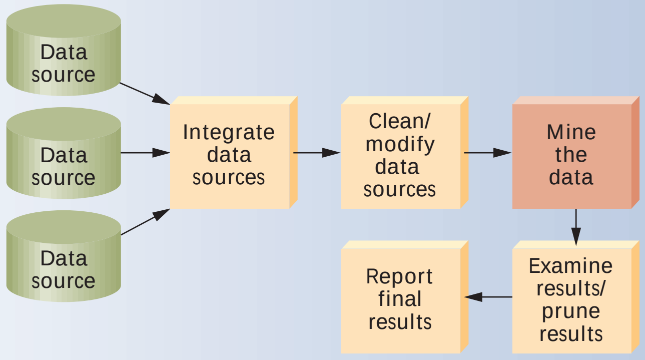
In IT applications, data mining techniques help ensure that you have good data. For example, in

1520-9202/00/$10.00 © 2000 IEEE

logistics applications, data mining helps to make sure that the right people are put on the right projects. Banks use data mining to get the data that helps them attract customers.

Although these kinds of applications have been used for quite some time, they relied on statistical analysis by hand and have only recently begun employing data mining technologies to analyze data and make correlations and predictions.

Figure 1. Steps in the data mining process.



AN INTEGRATION OF SEVERAL TECHNOLOGIES

Successful data mining requires integrating several technologies. Probably the first technologies that come to mind are database management systems (DBMSs) from companies such as Oracle, Sybase, IBM, and Informix. These systems play a major role in organizing and providing effective access to large quantities of data. Today, database system vendors are integrating data mining techniques into their database engines. Although this approach has many advantages, it also ties data mining to a particular DBMS. For example, IBM's Intelligent Miner is tightly integrated with IBM's DB2 product. Another alternative is to use data mining tools that work with multiple database management systems. Examples include tools like Lockheed's Recon that work on different databases and have loose integration with the database system.

Many consider data warehousing to be a prerequisite for mining. A data warehouse integrates data from multiple sources and ensures data quality. Without good data, it is impossible to get accurate results from mining. By "good data" I mean that it is possible to make complex associations between data. Therefore, even if you don't build a warehouse, you still have to integrate and clean the data before you carry out data mining.

Statistics are another key factor in data mining. Statistical packages such as Matlab by Mathworks only work on large databases. As data mining matures, it is becoming clear that effective mining requires more sophisticated and intelligent statistical reasoning techniques. For example, current techniques can't make complex associations between data.

Data mining uses various machine-learning techniques such as neural networks, decision trees, inductive logic programming, and the *k*-nearest-neighbor algorithm to extract key information from the data. These visualization techniques help users understand the data and get a clearer idea about which data mining techniques to apply. You use examples to train a neural network so that it learns about an environment. Then the network can recognize the normal occurrences and detect unusual patterns. Inductive-logic programming techniques deduce patterns and rules from the example data. You use training examples to construct decision trees, then you use the trees to classify data. With *k*-nearest-neighbor techniques, you use data values to compute points in space. Then you can group entities that have corresponding points close to each other into a cluster. High-performance parallel-processing techniques using multiple processors help to speed up the data mining activity.

PLANNING A DATA MINING PROJECT

Planning a data mining project starts by obtaining answers to several questions. As Figure 1 shows, the steps in a data mining project include integrating and cleaning or modifying the data sources, mining the data, examining and pruning the mining results, and reporting the final results.

Deciding whether to outsource or not

The first step is to determine whether you have the resources to carry out the mining or whether you should contract it out. Should you embark on a pilot project or should you start with a large-scale effort? The answer depends on your organization's expertise.

If you decide to do the data mining in-house, the next step is to identify the data. Data can be in databases, electronic files, or even on paper. You can use a data warehouse to integrate the data and convert it into the desired structure and format. You have to scrub (clean) the data and maintain its quality.

Choosing an outcome

Once you accumulate the data, you need to determine the desired outcomes. Data mining outcomes include classification, association, clustering, prediction, estimation, and deviation analysis. The "Data Mining Outcomes and Techniques" sidebar describes these different types of outcomes. Do you want to classify or cluster the data? What tools do you need to get the desired outcomes? How should you modify the tools? You need to work closely with the data mining tool vendor to accomplish this.

Once you determine which outcome you want, you need to consider whether you should use a top-down or bottom-up approach to data mining. In the top-down data mining method, you start with a hypothesis and then validate it. If any cases invalidate the hypothesis, you may need to revise it. In the bottom-up approach, you start with examples and deduce the hypothesis. Bottom-up data mining is appropriate if you know what you are looking for; it is less appropriate when you're using the data mining tool to find something interesting.

Interpreting the data

Once you apply the tools, you need the help of specialists to understand the data. In the future, we will have automated tools for extracting the nuggets hidden in data, but for now, understanding the mining results is still a human task. You need to implement some kind of action to determine whether the data mining results are useful. Getting valid data requires multiple iterations of this action. For example, a supermarket can place some products next to each other to see if sales improve. This produces data that the supermarket uses in the next iteration of the data mining cycle.

FUTURE TRENDS: DATA MINING RESEARCH

While many commercial data mining products work on relational databases, researchers are currently investigating how to mine multimedia databases and text databases. Text retrieval finds documents based on keywords; text mining makes associations between documents. Embedding mining techniques into the information retrieval engine would make information retrieval systems more sophisticated because it would provide a tight integration between the data mining tools and the database management system.

Data Mining Outcomes and Techniques

Different kinds of data mining techniques produce different outcomes.

- ***Classification*** **groups items based on a predefined attribute. For example, "people who live in Manhattan own apartments costing more than $500, 000" is a classification.**
- ***Association*** **makes correlations between items and individuals, deducing rules that define relationships. Examples include "fish and wine are purchased together" or "John and Mary travel together."**
- ***Clustering*** **groups items based on a previously undefined attribute such as, "workers in cluster A make less than $20,000, those in cluster B make between $20,000 and $50,000, and those in cluster C make more than $50,000."**
- ***Prediction*** **forecasts trends such as "in 2010, physicians will earn $300,000."**
- ***Estimation*** **examines trends for clues to deduce another characteristic. For example, you probably can analyze spending patterns to determine how many children a person has.**
- ***Deviation analysis*** **compares current data to a preestablished norm to detect anomalies. Network management tools use this technique to alert system administrators to unusual user behavior.**

Mining images is a challenging area in which little work has been reported. One approach trains neural networks on images, establishes baseline patterns, and uses them to detect any abnormality. For example, geographic data training detects seasonal and other normal changes in, say, vegetation. If an unusual type of growth occurs, the mining tool should alert the user. You could also use this technique for video or audio mining, using, for example, video data to train the neural network so that it detects unusual patterns in video frames.

Recently, Web mining has become critical for e-commerce applications. Customizing a Web site depending on the user is a major direction in data mining. Web sites mine usage patterns to determine who accesses the data and then develop profiles and customize information about each user. Web e-commerce sites also use mining to develop user profiles for use in targeted marketing and sales.

Mining distributed and heterogeneous databases is also a challenge. Currently, data mining tools work mainly on centralized databases. The question is how to apply them to distributed databases. Using data mining tools on each database produces partial associations. Then, either a central coordinator or various data mining tools have to integrate these associations to derive complete results. The various tools have to communicate with each other to provide a complete picture.

Researchers are also investigating how to use metadata to help sort and query data. You generate metadata from the data, then you mine the metadata. For example, metadata for a text document might be the name of the document's creator, its creation date, and its format. For video data, the metadata could be annotations. It's easier to mine the annotations than to mine the video data directly. Researchers are seeking standard ways to encode metadata for easy retrieval and comparison. They are also looking into efficient ways to automatically generate metadata from an existing database.

Improving the scalability of data mining techniques is another major challenge. Research areas include new parallel algorithms and architectural considerations such as componentizing data mining techniques. For example, it might be

Resources

Books about commercial data mining products and research prototypes

- **P. Adriaans and D. Zantinge, *Data Mining*, Addison-Wesley, Reading, Mass., 1996.**
- **M. Berry and G. Linoff, *Data Mining Techniques for Marketing, Sales, and Customer Support*, John Wiley & Sons, New York, 1997.**
- **B. Thuraisingham, *Data Mining: Technologies, Techniques, Tools and Trends*, CRC Press, Boca Raton, Fla., 1998.**

Other data mining information sources

- ***http://www.acm.org/sigkdd/:* The ACM Special Interest Group on Knowledge Discovery in Data and Data Mining provides information on basic KDD research, the adoption of standards for KDD terminology, evaluation, and methodology, and interdisciplinary education among KDD researchers, practitioners, and users.**
- ***http://www.kdnuggets.com:* *KDNuggets News* is an e-newsletter on data mining and knowledge discovery.**
- ***http://www.datamining.org:* The Data Mining Institute promotes a broader awareness of data mining technology by explaining its benefits and applications.**

possible to develop reusable object-based data mining components for use in different applications.

A LINGERING QUESTION: PRIVACY AND SECURITY

While you can use data mining in applications such as intrusion detection and auditing for computer security, its misuse also has a serious potential for compromising security and privacy. Without data mining tools, the user has to be fairly sophisticated to make intelligent deductions by posing queries and deducing unauthorized information from responses to legitimate queries. Mining tools make it easier for even naïve users to obtain sensitive information, which could compromise an individual's privacy.

This is an extremely difficult problem that has both legal and social aspects. Technology alone can't provide a sufficient solution to this dilemma. We also need appropriate privacy laws.

One approach is to give out only data samples so that the user can't come up with good data mining results. The question is, which sample is it safe to give out? It's nearly impossible to prevent users from collaborating and using many tools to get results from multiple samples. Another approach is giving out false information. But this would compromise the data's integrity and could cause problems in valid situations.

To date, there are no good solutions to concerns about data mining privacy issues. But with increased interest, we can expect some solutions in the near future.

Data mining is an area that will continue to explode during the next decade, presenting endless opportunities and challenges for developers who are finding practical ways to use this emerging technology. Only with a solid grounding in the basics, however, can IT professionals hope to make the best use of the answers they get from data mining. ■

Bhavani Thuraisingham *is a chief scientist in data management at Mitre's Information Technology Directorate. Contact her at thura@mitre.org.*