Data Mining: A Mature Technology In A Modern Economy

By: **Dwayne N. McSwain** and Betty Harper

Abstract

Computer technology that enhances the intuitive and predictive quality of information is highly attractive to any enterprise seeking a sustainable competitive advantage. Data mining products have become an integral part of every business plan, providing a broad spectrum of reporting and statistical data analysis. The flexibility and applicability make for an appealing return on investment. This paper examines the dynamic nature of the technology along with its current and developing utility in the business environment.

Data Mining: A Mature Technology in a Modern Economy

by

Dwayne N McSwain
Middle Tennessee State University

Betty Harper
Middle Tennessee State University

ABSTRACT

Computer technology that enhances the intuitive and predictive quality of information is highly attractive to any enterprise seeking a sustainable competitive advantage. Data mining products have become an integral part of every business plan, providing a broad spectrum of reporting and statistical data analysis. The flexibility and applicability make for an appealing return on investment. This paper examines the dynamic nature of the technology along with its current and developing utility in the business environment.

INTRODUCTION

The dynamic nature of technology and the rapid development of communication and information technologies over the last two decades have given rise to many new challenges for today's managers. Web surfers have developed a wary skepticism of technology designed to extrapolate data drawn from Internet surfing patterns. Teens and their parents are traumatized by accounts of Internet stalkers—and rightly so with teens hanging out at sites like Facebook and MySpace. MySpace alone has more than 100 million profiles, with 230,000 new members signing up each day (Andrews, 2006). In their quest for anonymity, people forget that they are giving up personal information when they apply for discount cards, contract for cell phone usage, or rent movies. Certainly, there is no dearth of data. Pluripotent are the possibilities for use and misuse of data! Perhaps the most pressing challenge is the development of innovative techniques to extrapolate and manipulate pertinent information from today's massive databases. Individuals and business entities are not necessarily opposed to sharing information; the chief concern is maintaining some control over what they share and how it is used. From this notion, the concept of *data mining* has evolved.

This discussion of data mining will focus on the various aspects employed in the process. Attention will be directed toward current use of the technique as well as on limitations, the future and challenges of data mining.

DATA MINING: WHO IS USING IT?

Albeit in somewhat limited data sets, basic techniques aimed at finding the relationships between and among datum have been around for decades and include regression, cluster, correlation, and discriminate analysis. Use of modern-day computer algorithms has enhanced the ability to discover hidden patterns and unsuspected relationships in massive data sets. These data mining technologies have application in every academic discipline. Economists can explore relationships between data points to measure economic benefits for local populations. For example, in the State of Tennessee an initiative passed that allows local governments to freeze assessed property values for homeowners of a specific age or income level ("Site Promotes Senior Tax Break," 2006). Accountants may mine a Securities and Exchange Commission database and easily gather industry-specific data for a historical profile or to develop a sales trend for a specific company.

Departments of Revenue in several states have formed an initiative to mine for abuse of use tax and Internet sales tax regulations. Data mined by this group is used in a lobbying effort to have the U.S. Supreme court address the issue as a matter of interstate commerce (Streamlinedsalestax.org). Both the Pentagon and the Department of Homeland Security benefit from use of data mining techniques in their effort to thwart terrorist activity. An upsurge in Medicare fraud and income tax evasion may prompt an increased reliance on data mining.

IMPLEMENTING THE PROCESS

Collecting and aggregating data has become so sophisticated that the term "data mining" encompasses many aspects of data extrapolation. There is no single technique, but rather a highly focused data transformation framework. The term loosely describes tools used to extrapolate, analyze, and exhaustively explore data with the objective of identifying complex relationships.

Adopting data mining methods could dramatically increase an entity's ability to derive meaningful relationships in existing data. In essence, the goals of data mining are prediction and description. Variables in the existing data are used to help predict future values of interest, and patterns found in the data are used to help describe the data for user interpretation (Arunasalam, 1999).

Data mining identifies discernable patterns in a far more sophisticated fashion than reliance on any one statistical method. Moreover, successful data mining does not hinge on the user's ability to pose the proper questions and interpret those outcomes correctly.

Figure 1
Categories of Algorithms Used in Data Mining Processes

Algorithms that collect data by association look for the presence of one set of items in a group [that imply] the existence of another set of items in the same group.

Those that gather information by clustering, group together items with similar characteristics using such methods as conceptual clustering and neural clustering.

Algorithms that classify data look for a common profile among cases with diverse attributes.

Those that search for sequential patterns look for the characteristics present in subjects when one instance occurs that may predict an ensuing occurrence.

The algorithms that search for similar time sequence discover patterns in elapsed time from one event to another over a period of time.

(Adapted from Axelrod, 1996)

Data mining technology is a conglomerate of several research disciplines including statistics, cybernetic and genetics, computer science, and artificial intelligence. From this conglomerate of technologies, highly advanced algorithms have been developed to solve specific objectives. Such algorithms drive the mining of data and can be divided into the following categories: association, clustering, classification, sequential patterns, and similar time sequence. These algorithm classifications are described in Figure 1.

Each algorithm is formulated to focus on three components of a successful model—the model's representation, evaluation, and search methods. The algorithms should produce a model that clearly represents limits and assumptions such that patterns can be identified. The model must indicate predictive validity—that can be verified by cross validation and should contain search methodology to optimize the evaluation criteria given the observed data and the model representation.

DATA UTILIZATION

The utility of data is not industry specific. Applications common to all businesses focus on data mining's role in financial and accounting decision making. Over the past 15 to 20 years, computers have captured detailed transaction information in a variety of corporate environments. Now with data mining, a company can capture these detailed transactions in a variety of ways and report on financial and analytical information based on the detail. The need for information has accelerated with data warehouses being the primary source. Data mining can integrate the data warehouse information from multiple operational systems to support decision-making.

The allure of data mining is enhanced by the shortcomings of traditional information systems. Traditional systems have often been plagued by disjointed data storage, often crossing incompatible platforms, making it very difficult if not impossible for managers to identify, gather, and report relevant, accurate, and timely information. To meet this need, firms are moving toward enterprise accounting, where a plethora of information can be mined to aid managers in the decision making process.

Information is extracted from the firm's existing information systems on a scheduled basis and predefined cubes are built. Interested parties can then use a high-level application to twist the cube of information into a useful report that meets their specific needs. The user can drill down from an object in a summarized report to the underlying detailed information.

If there is an inkling that data mining technology simulates the thought processes of a human being and combines the expediency and consistency of a computer, it is due to the direct influence of artificial intelligence technology in the data-mining scheme. Much of the data mining process is driven by neural networks, which artificially mimic the actions of the human brain. According to Berry and Linoff (1997),

> *Memory-based reasoning* (MBR) is a mainstay of data mining. MBR modulates known instances to make predictions about unknown events. For instance, one might maintain a database of claims and whether they were adjusted after investigation. To determine whether a new claim warrants further investigation, one would find similar claims—neighbors—in the database and make the "investigate-further" or "pay immediately" decision based on the status of the neighbors. The two key elements in MBR are the distance function used to find the nearest neighbors and the combination function that combines values at the nearest neighbors to make a prediction.

APPLYING DATA MINING IN THE REAL WORLD

Computer technology that enhances the intuitive and predictive quality of the information is highly attractive to any enterprise seeking a sustainable competitive advantage. For example, sales professionals may employ data mining to probe customer-buying habits, credit risks and other strategic issues. Association algorithms are of great use to those sales professionals who were accustomed to relying on classical market-basket analysis. A typical example of results from market basket analysis may show that customers who purchase 2-by-4s and nails also purchase paint and paintbrushes. As a clustering technique, market basket analysis is useful in identifying specific sequencing of events that might appear to be related occurrences.

Similarly, association algorithms build models that give the likelihood of different products being purchased together and can express these patterns as rules (Berry and Linoff, 1997). The ultimate objective is to determine trends in volumes of transactions that allow insight into buying behavior. This insight may be employed to regulate inventory, modify merchandise layout, or adjust promotional campaigns targeted at specific market segments. Data mining can also include information from external sources such as customer demographics and household information.

Enterprises may also adopt data mining to aid in pricing products and services. For example, a bank may search for customers who maintain low account low balances and who also frequently use their ATM cards at non-home bank teller machines. This would be a useful way of differentiating among customers. This discernment would permit a cost versus benefit focus on those depositors who are advantaged by such services, but yield little or no monetary benefit.

Regulators and auditors maintain a low profile in employing this technology to maintain market integrity. Data mining technology identifies suspect pre/post market opening information related to SEC violations and irregular trading patterns. Just as corporations use data mining to discern patterns in consumers' buying behaviors, to forecast return on investments, and to predict cash flow and revenues, regulators use data mining to identify suspicious trades. (Bucatinsky, 1998). If suspicious trading patterns are proven, those who violate regulation cannot easily escape the penalty as evidenced by the Martha Stewart fiasco (Panza, 2005).

LIMITATIONS OF DATA MINING

Data mining does have limitations. One major limitation is the cost of implementing and maintaining the necessary data mining tools; another is the time and effort. The actual writing of queries and algorithms is a complex and difficult task. Extensive training and practice are still needed for most users to take full advantage of data mining and its capabilities.

Available to consumers are some low-end data mining software programs. The problem with these programs is that their development is somewhat piece-meal (not enterprise-wide modules) with limited query capabilities. With these programs it is impossible to perform multidimensional analysis or to use open-ended questions in finding associations between data items. Making changes, additions, and other replacements to these low-end tools creates integration and implementation problems. In the end, many of the current data mining methods are not truly interactive and can only react in simple ways.

Massive business databases are impressive, but they also present problems in terms of finding efficient algorithms for association rules. Due to the large number of fields, search space needs increase, compounding the chances that the algorithm will find patterns that are invalid. The dynamic nature of data may prompt variables to be significantly modified, deleted, or broadened.

IN THE FUTURE

Data mining will continue to be used in a variety of areas: crime detection, banking, and quality control, forecasting and direct marketing. Retail stores can obtain customer profiles and discover buying patterns allowing them to optimize marketing campaigns. Certified Public Accountants will use the tools in the business valuation process and attorneys will increase their reliance on data mining in the discovery and deposition process. Data mining will become a reliable tool in other industries such as health care and insurance. In today's health care environment, it is not unusual to see a medical professional accessing a computer database instead of the traditional patient file. Data mining complements data warehousing by proposing a new form of analysis based on the discovery of hidden relationships in data.

Like any new technology, once the newness wears off and reality sets in, users tend to become less than enamored with its potential. Unrealistic expectations can be somewhat assuaged by a plan that considers the following:

- The need for high performance computer systems to integrate within the data mining process. Data mining algorithms perform more efficiently and effectively with the use of parallel and multi-processor hardware,
- The need for training and education. Because most query functions are not user–friendly, training must focus on understanding the hypothesis behind the table views generation, statistical inference and multidimensional databases.

- The need for support that will drive the market for new mining techniques. Users need to be focused on application and modification rather than demanding a unique product each time a new query is contemplated.

Focus groups are readily embracing data mining as a tool to develop consumer-specific information. The ever-increasing availability of Internet information has spurred a ready market for techniques to analyze data. Many major vendors are designing their platforms to include data mining techniques. These include vendors such as Microsoft's SQL Server and advanced statistical packages such as SPSS, SAS and S-Plus. Data mining's flexibility and broad applicability make for an appealing return on investment. Considered a complement to data warehousing, data mining can discover hidden relationships within the data.

Paramount to any technological benefit is the concern over violation of civil liberties. Citizens cannot be ambivalent about the use of private information. In this litigious society, the very concept of data mining will encourage privacy experts to establish watchdog groups focused on limiting identifiable personal information.

Planned growth continues to be precursor to a successful business. Data mining products are rapidly becoming an integral part of every business plan, providing a broad spectrum of reporting and statistical data analysis. Providing quality control over shared data will serve to enhance the availability of beneficial information.

REFERENCES

Andrews, Michelle., "Decoding MySpace," *U.S. News & World Report*, September 18, 2006, 46–60.

Arunasalam, Mark., "Data Mining," <http://www.rpi.edu/~arunmk/dm1.html>, July 1999.

Axelrod, C. Warren., "Cashing in on Data Mining," *Wall Street and Technology*, December 1996, 14, 60–62.

Berry, Michael J.A., and Linoff, Gordon., Data Mining Techniques: For Marketing, Sales and Customer Support, New York: John Wiley & Sons, 1997.

Bucatinsky, Julio., "Evading regulation gets a little tougher," *Wall Street and Technology*, August 1998, 16, 56–57.

Panza, Jessica., "Has Prison Helped Martha," *The Journal of the Business Law Society*, March 2005, http://iblsjournal.typepad.com/illinois_business_law_soc/2005/03/has_prison_help.html.

"Site Promotes Senior Tax Break," *The Tennessean*, September 26, 2006, 2B.

Streamlinedsalestax.org, <http://www.streamlinedsalestax.org>.