Data Mining and Homeland Security

Jeffrey W. Seifert
Congressional Research Service, USA

INTRODUCTION

A significant amount of attention appears to be focusing on how to better collect, analyze, and disseminate information. In doing so, technology is commonly and increasingly looked upon as both a tool, and, in some cases, a substitute, for human resources. One such technology that is playing a prominent role in homeland security initiatives is data mining. Similar to the concept of homeland security, while data mining is widely mentioned in a growing number of bills, laws, reports, and other policy documents, an agreed upon definition or conceptualization of data mining appears to be generally lacking within the policy community (Relyea, 2002). While data mining initiatives are usually purported to provide insightful, carefully constructed analysis, at various times data mining itself is alternatively described as a technology, a process, and/or a productivity tool. In other words, data mining, or factual data analysis, or predictive analytics, as it also is sometimes referred to, means different things to different people.

Regardless of which definition one prefers, a common theme is the ability to collect and combine, virtually if not physically, multiple data sources, for the purposes of analyzing the actions of individuals. In other words, there is an implicit belief in the power of information, suggesting a continuing trend in the growth of "dataveillance," or the monitoring and collection of the data trails left by a person's activities (Clarke, 1988). More importantly, it is clear that there are high expectations for data mining, or factual data analysis, being an effective tool.

Data mining is not a new technology but its use is growing significantly in both the private and public sectors. Industries such as banking, insurance, medicine, and retailing commonly use data mining to reduce costs, enhance research, and increase sales. In the public sector, data mining applications initially were used as a means to detect fraud and waste, but have grown to also be used for purposes such as measuring and improving program performance. While not completely without controversy, these types of data mining applications have gained greater acceptance. However, some national defense/homeland security data mining applications represent a significant expansion in the quantity and scope of data to be analyzed. Moreover, due to their security-related nature, the details of these initiatives (e.g., data sources, analytical techniques, access and retention practices, etc.) are usually less transparent.

BACKGROUND

What is Data Mining?

Data mining involves the use of sophisticated data analysis tools to discover previously unknown, valid patterns and relationships in large data sets (Adriaans & Zantinge, 1996). These tools can include statistical models, mathematical algorithms, and machine learning methods (algorithms that improve their performance automatically through experience, such as neural networks or decision trees). Consequently, data mining consists of more than collecting and managing data, it also includes analysis and prediction.

Data mining can be performed on data represented in quantitative, textual, or multimedia forms. Data mining applications can use a variety of parameters to examine the data. They include association (patterns where one event is connected to another event, such as purchasing a pen and purchasing paper), sequence or path analysis (patterns where one event leads to another event, such as the birth of a child and purchasing diapers), classification (identification of new patterns, such as coincidences between duct tape purchases and plastic sheeting purchases), clustering (finding and visually documenting groups of previously unknown facts, such as geographic location and brand preferences), and forecasting (discovering patterns from which one can make reasonable predictions regarding future activities, such as the prediction that people who join an athletic club may take exercise classes) (Han & Kamber, 2001; Taipale, 2003; Two Crows Corporation, 1999).

As an application, compared to other data analysis applications, such as structured queries (used in many commercial databases) or statistical analysis software, data mining represents a *difference of kind rather than degree*. Many simpler analytical tools utilize a verification-based approach, where the user develops a hypothesis and then tests the data to prove or disprove the hypothesis. For example, a user might hypothesize that a

customer who buys a hammer, will also buy a box of nails. The effectiveness of this approach can be limited by the creativity of the user to develop various hypotheses, as well as the structure of the software being used. In contrast, data mining utilizes a discovery approach, in which algorithms can be used to examine several multidimensional data relationships simultaneously, identifying those that are unique or frequently represented. For example, a hardware store may compare their customers' tool purchases with home ownership, type of automobile driven, age, occupation, income, and/or distance between residence and the store. As a result of its complex capabilities, two precursors are important for a successful data mining exercise; a clear formulation of the problem to be solved, and access to the relevant data (Makulowich, 1999).

Reflecting this conceptualization of data mining, some observers consider data mining to be just one step in a larger process known as knowledge discovery in databases (KDD). Other steps in the KDD process, in progressive order, include data cleaning, data integration, data selection, data transformation, *data mining*, pattern evaluation, and knowledge presentation (Han & Kamber, 2001).

A number of advances in technology and business processes have contributed to a growing interest in data mining in both the public and private sectors. Some of these changes include the growth of computer networks, which can be used to connect databases; the development of enhanced search-related techniques, such as neural networks and advanced algorithms; the spread of the client/server computing model, allowing users to access centralized data resources from the desktop; and an increased ability to combine data from disparate sources into a single searchable source (Adriaans & Zantinge, 1996).

In addition to these improved data management tools, the increased availability of information, and the decreasing costs of storing it, have also played a role. Over the past several years, there has been a rapid increase in the volume of information collected and stored, with some observers suggesting that the quantity of the world's data approximately doubles every year (Adriaans & Zantinge, 1996). At the same time, the costs of data storage have decreased significantly from dollars per megabyte to pennies per megabyte. Similarly, computing power has continued to double every 18-24 months, while the relative cost of computing power has continued to decrease.

Data mining has become increasingly common in both the public and private sectors. Organizations use data mining as a tool to survey customer information, reduce fraud and waste, and assist in medical research. However, the proliferation of data mining has raised some implementation and oversight issues as well. These include concerns about the quality of the data being analyzed, the interoperability of the databases and software between agencies, and potential infringements on privacy. Also, there are some concerns that the limitations of data mining are being overlooked as agencies work to capitalize on new means to collect data.

Limitations of Data Mining

While data mining products can be very powerful tools, they are not self-sufficient applications. To be successful, data mining requires skilled technical and analytical specialists who can structure the analysis and interpret the output that is created. Consequently, the limitations of data mining are primarily data or personnel-related, rather than technology-related (Klosgen & Zytkow, 2002; Martinez-Solano, Giblin, & Walshe, 2005).

Although data mining can help reveal patterns and relationships, it does not tell the user the value or significance of these patterns. These types of determinations must be made by the user. Similarly, the validity of the patterns discovered is dependent on how they compare to real-world circumstances. For example, to assess the validity of a data mining application designed to identify potential terrorist suspects in a large pool of individuals, the user may test the model using data that includes information about known terrorists. However, while possibly re-affirming a particular profile, it does not necessarily mean that the application will identify a suspect whose behavior significantly deviates from the original model.

Another limitation of data mining is that, while it can identify connections between behaviors and/or variables, it does not necessarily identify a causal relationship. For example, an application may identify that a pattern of behavior, such as the propensity to purchase airline tickets just shortly before the flight is scheduled to depart, is related to characteristics such as income, level of education, and Internet use. However, that does not necessarily indicate that the ticket-purchasing behavior is caused by one or more of these variables. In fact, the individual's behavior could be affected by some additional variable(s), such as occupation (the need to make trips on short notice), family status (a sick relative needing care), or a hobby (taking advantage of last minute discounts to visit new destinations) (Two Crows Corporation, 1999).

Data Mining Challenges

As data mining initiatives continue to evolve, there are several issues for the policy community to consider related to implementation and oversight. These issues include, but are not limited to, data quality, interoperability,