Order Code RL31798

Data Mining: An Overview

Updated May 3, 2004

Jeffrey W. Seifert
Analyst in Information Science and Technology Policy
Resources, Science, and Industry Division

Data Mining: An Overview

Summary

Data mining is emerging as one of the key features of many homeland security initiatives. Often used as a means for detecting fraud, assessing risk, and product retailing, data mining involves the use of data analysis tools to discover previously unknown, valid patterns and relationships in large data sets. In the context of homeland security, data mining is often viewed as a potential means to identify terrorist activities, such as money transfers and communications, and to identify and track individual terrorists themselves, such as through travel and immigration records.

While data mining represents a significant advance in the type of analytical tools currently available, there are limitations to its capability. One limitation is that 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. A second limitation is that while data mining can identify connections between behaviors and/or variables, it does not necessarily identify a causal relationship. To be successful, data mining still requires skilled technical and analytical specialists who can structure the analysis and interpret the output that is created.

Data mining is becoming increasingly common 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. However, some of the homeland security data mining applications represent a significant expansion in the quantity and scope of data to be analyzed. Two efforts that have attracted a higher level of congressional interest include the Terrorism Information Awareness (TIA) project and the Computer-Assisted Passenger Prescreening System II (CAPPS II) project.

As with other aspects of data mining, while technological capabilities are important, there are other implementation and oversight issues that can influence the success of a project's outcome. One issue is data quality, which refers to the accuracy and completeness of the data being analyzed. A second issue is the interoperability of the data mining software and databases being used by different agencies. Interoperability is a critical part of the larger efforts to improve interagency collaboration and information sharing through e-government and homeland security initiatives. A third issue is privacy. Questions that may be considered include the degree to which government agencies should use and mix commercial data with government data, whether data sources are being used for purposes other than those for which they were originally designed, and possible application of the Privacy Act to these initiatives. It is anticipated that congressional oversight of data mining projects will grow as data mining efforts continue to evolve. This report will be updated as events warrant.

Contents

Data Mining: An Overview

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.[1] 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).[2]

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

---

[1] Two Crows Corporation, *Introduction to Data Mining and Knowledge Discovery, Third Edition* (Potomac, MD: Two Crows Corporation, 1999); Pieter Adriaans and Dolf Zantinge, *Data Mining* (New York: Addison Wesley, 1996).

[2] For a more technically-oriented definition of data mining, see [http://searchcrm.techtarget.com/gDefinition/0,294236,sid11_gci211901,00.html].

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.[3]

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.[4]

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.[5]

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.[6] 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.[7]

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 emphasize their homeland security initiatives.

---

[3] John Makulowich, "Government Data Mining Systems Defy Definition," *Washington Technology*, 22 February 1999, [http://www.washingtontechnology.com/news/13_22/tech_features/393-3.html].

[4] Jiawei Han and Micheline Kamber, *Data Mining: Concepts and Techniques* (New York: Morgan Kaufmann Publishers, 2001), p. 7.

[5] Pieter Adriaans and Dolf Zantinge, *Data Mining* (New York: Addison Wesley, 1996), pp. 5-6.

[6] Ibid., p. 2.

[7] Two Crows Corporation, *Introduction to Data Mining and Knowledge Discovery, Third Edition* (Potomac, MD: Two Crows Corporation, 1999), p. 4.

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.[8]

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).[9]

Data Mining Uses

Data mining is used for a variety of purposes 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. For example, the insurance and banking industries use data mining applications to detect fraud and assist in risk assessment (e.g., credit scoring). Using customer data collected over several years, companies can develop models that predict whether a customer is a good credit risk, or whether an accident claim may be fraudulent and should be investigated more closely. The medical community sometimes uses data mining to help predict the effectiveness of a procedure or medicine. Pharmaceutical firms use data mining of chemical compounds and genetic material to help guide research on new treatments for diseases. Retailers can use information collected through affinity programs (e.g., shoppers' club cards, frequent flyer points, contests) to assess the

---

[8] Ibid., p. 2.

[9] Ibid., p. 1.

effectiveness of product selection and placement decisions, coupon offers, and which products are often purchased together. Companies such as telephone service providers and music clubs can use data mining to create a "churn analysis," to assess which customers are likely to remain as subscribers and which ones are likely to switch to a competitor.[10]

In the public sector, data mining applications were initially used as a means to detect fraud and waste, but they have grown also to be used for purposes such as measuring and improving program performance. It has been reported that data mining has helped the federal government recover millions of dollars in fraudulent Medicare payments.[11] The Justice Department has been able to use data mining to assess crime patterns and adjust resource allotments accordingly. Similarly, the Department of Veterans Affairs has used data mining to help predict demographic changes in the constituency it serves so that it can better estimate its budgetary needs. Another example is the Federal Aviation Administration, which uses data mining to review plane crash data to recognize common defects and recommend precautionary measures.[12]

Recently, data mining has been increasingly cited as an important tool for homeland security efforts. Some observers suggest that data mining should be used as a means to identify terrorist activities, such as money transfers and communications, and to identify and track individual terrorists themselves, such as through travel and immigration records. Two initiatives that have attracted significant attention include the now-discontinued Terrorism Information Awareness (TIA) project[13] conducted by the Defense Advanced Research Projects Agency (DARPA), and the Computer-Assisted Passenger Prescreening System II (CAPPS II) being developed by the Transportation Security Administration (TSA).

DARPA described TIA as three-part project, anticipated to be conducted over five years, to develop technologies that can assist in the detection of terrorist groups planning attacks against American interests, both inside and outside the country. The

---

[10] Two Crows Corporation, *Introduction to Data Mining and Knowledge Discovery, Third Edition* (Potomac, MD: Two Crows Corporation, 1999), p. 5; Patrick Dillon, *Data Mining: Transforming Business Data Into Competitive Advantage and Intellectual Capital* (Atlanta GA: The Information Management Forum, 1998), pp. 5-6.

[11] George Cahlink, "Data Mining Taps the Trends," *Government Executive Magazine*, 1 October 2000, [http://207.27.3.29/tech/articles/1000managetech.htm].

[12] Ibid.

[13] This project was originally identified as the Total Information Awareness project until DARPA publicly renamed it the Terrorism Information Awareness project in May 2003.

Section 8131 of the FY2004 Department of Defense Appropriations Act (P.L. 108-87) prohibited further funding of TIA as a whole, while allowing unspecified subcomponents of the TIA initiative to be funded as part of DOD's classified budget, subject to the provisions of the National Foreign Intelligence Program, which restricts the processing and analysis of information on U.S. citizens. For further details regarding this provision, see CRS Report RL31805 *Authorization and Appropriations for FY2004: Defense*, by Amy Belasco and Stephen Daggett.

three parts of the experimental prototype system included "language translation capabilities, data search and pattern recognition technologies, and advanced collaborative and decision support tools."[14] Each part had the potential to improve the data mining capabilities of agencies that adopt the technology.[15] Automated rapid language translation could allow analysts to search and monitor foreign language documents and transmissions more quickly than currently possible. Improved search and pattern recognition technologies may enable more comprehensive and thorough mining of transactional data, such as passport and visa applications, car rentals, driver license renewals, criminal records, and airline ticket purchases. Improved collaboration and decision support tools might facilitate the search and coordination activities being conducted by different agencies and levels of government.[16]

CAPPS II is described by TSA as "an enhanced system to confirm the identities of passengers and to identify foreign terrorists or persons with terrorist connections before they can board U.S. aircraft."[17] Using information from commercial databases, combined with information provided by the passenger, including full name, address, phone number, and date of birth, the CAPPS II system will assign each passenger a risk score corresponding to a three-color scale.[18] Passengers with a "green" score will undergo "normal screening," while passengers with a "yellow" score will undergo additional screening. Passengers with a "red" score will not be allowed to board the flight and will receive "the attention of law enforcement."[19] While drawing on information from commercial databases, TSA has stated that it will not see the actual information used to calculate the scores, and that it will not retain the traveler's information. TSA plans to test the system at selected airports during spring 2004.[20]

---

[14] DARPA, "Defense Advanced Research Projects Agency's Information Awareness Office and Total Information Awareness Project," [website no longer available].

[15] It is important to note that while DARPA's mission is to conduct research and development on technologies that can be used to address national-level problems, it would not be responsible for the operation of TIA, if it were to be adopted.

[16] For more details about the Terrorism Information Awareness program and related information and privacy laws, see CRS Report RL31730 *Privacy: Total Information Awareness Programs and Related Information Access, Collection, and Protection Laws*, by Gina Marie Stevens, and CRS Report RL31786, *Total Information Awareness Programs: Funding, Composition, and Oversight Issues*, by Amy Belasco.

[17] Transportation Security Administration, "TSA's CAPPS II Gives Equal Weight to Privacy, Security," Press Release, 11 March 2003, [http://www.tsa.gov/public/display?content=09000519800193c2].

[18] Robert O'Harrow, Jr., "Aviation ID System Stirs Doubt," *The Washington Post*, 14 March, 2003, p. A16.

[19] Transportation Security Administration, "TSA's CAPPS II Gives Equal Weight to Privacy, Security," Press Release, 11 March 2003, [http://www.tsa.gov/public/display?content=09000519800193c2].

[20] Sara Kehaulani Goo, "U.S. to Push Airlines for Passenger Records," *The Washington Post*, 12 January 2004, p. A1.

Data Mining Issues

As data mining initiatives continue to evolve, there are several issues Congress may decide to consider related to implementation and oversight. These issues include, but are not limited to, data quality, interoperability, and privacy. As with other aspects of data mining, while technological capabilities are important, other factors also influence the success of a project's outcome.

Data Quality

Data quality is a multifaceted issue that represents one of the biggest challenges for data mining. Data quality refers to the accuracy and completeness of the data. Data quality can also be affected by the structure and consistency of the data being analyzed. The presence of duplicate records, the lack of data standards, the timeliness of updates, and human error can significantly impact the effectiveness of the more complex data mining techniques, which are sensitive to subtle differences that may exist in the data. To improve data quality, it is sometimes necessary to "clean" the data, which can involve the removal of duplicate records, normalizing the values used to represent information in the database (e.g., ensuring that "no" is represented as a 0 throughout the database, and not sometimes as a 0, sometimes as a N, etc.), accounting for missing data points, removing unneeded data fields, identifying anomalous data points (e.g., an individual whose age is shown as 142 years), and standardizing data formats (e.g., changing dates so they all include MM/DD/YYYY).

Interoperability

Related to data quality, is the issue of interoperability of different databases and data mining software. Interoperability refers to the ability of a computer system and/or data to work with other systems or data using common standards or processes. Interoperability is a critical part of the larger efforts to improve interagency collaboration and information sharing through e-government and homeland security initiatives. For data mining, interoperability of databases and software is important to enable the search and analysis of multiple databases simultaneously, and to help ensure the compatibility of data mining activities of different agencies. Data mining projects that are trying to take advantage of existing legacy databases or that are initiating first-time collaborative efforts with other agencies or levels of government (e.g., police departments in different states) may experience interoperability problems. Similarly, as agencies move forward with the creation of new databases and information sharing efforts, they will need to address interoperability issues during their planning stages to better ensure the effectiveness of their data mining projects.

Privacy

As additional information sharing and data mining initiatives have been announced, increased attention has focused on the implications for privacy. Concerns about privacy focus both on actual projects proposed, as well as concerns about the potential for data mining applications to be expanded beyond their original

purposes. For example, some experts suggest that anti-terrorism data mining applications might also be useful for combating other types of crime as well.[21] So far there has been little consensus about how data mining should be carried out, with several competing points of view being debated. Some observers contend that tradeoffs may need to be made regarding privacy to ensure security. Other observers suggest that existing laws and regulations regarding privacy protections are adequate, and that these initiatives do not pose any threats to privacy. Still other observers argue that not enough is known about how data mining projects will be carried out, and that greater oversight is needed. There is also some disagreement over how privacy concerns should be addressed. Some observers suggest that technical solutions are adequate. In contrast, some privacy advocates argue in favor of creating clearer policies and exercising stronger oversight. As data mining efforts move forward, Congress may consider a variety of questions including, the degree to which government agencies should use and mix commercial data with government data, whether data sources are being used for purposes other than those for which they were originally designed, and the possible application of the Privacy Act to these initiatives.

Legislation in the 108th Congress

During the 108th Congress, some legislative proposals have been introduced that would restrict data mining activities by some parts of the federal government, and/or increase the reporting requirements of such projects to Congress. On January 16, 2003, Senator Feingold introduced S. 188 the Data-Mining Moratorium Act of 2003, which would impose a moratorium on the implementation of data mining under the Total Information Awareness program (now referred to as the Terrorism Information Awareness project) by the Department of Defense, as well as any similar program by the Department of Homeland Security. S. 188 has been referred to the Committee on the Judiciary.

On January 23, 2003, Senator Wyden introduced S.Amdt. 59, an amendment to H.J.Res. 2, the Omnibus Appropriations Act for Fiscal Year 2003. As passed in its final form as part of the omnibus spending bill (P.L. 108-7) on February 13, 2003, and signed by the President on February 20, 2003, the amendment requires the Director of Central Intelligence, the Secretary of Defense, and the Attorney General to submit a joint report to Congress within 90 days providing details about the TIA program.[22] Some of the information required includes spending schedules, likely effectiveness of the program, likely impact on privacy and civil liberties, and any laws and regulations that may need to be changed to fully deploy TIA. If the report had not submitted within 90 days, funding for the TIA program could have been

---

[21] Drew Clark, "Privacy Experts Differ on Merits of Passenger-Screening Program," *Government Executive Magazine*, 21 November 2003, [http://www.govexec.com/dailyfed/1103/112103td2.htm].

[22] The report is available at [http://www.eff.org/Privacy/TIA/TIA-report.pdf].

discontinued.[23] Funding for TIA was later discontinued in Section 8131 of the FY2004 Department of Defense Appropriations Act (P.L. 108-87), signed into law on September 30, 2003.[24]

On March 13, 2003, Senator Wyden introduced an amendment to S. 165 the Air Cargo Security Act, requiring the Secretary of Homeland Security to submit a report to Congress within 90 days providing information about the impact of CAPPS II on privacy and civil liberties. The amendment was passed by the Committee on Commerce, Science, and Transportation, and the bill was forwarded for consideration by the full Senate (S.Rept. 108-38). In May 2003, S. 165 was passed by the Senate with the Wyden amendment included and was sent to the House where it was referred to the Committee on Transportation and Infrastructure.

Funding restrictions on CAPPSII were included in section 519 of the FY2004 Department of Homeland Security Appropriations Act (P.L. 108-90), signed into law October 1, 2003. This provision included restrictions on the "deployment or implementation, on other than a test basis, of the Computer-Assisted Passenger Prescreening System (CAPPSII)," pending the completion of a GAO report regarding the efficacy, accuracy, and security of CAPPSII, as well as the existence of a system of an appeals process for individuals identified as a potential threat by the system.[25] In its report delivered to Congress in February 2004, GAO reported that "As of January 1, 2004, TSA has not fully addressed seven of the eight CAPPSII issues identified by the Congress as key areas of interest."[26] The one issue GAO determined that TSA had addressed is the establishment of an internal oversight board. GAO attributed the incomplete progress on these issues partly to the "early stage of the system's development."[27]

On March 25, 2003, the House Committee on Government Reform Subcommittee on Technology, Information Policy, Intergovernmental Relations, and the Census held a hearing on the current and future possibilities of data mining. The witnesses, drawn from federal and state government, industry, and academia,

---

[23] For more details regarding this amendment, see CRS Report RL31786, *Total Information Awareness Programs: Funding, Composition, and Oversight Issues*, by Amy Belasco.

[24] For further details regarding this provision, see CRS Report RL31805 *Authorization and Appropriations for FY2004: Defense*, by Amy Belasco and Stephen Daggett.

[25] Section 519 of P.L. 108-90 specifically identifies eight issues that TSA must address before it can spend funds to deploy or implement CAPPSII on other than a test basis. These include 1. establishing a system of due process for passengers to correct erroneous information; 2. assess the accuracy of the databases being used; 3. stress test the system and demonstrate the efficiency and accuracy of the search tools; 4. establish and internal oversight board; 5. install operational safeguards to prevent abuse; 6. install security measures to protect against unauthorized access by hackers or other intruders; 7. establish policies for effective oversight of system use and operation; and 8. address any privacy concerns related to the system.

[26] General Accounting Office, *Aviation Security: Computer-Assisted Passenger Prescreening System Faces Significant Implementation Challenges*, GAO-04-385, February 2004, p. 4.

[27] Ibid.

highlighted a number of perceived strengths and weaknesses of data mining, as well as the still-evolving nature of the technology and practices behind data mining.[28] While data mining was alternatively described by some witnesses as a process, and by other witnesses as a productivity tool, there appeared to be a general consensus that the challenges facing the future development and success of government data mining applications were related less to technological concerns than to other issues such as data integrity, security, and privacy. On May 6 and May 20, 2003 the Subcommittee also held hearings on the potential opportunities and challenges for using factual data analysis for national security purposes.

On July 29, 2003 Senator Wyden introduced S. 1484 The Citizens' Protection in Federal Databases Act, which was referred to the Committee on the Judiciary. Among its provisions, S. 1484 would require the Attorney General, the Secretary of Defense, the Secretary of Homeland Security, the Secretary of the Treasury, the Director of Central Intelligence, and the Director of the Federal Bureau of Investigation to submit to Congress a report containing information regarding the purposes, type of data, costs, contract durations, research methodologies, and other details before obligating or spending any funds on commercially available databases. S. 1484 would also set restrictions on the conduct of searches or analysis of databases "based solely on a hypothetical scenario or hypothetical supposition of who may commit a crime or pose a threat to national security."

On July 31, 2003 Senator Feingold introduced S. 1544 the Data-Mining Reporting Act of 2003, which was referred to the Committee on the Judiciary. Among its provisions, S. 1544 would require any department or agency engaged in data mining to submit a public report to Congress regarding these activities. These reports would be required to include a variety of details about the data mining project, including a description of the technology and data to be used, an assessment of the expected efficacy of the data mining project, a privacy impact assessment, an analysis of the relevant laws and regulations that would govern the project, and a discussion of procedures for informing individuals their personal information will be used and allowing them to opt out, or an explanation of why such procedures are not in place.

For Further Reading

CRS Report RL31408, *Internet Privacy: Overview and Pending Legislation*, by Marcia S. Smith.

CRS Report RL30671, *Personal Privacy Protection: The Legislative Response*, by Harold C. Relyea. Archived.

---

[28] Witnesses testifying at the hearing included Florida State Senator Paula Dockery, Dr. Jen Que Louie representing Nautilus Systems, Inc., Mark Forman representing OMB, Gregory Kutz representing GAO, and Jeffrey Rosen, an Associate Professor at George Washington University Law School.

CRS Report RL31730, *Privacy: Total Information Awareness Programs and Related Information Access, Collection, and Protection Laws,* by Gina Marie Stevens.

CRS Report RL31786, *Total Information Awareness Programs: Funding, Composition, and Oversight Issues*, by Amy Belasco.

DARPA, *Report to Congress Regarding the Terrorism Information Awareness Program*, May 20, 2003, [http://www.eff.org/Privacy/TIA/TIA-report.pdf].

Department of Defense, Office of the Inspector General, *Information Technology Management: Terrorism Information Awareness Program (D-2004-033)*, December 12, 2003, [http://www.dodig.osd.mil/audit/reports/FY04/04-033.pdf].