

Data Mining

Because of advances in computer technology, massive amounts of information are collected every day by businesses and other entities. Sources of information can be records of website visits, purchases made on-line, or at retail outlets, groceries, bank transactions and credit card purchases. Massive databases, called “data warehouses” are used to store and process these data (Tan 2005).

As computers have become cheaper and more powerful in recent years, processing and storage capacity have increased exponentially. Commercial entities and government agencies have had both the incentive and the ability to collect large amounts of information about their customers and clients. However, the massive amount of data now being collected is beyond the ability of a single data analyst or even a team of analysts to classify and comprehend.

New technologies and more powerful computers have made large-scale data mining possible (GAO, 2004). “Hidden” information and relationships can now be found in the data that would not be readily apparent if the analysis were being done by human analysts (Tan 2005).

“Every single search you’ve ever conducted – **ever** – is stored on a database somewhere...” - Tim Wu, a professor at Columbia Law School in New York City, as quoted on the CommonDreams.org website (Boyd, 2008)

Data mining is the process of sorting through collections of data to discover trends and meaningful relationships. Typically, this sorting is now done

automatically or semi-automatically by software programs. Although the term “Data Mining” is new, many of the statistical techniques used are not.

Data Mining can be used to predict future events, using some variables to predict unknown or future values of other variables. Data mining can also be used to describe patterns of events, and find human-interpretable patterns that describe the data (Tan 2005). Before the development of data aggregation and data mining techniques, it was much more difficult to collated and analyze massive amounts of data. Many types of data can now be analyzed, including structured, textual, spatial, web and multimedia forms. Data mining tools are now generally available either as features of major commercial databases or as add-ons (GAO 2004).

In traditional database applications, the user is already familiar with the types of information available (e.g., name, address, etc.). Data mining techniques, on the other hand, extract information from a database that the user did not previously know existed. (GAO, 2004) An example commonly cited of the usefulness of data mining is the correlation that has been found to exist between the purchase of diapers or milk by men on Thursday or Sunday, and sales of beer to the same customers. (Tan, 2005)

Large scale data mining has been made possible by the development of new algorithms and techniques (including Statistical Analysis/Artificial Intelligence, Machine Learning/Pattern Recognition and Database Systems) which allow computers to “learn” and enable computers to do things such as detect credit card fraud, perform stock market analysis, or recognize speech or

handwriting. Machine Learning is a subfield of Artificial Intelligence, and Pattern Recognition is a subfield of Machine Learning. (Tan, 2005) Internet users make the collection of data possible by providing information about themselves when they visit websites in several ways:

- Intentionally, by providing personal information or completing a lifestyle survey
- Unintentionally, because websites can track which links are clicked, how long visitors viewed the link, the search terms they used, and the time of day.
- Unwittingly, because web tracking software can access records of the URLs that the visitor viewed before and after their visit to the business' website.

Data mining can be used to predict future events, using some variables to predict unknown or future values of other variables, and also to describe patterns or events (GAO 2004). Rudimentary examples of the types of data mining include the grouping of certain names which are common in a specific **location**, e.g., O'Brien and O'Rourke in the Boston area or a **contextual** grouping, such as might be returned by a search engine (e.g., Amazon rainforest and Amazon.com) (Tan 2005)

Limitations of Data Mining

Data Mining can detect patterns and relationship, but it cannot indicate the value or significance of the patterns. Although data mining can identify connections between behaviors and/or variables, it does not necessarily identify a causal relationship. Skilled technical and analytical specialists are still required to structure the analysis and interpret the output that is created (Seifer 2004).

Data quality is an important consideration in data mining efforts.

Important considerations in evaluating the quality of data to be used in data mining efforts and the accuracy of their results include:

- Accuracy and completeness of data being analyzed.
- Interoperability of the data mining software and the databases being use by various entities.
- Mission Creep, i.e., the tendency to use data for purposes for which it was not originally collected. (Seifer, 2004)
- Problems with interpretation if the data is “dirty”

Availability of Personal Data

An amazing quantity of personal data is now available to anyone with a computer and an internet connection.

- Public Records, birth, death, property ownership and more
- www.zabasearch.com can search names, address and more for a minimal fee. (In my experience, much of the data available online is either old or inaccurate)
- www.google.com Searching by name it is possible to find Letters to the Editor and other published information.
- With an address, it is possible to get fairly accurate information about selling prices of properties. www.zillow.com
- Businesses who subscribe to a credit bureau service, can access much additional information, unless, as in California, consumers can block unauthorized access to their personal information with a “security freeze”.

Users of data mining programs have similar objectives:

- Customer Service/Satisfaction
- Improving Efficiency/Reducing Costs
- Inventory Control
- Network Intrusion Detection
- Monitoring Illegal Activities
- Detecting Deviations from normal behavior

Customer Service/Satisfaction

For businesses, the goal is Customer Relationship Management (CRM), which means tracking customer preferences, targeting the correct group of customers for each promotion, and providing customer satisfaction. At the same time, businesses must provide a reasonable return to their shareholders. Government agencies also have “customers”, although the details of the relationship differ. For a government agency, the customer might be a Social Security beneficiary or patient at a Veterans Administration Hospital. The agency must also show congress that it’s programs are achieving service and performance goals. Business customers want efficient service at the lowest possible cost, the government agency’s customers want fast service, and Congress wants to control spending, so businesses and government agencies have substantially similar goals.

Inventory Control

For businesses, this means have the “just right” amount of stock on hand to meet customer needs, on a given day or season. It also means have the “just right” level of staffing at all times, and the “just right” size of facilities and equipment to do its work. Government agencies face the same balancing act, and if careful planning is not done, the result can be disastrous, not just for the agency, but also for the people it serves (e.g., FEMA’s mishandling of relief events in the wake of Katrina). Both businesses and government need to be able accurately forecast their needs for supply, staff and equipment as accurately as possible so that managers can make good decisions about the use of resources.

Network Intrusion Detection

Hardly a week goes by without the announcement that a well-known business or government agency has suffered a loss of computer data or a hacker attack on its system. In California, such losses and attacks must be reported, but this is not always the case in other states, and even if it is, a company or agency representative might decide that it had a better way of handling the incident. (Just one of the many things that went wrong when a VA employee had his laptop stolen a couple of years ago.)

Businesses want their customers to trust them with their personal information to speed transactions and improve marketing efforts. A business which loses the trust of its customers also faces the loss of their business. Likewise, for government agencies, theft of data, loss of confidence in the agency's security procedures, and threats to national security are only a few problems that a breach of computer security can bring. Other fallout from lax computer security efforts can include Inspector General Investigations, lawsuits, Congressional hearings, taxpayer complaints, and budget cuts (GAO, 2004).

Detecting Fraud, Waste and Abuse

In addition to hacker attacks, businesses must be on guard for credit card fraud. A credit card issuer can reverse a payment to a company if it can establish that specific procedures were not followed and many companies are not careful about following procedures. Transactions conducted over the internet are particularly risky, because the business does not receive a physical copy of the signed receipt. Businesses must also have good financial controls in place,

and monitor the activities of employees to insure that their activities do not place the company in financial or legal peril. Many companies are reluctant to acknowledge this risk because of the adverse publicity it can bring. In a government agency setting, the risk may be improperly paid client claims, or inaccurate travel documents. Data mining can help government agency managers control fraud, waste and abuse of resources by providing relevant information in a timely manner (GAO 2004).

Analysis of Scientific and Research Information

Data mining can be beneficial to both businesses and government agencies. Businesses can use data mining to analyze scientific and research information. Banks, insurance companies and drug companies rely on data mining to analyze statistics related to their operations, claims payments, underwriting data and financial data. Drug companies track the results of drug trials, and records of adverse reactions. (Tan, 2005) The National Aeronautics and Space Administration (NASA) makes extensive use of data mining to analyze scientific and research data.

Deviations from Normal Behavior

Both businesses and government agencies use data mining to detect deviations from normal behavior. Financial institutions such banks have a financial interest in detecting fraudulent account activity. Banks use specialized software which monitors behavior patterns and flags unusual movements of money. Banks are interested in preventing fraudulent credit card transactions and the use of stolen and/or forged checks. In the case of credit cards, the seller

sometimes bears the responsibility for the loss, and the card holder is usually limited to a \$50 loss (assuming the loss was reported in a timely manner). In a bank which accepts a forged check, it generally bears the financial responsibility for the loss. As a stop-loss measure, banks maintain active fraud investigation departments. Government agencies can use data mining to monitor electronic traffic for signs of illegal or criminal activity or patterns, money laundering, national security threats and terrorists activities (GAO 2004). Elliot Spitzer's trysts with a prostitute were discovered during an operation to monitor terrorist activities, which interestingly enough, had been approved by Spitzer himself when he was Attorney General of New York State. (New York Times 2008) While you don't need to be dumb or careless to get caught doing something illegal, it certainly helps.

Data mining by government agencies

According to the GAO report, *Data Mining: Federal Efforts Cover a Wide Range of Uses*, out of 128 federal agencies surveyed, 52 are using or plan to use data mining. Of the 199 data mining efforts reported, 68 are planned, and 131 are operational (GAO, 2004). Of the data mining efforts reported by agencies to the GAO, 36 out of 54 involved the use of personal information from private sector sources, e.g., credit reports and credit card transactions. Also, of 77 data mining efforts reported to the GAO, in 46 instances, data mining involved personal information from other government agency sources, including student loan application data, bank account numbers, credit card information and taxpayer identification numbers (GAO, 2004).

The section on Privacy Challenges in the GAO report includes information the government agencies at all levels are now interested in collecting large from commercial sources, and using the data not only to investigate the activities of know terrorists, but also analyze large amounts of data looking for signs of possible terrorist activity by unknown individuals (GAO 2004).

New Data Mining Techniques: Web Mining and Web Structure Mining

Web Mining is the extraction of useful patterns and information from data or activity related to the WorldWide Web. *Web Content Mining* is an automated process that goes beyond keyword extraction, and is generally done by using wrappers to map documents to a data model. Some web mining strategies directly mine content, and some use techniques to improve the content search of other tools like search engines (Galeas 2008).

Using a different methodology, *Web Structure Mining* uses counters to track the number of hyperlinks accessed, and involves analyzing the web access logs of web sites, which can be done in two ways: General Access Pattern Tracking analyzes web logs and Customized Usage Tracking analyzes individual trends. (Galeas, 2008) Both techniques collect information about the user's activity on the web.

Web mining accesses sources of data that were previously unusable, including textural comments from survey research, and log files from web servers. Mining strategies include directly mining the content of documents and techniques that improve on the content search of other tools like search engines. (Galeas, 2008) From the website of SPSS, a commercial software program used

to analyze data: "...Applying data mining to these data adds a richness and depth to the patterns already uncovered by your mining efforts..." (SPSS, 2008)

The Dark Web Project

Hinchun Chen, director of the University of Arizona's Artificial Intelligence Lab notes that since 2001, "...the terrorist presence online has multiplied tenfold...", adding that in 2000, there were approximately 70 to 80 core terrorists sites online, and "...now there are at least 7000 to 8000..."(Fox, 2007) The goal of Chen's Dark Web Project at UA is to systematically collect and analyze all terrorist-generated content on the Web. The project is funded by the National Science Foundation and other federal agencies. Terrorists have found the internet to be fertile ground for recruitment and other activities. Chen and a team of computational scientists are using advanced technological tools such as web spidering, link analysis, content analysis authorship analysis, sentiment analysis and multimedia analysis to study suspect websites (NSF 2007).

Chen's Dark Web Project is funded by the National Science Foundation and other government agencies, and uses web spiders to find and monitor terrorist web sites and forums. According to Chen, sometimes the terrorists fight back. "They put booby-traps in their Web forums," Chen explains, "and the spider can bring back viruses to our machines" (NSF, 2007). According to Chen, scenarios involving vast amounts of information and data points are ideal challenges for computational scientists, who use the power of advanced computers and applications to find patterns and connections where humans can not (NSF 2007).

The Dark Web project team recently completed a study of terrorists training videos, which provided information on how to construct Improvised Explosive Devices (IED's). Understanding what information is being provided, and being able to find out where it is being downloaded, can improve existing countermeasures (NSF 2007).

Conclusion

Data collection has morphed from the simple collection of names and addresses for a mailing list to a vast collection of identifiable personal data which can be used by anyone with a computer, an internet connection and the right kind of commercially available data mining software to access personal information. Many Americans are uncomfortable with this aspect of the web, but as one Silicon Valley executive pointed out recently, it's too late to worry about that now, "The cat is already out of the bag..."

Sources Cited

Boyd, R., (2008), *Data Mining Tells Government and Business a Lot About You*. Retrieved May 10, 2008 from

http://www.commondreams.org/headlines_06/0202-01.htm

Galeas, P. (2007). Untitled, Patricio Galeas.Org, Retrieved May 3, 2008 from

<http://www.galeas.de/webmining.html>

Kottler, S (2007). *'Dark Web' Project Takes on Cyber-Terrorism*, Fox News.com, Retrieved May 10, 2008 from

http://www.foxnews.com/printer_friendly_story/0,3566,300956,00.html

Government Accountability Office (GAO), (2004), *DATA MINING: Federal Efforts Cover a Wide Range of Uses*, GAO-04-548. (This report prepared at the request of Senator Daniel K. Akaka), Retrieved April 20, 2008 from

<http://searching.gao.gov/query.html?charset=iso-8859-1&ql=&rf=2&qt=GAO-04-548&Submit=Search>

National Science Foundation, *The Dark Web project team catalogues and studies places online where terrorists operate* (2007). Retrieved May 10, 2008 from

http://www.nsf.gov/news/news_summ.jsp?cntn_id=110040&org=NSF

New York Times, *As Spitzer Cleared, Experts Cite Precedent*, (11/7/2008). Retrieved April 20, 2008 from

http://topics.nytimes.com/top/reference/timestopics/people/s/eliot_i_spitzer/index.html?inline=nyt-per

Palace, B. (1996). *Data Mining*, Technology Note prepared for Management 274A, Retrieved from

<http://www.anderson.ucla.edu/faculty/jason.frand/teacher/technologies/palace/index.htm>

Seifert, J., *Data Mining: An Overview*, Order Code RL31798, (2004) Congressional Research Service, Washington, DC, The Library of Congress, Retrieved April 20, 2008 from www.crs.gov

SPSS.com, (2008). *Data Mining Improves Decision Making*, Retrieved May 3, 2008 from http://www.spss.com/data_mining/

Taipale, K. (2008). *Data Mining and Domestic Security; Connecting the Dots to Make Sense of Data*, *The Columbia Science and Technology Law Review*, Retrieved May 3, 2008 from www.stir.org

Tan, P., Steinbach, M., Kumar, V, (2005), Introduction to Data Mining. Addison-Wesley. Retrieved May 3, 2008 from
<http://www-users.cs.umn.edu/~kumar/dmbook/index.php>

Wikipedia, *Machine Learning*, Retrieved May 12, 2008 from
http://en.wikipedia.org/wiki/Machine_Learning