2002

Ethical data mining

Laura Cook

Jack Cook

Ethical Data Mining

Laura L. Cook, Rochester Institute of Technology, 38 Haley Ave., Geneseo, NY 14454, (585) 414-7334
Jack S. Cook, Rochester Institute of Technology, 38 Haley Ave., Geneseo, NY 14454, (585) 414-7334

Abstract: The advent of data mining opens up a number of interesting prospects to increase competitiveness. To remain competitive, a corporation must strategically manage its information and react quicker than its competitors. However, this information must be kept secure, but accessible. Every organization must be held responsible for ensuring that their data is being used in a legal and ethical manner. This paper highlights both the positive and negative aspects of data mining. In addition, it provides insight into how information systems (IS) professionals and businesses may protect themselves from the negative ramifications associated with improper use of data.

INTRODUCTION

How would you feel if someone had been following you for months? They know your name, address, birthday, and personal information about your family. They know what medications you take and that you will be taking a vacation in Florida soon. They know the music you listen to, how old your children are, and your favorite books. This is not a stalker. This is just a sample of what marketers and websites can gather about you from tracking you online. As more sophisticated technology is developed, it will challenge the protection of personal privacy and threaten the boundaries of private information usage. In analyzing all the information that can be gathered about a person from online activities, data mining has become increasingly important to businesses looking for a competitive advantage.

Customers today expect businesses to not only meet their needs but also anticipate them. Through giveaways, discounts, and better service, companies have enticed consumers to reveal more and more information about themselves. With technological advances, companies can now regularly employ data mining techniques to explore the contents of data warehouses looking for trends, relationships, and outcomes to enhance their overall operations and discover new patterns that theoretically allow companies to better serve their customers. However, Fayyad and Uthurusamy (2002, p. 28) point out that "Our ability to capture and store data has far outpaced our ability to process and utilize it." There exists a gap in the amount of information that is gathered and the amount of information that is mined. In addition, how this data is mined and used has many implications for personal privacy. The uses of data mining are numerous and implemented by many organizations including government agencies and not-for-profit organizations. However, with this great ability, concerns exist regarding ethical issues associated with data mining.

It should be pointed out that data mining is not a new phenomenon. However, with the increased and widespread use of easily managed technologies, interest in data mining has increased rapidly. There are many applications where data mining is ethical. For example, financial and insurance companies have mined their data to detect patterns of fraudulent credit card usage, find hidden correlations between financial indicators, identify behavior patterns of risky customers, and analyze claims. Utility companies have for decades used it to predict when a generator might fail.

Can you separate technology and its uses and still be ethically and socially responsible? As a professional, you must explore how the technology you create will be used. In this paper, we will explore what constitutes ethical data mining, future directions of data collection, what can be inferred from data, and ethical and social implications.

ETHICS AND DATA MINING

Numerous definitions of data mining exist. One definition is that it is the "mechanized process of identifying and discovering useful structures in data." (Fayyad, 2001, p. 62) In this context, "structure" refers to patterns, models, and relationships in the data. Data mining techniques draw upon such diverse areas as probability theory, information theory, estimation, uncertainty, graph theory, and database techniques (Fayyad, 2001). In addition, artificial intelligence techniques such as neural networks, expert systems, and classification algorithms are also used. Another definition states that data mining involves extracting hidden predictive information from databases to solve business problems (Brandel, 2001). Often the term is misused to describe new ways to present data. Data mining does more than just present existing data in new ways but rather it facilitates the discovery of previously unknown relationships among the data. To further illustrate the point, consider most standard database operations. These operations present results that users already intuitively knew existed in the database. Data mining extracts information from the database the user did not anticipate. Data mining creates information that can be leveraged by the organization to create a competitive advantage. However, it is just as likely to identify meaningless patterns or trends, which in turn wastes time and resources.

Before we further explore ethical data mining, we must first provide a basic understanding of the purpose of data mining. Fayyad and Uthurusamy (2002, p. 30) state it concisely: "Data mining is primarily concerned with making it easy, convenient, and

practical to explore very large databases for organizations and users with lots of data but without years of training as data analysts." With that said, who is responsible for the integrity of this data? Who makes decisions about how this data will be managed and used?

Often, when professionals in the data industry are asked what they think of their ethical obligation concerning databases and data mining, their response is that they must follow the letter of the law. However, in order to be ethical, one must go beyond the minimum restrictions imposed by law. Sadly, some IS professionals believe they are absolved from responsibility since applications are the responsibility of management. What management does with that data is not their concern.

Are there guidelines or an ethical code to help database experts as they grapple with ethical dilemmas? Should there be such guidelines? CIO Magazine put together a list of what they call The Six Commandments of Ethical Data Management [Heller, 2002]. This information was gathered through their IT executive forum with input from over 100 Chief Information Officers (CIOs). To shed light on what is considered ethical data management, below are their six commandments.

1. Data is a valuable corporate asset and should be managed as such, like cash, facilities or any other corporate asset.
2. The CIO is steward of corporate data and is responsible for managing it over its life cycle – from its generation to its appropriate destruction.
3. The CIO is responsible for controlling access to and use of data, as determined by governmental regulation and corporate policy.
4. The CIO is responsible for preventing inappropriate destruction of data.
5. The CIO is responsible for bringing technological knowledge to the development of data management practices and policies.
6. The CIO should partner with executive peers to develop and execute the organization's data management policies.

Particularly in light of recent reports of document shredding and data deletion, every corporation should explicitly state the Chief Information Officer's responsibilities with respect to his or her stewardship of the corporate data. In the next section, information about future trends will be discussed.

WHERE ARE WE HEADING?

The world is becoming more connected and the lines of what is ethical are becoming blurred. It will someday be possible to have detailed profiles on just about anyone. The line between private and public information continues to shift with time. It has become easier to search for information about everything on the Internet. As searching becomes easier, what people consider private information is diminishing as well, especially with the younger generation. They have fewer inhibitions about the Internet and are willing to provide a great deal of information about themselves. As shown in Figure 1, what was once considered private information is now more than likely contained in the public domain. The shift can greatly be attributed to the increased use of databases and the more recent use of the Internet as a way to collect and manipulate information (Caudill and Murphy, 2000). In a recent article legal scholar Jeffrey Rosen stated that, "At the beginning of the 21st century, however, thanks to the Internet, and the Sept. 11 attack on the World Trade Center – which gave rise to the new antiterrorism law that gives authorities far more latitude to inspect logs of Internet use – we have vastly expanded the aspect of private life that can be monitored and recorded. As a result, there is increased danger that personal information originally disclosed to friends and colleagues may be exposed to, and misinterpreted by, a less-understanding audience." (Rosen, 2001, p. 19) As shown in Figure 1, the events of September 11, 2001 (which prompted new legislation and concern) have greatly excelled the rate at which information is changing from the private domain to the public domain.

At some point enough individuals will be negatively impacted so that society will react by implementing legal countermeasures that prohibit and restrict all aspects of the data industry. It is in businesses' best interest to behave in the most ethical manner possible, so as not to create an environment in which lawmakers feel compelled to act.

THE IMPACT OF DATA MINING ON CONSUMERS

Fear often accompanies progress. Historically, the threat of invading ones personal privacy was more of a potential than a reality. However, with the increased use of electronic communications and the World Wide Web (WWW), it has become quite easy and inexpensive to share information among trading partners. Prior to the mid-1990s, there were technical barriers as well as economic disincentives to the sharing of information. As these barriers have fallen, the potential for data mining use and abuse has increased.

At one time, society was very concerned about "Big Brother" (the government) gathering data and determining what they were doing in their personal lives. Interestingly, as a new century begins, it appears as if the organizations most likely to invade your privacy are local businesses. When one considers the quantity of data collected about consumers, it is mind-boggling. Just consider the level of detail contained in the purchasing history of individuals who use VIP, shopper club cards or credit cards to obtain store discounts. Through these membership cards, companies are able to track your purchases, possibly deducing your interests. Furthermore, data may be gathered about you in the most unlikely of places. For example, imagine working out at your local gym on a computerized stair stepper or stationary bike. A computer tracks your heartbeat, or the number of steps taken per minute. Netpulse Communications Incorporated (www.netpulse.com) does just that. They link their exercise equipment to a national database of

healthcare member profiles. "By surveying members, Netpulse plans to flesh out the profiles to include the person's age, weight, gender, birth date, address, and product-buying preferences." (Markoff, 1999, p. 96) Their intention is to provide online advertising based on individual profiles.

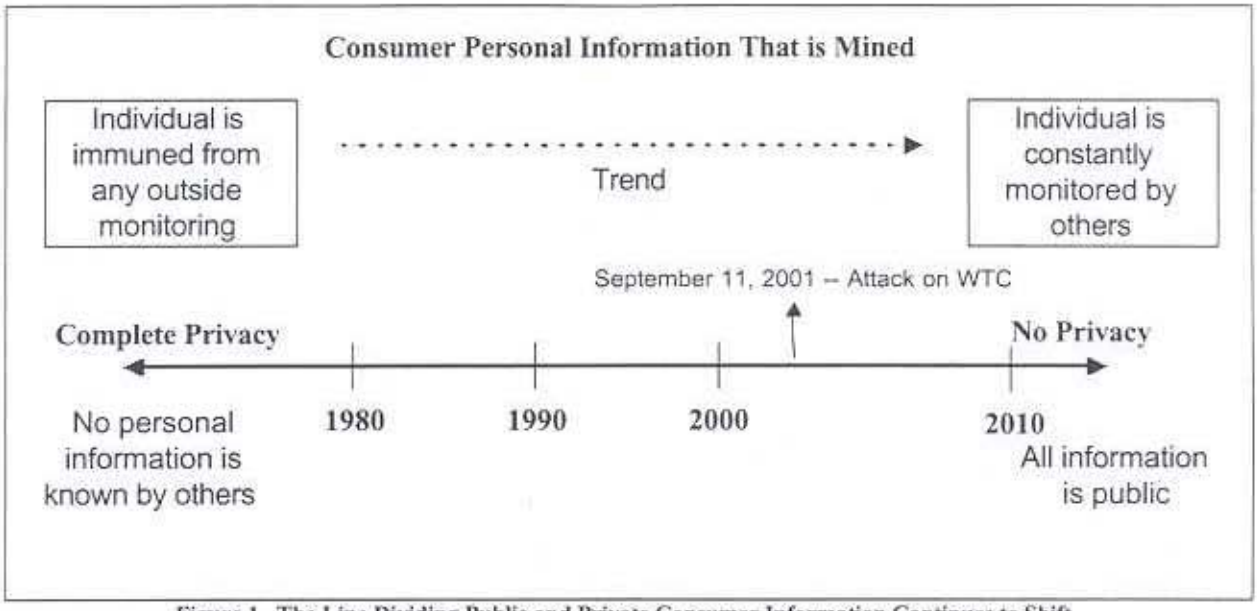

Figure 1. The Line Dividing Public and Private Consumer Information Continues to Shift

Data, information, and knowledge vary in their stability. For example, knowing a customer bought Scooby Doo fruit snacks is less important than the fact the customer has children. "The fact that a customer has diabetes is more stable than a particular pattern of food purchases that may allow inferring he or she has diabetes." (McCarthy, 2000, p. 75) More stable facts such as a person has children or diabetes are more predictive of future behaviors than simple observational facts such as diapers were purchased on the 12th of last month.

Needless to say, no matter how you categorize data, the quantity of data collected about an individual is substantial -- demographic information, customer satisfaction, legal history, insurance records, purchase preferences, financial and banking information, as well as medical profiles. One thing that IS and business professionals must realize is that following ethical practices and respecting the privacy of individuals makes good business sense. Bad publicity associated with a single incident can taint a company's reputation for years, even when that company has followed the law and done everything that it perceives possible to ensure the privacy of those from whom the data was gathered. An example of a company that knows all too well the politics of the privacy debate is N2H2. Their Internet filtering software is used by 40 percent of U.S. schools. They decided last year to sell their aggregated data. They followed the rules set forth by the Children's Online Privacy Protection Act and the data did not contain names or personal information (Wilder and Soat, 2001). However, they had so many people up in arms over the selling of this data that they scrapped the project. Thus, even though N2H2 was well within their legal rights to sell their aggregate data, the public viewed this as unethical.

THE IMPACT OF DATA MINING ON SOCIETY

How can governments take data from a number of different sources and mine it for possible suspects associated with terrorist acts? In the aftermath of the World Trade Center tragedy, governments will be seeking new broad powers to monitor electronic communications in an effort to identify and locate potential terrorists. One of the means at their disposal is to gather data from a number of different sources, and use data mining to identify suspect financial transactions, which ultimately will hopefully lead back to those who sponsor terrorism.

Shari Steele, an executive director of the San Francisco-based Electronic Frontier Foundation, which is a special interest group that aims to keep the Web as free and democratic as possible, is concerned about potential new powers for government to invade personal privacy (Swisher, 2001). "Now it looks like the government will be able to know that and a whole lot more, such as from where you surf, patterns of your e-mail use, what you buy. The ability to learn these patterns has been the dream of marketers and the rallying point for privacy advocates, who have fought successfully since cyberspace's earliest days to prevent such snooping." (Swisher, 2001, B1) Prior to the World Trade Center disaster, it would have been impossible for government to put legislation in place that would allow them to monitor behavior on such an individual basis. In the aftermath, it is not clear whether those who are in

society who are clamoring for restraint and a reasonable response will be heard. More than likely, their calls for a reasoned response will be drowned out by the cries of the masses for measures to safeguard society.

Another drawback is that there can be flaws in the data mining process. There are a number of reasons why records in different databases, even though they actually contain information about the same person or entity, may not "match up" through the data mining process. Some of the difficulties that arise might be through letter, field, or word-related mismatches. Bell and Sethi, in their article "Matching Records in a National Medical Patient Index" examine some of the reasons why data, from a number of different sources, may have difficulty matching records. Under the letter related mismatches, transposition of letters, omission of letters, misspellings, and typing errors can occur during data entry. With respect to word or field related mismatches, maybe the person whose data is contained in two different databases, may have had a change of address or a change of name. There is always the possibility of fraud, and change of zip code, or change of phone number. So one of the concerns that consumers must have is that when information is mined, that it is accurate. And that data that is from another person's history is not accidentally, through improper matching, coupled with legitimate data concerning themselves. In addition to the letter and field related mismatches, there are a number of other issues that might create inaccuracies. For example, Asian names often have surname and given name reversal. If someone is not aware that this is the case, then upon data entry, the names may be switched. Another example where names may appear differently for the same individual is when a person may be named William, but is going by Bill. Hence, through the data mining process, these types of possibilities must be taken into account. Another problem is when a last name might be Cook and is entered as Cooke. Another name that you might find a mistake with is Smith, which can be spelled Smith, Smithe, or Smyth.

After mounting complaints about excessive force, false arrests, and racial profiling, the Los Angeles Police Department (LAPD) is being forced under a federal consent decree to implement a computerized risk-management system that uses data mining to track officers' conduct and performance (Mearian and Rosencrance, 2001). In the year 2000, LAPD paid out more than $28 million as a result of civil lawsuits (Mearian and Rosencrance, 2001). The New Jersey State Police as well as the Pittsburgh Police Department have installed similar systems to track officers' use of force, search and seizure, citizen complaints, criminal charges, civil lawsuits, and commendations and awards earned. However, one cannot forget that these systems are only as good as the data entered. Here are some other examples of data mining software used in well-known industries. "Advanced Scout" is a data mining application developed by an IBM researcher. NBA coaches use it. It can provide information such as "Under what circumstances do the Chicago Bulls outscore the New York Knicks?" (Glode, 1997) This application can also be used by television announcers, fans at NBA web sites, and also used with other sports. In the automobile industry, manufacturers search huge databases of car repairs using pattern recognition algorithms to identify patterns of breakdowns (Waldrup, 2001). The University of California at Berkeley's LINDI system is "used to help geneticists search the biomedical literature and produce plausible hypotheses for the function of newly discovered genes." (Sheier, 2001) Carnegie Mellon University's Informedia II system produces a computer-searchable index of say, for example, CNN news clips "by automatically dividing each clip into individual scenes accompanied by transcripts and headlines." (Sheier, 2001) Data mining is also used for national security and military applications – examples of which are provided in the next two paragraphs.

Advances in networking, medical remote sensing, and data mining will, in the future, be combined to detect the presence and origin of chemical weapons on the battlefield (Ceruti, 2000). This will be accomplished by "mining geographic patterns from networks of devices worn by troops in the field designed to record and transmit a soldier's or a marine's vital health data and environmental data. These geographic patterns will help to identify the origin of the attack. It will also affect the early response and treatment of wartime casualties with a result of more lives saved on the battlefield." (Ceruti, 2000, p. 1875) Another military application of data mining revolves around intrusion detection systems for military networks. Unlike their commercial counterparts, military networks "often face unique constraints – operation over wireless media, unique message traffic, different perceived threats, limited bandwidth, mobile and dynamic environment, robustness in the face of direct attacks on infrastructure – that lead to "normal" operation that is different from civilian networks. This results in an unacceptably high false-alarm rate from Intrusion Detection systems." (Clifton and Gengo, 2000, p. 440) Data mining can be used to identify patterns of false alarms created during battlefield conditions that are substantially different from commercial traffic.

ETHICAL ISSUES

Article 31 of *The Direct Marketing Association's Guidelines for Ethical Business Practice* states that "Marketers should be sensitive to the issue of consumer privacy and should only collect, combine, rent, sell, exchange or use marketing data. Marketing data should be used only for marketing purposes." (DMA Ethical Guidelines, 2001) Essentially, what is of the utmost importance to consumers is that information collected for one purpose should not be analyzed for an unrelated secondary purpose unless it is clearly compatible with the original purpose. Michael Turner, executive director of the Information Services Executive Council, a New York-based affiliate of the Direct Marketing Association, states, "For instance, detailed consumer information lets apparel retailers market their products to consumers with more precision. But if privacy rules impose restrictions and barriers to data collection, those limitations could increase the prices consumers pay when they buy from catalog or online apparel retailers by 3.5% to 11%." (Thibodeau, 2001, p. 36) Obviously, if retailers cannot target their advertising, then their only option is to mass advertise, which drives up costs.

Technological advances make it possible to track in great detail what a person does in their personal life. With this profile of personal details comes a substantial ethical obligation to safeguard this data from disclosure to unauthorized individuals. Ignoring any legal ramifications, the ethical responsibility is firmly placed on IS professionals and businesses whether they like it or not.

Otherwise, they risk lawsuits and harm individuals in the process. "The data industry has come under harsh review. There is a raft of federal and local laws under consideration to control the collection, sale, and use of data. American companies have yet to match the tougher privacy regulations already in place in Europe, while personal and class-action litigation against businesses over data privacy issues is increasing." (Wilder and Soat, 2001, p. 38)

Suppose, based on your participation in a chat room or mailing list or your surfing behavior, someone predicts that either you or someone close to you has a terminal illness. First, what if you wanted this information to remain private? Second, how could such information be used in an unethical manner? Maybe you would start receiving solicitations for donation from organizations that are seeking cures for this illness. Even worse, what if you began to receive offers for cures, which at best have questionable chances of success? Data mining and profiling techniques that allow companies to identify their best customers could just as easily be used by unscrupulous businesses to zero in on vulnerable customers – the elderly, the poor, the sick, and the unsophisticated – offering them inferior or predatory deals (Consumer Reports, 2000). Luckily for the public, data mining abuses have been rarely reported so far. "Canadian Banking Ombudsman Michael Lauber reports that of 175 formal complaints he has handled in 3 ½ years in office, not one has involved a breach of privacy associated with data warehousing or data mining." (Canadian Banker, 2000, p. 18)

CONCLUSION

The benefits of data mining are numerous for businesses, not-for-profits, governments, and individuals as well. Why pursue data mining? Simply satisfying customers is not enough. Relationships need to be built based on loyalty fostered by employee enthusiasm and customized product and service offerings that delight customers. Data mining allows management to create and analyze customer profiles to appropriately customize marketing efforts. Data mining can be beneficial to both consumers and businesses.

REFERENCES

Brandel, M. (2001, March 26). Spinning Data Into Gold. ComputerWorld, 67.

Caudill, E. and Murphy, P. (2000, Spring). Consumer Online Privacy: Legal and Ethical Issues. Journal of Public Policy & Marketing, 7-19.

Ceruti, M. (2000). The Relationship Between Artificial Intelligence and Data Mining: Application to Future Military Information Systems. Proceedings of 2000 IEEE International Conference on Systems, Man, and Cybernetics, 3, 1875.

Clifton, C. and Gengo, G. (2000). Developing Custom Intrusion Detection Filters Using Data Mining. 21st Century Military Communications Conference Proceedings, 1, 440-443.

Fayyad, U. (2001, March). The Digital Physics of Data Mining. Communications of the ACM 44(3), 62-65.

Fayyad, U and Uthurusamy, R. (2002, August). Evolving Data Mining into Solutions for Insights, Communications of the ACM, 45(8), 28-31.

Glode, M. (1997, July 7). Most Valuable Programmer. Wired, archive 5.

Heller, M. (2002, June 1). The Six Commandments of Ethical Data Mining. CIO Magazine.

Markoff, J. (1999, April). The Privacy Debate: Little Brother and the Buying and Selling of Consumer Data. Upside, 95-106.

McCarthy, J. (2000, August). Phenomenal Data Mining. Communications of the ACM, 43(8), 75-79.

Mearian, L. and Rosencrance, L. (2001, April 4). Police Policed with Data Mining Engines. ComputerWorld, 6.

Rosen, J. (2001, Dec.1) Privacy, Reconsidered. An Interview with Jeffrey Rosen CIO Insight, 18-24.

Scheier, R. (2001, July). Finding Pearls in an Ocean of Data. ComputerWorld, 48-49.

Swisher, K. (2001, September 24). Will the Hunt for Terrorists Target Privacy? Wall Street Journal, B1, B6.

Thibodeau, P. (2001, March 26). FTC Examines Privacy Issues Raised by Data Collectors. ComputerWorld, 36.

Waldrop, M. (2001, Jan./Feb.). The Technology Review Ten: Data Mining. Technology Review.

Wilder, C. and Soat, J. (2001, May 14) The Ethics of Data. Information Week, 37-48.

(Nov. 2000). Selling is Getting Personal. Consumer Reports, 65(11), 16.

(Fall Quarter 2000). In Praise of Privacy. Canadian Banker, 14-19.