Knowledge is Power: Ethical Data Mining

B.K.U.R. Nawaratne, J.L.P.R. Perera, K.D.C. Perera, W.P.R.M. Perera, C. Mangaleswaran
Department of Computer Science & Engineering, University of Moratuwa

Abstract — *This paper discusses what Ethical Data mining is and how significant it is for the current world. It describes how the information and data available can be used efficiently while ensuring the privacy of the information and data without any privacy violation. Thus it identifies the problems and implications associated with data mining and gives suggestions to practice ethical data mining.*

Index Terms — Data Mining, Data Integrity, Privacy, Confidential Data

I. Introduction

Our desire for knowledge is really a desire for power. There is not an iota of doubt about it. Today's world runs on knowledge, information. It's vital to our lives; we cannot survive without it. Most of the decisions made are based on that information. Before 21st century we were not able to make such informed and wise decisions due to the limited technologies back then. Therefore currently, with a single piece of information one can change the behavior of an entire community.

Today we have computers, Internet, wireless technology and more enhanced technological entities that will provide unlimited information to its consumers. Thus, the ability of an individual to generate, classify, collect, and exploit data has grown exponentially with the advent of this countless technologies. Yet there exist a fact that never before we had so much information been available, so easily and inexpensively which has led to the biggest challenge of not only getting information, but searching through it to find connections and data that was not previously known.

The knowledge gathered from that data can be useful for different purposes. As an example, a super market can scan its customers' reward cards to gather data to find patterns on their usage of goods. An intelligence agency can extract personal data from government departments' databases and gather knowledge on personal data of specific individuals [1].

One of the techniques which used to gather knowledge is Data Mining. According to Kurl Therling's web site, Data Mining is the extraction of hidden predictive information from the extraction of hidden predictive data. These data can be extracted from large collections of data that are highly complicated for unaided human reasoning to cope with. Thus this can be viewed as a mechanism for acquiring power.

In a general sense, power has the tendency to corrupt. Therefore, a tool like data mining, when applied to human beings, must have certain ethical implications.

According to the paper "Balance of Power" by L. Tuovinen and J. Roning;

"The ethical dilemma of data mining technology stems from the fact that when data representing qualities and activities of individuals is mined, the result may be the disclosure of information that the subjects of the data never intended to be disclosed, even though the data may have been gathered with their consent. This fundamental problem of privacy violation is further amplified if the data is insecure or incorrect (Wahlstrom and Roddick, 2000). In the worst case scenario the source data or the discovered knowledge is deliberately corrupted or otherwise used to attack an individual or group. The knowledge discovery and computer ethics communities have thus been compelled to direct their attention toward establishing ethical standards for Data mining and technical means for ensuring that the standards are not (advertently or inadvertently) disobeyed" [2].

Currently "Data Mining" is not merely an academic example or a buzzword among the geeks and researchers. Moreover it has become a matter of public interest. The general public has a concern on the question, how and for what the data, knowledge is being used.

Current studies shows that many individuals are against and trying to implement rules against the use of data mining to prevent it becoming a handy tool in the wrong hands. As an example; In 2006, New Hampshire passed the Prescription Privacy Law, which prevents patient and prescriber identifying data from being sold or used for advertising, marketing, promotion or any activity intended to influence sales or market share of a pharmaceutical product [3].

There is, however, another side to the equation. If data mining can be used, for example, to fight crime or find solutions for prevailing problems: is it unreasonable to call off the legitimacy of Data Mining. Besides, the positive ethical side of data mining includes applications, such as medical diagnostics, to which it is very difficult to object in the context of any widely accepted value system. However, how the equation should be balanced stands open to debate.

This paper is organized as follows. Section 2 looks in to the ethical problems exist in data mining and illustrates possible implication using examples. Section 3 describes about the legal background with respect to ensuring privacy preservation in data mining and suggests some resolutions to overcome the existing ethical problems. Section 4 provides a summary of the overall discussion.

II. Possible Problems and Implications

With new technologies data mining has improved greatly in the past years, there were challenges and barriers. However with technological development most of them are solved by now. With this development people and organizations have started to identify the possible threats hidden in data mining, mainly ethical issues. Other than the

main ethical issues there are some more problems in this area. Below list is a part of current problems presented in the ICDM (IEEE International Conference on Data Mining), 2005 [12].

1. Developing a Unifying Theory of Data Mining: The current state of the art of data-mining research is too informal, so as a solution we need much unified process that supports deep research.
2. Scaling Up for High Dimensional Data and High Speed Streams: We should have techniques that can handle large quantities of data, and high speed steams (continuous data flows)
3. Sequential and Time Series Data: We should have accurate techniques, to identify trends in data. Moreover, there should be ways to identify them and do predictions on them.
4. Mining Complex Knowledge from Complex Data: There should be ways of identifying complex knowledge like pictures, graphs, multimedia other than simple text.
5. Data mining in a Network Setting: Data mining in network environments, Community and Social Networks has been a challenge for many years.
6. Security, Privacy and Data Integrity: This has been the main problem for all this time.

The mentioned are more of challenges whereas the last issue is the most recurring problem for a long time. Thus in this paper we will discuss more on the issues aroused from security, privacy and data integrity.

World has identified that some data mining activates run by some governmental and non-governmental organizations have crossed the line into private and confidential data of other parties. Dictionary, defines the word privacy as someone's right to keep their personal matters and relationships secret, Today, every form of commerce leaves an electronic trail, and acts that were once considered private or at least quickly forgotten, are stored for future reference. It is an important issue to consider both as individuals and groups, in the work we do that may intrude on the privacy of other.

As an example, according to the study by the United States Government Accountability Office [6], that took five of the dozens of federal agencies which use computerized data analysis showed some interesting facts on this issue. In May 2004, a GAO survey found that federal agencies were using or planning 199 data mining projects, including 122 that used personal information, including credit reports, credit card transactions, student loan application data, bank account numbers and taxpayer identification numbers. Once The New York Times stated that "It is not known precisely why searching the databases, or data mining, raised such a furious legal debate. But such databases contain records of the phone calls and e-mail messages of millions of Americans, and their examination by the government would raise privacy issues" [7].

The sensitive confidential data can vary in their confidentiality levels. Thus processing of certain types of data should be subject to more stringent controls than other personal data. In practice, it has been rather difficult to identify categories of sensitive data [8], which has led to the enclosure of highly sensitive private data to untrusted third parties. As a result, privacy preserving data mining has gained increasing popularity in the data mining research community and also among the people.

For example, in E-Commerce data mining Web server logs are commonly used as the source of data for mining. The open nature of the web logs makes almost everything we place on the web available to the world, including competitors, adversaries and governmental agencies. This technology makes it easy to monitor web usage of others. It also has created an industry of selling data, much of it obtained through web mining, usually without the permission of the subject [9].

People tend to fear about their personal details in the past years. Previous year, we heard a lot of buzz on Facebook cookies that collect huge amount of user data, which are clearly irrelevant to them. Similar fears were there with the Google databases with Gmail, YouTube, calendar, maps, etc. which contained almost half of private data of an average internet user. Even though Google thoroughly said that they will never disclose the information, fear in the minds of people still exists.

If we look from the angle of the data miners (companies, organizations, and governments), they always have reasons for what they are doing. As an example, Federal Bureau of Investigation (FBI) thinks that they can access anything because all small detail can relate to state security in some way. Companies use their customer information to plan the next product or may give them to an advertising company to identify the trends to plan the next advertisement. The consumers might for instance be aware of the fact that collected information about them is used for billing purposes, but that they did not necessarily implicitly agree to allow the organization to use the data in a data-mining scenario, thereby exceeding the original intent of the data collection. To this end, it is important to pay particular attention to how the data used in data mining was obtained in the first place, and whether it is used could result in a violation of privacy.

It is up to the organization employing data mining to ensure that their actions result in neither of the negative effects, namely, incurring legal liability or obtaining bad press as a result of privacy violations associated with their data mining effort [10]. Unethical data mining not only harm the customer but also the data miner. If it is discovered it can taint an organization's reputation for years. Therefore, information technology experts and business professionals must realize that ethical practices and respecting the privacy of individuals is important to a business [11].

III. Legal Background

Most of the proposed problems mentioned in the previous section arise mainly due to the availability of personal data and patterns on per specific user basis. For example an e-commerce website may gather data on buying patterns of its customers for better planning which only needs the aggregated data of all the users but instead it may also use the data to predict per user buying trends and recommendations on what to buy. There are some proposed

resolutions available in the current context.

Data mining activities are often limited by both mandatory and voluntary controls. Mandatory controls consist of legal restriction on access and use of personally identified information and/or judicial means of redress of individuals who are falsely identified and harmed by the data mining activity. Voluntary controls consist of technical, methodological, and institutional (policy) approaches to limit the opportunity for inappropriate access and to ensure that the data mining methodology is sound and produces the highest likelihood of achieving the desired outcome.

Three broad areas of technical/methodological controls have been used to protect the privacy of individuals when information about them is included in data bases subject to data mining are as follows:

1. "Anonymization" techniques that allow data to be usefully shared or searched without disclosing identity
2. "Permissioning" systems that build privacy rules and authorization standards into databases and search engines
3. Immutable audit trails that will make it possible to identify misuse or inappropriate access to or disclosure of sensitive data.

Several federal laws (as the paper [13] discusses on the USA context) afford protection to personally identified data. All the federal agencies are governed by these acts.

First Privacy Act [14] requires that a federal agency to inform the owner of the information, about any databases or information stored from which per individual basis records can be retrieved. In addition the use should be clearly stated. System should also allow the individuals to access or correct or delete any data or information about them being held by other organizations. Under Computer Matching and Privacy Protection Act [15], each agency must take an approval from Data Integrity Board (DIB) before matching of individually identified data from multiple data sources and report the Federal Register the type of the matches.

Most of the acts provides legal guidance to the federal agencies but the fact that the large amount of data on individuals are available on third party websites including their viewing patterns and the people they interact with has much value which may be used to sell those information without noticing the owner. This is not covered under the current laws as there is no viable way to enforce the law. However the Privacy act covers the illegal use of information without the approval from the owner.

Main issue with data mining is that the users are interacting with the organization directly and does not leave a room for a proxy to stand in-between to enforce the law. Suggest several proposals that can be used to enforce the Laws on organizations [16].

One option the paper suggests is pointing of a respected third party (preferably a government organization) .When user registers and disseminates, data; the organization is only allowed to access data from the third party and should not keep information about users more than a day. This model is used by Facebook which only allows the developers of applications that run on Face book to keep usage data about users with a maximum of 20 hours.

There are others proposals available in detail in the paper but an identified drawback is the lack of enforcement of the law on organizations. Many proposals allow room for a malicious organization to exploit the system and retain use of data/information about users. Following are some proposals we suggest to solve the problem.

Introducing standards and making organizations to adhere those standards are one of the main methods the current security system is using to ensure the integrity of web resources. When the users browse resources the users will be notified about the standards that the relevant organization follows. This method should be coupled with the security information about the resource which is currently using this model.

One other option is to have a centralized data store and the users can allow other organizations to access it on behalf of the users with permission. With this system the users have more flexibility and control over the information they share with.

IV. Conclusion

Availability of large data sets about behavioral patterns of general customers logged by companies that they work with, tends to have more specific and more realistic information about the behavioral and personal interests about the customers whom may not be willing to share this information [18].

Based on the past records a company can provide services to customers more specifically and quickly as the system will be capable of suggesting products based on the customers' preference. Similarly the company may sell or use this data in a manner the customer has not provided permission to or does not intend to. This has become a major issue concerning privacy issues.

There are several Acts governing the fair usage of data availability and there are proposed guidelines and standards for the companies to follow the ethical aspects of data mining. But since the prevailing system cannot enforce ethical data mining, there exist a need to discover a new privacy preserving model where the ethical aspects of data mining can be enforced to the organizations [17].

References

[1] Miriam Schulman (2009), "The Uses of Knowledge" [Online]. Available: http://www.scu.edu/ethics/publications/iie/v9n2/

[2] Y. Tuovinen and J. Roning, "Balance of power: the social-ethical aspect of data mining", Intelligent Systems Group, Department of Electrical and Information Engineering University of Oulu, Finland.

[3] Sheppard Mullin (2009, June 15), "Data Mining Bans and Restrictions" [Online]. Available:http://www.fdalawblog.com/2009/06/articles/data-mining/data-mining-bans-and-restrictions/

[4] W. Seltzer, "The Promise and Pitfalls of Data Mining: Ethical Issues ", Fordham University, Copyright of ASA Section on Government Statistics.

[5] Kurt Thearling (2009), "An Introduction to Data Mining" [Online]. Available: http://www.thearling.com/text/dmwhite/dmwhite.htm

[6] Accountability, United States Government, "Agencies Have Taken Key Steps to Protect Privacy in Selected Efforts, but Significant Compliance Issues Remain", United States Government Accountability Office. [Online] August. 2005. Available: http://www.gao.gov/products/GAO-05-866.

[7] S. Shane, D. Johnston, "Mining of Data Prompted Fight Over U.S. Spying" [Online] New York Times. [Online] 29, July, 2007. Available:http://www.nytimes.com/2007/07/29/washington/29nsa.html?ei=5065&en=f770108f5d23c8b4&ex=1186286400&partner=MYWAY&pagewanted=print.

[8] Al-Fedaghi, Sabah, "How Sensitive is Your Personal Information?" Computer Engineering Department, Kuwait University, March. 2007.
[9] L. Olsan, David, "Ethical aspects of web log data mining". Department of Management, University of Nebraska, August 2006.
[10] Pahl, Chen Li and Claus, "Security in the Web Services Framework".s.l.: Dublin City University, 2004.
[11] J. Wang, "Challenges and Opportunities", London: IRM Press, 2003.
[12] Qiang Yang, Xindong Wu (2005, October) IEEE International Conference on Data Mining. [Online].
Available: http://www.cs.uvm.edu/~icdm/10Problems/index.shtml.
[13] J. X. A. R. P. Dempsey, "Technologies that Can Protect Privacy as Information is Shared to Combat Terrorism," Center for Democracy and Technology, May 26, 2004.
[14] "The Privacy Act of 1974", United States Department of Justice.. [Online].
Available: www.usdoj.gov/opcl/privacyact1974.htm.
[15] D. O. o. M. a. B. Jacob J. Lew (2000, December). Memoranda 01-05 - - Guidance on Inter-Agency Sharing of Personal Data - Protecting Personal Privacy. [Online].
Available: http://www.whitehouse.gov/omb/memoranda_m01-05.
[16] Data-Mining Proposals [Online].
Available:
http://www.pfaw.org/pfaw/general/default.aspx?oid=15200.
[17] Peter Fule and John Roddick, "Detecting Privacy and Ethical Sensitivity in Data Mining Results", Copyright of School of Informatics and Engineering, 2004.
[18] Qiang Yang and Xindong Wu, "10 Challenging Problems in Data Mining Research", Copyright of International Journal of Information Technology & Decision MakingVol. 5, No. 4 (2006) 597–604