Review and Analysis of Data Security in Data Mining

Dileep Kumar Singh
IT Resource Centre
Madan Mohan Malaviya Engineering College
Gorakhpur, India
Email : gkp.dks@gmail.com

Vishnu Swaroop
Computer Science and Engineering College
Madan Mohan Malaviya Engineering College
Gorakhpur, India
Email: rsvsgkp@rediffmail.com

Abstract— **In new era the information and data communication technologies are highly used in the Business Industry. The data warehouse is used in the significant business value by improving the effectiveness of managerial decision-making. Naturally such a process may open up new assumption dimensions, detect new invasion patterns, and raises new data security problems. Recent developments in information technology have enabled collection and processing of enormous amount of personal data, such as criminal records, shopping habits, banking, credit and medical history, and driving records. This information is undoubtedly very useful in many areas, including medical research, law enforcement and national security. Data Storage, data access efficiently and speedily is not only the key for competitiveness but the date security and privacy is important and is the new research challenges for Researchers as well as for any business industry.**

Keywords-component Database mining, Database security, Data Privacy, Data Warehouse, Data mining cycle.

I. Introduction

The advent of information technology in various fields of human life has lead to the large amount of data storage in various formats like records, documents, images, sound recording, videos, scientific data and many new data formats. The Data collected from different applications require proper mechanism of extracting knowledge/ information from large repositories for better decision making Knowledge discovery in Databases (KDD), often called data mining, aims at the discovery of useful information from large collection of Data.[1] The field of data mining is gaining significance recognition to the availability of large amounts of data, easily collected and stored via computer systems. Recently, the large amount of data, gathered from various channels, contains much personal information. When personal and sensitive data are published and/or analyzed, one important question to take into account is whether the analysis violates the privacy of individuals whose data is referred to. The importance of privacy is growing constantly. For this reason, many research works have focused on privacy-preserving data mining, proposing novel techniques that allow extracting knowledge while trying to protect the privacy of users. Some of these approaches aim at individual privacy while others aim at corporate privacy.

Data mining, popularly known as Knowledge Discovery in Databases (KDD), it is the nontrivial extraction of implicit, previously unknown and potentially useful information from data in databases. Though, data mining and knowledge discovery in databases (or KDD) are frequently treated as synonyms, data mining is actually part of the knowledge discovery process. [2,3,4]

Usually, data mining e.g. data or knowledge discovery is the process of analyzing data from different perspectives and summarizing it into useful information - information that can be used to increase revenue, cuts costs, or both. Data mining software is one of a number of analytical tools for analyzing data. It allows users to analyze data from many different dimensions or angles, categorize it, and summarize the relationships identified. Technically, data mining is the process of finding correlations or patterns among dozens of fields in large relational databases.[5] Although data mining is a comparatively new term but the technology is not. Companies have used powerful computers to filter through volumes of superstore scanner data and analyze market research reports for many years. However, continuous innovations in computer processing power, disk storage, and statistical software are dramatically increasing the accuracy of analysis while driving down the cost. Data mining, the discovery of new and interesting patterns in large datasets, is an exploding field. One aspect is the use of data mining to improve security, e.g., for intrusion detection. A second aspect is the potential security hazards posed when an adversary has data mining capabilities.

The databases and data warehouses become more and more popular and imply huge amount of data which need to be efficiently analyzed. Knowledge Discovery in Databases can be defined as the discovery of interesting, implicit, and previously unknown knowledge from large databases [5,6]. The collected data exceeds human's capacity to analyze and extract interesting knowledge from large databases. Many methods of data mining have been introduced (methods using classification, discovery of association, characterization. Different user profiles can be implied in data mining system.

Data mining involves the use of sophisticated data analysis tools to discover previously unknown, valid patterns and relationships in large data sets. These tools can include statistical models, mathematical algorithms, and machine learning methods. Consequently, data mining consists of more than collecting, organizing and managing data; it also includes analysis and prediction. Data mining can be performed on data represented in quantitative, textual, graphical, image or multimedia forms. Data mining applications can use a variety of parameters to examine the data. They include association sequence or path analysis, classification, clustering, and forecasting. Most companies already collect and refine massive quantities of data. Data mining techniques can be implemented rapidly on existing software and hardware platforms to enhance the value of existing information resources, and can be integrated with new products and systems as they are brought on-line. The databases and data warehouses become more and more popular and imply huge amount of data which need to be efficiently analyzed. Knowledge Discovery in Databases can be defined as the discovery of interesting, implicit, and previously unknown knowledge from large databases.[7,8]

The huge size of the available data-sets and their high-dimensionality make large-scale data mining applications computationally very demanding, to an extent that high-performance parallel computing is fast becoming an essential component of the solution. Moreover, the quality of the data mining results often depends directly on the amount of computing resources available. In fact, data mining applications are poised to become the dominant consumers of supercomputing in the near future. There is a necessity to develop elective parallel algorithms for various data mining techniques. However, designing such algorithms is challenging. [9]

The data mining database may be a logical rather than a physical subset of your data warehouse, provided that the data warehouse DBMS can support the additional resource demands of data mining. If it cannot, then you will be better off with a separate data mining database. [10]

II. Data Mining Methods Types

A knowledge discovery process involves preprocessing data, choosing a data-mining algorithm, and post processing the mining results. There are many choices for each of these stages, and non-trival interactions between them. [11]

- Relational Learning Methods
- Probabilistic Graphical Dependency Methods
- Example Based Methods
- Non-Linear Regression and Classification Methods
- Decision Tree and Rules Methods.

III. Data Security Issues

One of the key issues raised by data mining technology is not a business or technological one, but a social one. It is the issue of individual privacy. Data mining makes it possible to analyze routine business transactions and glean a significant amount of information about individuals buying habits and preferences.

Another issue is that of data integrity. Clearly, data analysis can only be as good as the data that is being analyzed. A key implementation challenge is integrating conflicting or redundant data from different sources. For example, a bank may maintain credit cards accounts on several different databases. The addresses (or even the names) of a single cardholder may be different in each. Software must translate data from one system to another and select the address most recently entered.

Finally, there is the issue of cost. While system hardware costs have dropped dramatically within the past five years, data mining and data warehousing tend to be self-reinforcing. The more powerful the data mining queries, the greater the utility of the information being gleaned from the data, and the greater the pressure to increase the amount of data being collected and maintained, which increases the pressure for faster, more powerful data mining queries. This increases pressure for larger, faster systems, which are more expensive.[12]

Data mining, the extraction of hidden predictive information from large databases, is a powerful new technology with great potential to help companies focus on the most important information in their data warehouses. Data mining tools predict future trends and behaviors, allowing businesses to make proactive, knowledge-driven decisions. The automated, prospective analyses offered by data mining move beyond the analyses of past events provided by retrospective tools typical of decision support systems. Data mining tools can answer business questions that traditionally were too time consuming to resolve. They scour databases for hidden patterns, finding predictive information that experts may miss because it lies outside their expectations.[13,14,15]

IV. Exploitation of Data Mining

Define 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. 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 to assess the 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.

V. Challenges in Data Mining

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, mission creep, and privacy**,** [16**]** As with other aspects of data mining, while technological capabilities are important, other factors also influence the success of a project’s outcome.

A. 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.

B. 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 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.

C. 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.[17] 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 initiatives.

Data mining has attracted significant interest especially in the past decade with its vast domain of applications. From the security perspective, data mining has been shown to be beneficial in confronting various types of attacks to computer systems. However, the same technology can be used to create potential security hazards. In addition to that, data collection and analysis efforts by government agencies and businesses raised fears about privacy, which motivated the privacy preserving data mining research. One aspect of privacy preserving data mining is that, we should be able to apply data mining algorithms without observing the confidential data values. [18,19] This challenging task is still being investigated. Another aspect is that, using data mining technology an adversary could access confidential information that could not be reached through querying tools jeopardizing the privacy of individuals. Some initial research results in privacy preserving data mining have been published. However, there are still many issues that need further investigation in the context of data mining from both privacy and security perspectives. This workshop aims to provide a meeting place for academicians to identify problems related to all aspects of privacy and security issues in data mining together with possible solutions. Researchers and practitioners working in data mining, databases, data security, and statistics are invited to submit their experience, and/or research results.

VI. Interesting Challenges

- Threats imposed by data mining techniques to privacy/security and possible remedies.
- Statistical approaches to ensure privacy in data mining.
- Statistical disclosure control applied to privacy preserving data mining.
- New methodologies for privacy preserving data mining.
- Security leaks in existing privacy preserving data mining techniques.
- Privacy preserving data mining for specific applications particularly e-commerce.
- Effect of distributed data sources to privacy and security.
- Data quality, privacy, and security measures.

There has been much interest recently on using data mining for counter-terrorism applications. For example, data mining can be used to detect unusual patterns, terrorist activities and fraudulent behavior. While all of these applications of data mining can benefit humans and save lives, there is also a negative side to this technology, since it could be a threat to the privacy of individuals. This is because data mining tools are available on the web or otherwise and even naïve users can apply these tools to extract information from the data stored in various databases and files and consequently violate the privacy of the individuals. Recently we have heard a lot about national security vs. privacy in newspapers, magazines and television talk shows. This is mainly due to the fact that people are now realizing that to handle terrorism; the government may need to collect information about individuals. This is causing a major concern with various civil liberties unions.

We are beginning to realize that many of the techniques that were developed for the past two decades or soon the inference problem can now be used to handle privacy. One of the challenges to securing databases is the inference problem. Inference is the process of users posing queries and deducing unauthorized information from the legitimate responses that they receive. This problem has been discussed quite a lot over the past two decades. However, data mining makes this problem worse. Users now have sophisticated tools that they can use to get data and deduce patterns that could be sensitive. Without these data mining tools, users would have to be fairly sophisticated in their reasoning to be able to deduce information from posing queries to the databases. That is, data mining tools make the inference problem quite dangerous. While the inference problem mainly deals with secrecy and confidentiality we are beginning to see many parallels between the inference problem and what we now call the privacy problem.

VII. CONCLUSION:

Data mining has become 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 can be 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. Successful data mining still requires skilled technical and analytical specialists who can structure the analysis and interpret the output. 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. Some efforts that have attracted a higher level of congressional interest include the Terrorism 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. A third issue is mission creep, or the use of data for purposes other than for which the data were originally collected.

REFERENCES

[1] "Pragyaban Mishra, Neelamadhab Pandhy and Rasmita Panigrahi, " The Survey of Data Mining Applications and Feature Scope", Asian Journal of Computer Science And Information Technology 2-4, 68-77, 2012

[2] Introduction to Data Mining and Knowledge Discovery, Third Edition ISBN: 1-892095-02- 5, Two Crows Corporation, 10500 Falls Road, Potomac, MD 20854 (U.S.A.), 1999.

[3] Dunham, M. H., Sridhar S., "Data Mining: Introductory and Advanced Topics", Pearson Education, New Delhi, ISBN: 81-7758-785-4, 1st Edition, 2006

[4] Fayyad, U., Piatetsky-Shapiro, G., and Smyth P., "From Data Mining to Knowledge Discovery in Databases," AI Magazine, American Association for Artificial Intelligence, 1996.

[5] Fayyad U.M., Piatetsky-Shapiro G., Smyth P. (1996), « From Data Mining to KDD : an overview », AAAI/MIT Press, 1996.

[6] Han J. et Kamber M. (2002), Data Mining: Concepts and Techniques, Morgan Kaufmann Publishers, Canada, 2002.

[7] Larose, D. T., "Discovering Knowledge in Data: An Introduction to Data Mining", ISBN 0-471-66657-2, Wiley & Sons, Inc, 2005.

[8] Fayyad U.M., Piatetsky-Shapiro G., Smyth P. (1996), « From Data Mining to KDD : an overview », AAAI/MIT Press, 1996.

[9] Vipin Kumar, Mahesh V. Joshi, Eui-Hong (Sam) Han, Pang-Ning Tan, and Michael Steinbach, "High Performance Computing for Computational Science - VECPAR 2002", Palma, J. M.L.M., Dongarra, J., Hernndez, V., and Sousa, A. A. (Eds.) 5th International Conference, Porto, Portugal, June 26-28, 2002.

[10] Introduction to Data Mining and Knowledge Discovery, Third Edition ISBN: 1-892095-02-5, Two Crows Corporation, 10500 Falls Road, Potomac, MD 20854 (U.S.A.), 1999

[11] Bernstein, A and Provost , F., "An Intelligent Assistant for the Knowledge Discovery Process", IEEE Transactions on Knowledge and Data Engineering 17(4), pp. 503-518, 2005.

[12] Clifton, C. and D. Marks, "Security and Privacy Implications of Data Mining", Proceedings of the ACM SIGMOD Conference Workshop on Research Issues in Data Mining and Knowledge Discovery, Montreal, June 1996.

[13] Morgenstern, M., "Security and Inference in Multilevel Database and Knowledge Base Systems," Proceedings of the ACM SIGMOD Conference, San Francisco, CA, June 1987.

[14] Database Security IX Status and Prospects Edited by D. L. Spooner, S. A. Demurjian and J. E. Dobson ISBN 0 412 72920 2, 1996, pp. 391-399.

[15] Lin, T. Y. (1994), "Anamoly Detection -- A Soft Computing Approach", Proceedings in the ACM SIGSAC New Security Paradigm Workshop, Aug 3-5, 1994,44-53. This paper reappeared in the Proceedings of 1994 National Computer Security Center Conference under the title "Fuzzy Patterns in data.

[16] Scott W. Ambler (2001) "Challenges with legacy data: Knowing your data enemy is the first step in overcoming it", Practice Leader, Agile Development, Rational Methods Group, IBM, 01 Jul 2001.

[17] Agrawal, R, and R. Srikant, "Privacy-preserving Data Mining," Proceedings of the ACM SIGMOD Conference, Dallas, TX, May 2000.

[18] Clifton, C., M. Kantarcioglu and J. Vaidya, "Defining Privacy for Data Mining," Purdue University, 2002 (see also Next Generation Data Mining Workshop, Baltimore, MD, November 2002).

[19] Evfimievski, A., R. Srikant, R. Agrawal, and J. Gehrke, "Privacy Preserving Mining of Association Rules," In Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. Edmonton, Alberta, Canada, July 2002.

AUTHORS PROFILE

Dileep Kumar Singh received his Master degree in Computer Application in year 2010 presently he is working as SLT in IT Resource Center, Madan Mohan Malaviya Engineering College, Gorakhpur. He has more than 6 years professional experience. His area of interest includes DBMS, & Networks; he is going to register Ph.D. in Computer Science.

Vishnu Swaroop received his Master degree in Computer Application in year 2002 presently he is working as Computer Programmer in Computer Science and Engineering Department, Madan Mohan Malaviya Engineering College, Gorakhpur. He has more than 22 years teaching and professional experience. His area of interest includes DBMS, & Networks; he is pursuing his Ph.D. in Computer Science on Data Management in Mobile Distributed Real Time Database. He has published several papers in several National, International conferences and Journals.