I.J. Information Engineering and Electronic Business, 2012, 3, 34-39
Published Online July 2012 in MECS (http://www.mecs-press.org/)
DOI: 10.5815/ijieeb.2012.03.05

Security Measures in Data Mining

Anish Gupta, Vimal Bibhu, Md. Rashid Hussain
Department of Computer Science & Engineering,
DIT School of Engineering, Plot -48A, Knowledge Park – III, Greater Noida, Uttar Pardesh, India
gupta.anish01@rediffmail.com
vimalbibhu@gmail.com
md.rh16@gmail.com

Abstract **—** Data mining is a technique to dig the data from the large databases for analysis and executive decision making. Security aspect is one of the measure requirement for data mining applications. In this paper we present security requirement measures for the data mining. We summarize the requirements of security for data mining in tabular format. The summarization is performed by the requirements with different aspects of security measure of data mining. The performances and outcomes are determined by the given factors under the summarization criteria. Effects are also given under the tabular form for the requirements of different parameters of security aspects.

Index Terms **—** Artificial Neural Networks, CART – Classification and Regression Tree, CHAID – Chi Square Automatic Interaction, Detection, Genetic Algorithm

I. INTRODUCTION

Data mining is special technical term related with the discovery of new and interesting pattern of data from large data sets. The extraction of hidden predictive information from large databases is a new emerging technology having the huge potential for the help of companies to focus on the important information in the data warehouse. The tools of data mining predicts the future trends and behaviors. This future trends and behavior allow the businesses to make proactive analysis and decision making for the growth of different aspects of the companies. This data mining automates the system to search the relevant information from the databases of data warehouse of the given enterprise which maintains the data warehouse. The data mining tools can answer the business questions which are traditionally very complicated task and take too much time to analyze and produce the result. Most of the companies already collect and refine massive quantities of data. Data mining techniques can be implemented quickly on existing software and hardware platform to enhance the value of existing information resources, and can be integrated with newly products and systems as these are bought on-line. When the data mining tools are implemented on high performance client/server on parallel processing computers either on multiprocessor system or multicomputer system, the data mining tools can analyze massive databases to deliver answers to questions such as mentioned as

"Which clients ar most likely to respond to my next promotional mailing, and why?"[1].

The techniques of data mining are the result of a long process of research and product development. This evolution began when the business data were first stored in magnetic medium of the computer system. The computer system stores huge amount of data. By the way, these data and information require to dig to get the relevant information by the help of data mining and other tools. There are continuous improvements in the access tools of data from different types of databases. These days, generated technologies that allow users to navigate through their data in real time. Data mining takes this evolutionary process beyond retrospective data access and navigation to prospective and proactive information delivery. The prospective information delivery by the means of data warehousing and data mining needs quick and accurate processing of the pervasive data and information of the current and legacy systems. Data mining is ready for the application in the business field. There are basically three different aspects of the data mining [2][3] . These aspects and fields of data mining is given below.

1. Massive data collections.
2. Multiprocessor systems or multicomputer.
3. Data mining algorithms.

The databases those are used in the field of commercial applications growing with very high rate of growth. The recent survey by META group found that at least twenty percent respondents are beyond the fifty gigabytes. The need of powerful and improved computational engines now met with cost effective scale with parallel processing capabilities. The algorithms related with data mining are existing from ten years but have only recently been implemented as mature, reliable and understandable tools that consistently outperforms older statistical tasks and methods. Evolutionary phase of data mining and its associated tools are summarized in table 1.

Table 1. Evolutionary Chart for data mining tools developments

Evol Step	Busi. Ques.	Enab. Techn.	Prod. Provider	Property
Data Coll.	What was revenue in five yrs.	Comp. Tapes and discs	IBM, CDC,	Static data delivery
Data Acce	Sales in March	RDBMS, ODBC, SQL	Oracle, Sybase, IBM	Dynamic data delivery
Deci. Supp	Why low sales in March	OLAP	Pilot, Arbor, Congos	Dynamic data delivery
Data Min	Reason	Adva. Algo.	Pilot, IBM, SGI	Proactive info del.

Security issues and its measures for data mining is measure problem now a day. Data mining provides facts and this is not oblivious to the human beings to analyze the data. it also enables the inspection and analysis of huge amount of data. Due to this activity the analyst can leak the information and data of enterprise. Followings are the possible threats to the data and information of data mining [4].

Predict information about classified work from correlation with unclassified work.

Detect “hidden” information based on “conspicuous” lack of information.

Mining “Open Source” data to determine predictive events.

In this paper the first section is of introduction. Second section contains the scope of data mining. Third section holds the techniques related with data mining. In section four the security concerns are discussed. In section five security measures and performances are analysed. Finally, conclusion is given.

II. SCOPE OF DATA MINING

Data mining derives its name from the similarities between searching for valuable business information in a large database. The data mining processes require either sifting through an huge amount of material, or intelligently probing it to find exactly where the value resides. Given databases of sufficient size and quality, data mining technology can generate new business opportunities by providing these capabilities [5][6].

2.1 Automation in prediction of behavior and trends.

Data mining automates the process of finding predictive information in large databases. Traditionally methods of data mining required extensive analysis by humans hands and can now this becomes direct to answer the predictions and related terms. A typical example of a predictive problem is targeted marketing. Data mining uses data on past promotional mailings to identify the targets most likely to maximize return on investment in future mailings. Other predictive problems include forecasting, insurance analysis for prediction and decision making, income tax department of government for fraud discovery

2.2 Automated discovery of previously unknown patterns.

Data mining tools sweep through databases and identify previously hidden patterns in first step. An example of pattern discovery is the analysis of retail sales data to identify seemingly unrelated products that are often purchased together. Other pattern discovery problems include detecting fraudulent credit card transactions and identifying anomalous data that could represent data entry keying errors. Data mining techniques can produce the benefits of automation on existing software and hardware platforms. It can also be implemented on new systems as existing platforms are upgraded and new products developed [7]. When data mining tools are implemented on high performance parallel processing systems, they can analyze massive databases in minutes. Faster processing means that users can automatically experiment with more models to understand complex data. High speed makes it practical for users to analyze

huge quantities of data. Larger databases, in turn, yield improved predictions.

III. COMMON TECHNIQUES OF DATA MINING

There are many techniques of data mining. The most common techniques used in the field of data mining are followings.

3.1 Artificial neural networks

Non-linear predictive models that learn through training and resemble biological neural networks in structure. This predictive model uses neural networks and finds the patterns from large databases.

3.2 Decision trees

Set of decisions are represented by Tree-shaped structures. These decisions generate rules for the classification of a dataset under the large databases. Specific decision tree methods include Classification and Regression Trees (CART) and Chi Square Automatic Interaction Detection (CHAID).

3.3 Genetic algorithms

Optimization techniques that use processes such as genetic combination, mutation, and natural selection in a design based on the concepts of evolution.

3.4 Nearest neighbor method

A technique that classifies each record in a dataset based on a combination of the classes of the k record(s) most similar to it in a historical dataset (where k ³1). This is sometimes called the k-nearest neighbor technique.

3.5 Rule induction

The extraction of useful if-then rules from data based on statistical significance between different records of database.

Many of these technologies have been in use for more than a decade in specialized analysis tools that work with relatively small volumes of data. These capabilities are now evolving to integrate directly with industry-standard data warehouse and OLAP platforms [8]. The appendix to this white paper provides a glossary of data.

IV. SECURITY CONCERN IN DATA MINING

Databases are important and essential components of different government and private organizations. To protect the data of the databases used in data warehouse and then data mining is central theme of security system. The requirements of data mining security concerned with the following traits.

4.1 Physical Database Integrity

This physical database integrity related with the power failure of the system. When power fails the intermediate records are not posted or retrieved correctly. Due to this the data mining becomes unable to predict pattern by given applications.

4.2 Logical Database Integrity

This type of integrity indicates that modification of value of one field dos not affect other fields of the database records. Whenever this occurs the data mining algorithm can not be able to predict correct information due to logical integrity anomalies with given database for data mining.

4.3 Element Integrity

The integrity of each individual element is necessary for the database which is used for the data mining. If each element of database of data warehouse maintains the integrity, there is no chance for change by human mistake and by any other programs.

4.4 Auditability

The modification of records and fields of the database are taken with OLTP (On line transaction processing applications and by the human operators or by database administrator. The date, time, fileds, records and the previous value of the records should have to be recorded under a log file. This ensures that the proper modification is taken on the database implemented under the data warehouse.

4.5 Access Control

Database system has the capability for the access control. This access control ensures the access privileges of data items from the database. This means that who can read, modify, delete the records or individual fields of the database. This access control is

defined by the database administrator for the users of the enterprise. If a user has only privilege to read the data items of database then he or she can only see the records but can not do anything others. The database administrator can have all types of privileges on the database [9][10]. It means he or she is database administrator then he or she can read, delete, modify the records, tables and others elements of the database

4.6 User Authentication

Database management system requires the regrous user authentication. Without valid user identification number and password the database does not allow the user to do anything on data items of database. Each user has its own user authentication and identification entity. The user has to keep its user ID and password secret.

V. SECURIY MEASURE & PERFORMANCE

Data mining is associated term with database and data warehouse. A data warehouse is built by the help of relational database. There are so many different tools used for the finding meaningful things from databases used in the data warehouse. Database in data warehouse is the main component that provides the correct information which is taken by the tools. Data mining is one of the most popular combination of many tools for data abstraction and getting meaningful items.

Security concerns are related with database and tools. The security aspects deals many things for the data mining applications. The human related errors and mishandling is also a security concern for the data mining. General security concerns are related with the database. These type security measures are based on the characteristics of data mining [11].

5.1 Privacy

This is mandatory for the each individual who operates the data mining tools. Privacy is concerned with individual user. The individual duties is to keep the data items undisclosed to others [12]. The company should have to educate the employees about the privacy and its related aspect time to time according to attacks and breaches of current scenarios and past scenarios. Data privacy internally maintained with the help of different types of integrity constraints.

5.2 Sensitivity

A database of data warehouse keep whole information about the enterprise or company. Some data items of warehouse are sensitive and some are general. The sensitive or confidential information should be separated by other information of database. This separation can be maintained by the help of label or tag. The access right for sensitive information from database is not for all. There should be a policy regarding access of company sensitive information by any means of data mining.

5.3 Data Correctness

Data correctness is vital thing for the data mining. If a database contains incorrect data then mining tools will produce incorrect result. Thus, there would be a filter that filter out the data and correct the data which is not correct. Data correctness should be ensured before entry into the database. Correct data items always produces the correct output by extracting data by data mining tools or by any other tools.

5.4 Data Integrity

Integrity of data is also a security aspect. If data numeric field is in mode of character then it produces the incorrect result of mathematical operations during data mining. Integrity of data under database is managed by the help of various different types of integrity constraints of databases. Once a integrity constraint is enforced on data items then user should not have to right about removal of that integrity constraint.

5.5 Correction of Mistaken Data

The data and information stored in storage medium are not correct completely. Thus, there should be a mechanism that finds the mistaken and incorrect data to be corrected before the storing into the large databases. The correction should be automated not manual. Correction of mistaken data requires algorithms having considered integrity and availability. Manual correction takes too much time and there would be threat for disclosure of sensitive data. A proper mechanism should be implemented on behalf of the company policy to handle the correctness of data if manual procedure is applied for that.

5.6 Elimination of False Matches

In the process of data mining the extraction of information from databases may produce wrong matching output. This false information matching is eliminated by automated filtering. If manual system is applied then proper security aspects of leakage of information should be defined on behalf of the company policies. It is also mandatory to define the policies of the company to prevent the leak of information during the data processing.

The mentioned security measures for databases of data warehouse for data mining applications for extraction useful information summarized in table 2.

Table 2. Summarization of different security measures for data mining.

Security Measure	Requirement	Prformance Effect	Outcome
Privacy	Medium	Not Affected	No Disclosure
Sensitivity	High	Affected	No Disclosure
Correctness	Medium	Affected	High Availability
Integrity	High	Not Affected	Highly Correct
Mistaken Data	Low	Affected	False Output
False Matches	Medium	Affected	False Output

With the consideration of above table it is concluded that requirement of different security measures low, medium and high. If the high then that is mandatory and if medium then also, mandatory and if low then not mandatory.

Performance is affected by applying the requirements of security measures on databases for data mining. There are two terms under the performance factors after applying the different security measures. One is affected and second is not affected. It means different security measures affects accordingly. The table 2 shows all the criteria's.

Outcome is another term which indicates that by applying the different security measures onto the database of data warehouse for data mining. Outcome changes itself according to different aspects of security measure which are under the table 2.

VI. CONCLUSION

Data mining is very emergent technology in current scenarios of computer science and information technology. Data mining tools produces strategic information to the companies which maintain the database for whole company information. A data mining tool digs the information from databases. In this paper we present security aspects and measures related with the databases for data mining. Finally, we say that data mining security measures are very important for the data mining applications. A security measures should be implemented on behalf of the company policies.

REFERENCES

[1]Rakesh Agrawal, Tomasz Imieliski, and Arun Swami. Mining association rulesbetween sets of items in large databases. In *Proceedings of the 1993 ACM SIG-MOD international conference on Management of data*, pages 207{216. ACM Press, 1993.

[2]Varun Chandola and Vipin Kumar. Summarization { compressing data into an informative representation. In *Fifth IEEE International Conference on Data Mining*, pages 98{105, Houston, TX, November 2005.

[3] Levent ErtÄoz, Eric Eilertson, Aleksander Lazarevic, Pang-Ning Tan, Vipin Kumar, Jaideep Srivastava, and Paul Dokas. MINDS - Minnesota Intrusion Detection System. In *Data Mining - Next Generation Challenges and Future Directions*. MIT Press, 2004.

[4]Anil K. Jain and Richard C. Dubes. *Algorithms for Clustering Data*. Prentice Hall, Inc., 1988.

[5]Pawlak, Z. (1990). Rough sets. Theoretical Aspects of Reasoning about Data, Kluwer Academic Publishers, 1992

[6] Lin, T. Y. (1993), "Rough Patterns in Data-Rough Sets and Intrusion Detection Systems", *Journal of Foundation of Computer Science and Decision Support*, Vol.18, No. 3-4, 1993. pp. 225- 241. The extended version of "Patterns in Data-Rough Sets and Foundation of Intrusion Detection Systems" presented at the First Invitational Workshop on Rough Sets, Poznan-Kiekrz, September 2-4. 1992.

[7] Shariq J. Rizvi and Jayant R. Haritsa. Maintaining data privacy in association rule mining. In Proceedings of 28th International Conference on Very Large Data Bases. VLDB, August 20-23 2002. URL http://www.vldb.org.

[8] Oded Goldreich. Secure multi-party computation, September 1998. URL http: //www.wisdom.weizmann.ac.il/~oded/pp.html. (working draft).

[9]Yehuda Lindell and Benny Pinkas. Privacy preserving data mining. In Ad- vances in Cryptology { CRYPTO 2000, pages 36{54. Springer-Verlag, August 20-24 2000. URL http://link.springer.de/link/service/series/0558/bibs/1880/ 18800036.htm.

[10]Jaideep Shrikant Vaidya and Chris Clifton. Privacy preserving association rule mining in vertically partitioned data. In The Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, July 23-26 2002.

[11]Chris Clifton. Using sample size to limit exposure to data mining. Journal of Computer Security, 8(4):281{307, November 2000. URL http://iospress.metapress.com/openurl.asp?genre=artcle&issn=0926227X&volume=8&issue=4&spage=281.

[12]Xiaodong Lin and Chris Clifton. Distributed EM clustering without sharing local information. Journal of Information Science, February 2003. Submitted to Special Issue on Knowledge Discovery from Distributed Information Sources.

Author Profiles

Mr. Anish Gupta is Assistant Professor in department of information technology of DIT School of Engineering. He is B.Tech and M.Tech in Computer Science and Engineering. He is also Pursuing Ph.D in Computer Science & Engineering . He has 11 years of teaching experience.

He is currently working on the post of Assistant Professor at DIT School of Engineering, Greater Noida, Uttar Pardesh, India. Mr. Vimal Bibhu is currently pursuing Ph.D in Computer Science from B.R.A University, Muzaffarpur, Bihar, India, Md. Rashid Hussain is currently pursuing Ph.D in Information Technology from B.R.A University, Muzaffarpur, Bihar, India

Supported by Computer Science and Information Technology departments of DIT School of Engineerin, Greater Noida, Uttar Pardesh, India.

Mr. Vimal Bibhu is M.Tech in Computer Science and Engineering. He is also pursuing Ph.D in Computer Science and Information Technology. He has 8 Years of teaching experience and 2 Years of Research Experience. He is member of SERC, IACSIT and IEANG.

Md. Rashid Hussain is Bachelor and Master degree in Engineering. He is also pursuing Ph.D in Information Technology from B.R.A Bihar University, Muzaffarpur, Bihar, India. He has published many papers in the national and Internation Journals.

