International Journal of Research

International Journal of Research (IJR)
e-ISSN: 2348-6848, p- ISSN: 2348-795X Volume 2, Issue 4, April 2015
Available at http://internationaljournalofresearch.org

To Study the Data Security in Data Mining

Author Name: - Sukhdeep Kaur
Guide Name: - Ms. Ranjit Kaur
College: - University College of Computer Applications
Author Phone No: - 9988803345
Email Id: - Sukhmahal48@Gmail.Com

ABSTRACT

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 requirements 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

INTRODUCTION

1. 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 predict 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 are 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 are 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 exist from ten years but have only recently been implemented as mature, reliable and understandable tools that consistently outperform older statistical tasks and methods. Evolutionary phase of data mining and its associated tools are summarized in Table 1.1

Table 1.1: - 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

1.1 Data Security Technologies

- **Disk Encryption: -** Disk encryption refers to encryption technology that encrypts data on a hard disk drive. Disk encryption typically takes form in either software (see disk encryption software) or hardware (see disk encryption hardware). Disk encryption is often referred to as on-the-fly encryption (OTFE) or transparent encryption.

- **Software versus hardware-based mechanism for protecting data: -** Software-based security solutions encrypt the data to protect it from theft. However, a malicious program or a hacker could corrupt the data in order to make it unrecoverable, making the system unusable. Hardware-based security solutions can prevent read and write access to data and hence offer very strong protection against tampering and unauthorized access.

  Hardware based security or assisted computer security offers an alternative to software-only computer security. Security tokens such as those using PKCS#11 may be more secure due to the physical access required in order to be compromised. Access is enabled only when the token is connected and correct PIN is entered (see two-factor authentication). However, dongles can be used by anyone who can gain physical access to it. Newer technologies in hardware-based security solve this problem offering fool proof security for data.

  Working of hardware-based security: A hardware device allows a user to log in, log out and set different privilege levels by doing manual actions. The device uses biometric technology to prevent malicious users from logging in, logging out, and changing privilege levels. The current state of a user of the device is read by controllers in peripheral devices such as hard disks. Illegal access by a malicious user or a malicious program is interrupted based on the current state of a user by hard disk and DVD controllers making illegal access to data impossible. Hardware-based access control is more secure than protection provided by the operating systems as operating systems are vulnerable to malicious attacks by viruses and hackers. The data on hard disks can be corrupted after a malicious access is obtained. With hardware-based protection, software cannot manipulate the user privilege levels. It is impossible for a hacker or a malicious program to gain access to secure data protected by hardware or performs unauthorized privileged operations. This assumption is broken only if the hardware itself is malicious or contains a backdoor. The hardware protects the operating system image and file system privileges from being tampered. Therefore, a completely secure system can be created using a combination of hardware-based security and secure system administration policies.

- **Backups: -** Backups are used to ensure data which is lost can be recovered from another source. It is considered essential to keep a backup of any data in most industries and the process is recommended for any files of importance to a user.
- **Data Masking: -** Data Masking of structured data is the process of obscuring (masking) specific data within a database table or cell to ensure that data security is maintained and sensitive information is not exposed to unauthorized personnel. This may include masking the data from users (for example so banking customer representatives can only see the last 4 digits of a customer's national identity number), developers (who need real production data to test new software releases but should not be able to see sensitive financial data), outsourcing vendors, etc.
- **Data Eraser: -** Data erasure is a method of software-based overwriting that completely destroys all electronic data residing on a hard drive or other digital media to ensure that no sensitive data is leaked when an asset is retired or reused.

1.2 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].

1.2.1 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 human's 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

1.2.2 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.

1.3 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.

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.

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).

Genetic Algorithms

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

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 ³ 1). This is sometimes called the k-nearest neighbor technique.

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.

1.4 What is Knowledge Discovery?

Some people don't differentiate data mining from knowledge discovery while others view data mining as an essential step in the process of knowledge discovery. Here is the list of steps involved in the knowledge discovery process −

- **Data Cleaning** − in this step, the noise and inconsistent data is removed.
- **Data Integration** − in this step, multiple data sources are combined.
- **Data Selection** − in this step, data relevant to the analysis task are retrieved from the database.
- **Data Transformation** − in this step, data is transformed or consolidated into forms appropriate for mining by performing summary or aggregation operations.
- **Data Mining** − in this step, intelligent methods are applied in order to extract data patterns.
- **Pattern Evaluation** − in this step, data patterns are evaluated.
- **Knowledge Presentation** − in this step, knowledge is represented.

1.5 Challenges and problems

- **Error correction and loss of information**: The most challenging problem within data cleansing remains the correction of values to remove duplicates and invalid entries. In many cases, the available information on such anomalies is limited and insufficient to determine the necessary transformations or corrections, leaving the deletion of such entries as a primary solution. The deletion of data, though, leads to loss of information; this loss can be particularly costly if there is a large amount of deleted data.

- **Maintenance of cleansed data**: Data cleansing is an expensive and time-consuming process. So after having performed data cleansing and achieving a data collection free of errors, one would want to avoid the re-cleansing of data in its entirety after some values in data collection change. The process should only be repeated on values that have changed; this means that a cleansing lineage would need to be kept, which would require efficient data collection and management techniques.

- **Data cleansing in virtually integrated environments**: In virtually integrated sources like IBM's Discovery Link, the cleansing of data has to be performed every time the data is accessed, which considerably decreases the response time and efficiency.

Data-cleansing framework: In many cases, it will not be possible to derive a complete data-cleansing graph to guide the process in advance. This makes data cleansing an iterative process involving significant exploration and interaction, which may require a framework in the form of a collection of methods for error detection and elimination in addition to data auditing. This can be integrated with other data-processing stages like integration and maintenance.

PROBLEM FORMULATION

Before developing research we keep following things in mind so that we can develop powerful and quality research.

3.1 Security, Privacy and Data Integrity

Several researchers considered privacy protection in data mining as an important topic. That is, how to ensure the users' privacy while their data are being mined. Related to this topic is data mining for protection of security and privacy. One respondent states that if we do not solve the privacy issue, data mining will become a derogatory term to the general public. Some respondents consider the problem of knowledge integrity assessment to be important. We quote their observations: "Data mining algorithms are frequently applied to data that have been intentionally modified from their original version, in order to misinform the recipients of the data or to counter privacy and security threats. Such modifications can distort, to an unknown extent, the knowledge contained in the original data. As a result, one of the challenges facing researchers is the development of measures not only to evaluate the knowledge integrity of a collection of data, but also of measures to evaluate the knowledge integrity of individual patterns. Additionally, the problem of knowledge integrity assessment presents several challenges."

Related to the knowledge integrity assessment issue, the two most significant challenges are: (1) develop efficient algorithms for comparing the knowledge contents of the two (before and after) versions of the data, and (2) develop algorithms for estimating the impact that certain modifications of the data have on the statistical significance of individual patterns obtainable by broad classes of data mining algorithms. The first challenge requires the development of efficient algorithms and data structures to evaluate the knowledge integrity of a collection of data. The second challenge is to develop algorithms to measure the impact that the modification of data values has on a discovered pattern's statistical significance, although it might be infeasible to develop a global measure for all data mining algorithms.

3.2 OBJECTIVE

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 requirements 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.

RESEARCH METHODOLOGY

4.1 What is Algorithm in Computer Science?

Algorithm is a process of problem-solving in step by step to get result. Algorithm is very importance for programmers to do computer programming because it figures out the programming process. Algorithm is a part of problem-solving techniques. After the problem has been raise, we have to analyze the problem first then the inputs and outputs are defined. After that we start to design the algorithm that is a process to transform inputs into outputs.

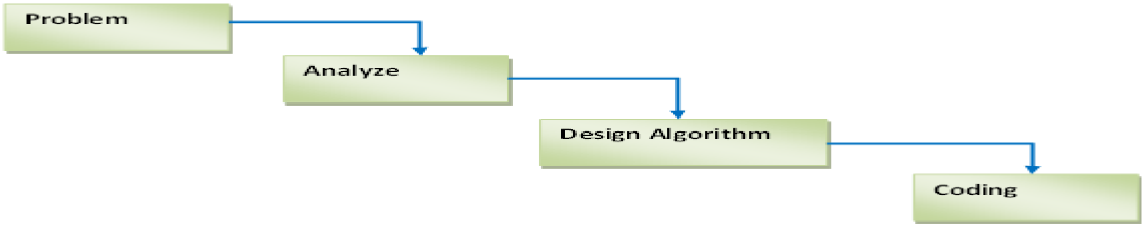

Figure 4.1 Problem – Solving Technique

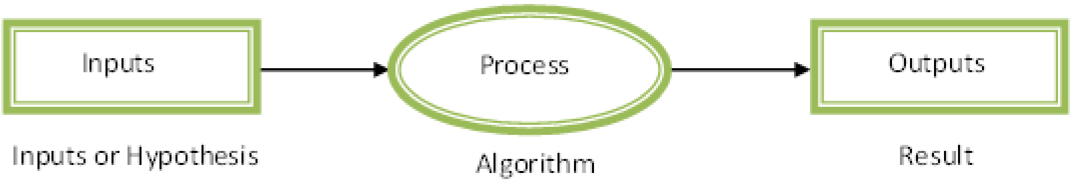

Figure 4.2 Flow – Chart of a Problem

4.2 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.

4.1.1 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.

4.1.2 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 cannot be able to predict correct information due to logical integrity anomalies with given database for data mining.

4.1.3 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.

4.1.4 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, fields, 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.

4.1.5 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 cannot 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

4.1.6 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.

XPERIMENTAL RESULT

5.1 Security 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 combinations of many tools for data abstraction and getting meaningful items.

Security concerns are related with database and tools. The security aspects deal 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].

Privacy

This is mandatory for the each individual who operates the data mining tools. Privacy is concerned with individual user. The individual duty 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.

Sensitivity

A database of data warehouse keeps 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.

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.

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.

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.

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.

5.2 Different Security Measures for Data Mining

The mentioned security measures for databases of data warehouse for data mining applications for extraction useful information summarized in Table 5.1.

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 5.1 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.

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

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

Outcome changes itself according to different aspects of security measure which are under the Table 5.1

CONCLUSION AND FUTURE WORK

6.1 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 rules between 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, and 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)