International Journal of Scientific Research in Computer Science, Engineering and Information Technology
© 2018 IJSRCSEIT | Volume 3 | Issue 3 | ISSN : 2456-3307

Data Mining For Security Purposes

V. Maria Antoniate Martin*[1], Dr. K. David[2], A. Paulin Jenifer[3]

*[1]Research Scholar, Department of Computer Science, Research and Development Centre, Bharathiar University, Coimbatore, Tamil Nadu, India

[2]Assistant Professor, Department of Computer Science, The Rajah's College, Pudukkottai, Tamil Nadu, India

[3]Student, Department of Information Technology, St. Joseph's College, Trichy, Tamil Nadu, India

ABSTRACT

The integrity of computer networks, both in relation to security and with regard to the institutional life of the nation in general, is a growing concern. Security and defense networks, proprietary research, intellectual property, and data based market mechanisms that depend on unimpeded and undistorted access, can all be severely compromised by malicious intrusions. We need to find the best way to protect these systems. In addition, we need techniques to detect security breaches. Data mining has many applications in security including in national security (e.g., surveillance) as well as in cyber security (e.g., virus detection). The threats to national security include attacking buildings and destroying critical infrastructures such as power grids and telecommunication systems. Data mining techniques are being used to identify suspicious individuals and groups, and to discover which individuals and groups are capable of carrying out terrorist activities. Data mining is also being applied to provide solutions such as intrusion detection and auditing. In this paper, we will focus mainly on data mining for security purpose.

Keywords: Data Mining, Security, Data Quality, Integrity

I. INTRODUCTION

DATA mining is the procedure of posing questions and taking out patterns, often in the past mysterious from huge capacities of data applying pattern matching or other way of thinking techniques. Data mining has several applications in protection together with for national protection as well as for cyber protection. The pressure to national protection includes aggressive buildings, demolishing dangerous infrastructures such as power grids and telecommunication structures. Data mining techniques are being examined to realize who the doubtful people are and who is competent of functioning revolutionary activities. Cyber security is concerned with defending the computer and network systems against fraud due to Trojan cattle, worms and viruses. Data mining is also being useful to give solutions for invasion finding and auditing. While data mining has several applications in protection, there are also serious privacy fears. Because of data mining, even inexperienced users can connect data and make responsive associations. Therefore we must to implement the privacy of persons while working on practical data mining. In this paper we will talk about the developments and instructions on privacy and data mining. In particular, we will give a general idea of data mining, the different types of threats and then talk about the penalty to privacy.

II. DATA MINING

Data Mining is defined as extracting information from huge sets of data. In other words, we can say that data mining is the procedure of mining knowledge from data. The information or knowledge extracted so can be used for any of the following applications –

- ✓ Market Analysis
- ✓ Fraud Detection
- ✓ Customer Retention
- ✓ Production Control
- ✓ Science Exploration

Data mining is the process of sorting through large datasets to identify patterns and establish relationships to solve problems through data analysis. Data mining tools allow enterprises to predict future trends.

In simple words, data mining is defined as a process used to extract usable data from a larger set of any raw data. It implies analyzing data patterns in large batches of data using one or more software. Data mining has applications in multiple fields, like science and research. As an application of data mining, businesses can learn more about their customers and develop more effective strategies related to various business functions and in turn leverage resources in a more optimal and insightful manner. This helps businesses be closer to their objective and make better decisions. Data mining involves effective data collection and warehousing as well as computer processing. For segmenting the data and evaluating the probability of future events, data mining uses sophisticated mathematical algorithms. Data mining is also known as Knowledge Discovery in Data (KDD).

III. ARCHITECTURE OF DATAMINING

Data mining is described as a process of discover or extracting interesting knowledge from large amounts of data stored in multiple data sources such as file systems, databases, data warehouses…etc. This knowledge contributes a lot of benefits to business strategies, scientific, medical research, governments and individual. The architecture contains modules for secure safe-thread communication, database connectivity, organized data management and efficient data analysis for generating global mining model. [1].

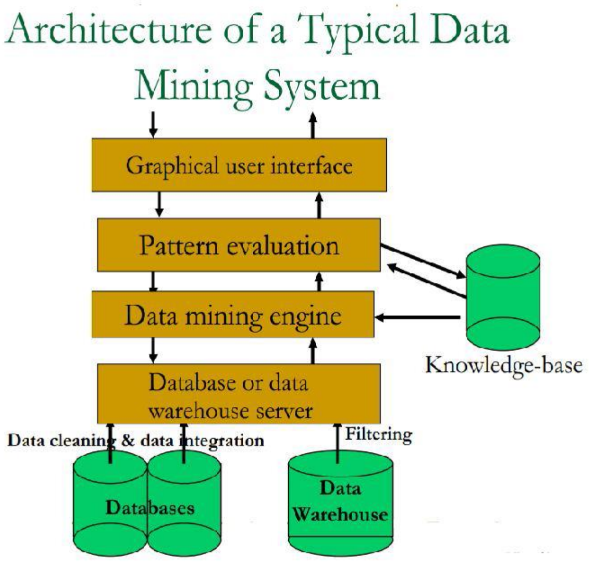
Architecture of a Typical Data Mining System
Graphical user interface
Pattern evaluation
Data mining engine
Knowledge-base
Database or data warehouse server
Data cleaning & data integration
Filtering
Databases
Data Warehouse

Figure 1

IV. DATA MINING FOR SAFETY APPLICATIONS

Data mining is fitting a key technology for identifying doubtful activities. In this section, data mining will be discussed with respect to use in both ways for non-real-time and for real-time applications. In order to complete data mining for counter terrorism applications, one wants to gather data from several sources. For example, the subsequent information on revolutionary attacks is wanted at the very least: who, what, where, when, and how; personal and business data of the possible terrorists: place of birth, religion, education, ethnic origin, work history, finances, criminal record, relatives, friends and associates, and travel history; unstructured data: newspaper articles, video clips, dialogues, e-mails, and phone calls. The data has to be included, warehoused and mined. One wants to develop sketches of terrorists, and activities/threats. The data has to be mined to take out patterns of possible terrorists and forecast future activities and goals. Fundamentally one wants to find the "needle in the haystack" or more suitably doubtful needles among probably millions of needles. Data integrity is essential and also the methods have to SCALE. For several applications such as urgent situation response, one needs to complete real-time data mining. Data will be incoming from sensors and other strategy in the form of nonstop data streams together with breaking news, videocassette releases, and satellite images. Some serious data may also exist in caches. One wants to quickly sift through the data and remove redundant data for shortly use and analysis (non-real-time data mining). Data mining techniques require to meet timing restriction and may have to stick the quality of service (QoS) tradeoffs among suitability, accuracy and precision. The consequences have to be accessible and visualized in real-time. Additionally, alerts and triggers will also have to be employed. Efficiently applying data mining for safety applications and to develop suitable tools, we need to first find out what our present capabilities are. For instance, do the profitable tools balance? Do they effort only on particular data and limited cases? Do they carry what they assure? We require a balanced objective study with display. At the same time, we also require to work on the large picture. For instance what do we desire the data mining tools to carry out? What are our end consequences for the predictable future? What are the standards for achievement? How do we assess the data mining algorithms? What test beds do we construct? We require both a near-term as well as longer-term resolutions. For the future, we require to influence present efforts and fill the gaps in a objective aimed way and complete technology transfer. For the longer-term, we require a research and development diagrams. In summary, data mining is very helpful to resolve security troubles. Tools could be utilized to inspect audit data and flag irregular behavior. There are many latest works on applying data mining for cyber safety applications, Tools are being examined to find out irregular patterns for national security together with those based on categorization and link analysis. Law enforcement is also using these kinds of tools for fraud exposure and crime solving.

V. 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. [2] 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. [3, 4, 5]

VI. PREPERATION 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, [6] as with other aspects of data mining, while technological capabilities are important, other factors also influence the success of a project's outcome. We generate an enormous amount of data as a by-product of our everyday transactions (purchasing goods, enrolling for courses, etc.), visits to Web sites and interactions with government (taxes, census, car registration, voter registration, etc.). Not only is the number of records we generate increasing, but the amount of data gathered for each type of record is increasing.

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. Data Mining Application Areas

There are many areas of data mining application in most popular are Science (astronomy, bioinformatics, drug discovery), Business (advertising, customer relationship management, investment, manufacturing, entertainment, telecom, e-commerce, banking, marketing, health), web (serach engines, bots), government (law enforcement, proofing tax chater, anti-terror).

C. 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.

D. 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.[7] So far there has been little consensus about how data mining should be Carrie 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. [8, 9] 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. [10, 11] 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.

VII. SECURITY ANALYSIS

Most of the web development companies does not follow industrial standard of developing and hosting the websites. Customer using a website is unaware of whether it is a trusted website or untrusted website. In this paper security on e-commerce website is provided with trust path intermediate algorithm, false hit database algorithm and similarity search. Multistep processing is carried on nearest neighbor and similarity search. C-AMNC- used to reduce the size of false hit database. Query is authenticated and server maintains the database of trusted user details to reduce hang or lag in server. Provides accurate data with NN result-set. Security analysis module for providing security on ecommerce web sites. Module 1: Authentication; Module 2: Query processing; Module 3: Similarity Search; Module 4: False hit reduction. These techniques are used to provide security for e-commerce websites.

Module 1: Authentication Member user access search facility in the job site. Admin updates the database from the false hit.

Module 2: Query Processing Server and the user interaction take place in the module .Client post the query and server respond it back from the criteria

Module 3: Similarity Search Retrieve of relevant information from the database based with similar key word.

Module 4: False hit reduction Admin constantly checks the false hit record. He then finally post necessary response with search database for future verification

Case 1: Search keyword is updated if it is not found in database

Case2: If the search keyword is already present in database then admin post necessary response to the search database for future verification.

Case3: User can access necessary search details from database. Admin checks false hit data and update database. Administrator has a set of privileges to modify or update website based on user details.

VIII. GROWTH IN PRIVACY

Different types of privacy problems have been considered by researchers. We will point out the various problems and the solutions projected.

a. **Problem:** Privacy contraventions that consequence due to data mining: In this case the way out is Privacy protecting data mining. That is, we perform data mining and give out the results without enlightening the data values used to perform data mining.
b. **Problem**: Privacy contraventions that result due to the Inference problem. Note that Inference is the procedure of realizing sensitive data details from the lawful answers received to user inquiries. The way out to this problem is Privacy Constraint Processing.
c. **Problem**: Privacy contravention due to un-encrypted data: the way out to this problem is to make use of Encryption at different levels
d. **Problem**: Privacy contravention due to poor system design. Here the way out is to build up methodology for designing privacy-enhanced systems. Below we will observe the ways out projected for both privacy constraint/policy processing and for privacy preserving data mining. Privacy limitation or policy processing research was carried out and is footed on some of her prior research on security restriction processing. Instance of privacy restrictions include the following.
e. **Simple Constraint:** an aspect of a document is private. Content footed constraint: If document holds information about X, then it is private.
f. **Association-based Constraint:** Two or more documents used together are private; individually each document is public.
g. **Free constraint:** After X is freed Y becomes private. The way out projected is to augment a database system with a privacy checker for constraint processing. During the inquiry process, the constraints are checked up and only the public information is freed unless certainly the user is approved to obtain the private information. Our approach also contains processing constraints during the database update and design operation.

Some early work on managing the privacy problem that consequence from data mining was performed by Clifton at the MITRRE Corporation. The suggestion here is to avoid useful outcomes from mining. One could initiate “cover stories” to provide “false” outcomes. Another approach is to only build a sample of data existing so that a challenger is not capable to come up with helpful rules and analytical functions. However these approaches did not impression as it beaten the idea of data mining. The objective is to perform effective data mining but at the same time guard individual data values and sensitive relations. Agrawal was the first to invent the word privacy preserving data mining. His early work was to initiate random values into the data or to bother the data so that the real data could be confined. The challenge is to initiate random values or agitate the values without touching the data mining results [12]. Another new approach is the Secure Multi-party Computation (SMC) by Kantarcioglu and Clifton [13]. Here, each party knows its individual contribution but not the others’ contributions. Additionally the final data mining outcomes are also well-known to all. Various encryption techniques utilized to make sure that the entity values are protected. SMC was demonstrating several promises and can be used also for privacy preserving scattered data mining. It is provably safe under some suppositions and the learned models are correct; It is assumed that procedures are followed which is a semi truthful model. Malicious model is also investigated in some current work by Kantarcioglu and Kardes [14]. Many SMC footed privacy preserving data mining algorithms contribute to familiar sub-protocols (e.g. dot product, summary, etc.). SMC does have any disadvantage as it's not competent enough for very large datasets. (E.g. petabyte sized datasets); Semi-honest model may not be reasonable and the malicious model is yet slower. There are some novel guidelines where novel models are being discovered that can swap better between efficiency and security. Game theoretic and motivation issues are also being discovered. Finally merging anonimization with cryptographic techniques is also a route. Before performing an evaluation of the data mining algorithms, one wants to find out the objectives. In some cases the objective is to twist data while still preserving some assets for data mining. Another objective is to attain a high data mining accuracy with greatest privacy protection. Our current work imagines that Privacy is a personal preference, so should be individually adjustable. That is, we want to make privacy protecting data mining approaches to replicate authenticity. We examined perturbation based approaches with real-world data sets and provided applicability learning to the existing approaches [15]. We found that the rebuilding of the original sharing may not work well with real-world data sets. We attempted to amend perturbation techniques and adjust the data mining tools. We also developed a new privacy preserving decision tree algorithm [16]. Another growth is the platform for privacy preferences (P3P) by the World Wide Web association (W3C). P3P is an up-and-coming standard that facilitates web sites to convey their privacy practices in a typical format. The format of the strategies can be robotically recovered and appreciated by user agents. When a user comes in a web site, the privacy policies of the web site are communicated to the user; if the privacy policies are dissimilar from user favorites, the user is notified; User can then make a decision how to continue. Several major corporations are working on P3P standards.

IX. CONCLUSION

In this paper we have talk about data mining for security purpose. We have examined the idea of privacy and then talked about the growth privacy in data mining. We then presented an agenda for research on security purpose in data mining. Here are our conclusions. There is no collective definition for privacy, each organization must clear-cut what it indicates by privacy and growth privacy policies. Technology only is not adequate for privacy; we require Technologists, Policy expert, Legal experts and Social scientists to effort on Privacy. Some well acknowledged people have believed 'Forget about privacy" Therefore, should we follow research on Privacy? We trust that there are attractive research problems; therefore we need to carry on with this research. Additionally, some privacy is better than nil. One more school of consideration is tried to avoid privacy destructions and if destructions take place then put on trial. We need to put into effect suitable policies and checkup the legal aspects. We need to undertake privacy from all directions.

X. REFERENCES

1. Mafruz Zaman Ashrafi, David Taniar, Kate A. Smith, "Data Mining Architecture for Clustered Environments" , Proceeding PARA '02 Proceedings of the 6th International Conference on Applied Parallel Computing Advanced Scientific Computing, Pages 89-98, Springer-Verlag London, UK ©2002
2. Z. Ferdousi, A. Maeda, "Unsupervised outlier detection in time series data", 22nd International Conference on Data Engineering Workshops, pp. 51-56, 2006

3. Morgenstern, M., "Security and Inference in Multilevel Database and Knowledge Base Systems," Proceedings of the ACM SIGMOD Conference, San Francisco, CA, June 1987.
4. S. A. Demurjian and J. E. Dobson, "Database Security IX Status and Prospects Edited by D. L. Spooner ISBN 0 412 72920 2, 1996, pp. 391- 399.
5. Lin, T. Y., "Anamoly Detection -- A Soft Computing Approach", Proceedings in the ACM SIGSAC New Security Paradigm Workshop, Aug 3-5, 1994,44-53.,1994
6. Scott W. Ambler, "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.
7. Agrawal, R, and R. Srikant, "Privacy-preserving Data Mining," Proceedings of the ACM SIGMOD Conference, Dallas, TX, May 2000.
8. 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.
9. 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.
10. Fung B., Wang K., Yu P. "Top-Down Specialization for Information and Privacy Preservation. ICDE Conference, 2005.
11. Wang K., Yu P., Chakraborty S., " Bottom-Up Generalization: A Data Mining Solution to Privacy Protection.", ICDM Conference, 2004.
12. Agrawal, R., Srikant, R.: Privacy-Preserving Data Mining. In: SIGMOD Conference, pp.439–450 (2000)
13. Kantarcioglu, M., Clifton, C.: Privately Computing a Distributed k-nn Classifier. In: Bou-licaut, J.-F., Esposito, F., Giannotti, F., Pedreschi, D. (eds.) PKDD 2004. LNCS, vol. 3202,279–290. Springer, Heidelberg (2004)
14. Kantarcioglu, M., Kardes, O.: Privacy-Preserving Data Mining Applications in the Mali-cious Model. In: ICDM Workshops, pp. 717–722 (2007)
15. Liu, L., Kantarcioglu, M., Thuraisingham, B.M.: The applicability of the perturbation based privacy preserving data mining for real-world data. Data Knowl. Eng. 65(1), 5–21 (2008)
16. Liu, L., Kantarcioglu, M., Thuraisingham, B.M.: A Novel Privacy Preserving Decision Tree. In: Proceedings Hawaii International Conf. on Systems Sciences (2009)
17. Fayyad, U., Piatetsky-Shapiro, G., and Smyth P., "From Data Mining toKnowledge Discovery in Databases," AI Magazine, AmericanAssociation for Artificial Intelligence, 1996.
18. Larose, D. T., "Discovering Knowledge in Data: An Introduction to DataMining", ISBN 0-471-66657-2, ohn Wiley & Sons, Inc, 2005.

AUTHOR DETAILS

V. Maria Antoniate Martin is a Research Scholar in Computer Science at Bharathiar University, Coimbatore, Tamil Nadu, India. He is also working as an Assistant Professor in Department of Information Technology at St. Joseph's College, Tiruchirappalli, Tamil Nadu, India. He received his Bachelor of Science degree in Computer Science from Bharathidasan University in 2003; He completed his Masters in Science in Computer Science from the same University in 2006. He also completed his Masters in Philosophy in Computer Science from the same University in 2011.He has seven years of teaching experience. He has published seven research articles in reputed International Journals. He is also the co-author of a publication in a National Conference of importance. His area of research is Data Mining.

Dr. K. David is an Assistant Professor in the Department of Computer Science at H.H. the Rajah's College, Pudukkottai, TamilNadu, 622001. He has over fifteen years of teaching experience. He has published scores of papers in peer reviewed journals of National and International repute and is currently guiding seven Ph.D., scholars. His research interests include, UML, OOAD, Knowledge Management, Web Services and Software Engineering.

A.Paulin Jenifer is a student of M.Sc. Computer Science, St. Joseph's College, Trichy-620002. She received her Bachelor of Science degree in Computer Science from Madurai Kamarajar University in 2016.