RESEARCH PAPER | Computer Science | Volume : 5 | Issue : 8 | August 2015 | ISSN - 2249-555X

Securing Big Data in Privacy Preserving Data Mining

KEYWORDS: Data Mining, Privacy, Big Data

U.Indumathi
M.Phil Scholar, Department of Computer Science and Engineering, Alagappa University, Karaikudi

T.Meyyappan,Professor
Department of Computer Science and Engineering, Alagappa University, Karaikudi

ABSTRACT *Database mining can be defined as the process of mining for implicit, formerly unidentified, and potentially essential information from awfully huge databases by efficient knowledge discovery techniques. The privacy and security of user information have become significant public policy anxieties and these anxieties are receiving increased interest by the both public and government lawmaker and controller, privacy advocates, and the media. We propose a novel technique - for sharing private data securely among several parties an algorithm has been used. An anonymous ID assignment technique is used iteratively to assign the nodes with ID numbers ranging from 1 to N. This technique enhances data that are more complex to be shared securely. The nodes are assigned with the anonymous ID with the help of a central authority. We analysis the privacy in Big data with mining parameters.*

I INTRODUCTION

Security and Privacy protection have been a public policy concern for decades. However, rapid technological changes, the rapid growth of the internet and electronic commerce, and the development of more sophisticated methods of collecting, analyzing, and using personal information have made privacy a major public and government issues. The field of data mining is gaining significant 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 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 privacy that 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. Knowledge discovery is needed to make sense and use of data. 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. [1,2,3].

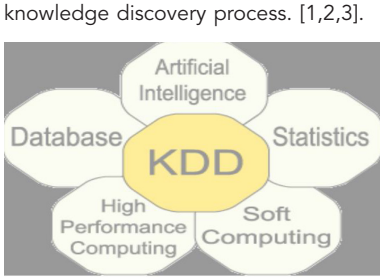

Fig: 1 Computing with Data base Structure

Usually, data mining e.g. data or knowledge discovery is the process of analyzing data from different perspectives and summarizing it into useful information 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.[4] 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.[5] 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. Privacy issues have attracted the attention of the media, politicians, government agencies, businesses, and privacy advocates

II DATA MING IN BIG DATA ENVINORNMENT

Data mining is an iterative and interactive process of discovering something innovative. The same as Novel- something we are not aware, Valid- generalize the future, Useful- some reaction is possible, Understandable- leading to insight, many step and process. Data mining is the process of discovering meaningful new correlations, patterns and trends by sifting through large amounts of data stored in repositories, using pattern recognition technologies as well as statistical and mathematical techniques." There are other definitions: Data mining is the analysis of (often large) observational data sets to find unsuspected relationships and to summarize the data in novel ways that are both understandable and useful to the data owner".[9]

Volume : 5 | Issue : 8 | August 2015 | ISSN - 2249-555X

Fig:2 Data mining blocks

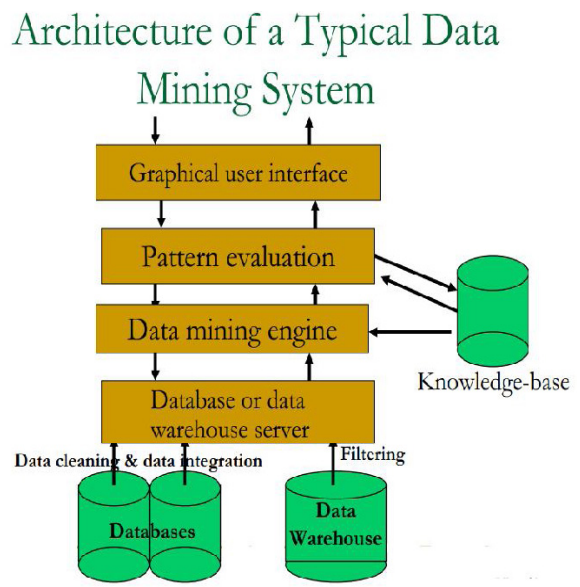

Fig: 3 Architecture

Data mining is an interdisciplinary field bringing together techniques from machine learning, pattern recognition, statistics, databases, and visualization to address the issue of information extraction from large data bases".[10] Evolution of database technology, data collection, database creation, IMS and Network DBMS, relational data model, Relational DBMS, advance database Models object oriented database, data collection centre, warehousing, multimedia database and recent web database needs to process the approach of data mining.

II RELATED WORKS

In existing system, the information about each node will be shared along with the data. It is usually encrypted along with the data. The usage of Newton's polynomial cannot be avoided as it increases the number of rounds of iterations that are used to compute the secure sum and power sum. Hence the performance of the system also decreases. It only focuses on the sum inputs whereas our project deals with the number of rounds. The usage of secure multiparty computation is being avoided with the usage of Sturm's theorem to make sure that the information about the nodes is not revealed. In the current system the main goal is to provide anonymous id for each node. Each node will have a secure communication of simple and complex data. Those data's may be from static data or dynamic data. By implementing secure sum hides permutations method and anonymous id assignment (AIDA) method the permutation methods are kept anonymous to each other. Hence, encoding technology is used here to create anonymous ID and the ID is being assigned to the user by the central authority and the receiver receives the data and decodes it with the key that is known only to the sender and the receiver which might not be known to the other semi honest node that might intrude.

Existing and new algorithms for assigning anonymous IDs are examined with respect to trade-offs between communication and computational requirements. Also, suppose that access to the database is strictly controlled, because data are used for certain experiments that need to be maintained confidential. Clearly, allowing Alice to directly read the contents of the tuple breaks the privacy of Bob; on the other hand, the confidentiality of the database managed by Alice is violated once Bob has access to the contents of the database. Thus, the problem is to check whether the database inserted with the tuple is still k-anonymous, without letting Alice and Bob know the contents of the tuple and the database respectively.

Disadvantage:

1. The database with the tuple data is not maintained confidentially.
2. The existing systems pave way for another person to easily access database.

Providing a secure and reliable data delivery over Internet is a challenging objective of the privacy preservation scheme and illustrates several presented ideas in the privacy design. More newly have employed cryptographic techniques to make data sharing with permissions in a common web services without revealing content to service providers (Baden et al., 2009). Another technique called One - Swarm (Isdal et al., 2010) that supports permissions as well as permitting users to distribute data widely without acknowledgment. A key characteristic of the One-Swarm design is that users have precise supportive power above the quantity of trust they put in peers and in the division model for their data: the similar data can be shared widely, secretly, or with admission power, with both trusted and untrusted peers. The scrutinizing of the privacy preservation is also being finished with the respective semantic policies for data sharing provides users much better privacy described in (Kagal and Pato, 2010). The privacy preserving can also be done in the form of document centric approach in dispersed environment. For document centric approach, users need to know about the location of relevant documents to access. The location of documents is identified through indexing facility discusssed by Zerr and Nejdl. Another useful tool for privacy preservation scheme is Support Vectors Machine (SVM) presented by Lin. SVM takes the data from training data set, discharging the data to SVM classifier for communal use to clients will reveal the confidential contented of support vectors.

The SVM for privacy preservation scheme violates the privacy preserving necessities for some authorized or viable reasons suggested by (Sun, 2010). A safe and privacy-preserving opportunistic framework, called SPOC (Lu et al., 2012), implemented in mobile Healthcare emergency for free distributed file sharing approach. With SPOC, elegant file transactions can be achieved to practice the computing-intensive Individual Health Information (PHI). In recent times, Vaidya and Clifton, have presented cryptographic techniques to facilitate data sharing using kth element over data set. Fong and Weber-Jahnke (2012), begins a privacy preserving approach with decision free learning, without associated thrashing of accuracy. But, conservation of the privacy for collected data samples has a chance of being vanished. The troubles of Privacy-Preserving are Addressed effectively with Duplicate Tuple Matching (PPDTM) (Sang et al., 2009a) and Privacy-Preserving Verge Attributes Matching (PPTAM). An analogous confront for privacy preservation is done with tuple matching (Sang et al., 2009b), endeavors to mask conscious participants. In BitTorrent

Volume : 5 | Issue : 8 | August 2015 | ISSN - 2249-555X

swarms (Zhang et al., 2010) achieved privacy of file sharing based on the data obtained with it. In this study, binary tree representation is presented to enhance the privacy preservation file sharing mechanisms.

III PROPOSE WORK WITH MODULES

AUTHENTICATION

The process of identifying an individual usually based on a userrname and password. In security systems, Authentication merely ensures that the individual is who he or she claims to be, but says nothing about the access rights of the individual.

I LOGIN:

In User and Admin login we are going to check whether the system is trusted machine or distrusted machine. If the machine is trusted, then the user or admin is allowed with **n** attempts. If the machine is distrusted machine then the user is allowed with single attempt. Process Involved is to Ø Check the login name and password Ø Then allows the authorized user to use these pages. Ø If the unauthorized user attempts to access user login then restrict that user and give the information.

A FORGET PASSWORD:

When the users forget their password then the user can access this forget password. It is used to create a new password. To ensure that user accessing forget password is a legitimate user, the user will be asked a question. These questions and their answers are created, while the user is registering to the site. If the user enters the answer then the entered text will be matched with the database. If the result is true, then the user will be allowed to enter the new password to access the site. If the result is false, user will not be allowed to enter the new password to access the site.

B REGISTRATION:

When a new user is creating an account he needs to register here by giving the sufficient information. Registration might also contain some private data that will be kept confidential so that the information about username and password is retrieved when it is forgotten.

2 ADMIN:

In this module when the admin attempts to login we need to find whether the machine is trusted or it is distrusted machine. It is found by user id and pass word. Admin can provide AID to all nodes. With the help of that AID each node can share the data in internet. Admin can generate that AID for individual nodes. So the sharable data can be kept in a sharable database. As a result, their own private data will be maintained secret.

A.GENERATE AID:

In this module admin wants to create the AID for individual nodes. Admin can make this unique AID for each user presented in network. With the help of this AID user can share his data and also he can keep his own private data as secret.

B.ASSIGN AID:

Admin can provide unique AID to all nodes. Nodes presented in network will be communicating by using AID. AID can not contain any private information. AID helps to keep personal information as more secret.

3. USERS

Users can login by entering the given username and password. Then he/she may go for corresponding page. User can keep his own information in a sharable data base. And also he can retrieve data shared by other users. User has to use his AID for sharing information.

A. SHARE DATA:

New user has to get AID from admin. Admin will assign that AID for every node. So data shared by user can be kept in a sharable database and it can be shared by all users. Each node will have unique AID and with the help of that unique AID any user can store and retrieve from sharable database.

B. RETRIEVE SHARED DATA:

In this module, user can retrieve the shared data. Shared data may be stored by him or by any other node. So it will be easy to make own private data as secret by implementing anonymous id algorithm.

IV RESULTS WITH FRAME WORK

Fig 6: Data Securing Form Design from Admin Panel

V CONCLUSION

We differentiate four different user roles that are commonly involved in data mining applications, i.e. data provider, data collector, data miner and decision maker. Each user role has its own privacy concerns; hence the privacy-preserving approaches adopted by one user role are generally different from those adopted by others. Other algorithms can also be used and compared for detailed study of the

Volume : 5 | Issue : 8 | August 2015 | ISSN - 2249-555X

behaviour of the system. Here encoding and decoding methodologies largely decrease the number of rounds required for ID assignment and thereby decreasing the overheads and enhancing the performance of the entire system. As all the nodes are assumed to be dishonest, using this algorithm helps to continue with the reliable data sharing using big data.

REFERENCE [1] 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. | [2] Dunham, M. H., Sridhar S., "Data Mining: Introductory and Advanced Topics", Pearson Education, New Delhi, ISBN: 81-7758-785-4, 1st Edition, 2006 | [3] Fayyad, U., Piatetsky-Shapiro, G., and Smyth P., "From Data Mining to Knowledge Discovery in Databases," AI Magazine, American Association for Artificial Intelligence, 1996. | [4] Larose, D. T., "Discovering Knowledge in Data: An Introduction to Data Mining", ISBN 0-471-66657-2, ohn Wiley & Sons, Inc, 2005. | [5] L. Getoor, C. P. Diehl. "Link mining: a survey", ACM SIGKDD,Explorations, vol. 7, pp. 3-12, 2005. | [6] Fayyad U.M., Piatetsky-Shapiro G., Smyth P. "From Data Mining toKDD : An Overview", AAAI/MIT Press, 1996. | [7] Han J. et Kamber M., "Data Mining: Concepts and Techniques",Morgan Kaufmann Publishers, Canada, 2002. | [8] 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 | [9] David Hand, Heikki Mannila, and Padhraic Smyth," Principles of Data Mining", MIT Press, Cambridge, MA, 2001. | [10] Peter Cabena, Pablo Hadjinian, Rolf Stadler, JaapVerhees, and Alessandro Zanasi, Discovering Data Mining: From Concept toImplementation, Prentice Hall, Upper Saddle River, NJ, 1998. | [11] Mafruz Zaman Ashrafi, David Taniar, Kate A. Smith, "Data Mining Architecture for Clustered Environments" , Proceeding PARA '02Proceedings of the 6th International Conference on Applied Parallel | Computing Advanced Scientific Computing, Pages 89-98, Springer-Verlag London, UK ©2002 | [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. |