A Review on Implementation Technique for Preserving Privacy in Big Data Mining

Sarangkumar S. Dubey, Prof. A. P. Thakare

Abstract— The term big data usually describe the exponential growth of data in every sector of the life for both structured and unstructured formatted data.To make use of this large amount of data, the data owners uses the different data mining techniques to extract knowledge at the same time they have to compromise the privacy of their client [4]. For this purpose, in this paper we have define the Privacy Preserving Big Data Mining in which we define the technique for extracting the useful information from the large set of Big Data and to protect it from accessing by unauthorized users we have define some security techniques. A number of methods and techniqueshave been developed for privacy preserving data mining. In this paper, we try to provide a complete review on PPDM for Big data anddifferent techniques such as data partition,distributed storageusing HadoopMapReduce technique using the HDFS distributed file system which could be effectively used to preventthe data access from unauthorized users. Our approach has become increasingly popular because itallows sharing of Sensitive data for analysis purposes. Several data mining algorithms along with incorporating privacypreserving mechanisms that have developed allows one to extract relevant knowledge from large amount of useful data. At the same time ithide sensitive data or information from disclosure or inference.

Index Terms— Privacy Preserving Data Mining (PPDM), Big Data, Hadoop Distributed File System (HDFS), Randomized Response (RR),Sequential Pattern Hiding, Authentication Authorization Accounting (AAA).

1 Introduction

WITH the rapid advancement in the technology and the use of internet increases the development and will create large databases, carrying large amount of information. The literature studies shows that availability of data is increase so rapidly that world's volume of data doubles every eight-teen months [1]. As the use of internet increases at the same time the threat against privacy is also increasing and it create serious problem. With this the privacy preserving data mining for the large amount of data called as Big Data is growing on increasing. The PPDM has two different considerations [3]: (i) modification of raw data such as attributes like id, name, address, age etc. in order for receiver not to compromise any person's privacy. (ii) By using data mining techniques, mine the sensitive knowledge from the databases without compromising the data privacy.

Data mining is a recently emerging field, which helps to connect the three different fields that is of Databases,Artificial Intelligence and Statistics. As in the today's age of information itgatherlarge volumes of data that becomes the Big Data. However, the data becomes useful if "meaningful or user intended information"or "knowledge" can be extracted from the large storage of the data. Data miningis called as knowledge discovery from Database[2] thatattempts to discover important knowledge from big data. In contrast to standard statistical methods, data mining techniques searchfor user interestedinformation without demanding any sort of priori hypotheses. As a field, it has introduced the newconcepts and algorithms such as association rule learning. It has also been applied to machine-learningalgorithms such as for inductive-rule learning, by decision trees to the setting where very large databasesare involved [4]. Data mining techniques are used in business, healthcare, research and many other are becoming more and morepopular with time.

We are using one of the framework of Hadoop [5] that is HDFS which is mainly useful for processing our big data. Hadoop provides a distributed file system and a framework for the analysis and performing transformation of very large data sets using the MapReduce paradigm. While the interface to HDFS provides the distributed storage of the different files available with us. An important characteristic of Hadoop is that, it performs the partitioning of data and computation across many number of hosts, and the execution of application computations should be in parallel and it should be close to their data. A Hadoop cluster scales computation capacity, storage capacity and I/O bandwidth by simply adding commodity servers. In this paper, we propose a cryptographic algorithm for preserving privacy of raw data. Now, we are in a data rich situation, if these available data are not analyzed or mined to gain knowledge then it will have no use. But while mining the important knowledge from the large datasets it is very important to preserve privacy of customers.

2 Big Data Mining

Big data are the huge volume of data in terabytes that cannot be handle easily or make the use of it easily. Big data also has some other important characteristic in addition along with volume, include variety, velocity and value [6].Big data include both thestructured and unstructured data which contains the multimedia format such as text, audio, video and website log files etc. This follows the real-time streams for analysis to maximizing the business value by making the deci-

- *Sarangkumar S. Dubey is currently pursuing masters degree program in computer science & engineering, SIPNA College of Engineering & Technology, Amravati ,India. Email: Sarangdubey97@gmail.com*
- *Prof. A. P. Thakare is working as an Associate Professo in SIPNA College of Engineering & Technology ,Amravati ,India. Email: apthakare40@rediffmail.com*

sion to real-time.It is always beneficial for companies to mining their big data stores without violating the clients' privacy.The process of mining big data can be viewed as a threat to the privacy preserving if our data get access by unauthorized parties. In data mining process, the algorithms require all participant in a distributed system should follow a privacy model and make assumption on behavior of participating node. This done more securely by using the HadoopMapReduce framework and the data is stored by partitioning in the distributed manner in the HDFS framework. With this we have developed a novel framework for privacy preserving data mining on big data, where all participants are divided into some category based on some conditions and after that will go for mining.

Now we are in data rich situation where data plays an important role and consider as "Gold". So security is the primary important thing to these data from unauthorized used. For any business or personal sensitive information, one has to take care of their customers and their personal details which if reveal it will create problems for its survival. Privacy of individuals is the primary important thing to be preserved within the organization. Data mining or other techniques are used on big data, but it will be taken care of these that will not been effect any individual.

3 Secure Mining Techniques

3.1 Randomized Response (RR)

Randomized Response (RR) techniques were developed inthe statistics community for the purpose of protecting individual'sprivacy and serves as the effective method in the PPDM. Randomized Response technique was first introduced by Warner [7] in 1965 as a technique to solve the following survey problem: to estimate the percentage of people in a population that has attributeA. This queries sent to a group of people, as the attribute A can be related to some confidential aspects of any human life, respondents may decide to give the randomly generated answer. For this the technique is not to reply at all or to reply with incorrect answers. The Two models namely Related-Question Model and Unrelated-Question Model have been proposed that solve this survey problem. The technique of randomization approach protects the customer's data by letting them arbitrarily alter theirrecords before sharing them, taking away some true information and introducing some noise content with it. Some methods ofrandomization are numerical randomization and item set randomization. Noise can be introduced either by adding ormultiplying random values to numerical records or by deleting real items and then adding "fake"values to the set of attributes.

3.2 Sequential Pattern Hiding

Sequence data are increasingly shared that enable mining different applications, related to various different domains such as marketing, telecommunications, and healthcare. This may exposes to sensitive sequential patterns, which may be lead to inferences about individuals or leak confidential and sensitive information about organizations. One of the technique of sequential pattern hiding helps to prevent this threat. Sequential pattern hiding method is necessary to conceal sensitive patterns that can possibly be extracted from published data, without critically affecting the data and the non-sensitive interesting patterns. [8] This approach hides sensitive patterns by replacing them with carefully selected the set of frequent non-sensitive patterns (side-effects). By this, it retains data utility in sequence mining and tasks based on item-set properties, the approach of sequential pattern hiding one can develop an efficient and effective algorithm for performing our task with minimal side-effects and distortion. This method also avoids implausible symbol orderings that may exist in certain applications and hide sensitive patterns from a sequence dataset. This method allows us significantly more accurate data analysis than the state-of-the-art approach. Sequential pattern hiding is seems to be a challenging problem, as sequences have more composite semantics than item sets, and calls for efficient solutions that offer high utility.

3.3 Statistical Anonymization Techniques

One of the main technique of our PPDM approach is anonymization which is process of removing or modifyingthe identifying variables contained in the micro datadataset [9]. An identifying variable can be said to one thatdescribes a characteristic of a person which is observable, for an attributes concern it is the registered (identification numbers, etc.), or in other it can be said to be known to other persons. There are mostly two types of identifiers of the personal data as direct identifiers and indirect identifiers. The direct identifiers,are those variables such as names,addresses, or identity card numbers, they permitdirect identification of a receiver but that are not neededfor statistical or research purposes, and should thus beremoved from the published dataset. Indirectidentifiers,have the characteristics that may beshared by several respondents, and by combining them it should lead to the re-identification of one of them. Forexample, the combination of variables such as residence, age, sex, and profession would be identifyingif only one individual of that particular sex, age andprofession lived in that particular district identifiers. Such direct variablesare needed mainly for statistical purposes, and should thus notbe removed from the published data files.

Anonymizingthe data consist of determining which variables are potential identifiers, and in modifying the level of precision variables to reduce the risk of re-identification toan acceptable level. In K-anonymity, it is difficult for an imposter to decide the identity of the individuals in collection of data set that has personal information.The challenge is to maximize thesecurity while minimizing the resulting information loss. This statistical anonymization technique serves most effective for performing PPDM with our big data.

3.4 Secure Multi-Party Computation

In the Distributed computing environment itconsiders the scenario where a number of distinct and some connected, computing devices or parties wish to carry out a joint computation of some function. The main goal of thesecure multiparty computation[10] is to enable parties to carry out distributed computing tasks in a manner with privacy preservation. The system of multi-party computation classically deals with questions of computing under the threat of machine crashes or some other inadvertent faults. The secure multiparty computation is concerned with the possibility of the malicious behavior by some adversarial entity. That is, it isassumed that a protocol execution may have to face with some attack by an external entity, or even by asubset of the participating parties. This attack is made with the view of learning the private informationor cause the result of the computation to be incorrect. It creates the requirement of two important things onany secure computation protocol asprivacy and correctness. The requirement of privacy statesthat nothing should be learned beyond what is absolutely necessary. In other ways, the involved partiesshould learn their output and nothing anything else. The requirement of correctnessis that each partyinvolves should receive its correct output. Therefore, the adversary must not be able to cause the resultof the computation for deviating from the function which the parties had set out to compute.

4 Supporting and Providing Privacy Preservation

The main objective of privacy preserving data mining (PPDM) is to provide security to data on which we are performing data mining. A key problem that arises in any of the mass collection of data is that of confidentiality. The need for privacy is sometimes very much important due to its storage of most important and confidential data that can be motivated by business interests. In the today's world, the key utility of large databases is research belongs to any field as scientific, economic, medical or market oriented. As we are dealing here with large amount of data that is big data we have to use the methods that reduces the risk of mishandling of the data [11]. And for dealing with the big data, there is novel method for providing the security to the data by partitioning and storing them separately. This mechanism can be done by using the technology of Hadoop, in which security is provided to the data by storing them in the HDFS file system in the distributed manner. In some other way, there are some technique of Authentication, Authorization and Accounting (AAA) are used for privacy preservation. Most of the techniques uses the cryptographic technique that perform some form of alteration on theoriginal data in order to attain the purpose of privacy preservation. The altered dataset is obtainable for mining and must meetprivacy requirements without losing the [12] benefit of mining.The brief description of all this techniques of privacy preservation is given below.

4.1 Hadoop Distributed File System (HDFS)

The Hadoop Distributed File System (HDFS) [13] is one of the module of the Hadoop framework designed to store very large data sets reliably.The HDFS also stream those data sets at high bandwidth with user applications. In a large cluster having vast amount of users, thousands of servers both host directly attached storage and execute user application tasks. By distributing storage and computation across many servers, the resource can grow with demand while remaining economical at every size.

The HDFS provides us the following important things:

•Provides awareness of a node's physical location, at the time of both allocating storage and scheduling tasks.

•MapReduceone of the most important feature moves compute processes to the data only on HDFS and works on the principle like minimal data motion. Any processing is performed on the physical node where the data resides.

•It reduces the network I/O patterns significantly and keeps most of the I/O on the local disk or within the same rack and provides very high aggregate read/write bandwidth.

•It properly balance file system according to the utility of data stored in it, and can rebalance the data on different nodes.

•The HDFS allows the operation of Rollback in case of human or system errors, which allows system operators to bring back the previous version of HDFS after an upgrade also.

•It has the Standby Name_Nodewhich provides redundancy and availability. Hadoop handles different types of cluster and hence this design allows a single operator to maintain a cluster of 1000s of nodes. That means it can stored large amount of data separately at one place.

4.2 The Security by Authentication, Authorization, and Accounting (AAA)

Authentication, Authorization, and Accounting (AAA) [14] is the process of identifying the identity of user, determining will the permissionsbe granted to that user or not, and keeping a log and eye on that user's activity. Authentication, Authorization, and Accounting, performs the work of providing user-level control of connections. Using this AAA technique, one can identify its users, determine their identity as the Memberships to their group. This information is used to implement access control policies that effectively control what peopleare allowed to do on the network for keeping a record of all the activities and transactions performed.

One of the basic functions of AAA technique is to provide control over network connections and resources. Itexercises the control on a user or group basis as it provides limiting access to resources that should beavailable only to the Executive team, restricting internet access to only certain groups, or implementing bandwidth controlbased on user or group membership. As the name suggests, there are three distinct components to AAA which perform three different task of -Authentication, Authorization, and Accounting. Authentication is the process of verifying, or determining, the identity of a user. In this phase, the server send a credential challenge in response to the user's initial request. The user responds to the challenge by sending credentials from the application interface, which passes them to the authentication server for verification. If the verification is successful the user identity is known and group and attribute information for that user can also be obtained. After authentication phase is completed, administrators can create user and group-basedpolicy by decision making process that allows administrators to have more granular control to manage access to actual content. In the Accounting process the information of access logs along with the transactions and activities performed by the user are accounted. Sometimes the component of AAA are required by enterprises to comply with local or international laws.

4.3 The Cryptographic Approach

In PPDM, cryptography is serves as the important tools. In this technique, the information is protected by transforming it or encrypting it which is not readable by humans called as cipher text. Only those users who have a secret key can decrypt the encrypted message or data into plain text. As the electronic communication become more prevalent it serves as increasingly important need. Cryptography is used to protect most important and personal information as e-mail messages, credit card information and corporate data [15]. Cryptography system can be classified into two main categories as symmetric-key system and public-key or asymmetric-key system. In the symmetric-key system, one single key is used by both sender and receiver. And in public-key system, as the name suggest the public key is known to everyone and private key to only the recipient users. For our purpose the asymmetric cryptography technique is most suitable for encrypting the customer's information which havegreat importance from different application point of view. Asymmetric key cryptography is a class of cryptographic algorithms,which require two different key-one is secret or private and other is public key according to their different uses. The public key is used to encrypt the plaintext and private key is used to decrypt the cipher text to plain text. The public key may be published without compromising security but private key used only after authentication and authorization phase only by the authorized people.

REFERENCES

[1] Yehuda Lindell, Benny Pinkas, “Privacy Preserving Data Mining”, *Weizmann Institute of Science and the Hebrew University of Jerusalem.*

[2] Gartner, Post event brief, Gartner IT Infrastructure, Operations and Management Summit 2009, Orlando, FL. available at www.gartner.com June 23–25 2009.

[3] IDC, Digital data to double every 18 months, worldwide marketplace model and forecast, Framingham, MA. May 2009. [Online] available: www.idc.com

[4] ArieFriedman,”Privacy preserving data mining”pp.4, January 2011.

[5] NasrinIrshadHussain, BharadwajChoudhury, SandipRakshit, “A Novel Method for Preserving Privacy in Big-Data Mining”, *International Journal of Computer Applications (0975 –8887)*, Volume 103–No16, October 2014.

[6] NasrinIrshadHussain, BharadwajChoudhury, SandipRakshit, “A Novel Method for Preserving Privacy in Big-Data Mining”, *International Journal of Computer Applications (0975 – 8887)*. Volume 103 – No 16, October 2014.

[7] Wenliang Du and Zhijun Zhan, “Using Randomized Response Techniques for Privacy-Preserving Data Mining”, *SIGKDD* ’03, August 24-27, 2003, Washington, DC, USA.

[8] ArisGkoulalas-Divanis, & GrigoriosLoukides, “Revisiting Sequential Pattern Hiding to Enhance Utility”, *ACM*, August 2011.

[9] Yehuda Lindell, Benny Pinkas, “Secure Multiparty Computation for Privacy-Preserving Data Mining”, *The Journal of Privacy and Confidentiality* 1, Number 1(2009).

[10] K.Anbazhagan, Dr.R.Sugumar, M.Mahendran, R.Natarajan, “An Efficient Approach for Statistical Anonymization Techniques for Privacy Preserving Data Mining”, *International Journal of Advanced Research in Computer and Communication Engineering*, Vol. 1, Issue 7, September 2012.

[11] Pingshui WANG,” Survey on Privacy Preserving Data Mining”, *International Journal of Digital Content Technology and its Applications,*Volume 4, Number 9, December 2010.

[12] TamannaKachwala, Dr. L. K. Sharma, “A Literature analysis on Privacy Preserving Data Mining” , *International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCE),* Vol. 3, Issue 4, April 2015.

[13] 20 essential Hadoop tools for crunching Big Data [Online] available:http://bigdata-madesimple.com/20-essential-hadoop-tools-for-crunching-big-data/

[14] Technology Primer: Authentication, Authorization, and Accounting [Online] available: www.bluecoat.com/ technology-primer-:-authentication-authorization-and-accounting.

[15] B. Pinkas, “Cryptographic techniques for privacy-preserving data mining”, *SIGKDD Explore*, 2002, 4(2): 12-19