Research Journal of Applied Sciences 11 (5): 215-220, 2016
ISSN: 1815-932X

To Reduce Data Leakage in Horizontally Distributed Database Using Association Rules

[1]E. Gokulakannan and [2]K. Venkatachalapathy
[1]Department of Computer Science and Engineering, Annamalai University,
Annamalai Nagar 608002, MRK Institute of Technology, 608301 Kattumannarkoil,
[2]Department of Computer Science and Engineering, Annamalai University,
608 002 Annamalai Nagar, Tamil Nadu, India

Abstract: Data mining is used to extract important knowledge from large datasets but sometimes these datasets are split among various parties. Association rule mining is one of the data mining technique used in distributed databases. This technique disclose some interesting relationship between locally large and globally large item sets and proposes an algorithm, fast distributed mining of association rules (FDM) which is an unsecured distributed version of the Apriori algorithm used to generates a small number of candidate sets and substantially reduces the number of messages to be passed at mining association rules. The main ingredient in proposed protocol is two novel secure multi party algorithm-one that computes the union of private subsets that each of the interacting player holds and another that test the inclusion of an element held by one player in a subset held by another. This protocol offers enhanced privacy with respect to the protocol. In addition, it is simpler and significantly more efficient in terms of communication rounds, communication cost and computational cost.

Key words: Association rules, privacy preserving data mining, distributed database, anonymous ID assignment, cost

INTRODUCTION

There are several sites (or players) that hold homogeneous databases, i.e., databases that share the same schema but hold information on different entities (Kantarcioglu and Clifton, 2004). The goal is to find all association rules with support at least s and confidence at least c, for some given minimal support size s and confidence level c that hold in the unified database, while minimizing the information disclosed about the private databases held by those players. The information that we would like to protect in this context is not only individual transactions in the different databases but also more global information such as what association rules are supported locally in each of those databases (Patidar *et al*., 2014; Mathews and Manju, 2014).

We propose a protocol for secure mining of association rules in horizontally distributed databases. Our protocol is based on the Fast Distributed Mining (FDM) algorithm (Tassa, 2014). The main ingredients in our protocol are two novel secure multi-party algorithms one that computes the union of private subsets that each of the interacting players hold and another that tests the inclusion of an element held by one player in a subset held by another. Our protocol offers enhanced privacy. In addition, it is simpler and is significantly more efficient in terms of communication rounds, communication cost and computational cost. Data mining is defined as the method for extracting hidden predictive information from large distributed databases. It is new technology which has emerged as a means of identifying patterns and trends from large quantities of data.

The final product of this process being the knowledge, meaning the significant information provided by the unknown elements (Kantarcioglu and Clifton, 2004). This paper study the problem of association rules mining in horizontally distributed databases. In the distributed databases, there are several players that hold homogeneous databases which share the same schema but hold information on different entities. The goal is to find all association rule with support s and confidence c to minimize the information disclosed about the private databases held by those players.

Kantarcioglu and clifton studied the problem where more suitable security definitions that allow parties to choose their desired level of security are needed, to allow

Corresponding Author: E. Gokulakannan, Department of Computer Science and Engineering, Annamalai University, Annamalai Nagar 608 002, MRK Institute of Technology, 608301 Kattumannarkoil, Tamil Nadu, India

effective solutions that maintain the desired security (Kantarcioglu and Clifton, 2004). So, they devised a protocol for its solution. The main part of that protocol is sub protocol for secure computation of the union of private subsets that are held by the different players. It makes the protocol costly and its implementation depends upon encryption primitive's methods, oblivious transfer and hash function also the leakage of information renders the protocol not perfectly secure (Tassa, 2014).

This study proposed an algorithm, PPFDM, privacy preserving fast distributed mining algorithm for horizontally distributed data sets and find interesting association or correlation relationships among a large set of data items and to incorporate cryptographic techniques to minimize the information which is going to shared with others, while adding little overhead to the mining task (Tassa, 2014). In the proposed scheme, the inputs are the partial databases and the required output is the list of association rules that hold in the unified database with support and confidence no smaller than the given thresholds s and c, respectively. The information that would like to protect in this study is not only individual transaction in the different databases but also more global or public information such as what association rules are supported locally in each of those databases. The proposed protocol improves upon that in (Kantarcioglu and Clifton, 2004) in terms of simplicity and efficiency as well as privacy.

Literature review: The problem of secure mining of association rules in horizontally partitioned databases. Tamir Tassa proposed here a protocol Fast Distributed Mining algorithm (FDM) for mining of association rules in horizontally distributed databases studied by Tassa (2014). The main idea is that the players finds their locally s-frequent itemsets then the players check each of them to find out globally s-frequent item set. Study assumes that the players are semi honest; they try to extract information. Hence, the player compute the encryption of their private database together by applying commutative encryption .Study shows that their protocol offers better privacy and is significantly more efficient in terms of communication cost and computational cost while the solution is still not perfectly secure cause it leaks excess information.

A very effective way to deal with multiple data sources is to mine the association rules at those sources. The study suggests Elliptic Curve based Digital Signature Algorithm (ECDSA) and Elliptic Curve Integrated Encryption Scheme (ECIES). The algorithm provides privacy against involving parties in distributed environment where first owner encrypts random number and sign then send it to another owner. As the message is passing in encrypted form so the owners cannot read the communication channel. The objective of these algorithms is to find association rules with minimum iterations and consume less time. For the small amount of databases the communication and computational cost are reasonable (Patidar *et al*., 2013).

Mathews and Manju (2014) proposed privacy preserving data mining using extended distributed rk secure sum protocol in combination with apriori algorithm where firstly mining of frequent items from individual parties is done with apriori algorithm then applied extended distributed Rk-secure sum protocol to obtain global result. Apriori Algorithm follows bottom up strategy to find frequent item sets. The distributed RK-secure sum protocol is secure multi party computation protocol, holds frequent itemsets globally without affecting privacy. Where the parties p1, p2, ... pn are arranged in bus network. p1 is protocol initiator and pn is last party. If two parties join together the network then possibility that they can know each other's data. So to reduce this drawback extended distributed Rk-secure sum protocol is used which is also a secure multi party computation protocol. The proposed algorithm provides more security and privacy but the complexity of protocol is high.

Mathews and Manju (2014), the study also uses uses apriori algorithm for generating association rules and playfair cipher technique is used to transfer that generated rules (Varma *et al*., 2014). This study defines two parts of association rule; Antecedent, is the item found in database and consequent, found in combination with the first. Unlike, all cipher technique, playfair cipher encrypts pair of letters. This technique uses a 5 by 5 table containing a key word. Firstly, table have to fill up with key word and remaining spaces with remaining spaces removing the duplicate letters. I and J are written in one column. encrypts pair of letters.

Rautaray and Kumar (2013) association rules are generated and global frequent item sets in distributed environment if found with the help of FP tree. FP tree is a compact data structure. It finds frequent item set without generating candidate item set by traversing frequent item set through FP tree. This study also provide privacy to the databases with Data Encryption Standard (DES). In DES two keys are used, first party encrypts dataset with key 1 and this encrypted data is again encrypted with key 2. The receiving party decrypts data with key 2 first then key1. This is also called as double encryption and it provides higher security to databases than other cryptographic technique. This study shows that global frequent item set is found with minimal communication and time complexity with zero percentage of data leakage. But, this is applicable for homogeneous databases.

Narang *et al.* (2013) deals with the problems of association rule mining. The problems can de divided as data hiding and knowledge hiding. data hiding is defined as the trial of removing confidential or private information from the data before its disclosure. Knowledge hiding, on the other hand, concerns the information or else the knowledge that a data mining method may discover after having analyzed the data. This study reviews the methods of privacy preserving and proposed an improvement of sensitive rule hiding to make it more accurate and secured. The Secure Multiparty Computation (SMC) is used to find global support and confidence without data leakage. To provide privacy to the database Tiny Encryption Algorithm (TEA) is used.

Apriori algorithm is used to generate candidate itemsets. Firstly, it scan the database for pruning and thus concluded that candidate itemsets is frequent itemset. But Apriori algorithm cannot meet their needs over large databases. This algorithm is improved and put forward Sampling algorithm (Liu and Sang, 2013). Sampling algorithm is used to sample the data form original databases and find frequent item sets by reducing mining time. This study also studied about sampling HT algorithm. This algorithm is the combination of sampling method and hash table technology. In this algorithm, firstly sample size and negative border is calculated then frequent 1-itemset is generated by Hash table and frequent 2-itemset is generated directly. Now to the generate candidate 2-itemset, Negative Border pruned it to frequent 2-itemset according to the minimum support. It results into reduction of running time of this algorithm.

The outsourcing of data and computing services is acquiring a novel relevance which is expected to skyrocket in the near future. The main prolem which can breaks the security is that that the server has access to valuable data of the owner and may learn sensitive information from it. Both the transactions and the mined data are the property of the data owner and should remain safe and private to him only. This problem of protecting important private information of organizations/companies is referred to as corporate privacy. Giannotti *et al.* (2013) studied this problem of outsourcing the association rule mining. This study also proposed an attack model based on background knowledge of all participants and devise a scheme for privacy preserving outsourced mining. The Rob Frugal encryption scheme is used in this proposed scheme which is based on 1-1 substitution ciphers for items and adds fake transactions to make each cipher item share the same frequency. This study proved that the proposed technique is effective, scalable and protect privacy. And also robust against an adversarial attack based on the original items and their exact support.

Asthana *et al.* (2013) studied about the five algorithms. Apriori algorithm, MSApriori, MCISI algorithm, aprori with systematic rules and HMT. Apriori algorithm generates the candidate item set and eliminates those candidates which are less than user support level. The MSApriori (minimum support apriori) method specifies the minimum support of the item and provide different minimum item support values for different items. The MCISI algorithm is used to find many imperfectly sporadic rule and also sporadic item sets. Systematic rules are also proposed in this study where user is restricted to specify minimum support value to find frequent item sets and timing algorithm is also used to save time with scanning of the entire transactional database. The Hash Mapping Table (HMT) is used to compress the given data sets. Result of survey is the time of support is more than time of compressing.

Vijay and coauthors also provides survey of association rule based techniques for privacy preserving where it studied on three methods, i.e. heuristic-based technique, Cryptography-based techniques and Reconstruction-based techniques. A heuristic-based technique depends on adaptive modification which modifies only selected values that minimize the utility loss with the help of centralized data perturbation.

Based association rule confusion and centralized data blocking-based association rule confusion. Cryptography-based techniques used for vertically partitioned data as well as horizontally partitioned data. This technique is depend on secure multiparty computation where we can say that a computation is secure if at the end of the computation, no database owner knows anything except its own input and the results. The last technique is reconstruction-based technique which on used for numerical data and Binary and Categorical data. This technique worked on the issue of of privacy preservation by perturbing the data. Reconstruction-based techniques are constructed original distribution of the data from the randomized data.

Now a day, privacy preserving for the data and the owner is increasingly becoming a problem in case of distributed server sharing. The solutions are exists but for central server model which is computationally expensive and because of low data security and huge bandwidth tradeoff it is not useful for distributed server model. Focuses on distributed model assigning IDs to nodes (user) which are anonymous. Each node chooses random values with the help of Anonymous ID Assignment (AIDA) algorithm. These IDs can used for sharing communication bandwidth as it uses network setup where number of clients can register and shares data and also for data storage. The advantage of this algorithm is at the transaction, no id being visible to any group member or person. AIDA is not a cryptographic algorithm hence it saves memory space. This study shows that the privacy preserving with the help of anonymous ID assignment is successful. (Kalaivani and coauthors and Shiny and Gayathri addresses an algorithm to share the simple integer data, it allows the secure sum to be collected with the guarantees of anonymity. This study addresses the complexities of the secure multiparty computation. The Anonymous IDs are used in sensor networks to secure the individual nodes.

Distributed database: A distributed database is database in which storage devices are not all attached to a common processing unit such as the CPU, controlled by a distributed database management system (together sometimes called a distributed database system). It may be stored in multiple computers, located in the same physical location; or may be dispersed over a network of interconnected computers. Unlike parallel systems, in which the processors are tightly coupled and constitute a single database system, a distributed database system consists of loosely-coupled sites that share no physical components. System administrators can distribute collections of data (e.g., in a database) across multiple physical locations. Two processes ensure that the distributed databases remain up-to-date and current: replication and duplication. Replication involves using specialized software that looks for changes in the distributive database. Once, the changes have been identified, the replication process makes all the databases look the same. The replication process can be complex and time-consuming depending on the size and number of the distributed database. This process also requires lot of time and computer resources. Duplication, on the other hand has less complexity. It basically identifies one database as a master and then duplicates that database. The duplication process is normally done at a set time after hours. This is to ensure that each distributed location has the same data. In the duplication process, users may change only the master database. This ensures that local data will not be overwritten. Both replication and duplication can keep the data current in all distributive locations.

MATERIALS AND METHODS

The study propose an alternative protocol Privacy Preserving Fast Distributed Mining (PPFDM) for the secure computation of the union of private subsets. This protocol improves upon, in terms of simplicity and efficiency as well as privacy. In particular, this protocol does not depend on commutative encryption and oblivious transfer. While solution is still not perfectly secure because it leaks the excess information which results in small number of possible coalitions, unlike that protocol which discloses information also to some single players. The PPFDM research better than (Tassa, 2014; Kantarcioglu and Clifton, 2004) in terms of privacy and does not leak excess information through communication channel.

The protocol that proposed here computes a parameterized family of functions which is called as threshold functions, in which the two cases correspond to the problems of computing the union and intersection of private subsets. Those are in fact is a general-purpose protocols that can be used in other contexts as well. The another problem of secure multiparty computation that this study tried to solve here is the set inclusion problem; the problem where Alice holds a private subset of some ground set and Bob holds an element in the ground set and they desire to determine whether Bob's element is within Alice's subset, without revealing the information about the other party's input to either of them.The Privacy preserving fast distributed mining (PPFDM) algorithm is a combination of Fast Distributed mining algorithm (FDM) and anonymous ID assignment (AIDA). FDM is an unsecured distributed version of the Apriori algorithm and AIDA is used for security of the databases. The Privacy preserving fast distributed mining (PPFDM) protocol involves following steps (Fig. 1).

Synthetic database generation: The generation of synthetic transactions is to evaluate the performance of the algorithms over a large range of data characteristics. The creation of synthetic data is an involved process of data anonymization; that is to say that synthetic data is a subset of anonymized data. This data is used in a variety of fields as a filter for information that would otherwise compromise the confidentiality of particular aspects of the data. Researchers, engineers and software developers used to test against a safe data set without affecting or even accessing the original data, insulating them from privacy and security concerns as well as letting them generate larger data sets than would be available using only real data. These transactions mimic the transactions in the retailing environment. Our model of the real world is that people tend to buy sets of items together. Each such set is potentially a maximal large item sets. A transaction may contain more than one large itemsets. Transaction sizes are typically clustered around a mean and a few transactions have many items. To create a dataset, our synthetic data generation program takes the parameters shown in Table 1.

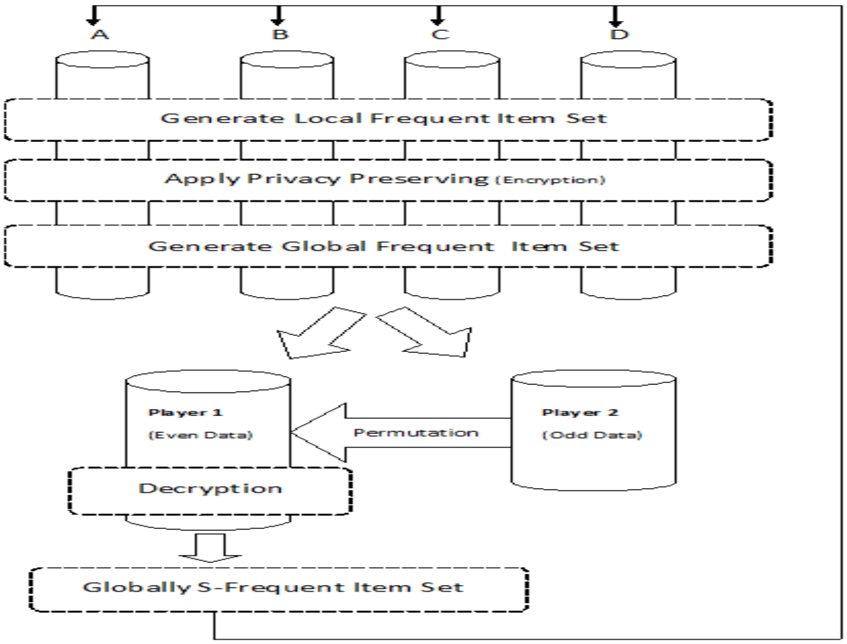
A
B
C
D
Generate Local Frequent Item Set
Apply Privacy Preserving (Encryption)
Generate Global Frequent Item Set
Player 1
(Even Data)
Permutation
Player 2
(Odd Data)
Decryption
Globally S-Frequent Item Set

Fig. 1: Architecture of PPFDM (privacy preserving fast distributed data mining)

Table 1: Parameters of interpretations

Parameters	Interpretations
N	Number of transactions in the whole database
L	Number of items
At	Transaction average size
Af	Average size of maximal potentially large itemsets
Nf	Number of maximal potentially large itemsets
CS	Clustering size
PS	Pool size
Cor	Correlation level
MF	Multiplying factor

RESULTS AND DISCUSSION

Apriori algorithm: The Apriori algorithm proposed to finds frequent items in a given data set. The name of Apriori is based on the fact that the algorithm uses a prior knowledge of frequent itemset properties. The purpose of the Apriori algorithm is to find associations between different sets of data. Each set of data has a number of items and is called a transaction. The first pass of this algorithm simply counts item occurrences to determines the frequent itemsets. A subsequent pass, K, consist of two phases. First, the frequent itemsets L_{K-1} found in $(K-1)^{th}$ pass are used to generate the candidate itemset C_K using the apriori candidate generation procedure. Next, the database is scanned and the support of candidate in C_K is counted. For last counting, we need to effciently determine the candidate in C_K contained in given transaction t. The set of candidate itemset is subjected to a pruning process to ensure that all the subsets of the candidate sets are already known to be frequent itemsets. The output of Apriori is sets of rules that tell us how often items are contained in sets of data.

Privacy preserving data mining: Privacy preserving data mining is defined as preserving the individual privacy and retaining the information in dataset to be released for mining. Analyzed the privacy offered by protocol UNIFI-KC. That protocol does not respect perfect privacy since it leaks player's information. This study used Anonymous ID Assignment (AIDA) for preserving privacy to the player's database. Currently, there are so many applications that require dynamic unique IDs. Such IDs can be used for data storage, sharing data and other resources anonymously and without conflict. In AIDA, random integers between 1 and S are chosen by each node (Dunning and Kresman *et al*., 2013).

Algorithm A:
Given nodes, n1, n2 ... nN uses distributed computation to find an anonymous indexing permutation. s: $\{1, ..., N\}\rightarrow\{1, ..., N\}$

Set the number of assigned nodes A = 0

Each unassigned node ni chooses a random number ri in the range 1 to S. A node assigned in a previous round chooses ri = 0

The random numbers are shared anonymously. Denote the shared values by q1, q2, ... qN

Let q1 ... qk denote a revised list of shared values with duplicated and zero values entirely removed where k is the number of unique random values. The nodes ni which drew unique random numbers then determine their index si from the position of their random number in the revised list as it would appear after being sorted:

```
si = A+Card {qj: qj< =ri}
    Update the number of nodes assigned: A = A+k
    If A<N then return to step (2)
```

Association rules: The association rule mining problem was formulated by Agrawal in 1993 and is often referred to as market-basket problem. In this problem, set of items is given and large collection of transaction is occurred which are subsets of these items. The task is to find relationship between the presence of various items within these baskets.

CONCLUSION

This system proposed a protocol for secure mining of association rules in horizontally distributed databases that improves significantly in terms of privacy and efficiency. One of the main ingredients in this proposed protocol is a novel secure multi-party protocol for computing the union of private subsets that each of the interacting players hold. Another ingredient is a protocol that tests the inclusion of an element held by one player in a subset held by another. Those protocols exploit the fact that the underlying problem is of interest only when the number of players is greater than two.

REFERENCES

Asthana, P., A. Singh and D. Singh, 2013. A survey on association rule mining using apriori based algorithm and hash based methods. Int. J. Adv. Res. Comput. Sci. Software Eng., 3: 599-603.

Dunning, L.A. and R. Kresman, 2013. Privacy preserving data sharing with anonymous ID assignment. Inf. Forensics Secur. IEEE. Trans., 8: 402-413.

Giannotti, F., L.V. Lakshmanan, A. Monreale, D. Pedreschi and H. Wang, 2013. Privacy-preserving mining of association rules from outsourced transaction databases. IEEE. Syst. J., 7: 385-395.

Kantarcioglu, M. and C. Clifton, 2004. Privacy-preserving distributed mining of association rules on horizontally partitioned data. IEEE Trans. Knowledge Data Eng., 16: 1026-1037.

Liu, Z. and T.S.G. Sang, 2013. An algorithm of association rules mining in large databases based on sampling. Int. J. Database Theory Appl., 6: 95-104.

Mathews, M.T. and E.V. Manju, 2014. Extended distributed RK-Secure sum protocol in apriori algorithm for privacy preserving. Int. J. Res. Eng. Adv. Technol., 2: 1-5.

Narang, G., A. Shaikh, A. Sonawane, K. Shegar and M. Andhale, 2013. Preservation of privacy in mining using association rule technique. Int. J. Sci. Technol. Res., Vol. 2,

Patidar, V.K., A. Raghuvanshi and V. Shrivastava, 2013. Literature survey of association rule based techniques for preserving privacy. Compusoft, 2: 59-59.

Rautaray, J. and R. Kumar, 2013. Privacy preserving in distributed database using data encryption standard. Int. J. Innovative Res. Sci. Eng. Technol., Vol. 2,

Tassa, T., 2014. Secure mining of association rules in horizontally distributed databases. IEEE. Trans. Knowl. Data Eng., 26: 970-983.

Varma, P.J., M. Amruthaseshadri, M. Priyanka, A. Kumar and B.L.V. Bharadwaj, 2014. Association rule mining with security based on playfair cipher technique. Int. J. Comput. Sci. Inf. Technol., 5: 744-747.