ISSN:0975-9646

Manali Rajeev Raut et al, / (IJCSIT) International Journal of Computer Science and Information Technologies, Vol. 5 (6) , 2014, 7540-7544

Association Rule Mining in Horizontally Distributed Databases

Ms. Manali Rajeev Raut
Mtech CSE (III sem)
Dept. of Computer Science and Engineering
G.H. Raisoni Institute of Engineering and Technology for Women, Nagpur
RTMNU, Nagpur

Ms. Hemlata Dakhore
Assistant Professor
Dept. of Computer Science and Engineering
G.H. Raisoni Institute of Engineering and Technology for Women, Nagpur
RTMNU, Nagpur

***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 are 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.**

Index Terms— association rules , Privacy preserving data mining, distributed database, anonymous ID assignment.

I. INTRODUCTION

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 [2]. 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 [1].

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 effective solutions that maintain the desired security [2]. 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 [1].

This paper 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 [1]. 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 paper 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 [2] in terms of simplicity and efficiency as well as privacy.

II. LITERATURE REVIEW

The paper [1] studied 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. 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. Paper 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 .Paper 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 [1].

A very effective way to deal with multiple data Sources is to mine the association rules at those sources. The paper [3] 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 [3].

The paper [4] 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.

Like paper [4], the paper [5] also uses uses apriori algorithm for generating association rules and playfair cipher technique is used to transfer that generated rules. This paper 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 keyword. firstly, table have to fill up with keyword and remaining spaces with remaining spaces removing the duplicate letters. I and J are written in one column. encrypts pair of letters.

In this paper [6], 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 paper 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 paper 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.

The paper [7] 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 paper 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. Hence in the paper [8], this algorithm is improved and put forward Sampling algorithm. Sampling algorithm is used to sample the data form original databases and find frequent item sets by reducing mining time. This paper also studied about SamplingHT 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. The paper [14] studied this problem of outsourcing the association rule mining. This paper also proposed an attack model based on background knowledge of all participants and devise a scheme for privacy preserving outsourced mining. The RobFrugal 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 paper 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.

The survey paper [9] 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 paper 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.

The paper [10] 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. Hence the paper [12] 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 paper shows that the privacy preserving with the help of anonymous ID assignment is successful. Like paper [12], the paper [13] addresses an algorithm to share the simple integer data, it allows the secure sum to be collected with the guarantees of anonymity. This paper addresses the complexities of the secure multiparty computation. The Anonymous IDs are used in sensor networks to secure the individual nodes.

III. PROPOSED METHODOLOGY

The paper 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 [2], in terms of simplicity and efficiency as well as privacy. In particular, this protocol does not depend on commutative encryption and oblivious transfer. While [1] 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 [2]. The PPFDM works better than [1] and [2] 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 [1] that this paper 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.

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.

Parameters	Interpretations
N	Number of transactions in the whole database
L	Number of items
A_t	Transaction average size
A_f	Average size of maximal potentially large itemsets
N_f	Number of maximal potentially large itemsets
CS	Clustering size
PS	Pool size
C_{or}	Correlation level
MF	Multiplying factor

2 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.

3. 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. The paper [1] analyzed the privacy offered by protocol UNIFI-KC. That protocol does not respect perfect privacy since it leaks player's information. This paper 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 [11].

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

1) Set the number of assigned nodes A=0.

2) 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 .

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

4) 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}

5) Update the number of nodes assigned: A=A+k.

6) If A<N then return to step (2).

4. 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. Association rule mining is to find out association rules that satisfy the predefined minimum support and confidence from given database. The problem is usually decomposed into two sub problems. One is to find those itemsets whose occurrences exceed a predefined threshold in the database; those item sets are called frequent or large itemsets with the constraint of minimal confidence.

Let I={I1,I2,….Im } be a set of items. Let D is the task relevant data and a set of database transaction where each transaction T is a set of items such that $T \subseteq I$. Each transaction is associated with an identifier, called TID. Let A be the set of items. A transaction T is contained A if and only if $A \subseteq T$. An association rule is an Implication of the form $A \Longrightarrow B$, where $A \subset I, B \subset I$ and $A \cap B = \emptyset$. The rule $A \Longrightarrow B$ holds the transaction set D with the help of support s, where s is called as the percentage of transaction in D that contains $A \cup B$. This is taken to be the probability, $P(A \cup B)$. The rule $A \Longrightarrow B$ has confidence c in the transaction set D, where c is called as the percentage of transaction in D containing A that also contain B. This is the conditional probability, p(B|A). That is,

$$\text{Support } (A \Longrightarrow B) = P(A \cup B)$$
$$\text{Confidence } (A \Longrightarrow B) = p(B|A)$$

In general, association rule mining can be viewed as a two step process:

1. Find all frequent itemsets : Here, each of this item sets will occur at least as frequently as a predetermined minimum support count, min_sup.
2. Generate small association rules from the frequent item sets : In this step, these rules mast satisfy minimum support and minimum confidence.

5 The Fast Distributed Mining Algorithm

This paper is based on the Privacy Preserving Fast Distributed Mining algorithm (PPFDM) which is a combination of preserving privacy and Fast Distributed mining algorithm which is an unsecured distributed version of the Apriori algorithm. Its main idea is that any s-frequent item set must be also locally s-frequent in at least one of the sites. Hence, in order to find all globally s-frequent item sets, each player reveals his locally s-frequent item sets and then the players check each of them to see if they are s-frequent also globally. The FDM algorithm proceeds as follows:

1. Candidate Sets Generation: Each player p_m computes the set of all (k-1) item sets, L_{K-1} that are locally frequent and also globally frequent. The intuition behind the candidate set generation is that if an itemset X has minimum support, so do all subsets of X. Hence the player then applies set the Apriori algorithm on L_{K-1} in order to generate the set of candidate k-item sets.
2. Local Pruning: The pruning step eliminates the extension of (K-1) itemsets which are not found to be frequent. Here, player p_m computes suppm(X). He the retains only those item sets that are locally s-frequent. We denote this collection of item set by $C_s^{k,m}$.
3. Computing local supports: All players compute the local supports of all item sets in $C_s^{k,m}$.
4. Broadcast mining results: Each player broadcasts the local supports that he computed. From that, everyone can compute the global support of every item set in $C_s^{k,m}$. Finally, F_s^k is the subset of $C_s^{k,m}$ that consists of all globally s-frequent k-item sets.

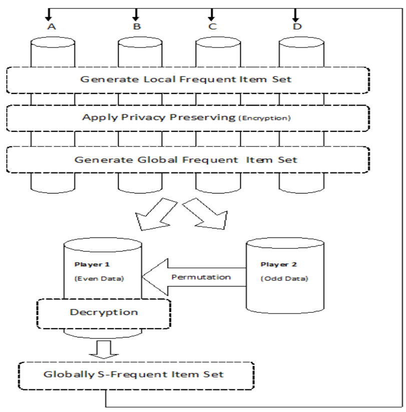

Fig: Architecture Of PPFDM (Privacy Preserving Fast Distributed Data Mining)

Above architecture of privacy preserving Fast distributed data mining (PPFDM) shows all steps discussed above. Cryptographic tools can enable data mining that would otherwise be prevented due to security concerns. The computing time in such a protocol is expected to dominate the cost of the secure computation of the union of all locally s-frequent item sets. Hence, the enhanced security offered by such a protocol is accompanied by increased implementation costs. Higher support thresholds entail smaller computation and communication costs since the number of frequent item sets decreases.

IV. 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

[1] Tamir Tassa,"Secure mining of association rule in horizontally distributed databases" ,IEEE trans. Konwledege and Data Engg. ,Vol. 26, no.2, April 2014

[2] M. Kantarcioglu and C. Clifton, "Privacy-Preserving Distributed Mining of Association Rules on Horizontally Partitioned Data," IEEE Trans. Knowledge and Data Eng., vol. 16, no. 9, pp. 1026-1037, Sept. 2004.

[3] Krishna Pratap Rao, Adesh chaudhary, Prashant johri "Elliptic Curve Cryptography Based Algorithm for Privacy Preserving in Data Mining", International Journal for research in Applied Science and Engineering Technology (IJRASET) ,Vol. 2 Issue V, May 2014

[4] Meera Treesa Mathews, Manju E.V," Extended Distributed RK-Secure Sum Protocol in Apriori Algorithm for Privacy Preserving", International Journal of Engineering and Advanced Technology (IJEAT), Volume-3, Issue-4, April 2014

[5] P. Jagannadha Varma, Amruthaseshadri,.M. Priyanka, M.Ajay Kumar, B.L.Bharadwaj Varma, " Association Rule Mining with Security Based on Playfair Cipher Technique" (IJCSIT) International Journal of Computer Science and Information Technologies, Vol. 5 (1) , 2014

[6] Jyotirmayee Rautaray, Raghvendra Kumar, "Privacy Preserving In Distributed Database Using Data Encryption Standard (DES) ", International Journal of Innovative Research in Science, Engineering and Technology Vol. 2, Issue 3, March 2013

[7] Prof. Geetika. Narang, Anjum Shaikh, Arti Sonawane, Kanchan Shegar, Madhuri Andhale," Preservation Of Privacy In Mining Using Association Rule Technique", International Journal of Scientific & Technology Research, Volume 2, Issue 3, March 2013

[8] Zhi Liu,Tianhong Sunand Guoming Sang," An Algorithm of Association Rules Mining in Large Databases Based on Sampling ", International Journal of Database Theory and Application Vol.6, No.6 , 2013

[9] Priyanka Asthana, Anju Singh , Diwakar Singh," A Survey on Association Rule Mining Using Apriori Based Algorithm and Hash Based Methods ", International Journal of Advanced Research in Computer Science and Software Engineering. Volume 3, Issue 7, July 2013

[10] Vijay Kumar Patidar, Abhishek Raghuvanshi, Vivek Shrivastava," Literature Survey of Association Rule Based Techniques for Preserving Privacy", An international journal of advanced computer technology, 2 (3), March-2013.

[11] Larry A. Dunning, Member, IEEE, and Ray Kresman,"Privacy Preserving Data Sharing With Anonymous ID Assignment", IEEE Transaction On Information Forensics and security, VOL. 8, NO. 2, FEBRUARY 2013

[12] Ms.R.Kalaivani, Ms.R.Kiruthika,"Automated Anonymous ID Assignment For Maintaining Data privacy", Proceedings of 2nd International Conference on Science,Engineering and Management, Srinivasan Engineering college, TamilNadu,India,March 28-29,2014

[13] Shiny. I.S , S. Gayathri,"Secure Multiparty Computation and Privacy Preserving Data Sharing with Anonymous ID Assignment", International Journal of Engineering Research and Applications (IJERA) ISSN: 2248-9622 International Conference on Humming Bird ,01st March 2014

[14] Fosca Giannotti, Laks V. S. Lakshmanan, Anna Monreale, Dino Pedreschi, and Hui (Wendy) Wang," Privacy-Preserving Mining of Association Rules From Outsourced Transaction Databases". IEEE SYSTEMS JOURNAL, VOL. 7, NO. 3, SEPTEMBER 2013.