ASSOCIATION RULE MINING WITH PRIVACY PRESERVATION IN HORIZONTALLY DISTRIBUTED DATABASES

[1] **Hajera Shireen,** [2] **Mohammad Naqueeb Ahmad**[3]**Dr. Syed Raziuddin,**
[1]PG Scholar, [2]Assistant Professor, [3]Professor and Head of Dept, [1,2,3]Dept of CSE
[1]hajerashireen022@gmail.com, [2]nasmtech@gmail.com, [3]hod_cse@deccancollege.ac.in
[1, 2, 3] DECCAN COLLEGE OF ENGINEERING & TECHNOLOGY Darussalam, Aghapura, Hyderabad, Telangana –India.

ABSTRACT

In horizontally distributed databases for mining of association rules a protocol has been proposed. This protocol is optimized than the Fast Distributed Mining (FDM) algorithm which is an unsecured distributed version of the Apriori algorithm. The main purpose of this protocol is to remove the problem of mining generalized association rules that affects the existing system. This protocol offers more enhanced privacy with respect to previous protocols. They are two rules, one that computes the union of private subsets that each of the interacting group of actors hold, and another that tests the inclusion of an element held by one player in a subset held by another. These set of rules uses the fact that the fundamental problem is of interest only when there is number of players greater than two. Besides, association rule mining has wide applications to discover interesting relationships relevant to task among attributes. In addition it is simpler and is optimized in terms of communication rounds, communication cost and computational cost than other protocols.

Index Terms— Association rules, Privacy preserving data mining, Apriori algorithm, Frequent Item sets, Distributed database.

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 be 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 terms of simplicity and efficiency as well as privacy.

II.LITERATUREREVIEW

We have to exploration of the Data Mining Literature Survey: Data mining derive its name from the similarities between searching for valuable business information in a large database —for example, finding linked products in terabytes of store scanner data —and mining a mountain for a vein of valuable ore. Both processes require either sifting through an immense amount of material, or intelligently probing it to find exactly where the value resides. Given databases of sufficient size and quality, data mining technology can generate new business opportunities by providing these capabilities:

Automatic prediction of trends and behaviors: Data mining automates the process of finding predictive information in large databases. Questions that traditionally required extensive hands-on analysis can now be answered directly from the data quickly [9]. A typical example of a predictive problem is targeted marketing. Data mining uses data on past promotional mailings to identify the targets most likely to maximize return on investment in future mailings. Other predictive problems include forecasting impoverishment and other forms of default, and identifying segments of a population likely to respond similarly to given events.

Automatic discovery of previously unknown patterns: Data mining tools sweep through databases and identify previously hidden patterns in one step. An example of pattern discovery is the analysis of retail sales data to identify seemingly unrelated products that are often purchased together [10]. Other pattern discovery problems include detecting fraudulent credit card transactions and identifying anomalous data that could represent data entry keying errors.

SQL extensions are defining aggregate functions for association rule mining. Their optimizations have the purpose of avoiding the joins to convey unit (cell) formulas, but are not optimized to perform partial Transposition for each group of result rows [2]. Conor Cunningalam proposed an optimization and Execution strategies in an RDBMS which uses two operators i.e ., PIVOT operator on tabular data that exchange rows and columns enabled data transformations are useful in data modeling, data analysis, and data presentation. They can quite easily be implemented inside a query processor system, much like select, project, and join operator. Such design provides the opportunities for better performance, both during query optimization and query execution. Pivot is an extension of Group By with a unique restrictions and optimization opportunities, and this makes it is very simple to

introduce incrementally on top of existing grouping implementations. H Wang.C.Zaniolo proposed asmall but Complete SQL Extension for Data Streams Data Mining and. This technique is a powerful database language and system that enables users develop complete data-intensive applications in SQL by writing new aggregates and table functions in SQL, rather than in procedural languages as in current Object -Relational systems .

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.

Fig. 1 System Architecture Diagram

The above system architecture explains the user module which enlists the secure data mining and has considered two connected settings. One, in which the data owner and the data miner are two different individual, and another, in which the data is scattered among several parties who aim to jointly perform data mining on the unified corpus of data that they grip. In the first location, the goal is to protect the data records from the data miner. Hence, the information holder aims at anonymizing the data prior to its release. The main approach in this framework is to apply data perturbation. The disconcerted data can be used to infer general trends in the data, without informative original documents information. In the second setting, the goal is to perform data mining while protectingthe data records of each of the data owners from the other data owners. The work of the administratoris to view user details. Direction to view the item set based on the user processing details using association rule with Apriori algorithm. Association set of laws are created

by analysing data for frequent if/then patterns and using the criteria support and confidence to identify the most important associations, Support is an indication of how regularly the items appear in the database. Confidence indicate the number of times the if/then statement have been found to be true [7].

The Privacy preserving fast distributed mining (PPFDM) algorithm is a combination of Fast Distributed mining algorithm (FDM) FDM is an unsecured distributed version of the Apriori algorithm. 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.

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 efficiently 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 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 presences 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 ⊆ 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 ⊆ T. An association rule is an Implication of the form A⟹B, where A ⊂ I, B ⊂ I and A ⋂ B = ∅ . The rule A⟹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 ⋃ B. This is taken to be the probability, P(A ⋃ B). The rule A⟹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.

4 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 then 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.

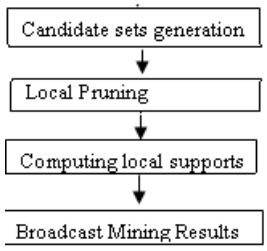

Fig. 2 Process of FDM Algorithm Diagram.

With the existence of many large transaction databases, the huge amounts of data, the high scalability of distributed systems, and the easy partition and distribution of a centralized database, it is important to investigate efficient methods for distributed mining of association rules. This study discloses some interesting relationships between locally large and glob-ally large itemsets and proposes an interesting distributed association rule mining algorithm, FDM (Fast Distributed Mining of association rules), which generates a small number of candidate sets and substantially reduces the number of messages to be passed at mining association rules. Our performance study shows that FDM has a superior performance over the direct application of a typical sequential algorithm.

IV.CONCLUSION

In this we have presented various techniques for secure mining of association rules in horizontally partitioned distributed databases. In this, the protocol used is more efficient than current leading K and C protocol. In this system, proposed 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.

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, 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] J. Vaidya and C. Clifton, ―Privacy preserving association rule mining in vertically partitioned data, in The Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Edmonton, Alberta, Canada, July 23-26 2002, pp. 639–644.

[10] R.Agrawal and R. Srikant.,―Privacy-preserving data mining‖, SIGMOD Conference, pages 439–450, 2000.