International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 4 Issue 02, February-2015

# Review on: Taut Extraction of Association Rule from Distributed Database

Miss. Mahale Mohini V.
Department of Computer Engineering,
SNDCOE and RC Yeola, Nashik-423401

Prof. Sheikh I. R.
Department of Computer Engineering,
SNDCOE and RC Yeola, Nashik-423401

***Abstract*— There are many techniques to extract association rules from large datasets, but sometimes these datasets are distributed horizontally which is called strew database. In the strew database there are several sites or players that hold homogeneous database this database shares the same schema but hold information on different entities. For extracting association rules from such database the existing system is not so secure and efficient. The proposed system given here provides a secure and efficient solution for the problem stated above. Here we are going to use Fast Distributed mining (FDM) which is an unsecured distributed version of the Apriori algorithm. The main ingredients of the proposed system 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. This protocol offers enhanced privacy with respect to the protocol in [18]. In addition, it is simpler and is significantly more efficient in terms of communication rounds, communication cost and computational cost.**

***Index Term- Privacy preserving data mining, distributed computation, frequent item sets, association rules***

## I. INTRODUCTION

In strew database where datasets are horizontally partitioned, there are several players that hold homogeneous databases, this databases share the same schema but hold information on different entities. In such scenario there is problem to find all association rules with support at least s and confidence at least c, for some given minimal support size sand confidence level c, that hold in the unified database. While doing this the information of private database should not disclosed to the participating players. The private information that we would like to protect 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.

In the problem of secure multi-party computation. there are *M* players that hold private inputs, y1, . . . , *yM*, and they wish to securely compute z = *f*(y1, . . . , *yM*) for some public function *f*. If there existed a trusted third party, the players could give him their inputs(private datasets) and he would perform the function evaluation and send to them the resulting output. In the absence of such a trusted third party, it is needed to devise a protocol that the players can run on their own in order to arrive at the required output z. Such a protocol is considered perfectly secure so that another player cannot learn the extra information in the absence of trusted third party. Kantarcioglu and Clifton gives solution to this problem [18]. The main part of the protocol is a sub-protocol for the secure computation of the union of private subsets that are held by the different players. This is also the only part in the protocol in which the players may extract from their view of the protocol information on other databases, beyond what is implied by the final output and their own input.

We propose here an alternative protocol for the secure computation of the union of private subsets. This protocol is more efficient than in [18] in terms of simplicity and efficiency as well as privacy. Our protocol does not depend on oblivious transfer and commutative encryption. This solution is still not perfectly secure, it leaks excess information only to a small number of possible coalitions, unlike the protocol of [18] that discloses information also to some single players. In addition, we claim that the excess information that our protocol may leak is less subtle than the excess information leaked by the protocol of [18].

## II LITERATURE SURVEY

There are many Privacy Preserving Association Rule Mining (PPARM) algorithms are proposed for different partitioning methods by satisfying privacy constraints. The various methods such as randomization, perturbation, heuristic and cryptography techniques are proposed by different authors to find privacy preserving association rule mining in strew databases and vertically partitioned databases. In the case of secure multiparty computation while computing the association rules, the data of participating parties should nod disclose to each other .There are many solutions to satisfy the above constrained. The first solution was propose by Yao [1] this technique was only suitable for two players.

Latter in paper [1] specifies the protocol for secure mining of association rules in horizontally partioned databse, where Fast Distributed Mining algorithm (FDM) is get uesd for mining of association rules . In this protocol players finds their locally s-locally frequent itemsets then the players check each of them to find out globally s-frequent item set. But the protocol 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 .This protocol offers better privacy and is significantly more efficient in terms of communication cost and computational cost than the pravious one .But this solution is not perfectly secure cause it leaks excess information.

In the problem of extracting association rules from strew database the goal is to perform data mining while protecting the data records of each of the data owners and from the other data owners. computation. The usual approach here is cryptographic rather than probabilistic. Lindell and Pinkas [22] gives the solution by implementing secure ID3 decision tree. Secure clustering using the EM algorithm was implemented by Lin et al. [21] over horizontally distributed data. In The problem of distributed association rule mining was studied in [31], and [33] but here the data was distributed vertically, where each party holds a different set of attributes. Also the work of [26] considered this problem in the horizontal setting, but they considered large-scale systems in which, on top of the parties that hold the data resources there are also managers which are computers that assist the resources to decrypt messages. There is another solution given in[20] this protocols uses the homomorphic encryption, while our protocol uses commutative encryption, the computational costs by using homomorphic encryption are significantly higher than using the commutative encryption.
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-

## III PROPOSED METHEDOLOGY

The proposed system gives the alternative protocol which will overcome from the problem which are occurred in Farst distributed Mining (FDM) proposed by Kantarcioglu and Clifton(18).The proposed system is more efficient than the existing system in terms of privacy, communication rounds, communication cost and computational cost. The existing and proposed system both are based on FDM [8],which is an unsecured version of the Apriori algorithm. The proposed system computes a parameterized family of function which is called as threshold function In which two cases correspond to the problems of computing the union and intersection of private subsets. The protocol used for this function can be used in other cases as well. The major problem of extraction of association rule is set inclusion problem; the problem where Bob holds a private subset of some ground set, and Alice holds an element in the ground set, and they wish to determine whether Alice's element is within Bob's subset, without revealing to either of them information about the other party's input beyond the above described inclusion.
Following Fig. shows the architecture of proposed system

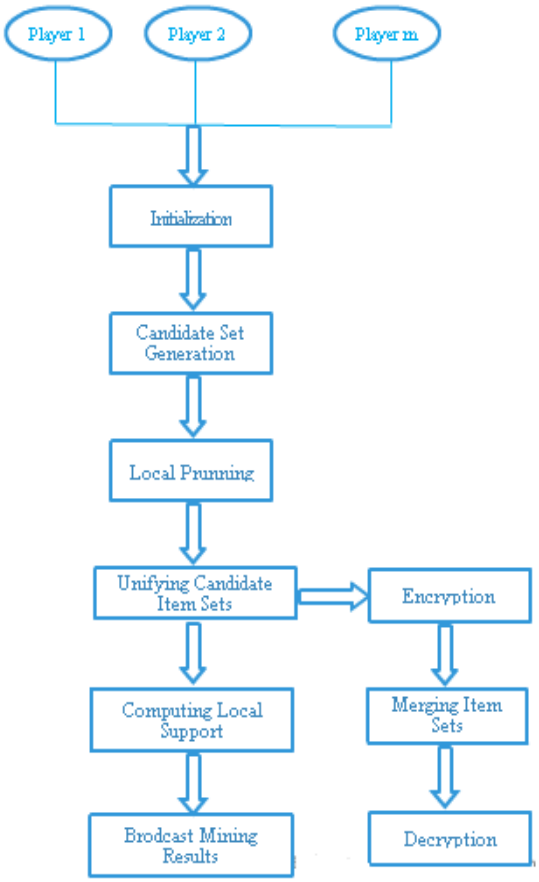


Fig. Architecture of proposed system

The detailed work of system stepwise is as given below

*1) Generation of synthetic database:*

To Evaluate performance of Algorithm over large range of data Patterns/characteristics synthetic transactions are generated. Synthetic transactions are subset of anonimized data which is actually a process of anonimization. The use of anonimized data is to prevent secrecy of data with using different fields as filter for information for particular aspect of data. It allows generating large data set using real data without affecting original data. This isolates privacy and security concerns for real data and allows researchers and data analyst's to test with safe data. These transactions are similar to real world –retail business scenario. In this model, customer have a tendency to buy specific set of items together, each of these set can be a maximal item set and a transaction might contain more than one of such item sets.

*2. Apriori Algorithm:*

The name Apriori suggests 'a prior Knowledge'. Purpose of this algorithm it to find association between different sets of data. It proceeds by identifying the frequent individual items in the database and extending them to larger and larger

item sets as long as those item sets appear sufficiently often in the database. The frequent item sets determined by Apriori can be used to determine association rules which highlight general trends in the database. Number of transaction is present in each set of data. Initial scan/pass of algorithm counts occurrence of each item in order to determine the frequent items set. Next scan K consists two phases .

1) In first phase , Candidate item set $C_K$ is generated using frequent item set $L_k$-1 found in $(K-1)^{th}$ pass .This is candidate generation process in Apriori Algorithm.
2) In second phase database is scanned to find support for Candidates $C_K$. In next step, it prunes the candidates which have an infrequent sub pattern and keep only subset of candidate sets which are already identified as frequent items sets.

Output of Apriori algorithm generates sets of rules which determine how often items are brought together in single set.

*3. Privacy Preserving in Data Mining:*

Privacy preservation in data mining is mining data without compromising data privacy and security. It is an algorithm that can mine the data in a distributed fashion while guaranteeing that the privacy of the data is not compromised. The paper of Tamir Tassa analyzes protocol UNIFI-KC for privacy. This protocol does not respect security and privacy as it discloses player's information. This paper uses AIDA - Anonymous ID Assignment for maintaining privacy to the player's database. There are many applications that require dynamic unique ID. These ID's can be used to share and store data and other resources anonymously and also do not generate conflict. AIDA uses random integer values between 1 and S for each node.

*4.Association Rules*

Agrawal quoted the Association rule problem for data mining in 1993 .It is sometimes also refereed as market basket problem. In this item set of items are defined and larger number of transactions are generated within give set. Task is to determine relationship within various items in basket. Association rule mining finds association rules to fulfill defined minimum support and confidence from given database. Normally the problem is divided into two sub-problems.

First is to determine items sets whose occurrences is greater than predefined threshold in database, these item sets are called frequent or large item sets with the restraint 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->B, where $A \subset I$, $B \subset I$ and $A \cap B = \emptyset$. 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 $A \cup B$. This is taken to be the probability, $P(A \cup B)$. 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 contains B. This is the conditional probability, p(B|A). That is, Support (A -> B) = $P(A \cup B)$ , Confidence (A->B) = p(B|A)

Association rule mining can be viewed as a two-step process:

1. Determine all frequent item sets, 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 guarantee minimum support and minimum confidence.

*5) FDM:*

The main notion of FDM 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 discloses his locally s-frequent item sets. Then the players check each of them to see if they are s-frequent also globally.

The FDM algorithm proceeds as follows:

1) Initialization: All the players should calculate all k-item sets that are s-frequent that is calculate $F^k_{s.}$
2) Generation of candidate set: The set of all local and global frequent item sets are get calculated by each player $P_{m.}$ Specifically Pm computes $F^{k-1,m}_s \cap F^{k-1}_s$. Then the Apriori algorithm is get performed to generate the set $B^{k,m}_s$ .
3) Local Pruning: Each player computes $supp_m(X)$. He then maintains only locally frequent item which is denoted by $Cs^{k,m}$ .
4) Unifying the candidate item sets: Each player broadcasts his own set of items $Cs^{k,m}$ which is calculated in above step. Then all players computes $C^k s$
5) Computing local supports: Local supports of all item sets that is $C^k s$ is get calculated.
6) Broadcast mining results: Each player broadcasts his own local support. So that everyone can compute the global support of every item set. Finally the set of all globally frequent item sets $F^k_s$ whish is subset of $C^k s$ is get produced.

## IV CONCLUSION

Extracting association rules from strew database involves the problem of secure multiparty communication. We proposed a protocol for secure mining of association. Rules from strew database that improves expressively upon the current leading protocol in terms of privacy and efficiency. The main ingredient is a protocol that tests the inclusion of an element held by one player in a subset held by another. n addition, this system is simpler and is significantly more efficient in terms of communication rounds, communication cost and computational cost.

REFERENCES:

1. 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.
2. A.C. Yao, "Protocols for Secure Computation," Proc. 23rd Ann. Symp. Foundations of Computer Science (FOCS), pp. 160-164, 1982.
3. Y. Lindell and B. Pinkas, "Privacy Preserving Data Mining," Proc. Crypto, pp. 36-54, 2000.
4. X. Lin, C. Clifton, and M.Y. Zhu, "Privacy-Preserving Clustering with Distributed EM Mixture Modeling," Knowledge and Information Systems, vol. 8, pp. 68-81, 2005.
5. J. Vaidya and C. Clifton, "Privacy Preserving Association Rule Mining in Vertically Partitioned Data," Proc. Eighth ACM SIGKDD Int'l Conf. Knowledge Discovery and Data Mining (KDD), pp. 639- 644, 2002.
6. J. Zhan, S. Matwin, and L. Chang, "Privacy Preserving Collaborative Association Rule Mining," Proc. 19th Ann. IFIP WG 11.3 Working Conf. Data and Applications Security, pp. 153-165, 2005.
7. A. Schuster, R. Wolff, and B. Gilburd, "Privacy-Preserving Association Rule Mining in Large-Scale Distributed Systems," Proc. IEEE Int'l Symp. Cluster Computing and the Grid (CCGRID), pp. 411-418,y 2004.
8. L. Kissner and D.X. Song, "Privacy-Preserving Set Operations," Proc. 25th Ann. Int'l Cryptology Conf. (CRYPTO), pp. 241-257, 2005.
9. T. ElGamal, "A Public Key Cryptosystem and a Signature Scheme Based on Discrete Logarithms," IEEE Trans. Information Theory, vol. IT-31, no. 4, July 1985