ISSN No. 0976-5697

Volume 6, No. 6, June - July 2015

International Journal of Advanced Research in Computer Science

REVIEW ARTICLE

Available Online at www.ijarcs.info

Association Rule Mining In Horizontally Distributed Database

Suvarna Hiwale, Smita Ponde
Department of computer science
Deogiri Institute of Engineering &management Technology
Aurangabad.

Abstract: The security of large database becomes a very serious issue now days while sharing data on the internet. However there are many algorithms to provide privacy preserving in data mining. The two leading protocols available are by the scientist Kantacioglu and Clifton this protocol is based on an unsecured distributed version of the Apriori algorithm named as Fast Distributed Mining (FDM) algorithm of Cheung et al. Our protocol is also based on the Fast Distributed Mining algorithm, having two secure multiparty algorithms. Theses protocols computes union and intersection of the private subsets held by each of the interacting site. In addition it is the simple and the efficient protocol.

Keywords*:* privacy preserving, Data mining, association rules private subsets.

I. INTRODUCTION

While considering database it may be distributed among the various sites. Most of the businesses and companies share their information along with their personal information for getting benefits. Sharing of this type of personal information can be risky as it arise the privacy issue. Though they share their private information but still have to be more careful that data remains as a private. This is known as secure mining

Data mining can be used to extract the knowledge from the large amount of database. Many times it will be difficult to directly share the data, because the data can be split on different sites with same schema. The large amount of data needs to be split on different sites. Users may want to use the data separated on different sites collectively for computing some function. This can be done but may have some difficulties with securing issues. There can be privacy issues for making the personal information secure. Privacy limits of these sites can prevent those to share this data. One of the approaches for privacy preserving data mining is the Association rule mining.[1] The security of large database on network becomes a serious issue. As there several sites that hold homogenous database with same schema but hold on the different sites. That means there are horizontally partitioned database.

To preserve the privacy in case of association rule mining can be called as privacy preserving association rule mining [2]. Database can be consisting of large amount of transactions which are extracted from a single source of data or from many sources. Depending upon the requirement of applications, database is maintained at single location called centralized database or the database may be distributed at multiple sites called distributed database. In distributed applications, the database is portioned into two types Horizontal and vertical database. The main objective of privacy preserving association rule mining in centralized database is mining process can be done by hiding sensitive data/information from users other than database owner. In distributed aim is finding the global mining results by preserving the individual sites private data/information from one another. Global results are determined only when the necessary results/information is captured based on all parties database individually like local frequent item sets and their support values of all sites are required to determine whether an item set is globally frequent or infrequent.[7]

The main aim of our protocol is to find all the association rules with minimum support S and confidence C, from the related database. While doing this, we have to secure the private database which is held by the interacting sites. So, there should be no or minimum leakage of private information.

The main ingredients of our protocol are two secure multiparty algorithms which defines secure multiparty computation. [5] Secure multiparty algorithm employs distributed algorithm in secure manner. Secure multiparty algorithm not only preserves individual privacy, but also prevents the leakage of private information other than results. In this problem multiple sites holding the homogenous database with same schema but on different sites. The aim is to take private inputs from multiple sites collectively and to compute some public function.

The inputs to our algorithm are the partial database and the required output will be the list of association rules with minimum support S and confidence C. The scientist kantacioglu and Clifton studied the problem and gives the protocol for the solution. [3] The solution is based on the undistributed version of Apriori algorithm which is Fast Distributed Mining algorithm (FDM). [4] The main thing of protocol is a sub protocol for secure computation of private subsets.

Our protocol is the alternative way for secure computation of union of private subsets which is held by each of the interacting site. This is the simpler, efficient and more secure. This protocol computes the parameterized family of functions called threshold function.

II. RELATED WORK

Previous work in privacy preserving data mining has considered two things, One in which data owner and miner are two different entities. And other in which the data is distributed among various sites. Our algorithm uses the secure multiparty algorithm, which is started when the scientist Yao proposed the millionaier's problem in which two parties wanted to know which one was the richer person without knowing the individual information of each parties. [6]

The protocol basically divides the whole database into number of different sites each of them have particular sequence number (s1,s2,..sn) so that if any of them want the information about other then it can be retrieve without disclosing private data. All the sites have their own data and none can disclose their private data.

A. Association rule mining:

Association rule mining is the important tool that is used to reveal the relationships in database. Association rule mining finds the interesting associations, correlations, frequent patterns and relationships among the data items.

The process of association rule mining includes two main sub problems. The first one is to discover all frequent itemsets and second is to use these discovered frequent itemsets to generate association rules. E.g. In a laptop store 70% of the customers who are buying laptop will also buy antivirus and pen drive for data portability.

We can formally define association rule mining as follows Let I= I1,I2,…Im be the set of m different attributes. T is the set of transactions.

i.e. $T \underline{C} I$. D be the database with different transactions. Then the association rule is the statement in the form of X→Y(X implies Y) where, X, Y C I are set of items called as itemsets. And X∩Y =ϕ. Here X is antecedent and Y is the consequent.

The two basic things that are required for association rule mining are support(S) and confidence (C). The terms support and confidence are defined as follows:

Support(S):

The rule X→Y has sup port S if S% of transaction in database D contains X U Y. If the support is greater than user specified support is said to have minimum support or threshold support. The support is defined as follows:

Support(X)= No of transactions that contain X / Total no of transactions.

Confidence(C):

The rule X→Y has confidence C if C% of transaction in database D that contains X also contain Y. If the confidence is greater than user specified confidence is said to have minimum confidence or threshold confidence. The confidence is defined as follows:

Confidence(X→Y)=Support(XUY)/Support(X)

B. Fast Distributed Mining (FDM):

Fast Distributed Mining (FDM) algorithm is an unsecured distributed version of the Apriori algorithm. Its main idea is that any s-frequent itemsets must be also locally s-frequent in at least one of the sites. Hence, in order to find all globally -frequent itemsets, each player reveals his locally s-frequent itemsets and then the players check each of them to see if they are s-frequent also globally.

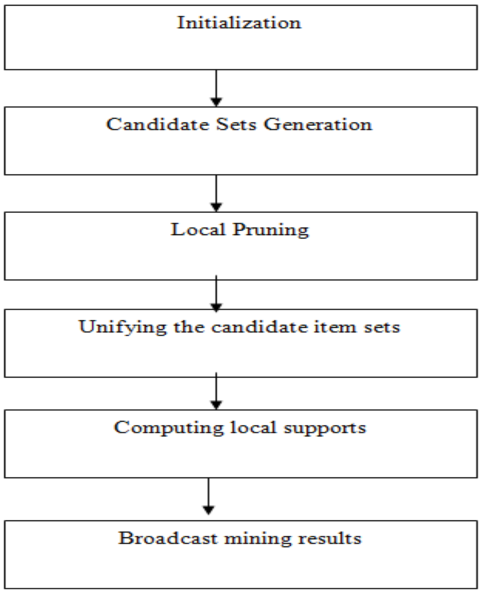

Figure 1: Steps of FDM

System architecture shows the flow of the steps inside the system. The figure shows the steps in the system. The first one is initialization step, in which the player is starting their role by holding some value in it. And then it will help to find out the next item. Next step will be the generating candidate set, in which we are finding the key which appears repeatedly or you may say it which is intersection or common for both sites and players. Next phase is local pruning, in which we are trying to eliminate the unwanted result or extra data which will in turn help in mining the data.

Next phase is Candidate key union, as word indicates it is based on the union of data of participating players. Next phase is local support computation, in which we are computing the local support that how much the participating player can support. Next phase is Broadcasting of the mining result in which we are going to display the result by merging the all result that we got from all participating player and then displaying it

III. PERFORMANCE EVALUATION

In the following section, we have described the database that we are using in our experimentation and also how the databases are split horizontally into partial databases.

A. Synthetic database generation:

The synthetic database that we are using in our experimentation is synthetic databases that were generated by the technique in [1]. The parameter N is used for number of transaction in whole database having value 500,000. L is the parameter for number of items with value 1000. A_t is the transaction average size with value 10. A_f is the average size of maximal potentially large itemsets with value 4. CS is the clustering size with value and PS is the pool size with value 60. Thses parameters are used for generating synthetic database.

B. Distributing the Database:

We have synthetic database D of N transactions and number of players M, we split D into M partial database, D_m

≤1≤M. such as for each 1≤m≤M we have drown a random number w_m with mean 1 and variance 0.1. Then the numbers are normalized $\sum_{m=1}^{M} w_m=1$. And finally we split D into M partial database of expected size such that each transaction tϵD is assigned at random to one of the partial database.

C. Experimental setup:

There is a comparison of the performance of two secure implementations of the FDM. In the first implementation (denoted FDM-KC), we executed the unification step (Step 4 in FDM) using Protocol UNIFI-KC, where the commutative cipher was 1024-bit RSA; in the second implementation (denoted FDM) we used our Protocol UNIFI, where the keyed-hash function was HMAC. In both implementations, there is implementation of Step 5 of the FDM algorithm in the secure manner that was described in Section 3. We have tested the two implementations with respect to three measures:

a. Total computation time of the complete protocols (FDMKC and FDM) over all sites. That measure includes the Apriori computation time, and the time to identify the globally s-frequent itemsets, as described in Section 3. (The latter two procedures are implemented in the same way in both Protocols FDM-KC and FDM.)
b. Total computation time of the unification protocols only (UNIFI-KC and UNIFI) over all sites.
c. Total message size. We ran three experimental sets, where each set tested the dependence of the above measures on a different parameter: N — the number of transactions in the unified database.

IV. CONCLUSION AND FUTURE WORK

Here, we have proposed the protocol for mining of association rules on horizontally distributed databases. The protocol is the improvisation of the leading protocol in terms of cost and simplicity and efficiency. The main part of our protocol is the secure multiparty algorithms which compute the union or intersection of private subsets of interacting parties. One research problem that this study suggests was to devise an efficient protocol for inequality verifications that uses the existence of a semi honest third party. Such a protocol might enable to further improve upon the communication and computational costs. The second and third stages of the protocol of, as described. Other research problems that this study suggests is the implementation of the techniques presented here to the problem of distributed association rule mining in the vertical setting , the problem of mining generalized association rules , and the problem of subgroup discovery in horizontally partitioned data.

V. REFERENCES

[1]. R. Agrawal and R. Srikant. Fast algorithms for mining association rules in large databases. In *VLDB*, pages 487–499, 1994.

[2]. R. Agrawal and R. Srikant. Privacy-preserving data mining. In *SIGMOD Conference*, pages 439–450, 2000.

[3]. M. Kantarcioglu; C. Clifton, “Privacy-Preserving Distributed Mining of Association Rules on Horizontally Partitioned Data”, IEEE Trans. Knowl. Data Eng. 16(9): 1026-1037, 2004.

[4]. D. W.-L. Cheung; J. Han, V. Ng; A. W.-C. Fu; Y. Fu, “A fast distributed algorithm for mining association rules"Proceedings of the 1996 International Conference on Parallel and Distributed Information Systems (PDIS'96), Miami Beach, Florida, USA, Dec. 1996.

[5]. W. Du; M. J. Atallah, “Secure multi-party computation problems and their applications: A review and open problems”, In Proceedings of the 2001 New Security Paradigms Workshop, Cloudcroft, New Mexico, 2001

[6]. A. Ben-David, N. Nisan, and B. Pinkas. Fairplay MP - A system for secure multi-party computation. In *CCS*, pages 257–266, 2008.

[7]. A. Evfimievski, R. Srikant, R. Agrawal, and J. Gehrke, “Privacy preserving mining of association rules,” in The Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Edmonton, Alberta, Canada, July 23-26 2002, pp. 217–228