International Journal of Interdisciplinary Innovative Research &Development (IJIIRD)
ISSN: 2456-236X
Vol. **06** Issue **01** | **2021**

Horizontally Distributed Databases with Secure Mining

Sharad Bhojane[1], G. P. Chakote[2]
[1] *Student, Computer Department, MSS's College of Engineering, Maharashtra, India*
[2] *Professor, Computer Department, MSS's College of Engineering, Maharashtra, India*

ABSTRACT

I implemented Fast Distribution Mining with Least Low Probability (FDM-LLP). The main aim of this algorithm is to implement a secure protocol. Security is of utmost importance in any kind of large scale data-mining, especially where the corporate is involved as parties. In this dissertation we introduce & implement a privacy-preserving protocol for horizontally partitioned data distributed over two or more parties. I implemented Fast Distribution Mining with Least Low Probability (FDM-LLP). The main aim of this algorithm is to implement a secure protocol which minimize the time require for union of locally declared candidate set without affecting on sensitivity of data. Our base paper looks at implementing a secure protocol for mining of association rules in horizontally distributed database. We aim to extend this work by developing mining using FDM-LLP. This could be an ideal approach for a scenario where mining is difficult in a distributed database system due to the lack of trust demonstrated by databases in each other's association rules, leading honest nodes to lose privacy.

Keyword: ***- Association Rule; Data mining; Horizontally Distributed Databases, Privacy-Preserving, Fast Distribution Mining with Least Low Probability (FDM-LLP).***

1. INTRODUCTION

A distributed database system is one in which each computer in the network can act as a client or server for the other computers in the network, allowing shared access to various resources such as files. Data-mining, is measure issue where the corporates are involved as parties. Most of the companies have to share their personal data for mutual benefits. As data is increasing day by day, we need to store it on different computer and whenever user want to access it, it works like a single unit though we are storing the data on different machines. The data on several computers can be simultaneously accessed and modified using a network. In a network each server is linked by its local database management system (DBMS), and each cooperates to maintain the consistency of the global database. To maintain the privacy of the data many scientists put their efforts so that we get data without losing the privacy of that related data. Data mining techniques can be implemented rapidly on existing software and hardware platforms to enhance the value of existing information resources, and can be integrated with new products and systems as they are brought on-line. When implemented on high performance client/server or parallel processing computers, data mining tools can analyze massive databases to deliver answers to questions such as, "Which clients are most likely to respond to my next promotional mailing, and why?"

Data mining techniques are the result of a long process of research and product development. This evolution began when business data was first stored on computers, continued with improvements in data access, and more recently, generated technologies that allow users to navigate through their data in real time. Data mining takes this evolutionary process beyond retrospective data access and navigation to prospective and proactive information delivery. Data mining is ready for application in the business community because it is supported by three technologies that are now sufficiently mature:

- Massive data collection
- Powerful multiprocessor computers
- Data mining algorithms

In the horizontal distributed databases system, the data are stored on different machines. And this information belongs to the one particular related subject. Consider one example for proper understanding of this concept. Let there is one table which is having lots of records in rows and columns format. Some system may store some records which belong to the same column and other columns along with their information may store on the next machines. So the data is distributed along the various machines. If the data stored on different machines and is divided by rows means while storing the data on different machines some rows are stored on one machine and other are stored on different machines. i.e. partitioning the data according to rows is called horizontally partitioned (Distributed Databases) and if that data is stored and partitioning according to the column wise then it is called as a vertical partitioning (Distributed databases). Some may prefer to store or partition of data according to vertical and some may prefer to use horizontal partition. Most of the scientist uses semi-honest model where node may follow or not all protocol for accessing the data among distributed system.

The protocol that we implement here computes a parameterized family of functions, which we call threshold functions, in which the two extreme cases correspond to the problems of computing the union and intersection of private subsets. Those are in fact general purpose protocols that can be used in other contexts as well. Another problem of secure multiparty computation that we solve here as part of our discussion is the set inclusion problem; namely, the problem where Alice holds a private subset of some ground set, and Bob holds an element in the ground set, and they wish to determine whether Bob's element is within Alice's subset, without revealing to either of them information about the other party's input beyond the above-described inclusion.

2. SYSTEM ARCHITECTURE

System architecture describes the way of data flow inside the system. It goes through various phases as shown in figure number 1. It is having initialization, 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 phase is 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.

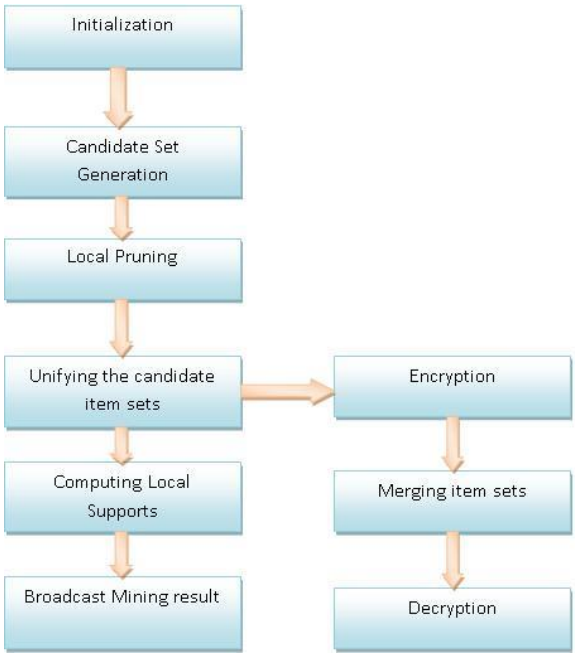

Fig -1: System Architecture

3. ALGORITHM

Step 1: Initialization
It is assumed that the players have already jointly calculated F_s^{k-1}. The goal is to proceed and calculate F_s^k.
Step 2: Candidate Sets Generation
P_m computes the set $F_s^{k-1,m} \cap F_s^{k-1}$. Then apply on that set the Apriori algorithm in order to generate the set of $B_s^{k,m}$ candidate k-itemsets.
Step 3: Local Pruning
For each $X \in B_s^{k,m}$, P_m computes support(X). Then retain only those itemsets that are locally s-frequent denoted as $C_s^{k,m}$.
Step 4: Unifying the candidate itemsets
Each player broadcasts his $C_s^{k,m}$ and then all players compute $C_s^k := U_{m=1}^M C_s^{k,m}$.
Step 5: Computing local supports
All players compute the local supports of all itemsets in C_s^k .
Step 6: Broadcast Mining Results
Each player broadcasts the local supports that he computed. From that, everyone can compute the global support of every itemset in C_s^k . Finally, F_s^k is the subset of C_s^k that consists of all globally s-frequent k-itemsets.
Step 7: Distribution of data
Data will distribute among the various site on the basis of load having on that site.

4. PROPOSED WORK

In previous system a smaller number of features is available and it's too much difficult to have accurate item. My proposed method overcomes this problem. In previous method we have to manually supply the number of participations for the distribution of data. And this is time consuming as all information have to distributed among the number of partitions given though its heavily loaded. Due to this performance of these site decreases.

By using FDMLLP algorithm, system itself will decide the number of partitions should have for better performance and fast calculation. So, time required for it is comparative less than previous one.

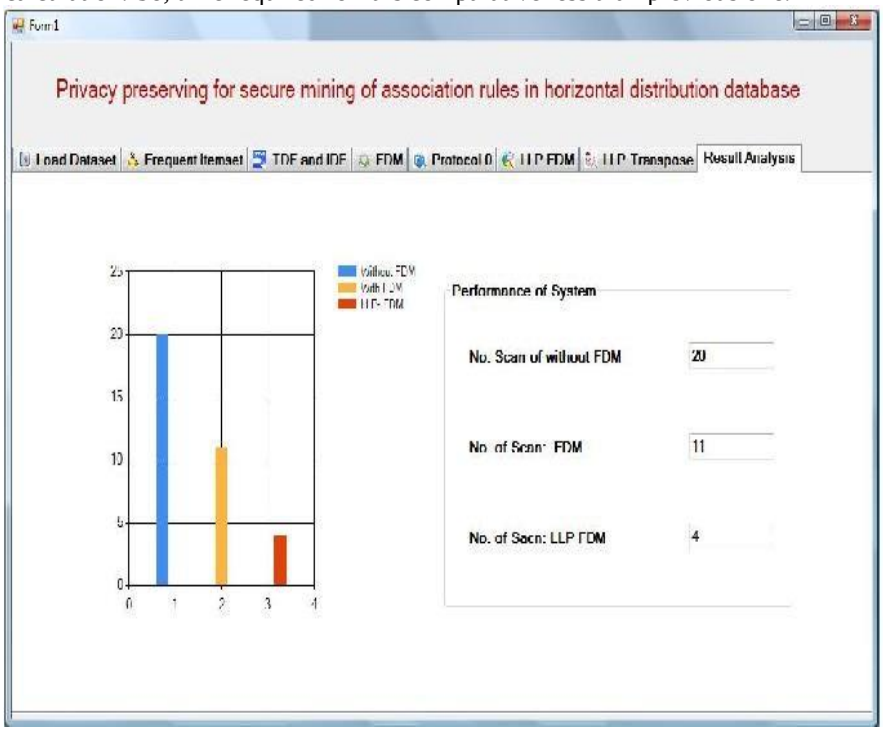

Fig -2: Comparative Result

Above figure shows the complete graph for the total number of iteration required for horizontal distribution. The number of scan required without FDM is 20 whereas 11 with FDM and 4 by using FDMLLP.

5. CONCLUSIONS

This paper aims to develop privacy preserving protocol by extending secure mining protocol used for distributed databases. We believe that this will improve security and privacy of mining operations in distributed database. We proposed a system that will help for secure mining of data in horizontally distributed databases. In our base paper we implement a secure protocol for mining of association rules in horizontally distributed database. We extend this work by developing FDM-LLP.

6. ACKNOWLEDGEMENT

Sincerely thank the all-anonymous researchers for providing us such helpful opinion, findings, conclusions and recommendations. I wish to thanks various people who contribute their work for privacy preserving and whose theory helped me to write this paper.

7. REFERENCES

[1] J. Vaidya, Clifton. “Privacy preserving association rule mining in vertically partitioned data “ In: Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. Edmonton, Alberta, Canada: ACM.2002:639-644.
[2] C. Yao, “Protocols for secure computations,” Proceedings of the 23rd Annual IEEE Symposium on Foundations of Computer Science, IEEE Press, New York, 1982.
[3] Gui Qiong, Cheng Xiao-Hui. Association Rule Mining Algorithm Based on Similarity Matrix of Transactions [J]. Journal of Guilin University of Technology, Vol. 28, No.4, Nov.2008, p p. 568-571.
[4] M. Kantarcioglu and C. Clifton. Privacy-preserving distributed mining of association rules on horizontally partitioned data. IEEE Transactions on Knowledge and Data Engineering, 16:1026–1037, 2004.
[5] Mahmoud Hussein, Ashraf El-Sisi, Nabil Ismail, “Fast Cryptographic Privacy Preserving Association Rules Mining on Distributed Homogenous DataBase”, Knowledge-Based Intelligent Information and Engineering Systems, Lecture Notes in Computer Science, Volume 5178/2008, pp. 607 -- 616 (2008).
[6] M. Kantarcioglu and C. Clifton. Privacy-preserving distributed mining of association rules on horizontally partitioned data. IEEE Transactions on Knowledge and Data Engineering, 16:1026–1037, 2004.
[7] Zhu Yu- quan, Tang Yang, Chen Geng, “A Privacy Preserving Algorithm for Mining Distributed Association Rules,” 19-21 May 2011.