Secure Mining of Association Rules in Horizontally Distributed Databases

Poojari Neelima[1] and B. Chakradhar[2]

1. M.Tech., Sri Sivani College of Engineering, Srikakulam, Andhra Pradesh
2. Assistant Professor, CSE Department, Sri Sivani College of Engineering, Srikakulam, Andhra Pradesh

Abstract: *We study here the problem of secure mining of association rules in horizontally distributed databases. In that setting, there are several sites (or players) that hold the databases that share the same schema but hold information on different entities. The goal is to find all association rules with support at least s and confidence at least c, for some given minimal support size and confidence level that hold in the unified database. This paper addresses the problem of computing association rules within a scenario of homogeneous database. We assume that all sites have the same schema, but each site does not have information on different entities. The goal is to produce association rules that hold globally while limiting the information shared about each site. Many proposals have been sited to implement in large scale databases which extends to preserve privacy to the private data of different sites. In this paper our focus is based on horizontal partitioned distributed data through a popular association rule mining technique.*

Keywords: Data Mining, User Interface, FDM Algorithm, ODBC, Acceptance Testing

1. INTRODUCTION

The security of the large database becomes a serious issue while sharing the data to the network against unauthorised access. However in order to provide the security many researchers cited the issue of Secured Multiparty Computation (SMC) that allows multiple parties to compute some function of their inputs without disclosing the actual input to one another. Secure sum computation method is popularly and widely accepted due to its simple and thorough solution. The outcomes of our proposed procedure provide a significant result so that it becomes impossible for semi honest party to know the private data of some other sites.

The developments of computed technology in last few decades are used to handle large scale data that includes large transaction financial data, bulletins, emails etc. Hence information has become a power that made possible for user to voice their opinions and interact. As a result revolves around the practice, data mining come into sites. Association rule mining is one of the Data Mining techniques used in distributed database. In distributed database the data may be partitioned into fragments and each fragment is assigned to one site. The issue of privacy arises when the data is distributed among multiple sites and no other party wishes to provide their private data to their sites but their main goal is to know the global result obtained by the mining process. However privacy preserving data mining came into the picture. As the database is distributed, different users can access it without interfering with one another. In distributed environment, database is partitioned into disjoint fragments and each site consists of only one fragment. The goal defines the problem of secure multi party computation. In such problems, there are many players that hold private inputs and they wish to securely compute for some public function. If there existed a trusted third party, the players

could surrender to him their inputs 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. Such a protocol is considered perfectly secure if no player can learn from his view of the protocol more than what he would have learnt in the idealized setting where the computation is carried out by a trusted third party.

2. PROBLEM DEFINITION

In our problem, 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 and confidence level. In generic solutions rely upon a description of the function as a Boolean circuit they can be applied only to small inputs and functions which are realizable by simple circuits. In more complex settings, such as ours, other methods are required for carrying out this computation. In such cases, some relaxations of the notion of perfect security might be inevitable when looking for practical protocols, provided that the excess information is deemed benign. And Fast Distributed Mining is not secure one as well as the Secure multi party algorithm which perform the mining slowly. So that the computation cost and communication cost are overhead.

3. FAST ALGORITHMS FOR MINING ASSOCIATION RULES IN LARGE DATABASES

Here we consider the problem of discovering association rules between items in a large database of sales transactions. We present two new algorithms for solving this problem that are fundamentally different from the known algorithms. Empirical evaluation shows that these algorithms outperform the known algorithms by factors ranging from three for small problems to more than an order of magnitude for large problems. We also show how the best features of the two proposed algorithms can be combined into a hybrid algorithm, called Apriori Hybrid. Scale-up experiments show that Apriori Hybrid scales linearly with the number of transactions. Apriori Hybrid also has excellent scale-up properties with respect to the transaction size and the number of items in the database.

4. PRIVACY-PRESERVING DATA MINING

Here we address the issue of privacy preserving data mining. Specifically, we consider a scenario in which two parties owning confidential databases wish to run a data mining algorithm on the union of their databases, without revealing any unnecessary information. Our work is motivated by the need to both protect privileged information and enable its use for research or other purposes. The above problem is a specific example of secure multi-party computation and as such, can be solved using known generic protocols. However, data mining algorithms are typically complex and, furthermore, the input usually consists of massive data sets. The generic protocols in such a case are of no practical use and therefore more efficient protocols are required. We focus on the problem of decision tree learning with the popular ID3 algorithm. Our protocol is considerably more efficient than generic solutions and demands both very few rounds of communication and reasonable bandwidth.

5. A FAST DISTRIBUTED ALGORITHM FOR MINING ASSOCIATION RULES

With the existence of many large transaction databases, the huge amounts of data the high scalability of distributed system 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 the interesting relationship between locally large and globally large itemset and proposes an interesting distributed association rule mining algorithm and Using FDM which has higher performance.

6. PRIVACY-PRESERVING ISTRIBUTED MINING OF ASSOCIATION RULES ON HORIZONTALLY PARTITIONED DATA

Data mining can extract important knowledge from large data collections – but sometimes these collections are split among various parties. Privacy concerns may prevent the parties from directly sharing the data, and some types of information about the data. This paper addresses secure mining of association rules over horizontally partitioned data. The methods incorporate cryptographic techniques to minimize the information shared, while adding little overhead to the mining task.

7. THE ROUND COMPLEXITY OF SECURE PROTOCOLS

In a network of n players, each player i having private input Xi, we show the players can collaboratively evaluate a function f(Xi,….Xn) in a way that does not compromise the privacy of the players inputs and yet requires only a constant number of round of interaction. The underlying model of computation is a complete network of private channels, with broadcast and a majority of the players must behave honestly. Here the solution assumes the existence of a one-way function.

8. ALGORITHM AND TECHNIQUES USED

• Fast Distributed Mining Algorithm (FDM).

• Secure Multi party Algorithm for Secure mining of frequent item set.

8.1 CREATION OF DATABASE

In this module we are creating a database by storing the various types of items. So that users can give request to the items and can find the item sets.

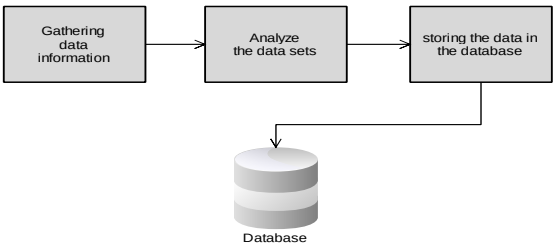

Fig 8.1.1

8.2 CREATION OF USER INTERFACE

Here we can create the user interface to give the request to the database. In the user interface there is a list of item sets will be listed so that users can select the item sets. Also they can give the personal data via user interface that can store into the database.

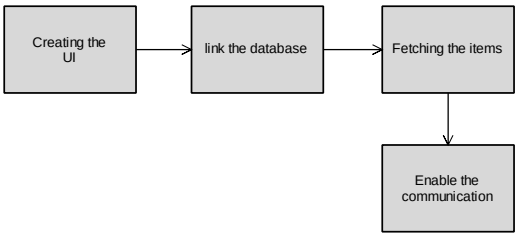

Fig 8.2.1

8.3 APPLYING FAST DISTRIBUTED ALGORITHM

The Fast Distributed Mining algorithm is an unsecured distributed version of the Apriori algorithm. Its main idea is that any s-frequent item set must be also locally 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.

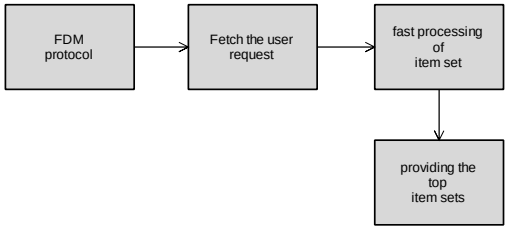

Fig 8.3.1

8.4 IMPLEMENTING SECURE MULTI-PARTY ALGORITHM

We describe our alternative implementation and proceed to analyse the two implementations in terms of privacy and efficiency and compare them. We show that our protocol offers better privacy and that it is simpler and is significantly more efficient in terms of communication rounds, communication cost and computational cost.

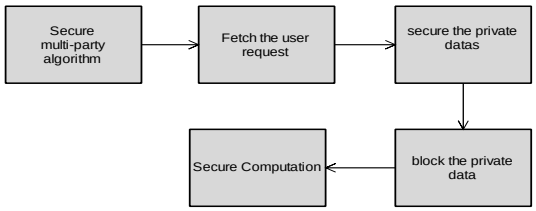

Fig 8.4.1

8.5 INTEGRATING FDM AND SECURE MULTIPARTY ALGORITHM

Here we discuss the implementation of the two remaining steps of the distributed protocol: The identification of those candidate item sets that are globally s-frequent, and then the derivation of all association rules. We describe shortly an alternative protocol, that is a full security at enhanced costs and evaluation which illustrates the significant advantages of our protocol in terms of communication and computational costs.

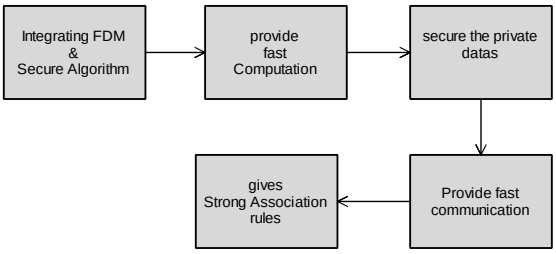

Fig 8.5.1

9. OUTPUT

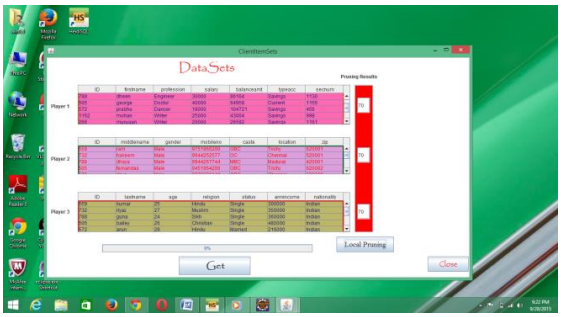

Fig 9.1

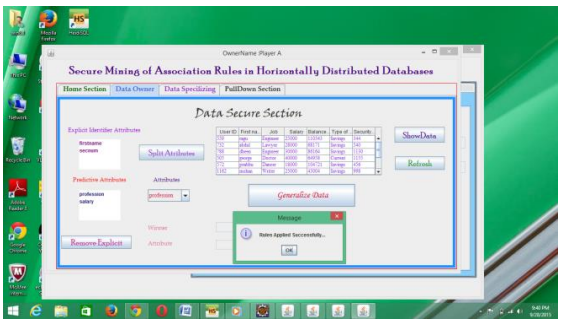

Fig 9.2

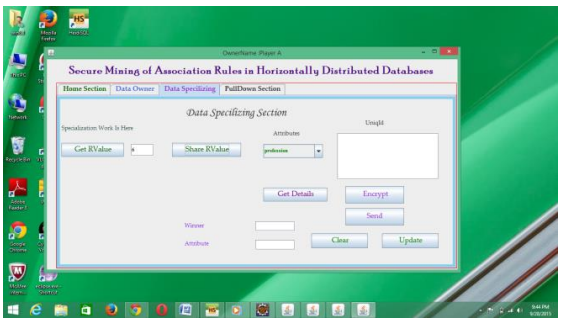

Fig 9.3

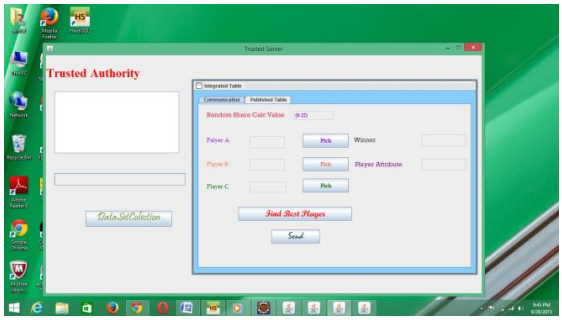

Fig 9.4

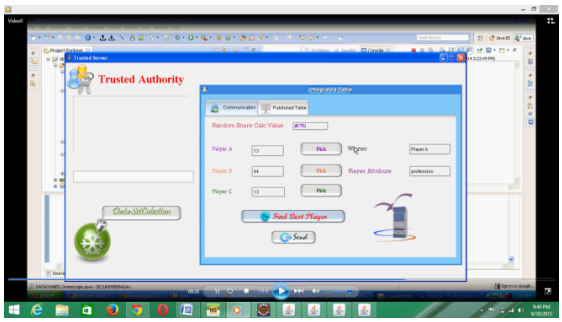

Fig 9.5

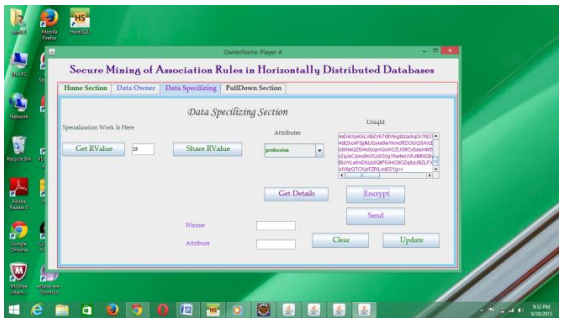

Fig 9.6

10. CONCLUSION

Data mining, the extraction of hidden predictive information from large databases, is a powerful new technology with great potential to help companies focus on the most important information in their data warehouses. Data mining tools predict future trends and behaviors, allowing businesses to make proactive, knowledge-driven decisions. The automated, prospective analyses offered by data mining move beyond the analyses of past events provided by retrospective tools typical of decision support systems. Data mining tools can answer business questions that traditionally were highly time consuming to resolve. They scour databases for hidden patterns, finding predictive information that experts may miss because it lies outside their expectations. 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. We proposed a protocol for secure mining of association rules in horizontally distributed databases that improves significantly upon the current leading protocol in terms of privacy and efficiency. One of the main ingredients in our proposed protocol is a novel secure multi-party protocol for computing the union (or intersection) of private subsets that each of the interacting players holds. Another ingredient is a protocol that tests the inclusion of an element held by one player in a subset held by another.

10. REFERENCES

[1] R. Agrawal and R. Srikant, “Fast Algorithms for Mining Association Rules in Large Databases,” Proc 20th Int’l Conf. Very Large Data Bases (VLDB), pp. 487-499, 1994.

[2] R. Agrawal and R. Srikant, “Privacy-Preserving Data Mining,”Proc. ACM SIGMOD Conf., pp. 439-450, 2000.

[3] D.W.L. Cheung, J. Han, V.T.Y. Ng, A.W.C. Fu, and Y. Fu, “A FastDistributed Algorithm for Mining Association Rules,” Proc. FourthInt’l Conf. Parallel and Distributed Information Systems (PDIS),pp. 31-42, 1996.

[4] 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.

[5] D. Beaver, S. Micali, and P. Rogaway, "The Round Complexity of Secure Protocols," Proc. 22nd Ann. ACM Symp. Theory of Computing (STOC), pp. 503-513, 1990.

[6] M. Bellare, R. Canetti, and H. Krawczyk, "Keying Hash Functions for Message Authentication," Proc. 16th Ann. Int'l Cryptology Conf. Advances in Cryptology (Crypto), pp. 1-15, 1996.

[7] A. Ben-David, N. Nisan, and B. Pinkas, "FairplayMP - A System for Secure Multi-Party Computation," Proc. 15th ACM Conf. Computer and Comm. Security (CCS), pp. 257-266, 2008.

[8] J.C. Benaloh, "Secret Sharing Homomorphisms: Keeping Shares of a Secret," Proc. Advances in Cryptology (Crypto), pp. 251-260, 1986.

[9] J. Brickell and V. Shmatikov, "Privacy-Preserving Graph Algorithms in the Semi-Honest Model," Proc. 11th Int'l Conf. Theory and Application of Cryptology and Information Security (ASIACRYPT), pp. 236-252, 2005.

[10] M. Kantarcioglu, R. Nix, and J. Vaidya, "An Efficient Approximate Protocol for Privacy-Preserving Association Rule Mining," Proc. 13th Pacific-Asia Conf. Advances in Knowledge Discovery and Data Mining (PAKDD), pp. 515-524, 2009.