ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 3, March 2015

Efficient Privacy Preserving Secure ODARM Algorithm in Horizontally Distributed Database

Priyanka.G[1], Premkumar.M[2]

PG Scholar, Department of Computer Science and Engineering, Sri Shanmugha College of Engineering and Technology, Sankari, India[1]

Assistant Professor, Department of Computer Science and Engineering, Sri Shanmugha College of Engineering and Technology, Sankari, India[2]

ABSTRACT: Data mining is the most fast growing area today which is used to extract important knowledge from large data collections but often these collections are divided among some parties. Privacy liability may prevent the parties from directly sharing the data and some types of information about the data. The security is major problem in larger database while sharing the data to the network against the unauthorized access. Proposed an Optimized Distributed Association Rule Mining (ODARM) algorithm. The ODARM algorithm helps to provide high security and generate a frequent itemset in distributed server. Proposed system also concentrate on privacy issues, when the data is distributed in multiple servers and no original server wishes to provide their own data to their server. This paper proposed a new model which utilizes the cryptographic hash techniques to produce a privacy and security in horizontally and vertically distributed database. Our proposed result shows that horizontally distributed database is secure than vertical distributed database and also computation and communication cost is more efficient.

KEYWORDS: Privacy Preserving data mining, Association Rule mining, ODARM Algorithm and Frequent Item Set.

I. INTRODUCTION

The developments of computed technology in last few decades are used to handle large scale data that includes large transaction financial data, 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. 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, the different users can access it without interfering with another. In distributed environment, database is partitioned into disjoint fragments and each site consists of only one fragment. Data can be partitioned in three ways, that is, horizontal partitioning, vertical partitioning and mixed partitioning. Again the details are discussed.

1.1 Partitioning of Database

Data can be partitioned in three ways that is, like horizontally partitioned data, vertically partitioned data or mixed partitioned data.

1.2 Horizontal partitioning:

The data can be partitioned horizontally where each fragment consists of a subset of the records of relation R. Horizontal partitioning divides a table into more tables. The tables have been partitioned in such a way that query references are done by using least number of tables else excessive UNION queries are used to merge the tables sensibly at query time that can affect the performance.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

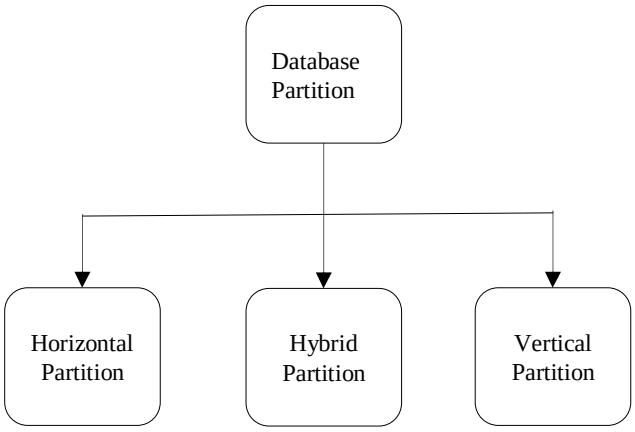
Vol. 3, Issue 3, March 2015

1.3 Vertical partitioning:

The data can be divided into a set of physical files each having the subset of the original relation, the relation is the database transaction that normally requires the subsets of the attributes.

1.4 Mixed partitioning:

The data is first partitioned into horizontally and each partitioned fragment is further partitioned into vertical fragments and vice versa.



Here propose an apriori algorithm for the secure computation of the union of private subsets. The proposed algorithm improves upon that in terms of simplicity and efficiency as well as privacy. In particular, our algorithm does not depend on commutative encryption and oblivious transfer.

The section II describes the related works about this project and section III describes proposed system and its implementation and finally section IV describes experimental result and its discussion.

II. RELATED WORKS

H. Grosskreutz, B. Lemmen, and S. R¨ uping proposed asupervised descriptive rule discovery techniques like subgroup discovery are quite popular in applications like fraud detection. Compared with other descriptive techniques, such as classical support/confidence association rules and subgroup discovery has the advantage that comes up with the top-k patterns and that it makes use of a quality function that avoids patterns uncorrelated with the target. These techniques are to be applied in privacy-sensitive scenarios involving distributed data, the precise guarantees are needed regarding the amount of information leaked during the execution of the data mining. Unfortunately, adaptation of secure multi-party protocols for classical support/confidence association rule mining to the task of subgroup discovery is impossible for fundamental reasons. Source is the different quality function and the restriction to a fixed number of patterns. Present new protocols which allow distributed subgroup discovery while avoiding the disclosure of the individual databases. Analyze the properties of the protocols; describe a prototypical implementation and present experiments that demonstrate the feasibility of the approach.

- ✓ The system is designed to discover subgroups in fraud detection and clinical studies
- ✓ Secure Top-I subgroup discovery protocol is used to fetch subgroups with security
- ✓ Privacy rate is improve in the system
- ✓ Vertical partition data model is not supported.

D.W.L Cheung, V.T.Y. Ng, A.W.C. Fu, and Y. Fu. Proposed a Many sequential algorithms have been proposed for the mining of association rules. Very little work has been done in mining association rules in distributed databases. Direct application of sequential algorithms to distributed databases is not effective, it requires a large amount of communication overhead. An efficient algorithm called DMA (Distributed Mining of Association rules), proposed. It generates a small number of candidate set and it requires only O (n) messages for support-count exchange for each candidate set, here n is the number of sites in a distributed database. The algorithm has been implemented on testbed, and its performance is studied. Results show that DMA has superior performance, when it compared with the direct application of popular consecutive algorithm, in distributed databases.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 3, March 2015

Rakesh Agrawaland Ramakrishnan Srikant Ibmalmaden Research Center proposed a problem of discovering association rules between items in a large database of sales transactions. Present two new algorithms for solving this problem that are fundamentally different from the known algorithms. The 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. 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.

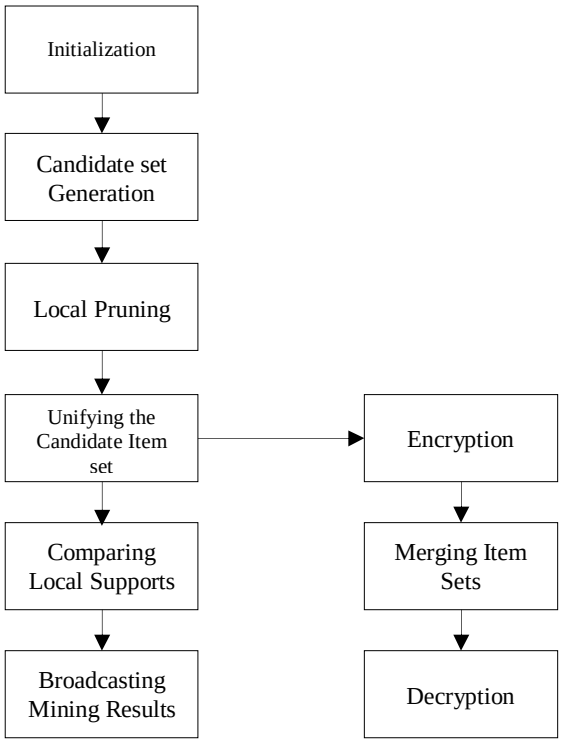
Presented two algorithms, Apriori and AprioriTidfor discovering significant association rules between items in a large database of transactions. Compared these algorithms to the previously known algorithms, the AIS and SETM algorithms. Presented experimental results, showing that proposed algorithms always outperform AIS and SETM. The performance gap increased with the size, and ranged from a factor of three for small problems to more than an order of magnitude for large problems.

III. PROPOSED SYSTEM

3.1 Architecture Diagram

System architecture describes the flow of data inside the system. It goes through various phases as shown in figure. It is having initialization, in which the user is starting their role by holding some value (money or balance) in it. And then it will help to find out the next item. Next phase is generating candidate set, in which are finding the key which appears repeatedly or may say it which is intersection or common for both sites and users.

Next phase is local pruning, in which 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 users. Next phase is local support computation, in which are computing the local support that how much the participating user can support. Next phase is broadcasting of the mining result in which are going to display the result by merging the all result that got from all participating user and then displaying it.



ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 3, March 2015

Propose an ODARM (Optimized Distributed Association Rule Mining)algorithm for the secure computation of the union of private subsets. The proposed algorithm improves upon that in terms of simplicity and efficiency as well as privacy. In particular,ODARMalgorithm does not depend on commutative encryption and oblivious transfer (what simplifies it significantly and contributes towards much reduced communication and computational costs).The algorithm that propose here computes a parameterized family of functions, which call threshold functions, in which two extreme cases correspond to the problems of computing the union and intersection of private subsets. Those are in fact general-purpose algorithm that can be used in other contexts as well. The ODARM also gives a high security and accuracy. For privacy here used general cryptographic functions.
The following modules are implemented in proposed system.

3.2 Modules:

3.2.1 User Module

Privacy preserving data mining has considered two related settings. Data owner and Data miner are two different entities, in which the data is distributed among several parties who aim to jointly perform data mining on the unified corpus of data that they hold.

In that first setting, the goal is to protect the data records from the data miner. The data owner aims at anonymizing the data prior to its release. The main approach is to apply data perturbation. Perturbed data can be used to conclude general trends in the data, without revealing unique record information.

In that second setting, the goal is to perform data miningwhile protecting the data records of each of the data ownersfrom the other data owners.

3.2.2 Admin Module

In this module, is used to view user details. Admin is used to view the item set based on the user processing details using association role with Apriori algorithm.

3.2.3Association Rule

Association Rule mining is one of the most important data mining tools used in many real life applications. It is used to reveal unexpected relationships in the data. Will discuss the problem of computing association rules within a horizontally partitioned database. Assume homogeneous databases. Sites have the same schema, but each site has different information on different entities. The main objective is to produce association rules that hold global information, while limiting the information shared about each site to preserve the privacy of data in each site.
Association rule is used if/then statements that help uncover relationships between seemingly unrelated data in a relational database or other information repository.
Association rules are created by analyzing data for frequent if/then patterns and using the criteria support and confidence to identify the most important relationships. Its support is an indication of how frequently the items appear in the database. The confidence indicates the number of times the if/then statements have been found to be true.
Association rule mining finds interesting associations and/or correlation relationships among large sets of data items. The Association rules show attributes value conditions that occur frequently together in a given dataset.

3.2.4Apriori Algorithm

Apriori is designed to operate on databases containing transactions. Apriori Algorithm is used to find associations between different sets of data. Apriori Algorithm referred to as "Market Basket Analysis". Each set of data has a number of items and is called a transaction. Output of Apriori is sets of rules that tell us how often items are contained in sets of data.

The Apriori Algorithm proposed to finds frequent items in a given data set using the ant monotone constraint. Apriori is an important algorithm in market basket analysis for mining frequent item sets for Boolean association rules. The name of Apriori Algorithm is based on the fact that the algorithm uses a prior knowledge of frequent itemset properties. Apriori employs an iterative approach known as a level wise search, where k item sets are used to explore (k+1) itemset .Apriori algorithm is an in fluential algorithm for mining frequent itemset for Boolean association rules. Apriorialgorithm contains a number of passes over the database. In pass k, the algorithm finds the set of frequent itemsetLk of length k that satisfy the minimum support requirement.

Apriori is designed to operate on databases containing transactions. The Apriori Algorithm is used to find associations between different set of data. Apriori Algorithm is referred to as "Market Basket Analysis". Each set of data has a number of items and is called a transaction. The result of Apriori is sets of rules that tell us how often items are

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 3, March 2015

contained in sets of data. Verification if the auditor is convinced with the data integrity; the auditor erases the local data.

3.3 ODARM Algorithm

There are following some steps that are going to use in this scheme.

Step 1:

All the users generate keys using following key generation method. Key Generation: Let k be the security parameter that chooses two randomly k-bit prime numbers p and q. then set N=pq. Choose random base g Є B.

Step 2:

User then jointly calculates Fsk-1.

Step 3:

Each user encrypts Fsk-1 using following encryption method.

Let c = cypher text.

Where c= gmrnmod N2

Where r= random value, r Є z*n

Step 4:

Each user Pm computes (k-1) item sets that are locally frequent in his site and also globally frequent Pm thencomputes Fsk-1, m^ Fsk,m. He then uses this to generate Bsk,mof candidate k item set and encrypt bits using step 3equation.

Step 5:

For each X Є Bsk,m, Pm computes suppm(X) and encrypt it using step 3 equation. He then retains only those itemsets that are locally s frequent.

Step 6:

Each user broadcast his encrypted Cs

k,m and then all user computes Cs

k :=Um=1

M Csk,m

Step 7:

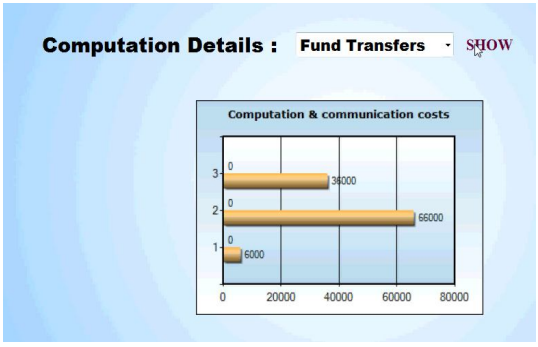
Computing local support is now done by all users

Step 8:

Each user broadcast the local support that he computed and encrypts it before sending from that everyone cancomplete global support of every item set Csk.

IV.RESULT AND DISCUSSION

- ODARM Algorithm provides high security in HDDB.
- The Horizontal distributed database is more secure than Vertical distributed database.
- To get Efficient item set based on the customer request.



Computation and Communication Cost is very efficient

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 3, March 2015

IV. CONCLUSION

The main threat in finding association rule mining in horizontally distributed database environment is privacy that is no site owner wish to provide database or local frequent item sets or support value to anyone. However every owner wishes to access mined result by participating indirectly in the mining process by providing partial results in disguised form. The problem of preserving privacy in association rule mining when the database is distributed horizontally among n (n> 2) number of sites with a trusted party is considered. The proposed system finds global frequent item.

The direction of future work is to devise an efficient protocol for inequality verifications that uses the existence of semi-honest third party and another in Implementation of the techniques to the problem of distributed association rule mining in vertical setting.

REFERENCES

[1] R. Agrawal and R. Srikant, "Fast Algorithms for Mining Associa- tion Rules in Large Databases," Proc 20th Int'l Conf. Very Large Data Bases (VLDB), pp. 487-499, 1994.
[2] 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.
[3] A. Ben-David, N. Nisan, and B. Pinkas, "FairplayMP - A System for Secure Multi-Party Computation," Proc. 15th ACM Conf. Com- puter and Comm. Security (CCS), pp. 257-266, 2008.
[4] 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.
[5] D.W.L. Cheung, J. Han, V.T.Y. Ng, A.W.C. Fu, and Y. Fu, "A Fast Distributed Algorithm for Mining Association Rules," Proc. Fourth Int'l Conf. Parallel and Distributed Information Systems (PDIS), pp. 31-42, 1996.
[6] D.W.L Cheung, V.T.Y. Ng, A.W.C. Fu, and Y. Fu, "Efficient Mining of Association Rules in Distributed Databases," IEEE Trans. Knowledge and Data Eng., vol. 8, no. 6, Dec. 1996.
[7] A.V. Evfimievski, R. Srikant, R. Agrawal, and J. Gehrke, "Privacy Preserving Mining of Association Rules," Proc. Eighth ACM SIGKDD Int'l Conf. Knowledge Discovery and Data Mining (KDD), pp. 217-228, 2002.
[8] 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.
[9] M. Kantarcioglu, R. Nix, and J. Vaidya, "An Efficient Approxi- mate Protocol for Privacy-Preserving Association Rule Mining," Proc. 13th Pacific-Asia Conf. Advances in Knowledge Discovery and Data Mining (PAKDD), pp. 515-524, 2009.
[10] Y. Lindell and B. Pinkas, "Privacy Preserving Data Mining," Proc. Crypto, pp. 36-54, 2000.
[11] J.S. Park, M.S. Chen, and P.S. Yu, "An Effective Hash Based Algorithm for Mining Association Rules," Proc. ACM SIGMOD Conf., pp. 175-186, 1995.
[12] R.L. Rivest, A. Shamir, and L.M. Adleman, "A Method for Obtaining Digital Signatures and Public-Key Cryptosystems," Comm. ACM, vol. 21, no. 2, pp. 120-126, 1978.