Optimization of Association Rule in Horizontally Distributed Database using Unique Key Value

Madhuri Mahajan
Department of Computer Engineering
Raisoni Collage of Engineering, Jalgaon,
Maharashtra, India

Sonal Patil
Department of Computer Engineering
Raisoni Collage of Engineering, Jalgaon,
Maharashtra, India

ABSTRACT
The proposed work describes the optimization of Association Rule for the distributed databases in terms of speed, memory used while transaction of distribution as well as extracting the data from various data sources in the network. The proposed work shall have two parts including distribution of data using Association rule and ensuring the search to be redirected to specific source based on key values used to create the sub set in association rule. The availability shall be tested by verifying if the specific source is ready or not if not the search for that part shall only be carried out on the server itself.

Keywords
Distributed Databases, Association rule, Database Mining, Database Security.

1. INTRODUCTION
A distributed database is a database in which storage devices are not all attached to a common processing unit such as the CPU. which is controlled by a distributed database management system (together sometimes called a distributed database system). Association rule learning is a method for discovering interesting relations between variables in large databases. It is intended to identify strong rules discovered in databases using different measures of interestingness. The proposed is the work including association rule in horizontally distributed databases in terms of homogeneous database. Prior to the existence of distributed databases the centralized databases were used in which the database is get stored in the centralized server and every node hits the query to the same database server and speed of data retrieval does get affected to solve the issue distributed databases comes into picture . The use of association rule helps to divide the entire database into different distinct computers so as to make it available all the time while the request shall be processed on a small set of data instead of entire large data. information secrete and secure. These databases help to improve the speed of retriaval and helps us to retrive the stored information from a very large databse with very less amount of time.

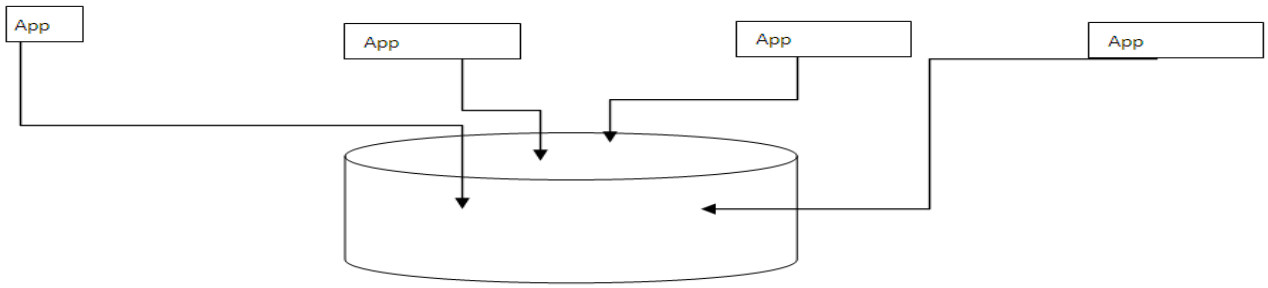

Fig. 1 Centralised Database

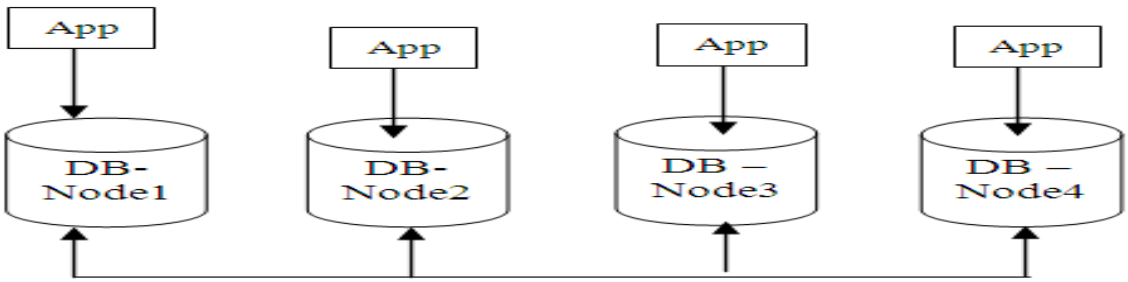

Fig. 2 Deistributted Databses

There are pros and cons of using the distributted databases. Main advantages of using the distributted databse includes more reliability and availability it also helps to keep information secrete and secure. These databases help to improve the speed of retriaval and helps us to retrive the stored information from a very large databse with very less amount of time. The major issues with these type of databses is that they are very complex in structure and they are difficult

to maintain the integrity. However using of distributted database is always a great help the only concern is to how the information shall be grouped and arranged on different machines and shall be helpful in fast retrival of information. In that regards Association rule can be used. The association rull will help to group the entire data into small segments(fragments) and those fragments can then be stored in the different computers. The use of association will try to maintain the normalisation of databses while making it in distrubutted. The proposed paper will help you to understand the use of association rule in distributted databases. The proposed method will optimise the entire process.

2. BACKGROUND

The problem of secure mining of association rules in horizontally partitioned databases. In that setting, there are several sites (or players) that hold homogeneous databases, i.e., databases that share the same schema but hold information on different entities. 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. Several papers have been already published in this area describing the use of data mining in the horizontally distributed databases. A paper published in April 2014 issue of international journal of engineering and advanced technology written by Sayad Shujaubuddin Sameer writes about secures sensitive data based on the association rule as Privacy preservation is important for data mining and other learning techniques. There is a need for different approaches required in this scenario. A fruitful direction for future data mining research will be the development of techniques that incorporate privacy. Rakesh Agrawal And Ramkrishnan Shrikant at IBM Almaden Research Center describes the fast algorithm for mining association rules to be used inside the distribution of data to various computers in the network. Progress in bar-code technology has made it possible for retail organizations to collect and store massive amounts of sales data, referred to as the basket data[2]. M. saraaswati and N. Kowsalya in recent paper about privacy preserving and data secure mining of association rule in distributed rule published in January 2015 issue of international journal of computer science and mobile computing Most existing parallel and distributed ARM algorithms are based on a kernel that employs the well-known Apriori algorithm. Directly adapting an Apriori algorithm will not significantly improve performance over frequent item sets generation or overall distributed ARM performance.

3. PROBLEM STATEMENT

In Proposed System, propose an alternative protocol for the secure computation of the union of private subsets. The proposed protocol improves upon that in terms of simplicity and efficiency as well as privacy. In particular, our protocol does not depend on commutative encryption and oblivious transfer (what simplifies it significantly and contributes towards much reduced communication and computational costs). While our solution is still not perfectly secure, it leaks excess information only to a small number (three) of possible coalitions, unlike the protocol of that discloses information also to some single players. In addition, we claim that the excess information that our protocol may leak is less sensitive than the excess information.

4. PROPOSED METHODOLOGY

As per the problem statement description the entire system can be viewed in 4 different modules to incorporate the entire division of database in to horizontally distributed database using association rule.
User Module.
Admin Module.
Association Rule.

Apriori Algorithm.

4.1 User Module

In this module, privacy preserving data mining has considered two related settings. One, in which the data owner and the data miner are two different entities, and another, 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 the initial setting, the goal is to protect the data records from the data miner. Hence, the data owner aims at anonymizing the data prior to its release. The main approach in this context is to apply data perturbation. He perturbed data can be used to infer general trends in the data, without revealing original record information. In the second setting, the goal is to perform data mining while protecting the data records of each of the data owners from the other data owners.

4.2 Admin Module

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

4.3 Association Rule

Association rules are if/then statements that help uncover relationships between seemingly unrelated data in a relational database or other information repository. An example of an association rule would be "If a customer buys a dozen eggs, he is 80% likely to also purchase milk." 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. Support is an indication of how frequently the items appear in the database. Confidence indicates the number of times the if/then statements have been found to be true.

4.4 Apriori Algorithm

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

Algorithm - Fast Distributed Mining (FDM)

The FDM algorithm proceeds as follows:

1) Initialization
2) Candidate Sets Generation
3) Local Pruning
4) Unifying the candidate item sets
5) Computing local supports
6) Broadcast Mining Results

5. EXPECTED RESULTS

The proposed system shall reduce the total computation time of data distribution. It shall be tested on the data with function evaluating to N while keeping the the size of data and elements inside each data shall be changing. The estimated

time shall not be changed inversely proportional to the change in size or element size.

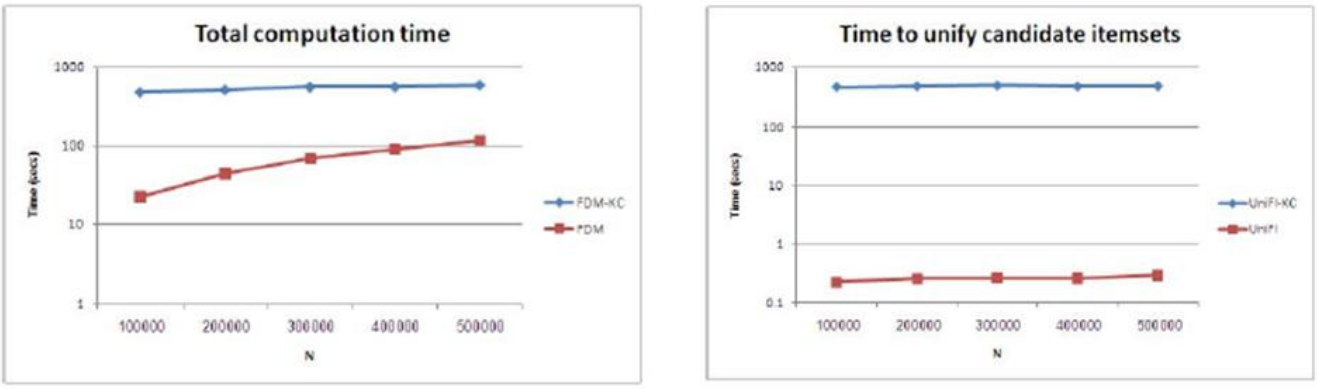

Fig.3 Expected Result Graph

6. CONCLUSION

In this paper I am presenting a method for secure mining of association rules in horizontally distributed databases for the improvement over the current leading protocol in terms of privacy and efficiency. One of the main modules in this methodology is a high secure multi-party protocol for computing the union (or intersection) of private subsets that each of the interacting players hold. Another sub module tests the inclusion of an element held by one player in a subset held by another. Those methodology exploit the fact that the underlying problem is of interest only when the number of players is greater than two.

7. ACKNOWLEDGEMENT

I wish to express my deep sincere of gratitude to my guide Miss Sonal Patil of guide for her grateful efforts to inoculate me & their direction encouraged me to complete this project report and only due to the illuminative, it became possible for me to study the whole process in detail. As an computer engineer student, it was very fulfill to take advantage studying various Web technology, Software and Networking tricks including troubleshooting of the overall system and software currently in use. On the grateful occasion, I am very thankful to acknowledgement to Mr. P.P. Rewagad (HOD, Computer Engineering) . I am also thankful to the staff members of Computer Science and Engineering Department for their highly co-operative & encouraging attitudes which have always boosted us. I am very thankful to all of them for inspiring me towards ensuring and retaining the quality of project report.

8. REFERENCES

[1] R. Agrawal and R. Srikant. Fast algorithms for mining association rules in large databases. 1994.

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

[3] D. Beaver, S. Micali, and P. Rogaway. The round complexity of secure protocols. In *STOC*, pages 503–513, 1990.

[4] M. Bellare, R. Canetti, and H. Krawczyk. Keying hash functions for message authentication. In *Crypto*, pages 1–15, 1996.

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

[6] J.C. Benaloh. Secret sharing homomorphisms: Keeping shares of a secret. In *Crypto*, pages 251–260, 1986.

[7] J. Brickell and V. Shmatikov. Privacy-preserving graph algorithms in the semi-honest model. In *ASIACRYPT*, pages 236–252, 2005.

[8] 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. In *PDIS*, pages 31–42, 1996.

[9] 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. Knowl. Data Eng.*, 8(6):911–922, 1996.

[10] T. ElGamal. A public key cryptosystem and a signature scheme based on discrete logarithms. *IEEE Transactions on Information Theory*, 31:469–472, 1985.