International Journal of Science and Research (IJSR)
ISSN (Online): 2319-7064
Index Copernicus Value (2013): 6.14 | Impact Factor (2013): 4.438

# Mining of Association Rules in Distributed Databases

**Nayana Marodkar**[1], **Manoj Chaudhari**[2]

[1]M.Tech Student of Computer Science and Engineering, RTMN University, PBCOE, Nagpur, Maharashtra, India

[2]Professor, H.O.D of Computer Science and Engineering, RTMN University, PBCOE, Nagpur, Maharashtra, India

**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 several parties. Association rule mining is one of the techniques in data mining. Here , we propose a protocol for mining of association rules in horizontally distributed databases and protocol is based on the Fast Distributed Mining (FDM) algorithm which is an unsecured distributed version of the Apriori algorithm. The main ingredients in protocol are two novel secure multi-party algorithms — one that computes the union of private subsets that each of the interacting players hold, and another that tests the inclusion of an element held by one player in a subset held by another. Our protocol offers enhanced privacy with respect to the protocol. In addition, it is simpler and is significantly more efficient in terms of communication rounds, communication cost and computational cost.***

**Keywords:** Apriori Algorithm, Association Rule, Distributed Database; FDM, secure multi-party algorithms

## 1. Introduction

Data mining can extract important knowledge from large data collections but sometimes these collections are split among various parties. Data mining is defined as the method for extracting hidden predictive information from large distributed databases. It is new technology which has emerged as a means of identifying patterns and trends from large quantities of data. The final product of this process being the knowledge, meaning the significant information provided by the unknown elements.Here we study the problem of mining of association rules in horizontally partitioned databases. There are several sites that hold homogeneous databases, i.e., databases that share the same schema but hold information on different entities [1]. With given minimal support and confidence levels that hold in the unified database the goal is to find all association rules, while minimizing the information disclosed about the private databases held by those players. That goal defines a problem of secure multi-party computation. The information that would like to protect in this proposed work, not only individuals transaction but also more global information such as association rules that are supported locally in each of these database .In such problems, there are M players that hold private inputs, x1, . . . , $x_M$, and they wish to securely compute y = f(x1, . . . , $x_M$) for some public function f. 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. It is needed to devise a protocol that in the absence of such a trusted third party the players can run on their own in order to arrive at the required output y [1]. Then such a devised protocol is considered 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.

In proposed system is, the inputs are the partial databases, and the required output is the list of association rules with given support and confidence. As the above mentioned generic solutions rely upon a description of the function f as a Boolean circuit, they can be applied only to small inputs and functions which are realizable by simple circuits. In more complex settings, 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.

Kantarcioglu and Clifton studied that problem where more suitable security definitions that allow parties to choose their desired level of security are needed, to allow effective solutions that maintain the desired security and devised a protocol for its solution[2]. The main part of the protocol is a sub-protocol for the secure computation of the union of private subsets that are held by the different players. That is the most costly part of the protocol and its implementation relies upon crypto-graphic primitives such as commutative encryption, oblivious transfer, and hash functions. This is also the only part in the protocol in which the players may extract from their view of the protocol information on other databases, beyond what is implied by the final output and their own input. While such leakage of information renders the protocol not perfectly secure, the perimeter of the excess information is explicitly bounded in and it is argued that such information leakage is innocuous, whence acceptable from practical point of view.

In this we 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, protocol does not depend on cryptographic primitive i.e. commutative encryption and oblivious transfer. While the solution is still not perfectly secure, it leaks excess information only to a small number of coalitions (three), unlike the protocol of that discloses information also to some single players. In addition, also claim that the excess information that our protocol may leak is less sensitive than the excess information leaked by the protocol.

The protocol that we propose 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.

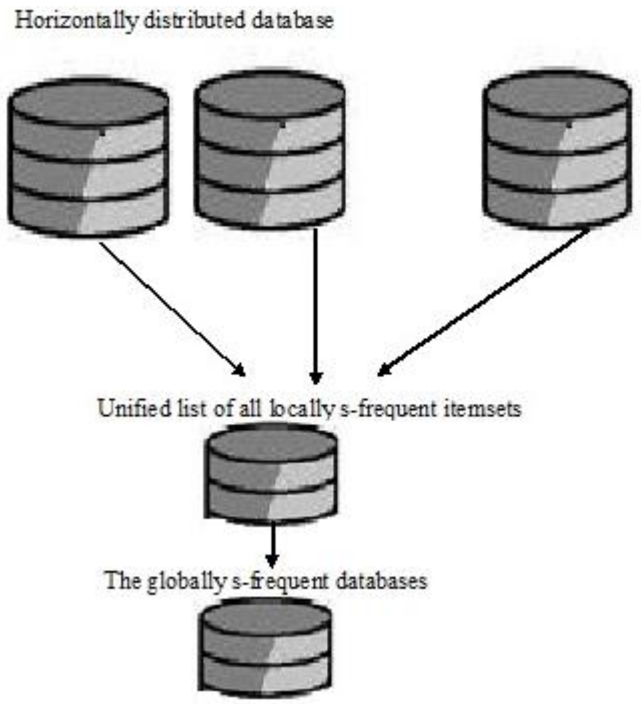
Horizontally distributed database
Unified list of all locally s-frequent itemsets
The globally s-frequent databases

**Figure 1:** Architecture of Distributed Database

## 2. Data mining

Data mining is the process of extracting hidden patterns from data. Data mining is becoming an increasingly important tool to transform this data into knowledge. Data mining can be applied to data sets of any size, and while it can be used to uncover hidden patterns, it cannot uncover patterns which are not already present in the data set. It is commonly used in a wide range of applications, such as marketing, fraud detection and scientific discovery[5]. Data mining extracts novel and useful knowledge from data and has become an effective analysis and decision means in corporation. Large repositories of data contain private data and sensitive rules that must be preserved before published. Motivated by the multiple conflicting requirements of data sharing, privacy preserving and knowledge discovery, privacy preserving data mining has become a research hotspot in data mining and database security fields. There are Two problems in PPDM: one is the protection of private data; another is the protection of sensitive rules (knowledge) contained in the data. The former settles how to get normal mining results when private data cannot be accessed accurately; the latter settles how to protect sensitive rules contained in the data from being discovered, while non-sensitive rules can still be mined normally [6].

Data mining methodology has emerged as a means of identifying patterns and trends from large quantities of data. Data mining go hand in hand: most tools operate by gathering all data into a central site, then running an algorithm against that data. Many people take data mining as a synonym for another popular term, Knowledge Discovery in Database (KDD). Alternatively other people treat data mining as the core process of KDD. Usually there are three processes. One is called pre-processing, which is executed before data mining techniques are applied to the right data. The pre- processing includes data cleaning, integration, selection and transformation [7]. The main process of KDD is the data mining process, in this process different algorithms are applied to produce hidden knowledge. After that comes another process called post-processing, which evaluates the mining result according to users requirements and domain knowledge. Regarding the evaluation results, the knowledge can be presented if the result is satisfactory, otherwise we have to run some or all of those processes again until we get the satisfactory result.

The most commonly used techniques in data mining are:

- Clustering:
- Associations Rule:
- Sequential patterns
- Artificial neural networks
- Genetic algorithms
- Decision trees:
- Nearest neighbour method
- Rule induction:
- Data visualization

## 3. Distributed Database

A distributed database system consists of loosely coupled sites that share no physical component n Database systems that run on each site are independent of each other n Transactions may access data at one or more sites. A distributed database management system (DDBMS) is the software that manages the DDB and provides an access mechanism that makes this distribution transparent to the users [9].

A distributed database is a database in which storage devices are not all attached to a common processing unit such as the CPU, controlled by a distributed database management system (together sometimes called a distributed database system)[11]. It may be stored in multiple computers, located in the same physical location; or may be dispersed over a network of interconnected computers. Unlike parallel systems, in which the processors are tightly coupled and constitute a single database system, a distributed database system consists of loosely-coupled sites that share no physical components. Types of distributed database

- Homogeneous distributed database
- Heterogeneous distributed database

In a homogeneous distributed database all sites have identical software and are aware of each other and agree to cooperate in processing user requests. Each site surrenders part of its autonomy in terms of right to change schema or software. A homogeneous DDBMS appears to the user as a single system. The homogeneous system is much easier to design and manage.

A heterogeneous database system is an automated (or semi-automated) system for the integration of heterogeneous, disparate database management systems to present a user with a single, unified query interface. Heterogeneous database systems (HDBs) are computational models and software implementations that provide heterogeneous database integration.

## 4. Related Work

Association rule mining finds interesting associations and/or correlation relationships among large sets of data items. Association rules show attributes value conditions that occur frequently together in a given dataset. The market basket analysis used association rule mining in distributed environment.

Association rule mining is used to find rules that will predict the occurrence of an item and based on the occurrences of other items in the transaction [10], search patterns gave association rules where the support will be counted as the fraction of transaction that contains an item X and an item Y and confidence can be measured in a transaction the item i appear in transaction that also contains an item X.

The Apriori Algorithm proposed to finds frequent items in a given data set using the ant monotone constraint. Apriori is an influential algorithm in market basket analysis for mining frequent item sets for Boolean association rules. This algorithm in [13] contains a number of passes over the database. During pass k, the algorithm finds the set of frequent itemsets Lk of length k that satisfy the minimum support requirement.

Apriori is designed to operate on databases containing transactions. The purpose of the Apriori Algorithm [11] 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. Association rule mining is used to find rules that will predict the occurrence of an item and based on the occurrences of other items in the transaction, search patterns gave association rules where the support will be counted as the fraction of transaction that contains an item X and an item Y and confidence can be measured in a transaction the item i appear in transaction that also contains an item X.

**Support (s): -** Fraction of transactions that contain both X and Y
Support (X->Y) = P (X∪Y)/T

**Confidence(c): -** Measure show often items in Y appear in transactions that contain X.
Confidence (X->Y) = Support (X∪Y) / Support (X)

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.

Privacy preserving distributed mining of association rule for a horizontally partitioned dataset across multiple sites are computed as follows where I = {i1, i2, .in} be a set of items and T = {T1, T2,…Tn} be a set of transactions where each *T I i* Í . A transaction Ti contains an item set *X* Í *I* only if *i X* Í *T* . An association rule implication is of the form *X*_*Y* (*X* Ç*Y* =0) with support S and confidence C if S% of the transactions in T contains *X* È*Y* and C% of transactions that contain X also contain Y. In a horizontally partitioned database, the transactions are distributed among n sites. Support (*X* _*Y* ) = probe (*X* È*Y* ) /Total number of transaction.

The global support count of an item set is the sum of all local support counts.
Support g(X)=Support1(x)+Support2(x)+……………………+ Support n(x).

Confidence of rule (*X* _*Y* ) = Support (*X* È*Y* ) / Support(X) The global confidence of a rule can be expressed in terms of the global support.

Confidence g (*X* _*Y* ) = Support g (*X* È*Y*) / Support g(X) The basis of this algorithm is the Apriori algorithm that uses K-1 frequent sets.

## 5. The Fast Distributed Mining Algorithm

The protocols are based on the Fast Distributed Mining (FDM) algorithm like in [2], which is an unsecured distributed version of the Apriori algorithm. Its main idea is that any s-frequent itemset must be also locally s-frequent in at least one of the sites. Hence, in order to find all globally s-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. The stages of the FDM algorithm are as follows:

1) Initialization: It is assumed that the players have already jointly calculated Fk−1 s .The goal is to proceed and calculate $F^{k}{}_{s}$ .
2) Candidate Sets Generation: Each Pm generates a set of candidate k- itemsets $B^{k,}{}^{m}{}_{s}$ out of $F^{k-1,m}{}_{s} \cap F^{k-1}{}_{s}$ — the (k − 1)-itemsets that are both globally and locally frequent, using the Apriori algorithm.
3) Local Pruning: For each X ∈ $B^{k,\ m}{}_{s}$. Pm computes supp m(X) and retains only those itemsets that are locally s-frequent. We denote this collection of itemsets by $C^{k,\ m}{}_{s}$ .
4) Unifying the candidate item sets: Each player broadcasts his $C^{k,\ m}{}_{s}$ and then all players compute $C^{k}{}_{s} := S^{M}{}_{m=1}\ C^{k,m}{}_{s}$
5) Computing local supports: All players compute the local supports of all itemsets in $C^{k}{}_{s}$ .

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 $C^k{}_s$ . Finally, $F^k{}_s$ is the subset of $C^k{}_s$ that consists of all globally s-frequent k-itemsets.

With the existence of many large transaction databases, the huge amounts of data, the high scalability of distributed systems, 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 some interesting relationships between locally large and glob-ally large itemsets and proposes an interesting distributed association rule mining algorithm, FDM (Fast Distributed Mining of association rules), which generates a small number of candidate sets and substantially reduces the number of messages to be passed at mining association rules. Our performance study shows that FDM has a superior performance over the direct application of a typical sequential algorithm. Further performance enhancement leads to a few variations of the algorithm.

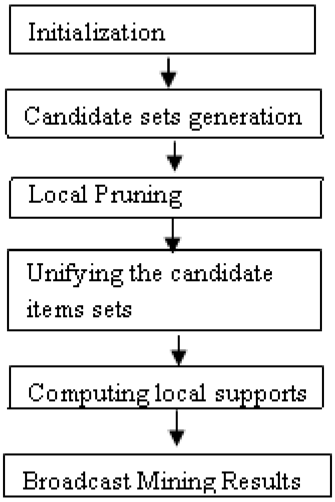

**Figure:** The Process of FDM Algorithm

## 6. Conclusion

The problem of computing association rules within a scenario of homogeneous database. 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 protocols have been implemented. In this, focus is based on horizontal partitioned distributed data through a popular association rule mining technique. Protocols exploit the fact that the underlying problem is of interest only when the number of players is greater than two.

## References

[1] Tamir Tassa,"Secure mining of association rule in horizontally distributed databases" ,IEEE trans. Konwledege and Data Engg.,Vol. 26, no.2, April 2014J.

[2] 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.

[3] Krishna Pratap Rao, Adesh chaudhary, Prashant johri "Elliptic Curve Cryptography Based Algorithm for Privacy Preserving in Data Mining", International Journal for research in Applied Science and

[4] P. Jagannadha Varma, Amruthaseshadri,.M. Priyanka, M.Ajay Kumar, B.L.Bharadwaj Varma, " Association Rule Mining with Security Based on Playfair Cipher Technique" (IJCSIT) International Journal of Computer Science and Information Technologies, Vol. 5 (1), 201

[5] Prof. Geetika. Narang, Anjum Shaikh, Arti Sonawane, Kanchan Shegar, Madhuri Andhale," Preservation Of Privacy In Mining Using Association Rule Technique", International Journal of Scientific & Technology Research, Volume 2, Issue 3, March 2013

[6] Zhi Liu,Tianhong Sunand Guoming Sang," An Algorithm of Association Rules Mining in Large Databases Based on Sampling ", International Journal of Database Theory and Application Vol.6, No.6 , 2013

[7] Priyanka Asthana, Anju Singh , Diwakar Singh," A Survey on Association Rule Mining Using Apriori Based Algorithm and Hash Based Methods ", International Journal of Advanced Research in Computer Science and Software Engineering. Volume 3, Issue 7, July 2013

[8] J. Vaidya and C. Clifton, "Privacy preserving association rule mining in vertically partitioned data," in The Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Edmonton, Alberta, Canada, July 23-26 2002,

[9] A.C. Yao. Protocols for secure computation. In FOCS, pages 160–164, 1982.

[10] Sotiris Kotsiantis, Dimitris Kanellopoulos Association Rules Mining: A Recent Overview GESTS International Transactions on Computer Science and Engineering, Vol.32 (1), 2006, pp. 71-82

[11] T.Kartikeyan and N.Ravikumar A Survey on Association Rule Mining International Journal of Advanced Research in Computer and Communication Engineering Vol. 3, Issue 1, January 2014

[12] Shiny. I.S , S. Gayathri,"Secure Multiparty Computation and Privacy Preserving Data Sharing with Anonymous ID Assignment", International Journal of Engineering Research and Applications (IJERA) ISSN: 2248-9622 International Conference on Humming Bird ,01st March 2014

**[13]** Chitteni Siva, Selvi Secure Mining of Association Rules in Horizontally Distributed Databases International Journal of Computer Science and Mobile Computing A Monthly Journal of Computer Science and Information Technology IJCSMC, Vol. 3, Issue. 4, April 2014, pg.1079 – 1082