Published by :
http://www.ijert.org

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 5 Issue 01, January-2016

A Cryptographic Approach to Mine Unified Frequent Itemsets in Distributed Databases

Shaheen Banu
Senior Grade Lecturer
Computer Science Department
DRR Government polytechnic
Davanagere, Karnataka, India

Asiya Banu B
Senior Grade Lecturer
Computer Science Department
Government polytechnic
Harihara, Karnataka, India

***Abstract*— Discovering meaningful patterns from huge quantities of data lead to an emerging technology called data mining. Many data mining techniques are available to find useful hidden information from huge database, Association Rule Mining is one such data mining technique. Existing protocols determine interesting relationships among itemsets of database without considering the privacy of data. When the data is distributed among multiple sites the data privacy is of key concern. Local data miners will never wish to share their private data to others and are concerned to know the global results found from mining process for decision making. The proposed model uses AES cryptography technique, Apriori algorithm and Secure Mining algorithm to mine distributed databases. Local data miners will pass their encrypted frequent itemsets to the assembler by avoiding leakage of sensitive information, resulting in a unified frequent itemsets. In the proposed model data is directly transferred to the assembler thus eliminating the oblivious data transfers involved among local data miners and hence improves communication time.**

Keywords*— *Mining; Distributed Datases; Unified Frequent Itemsets;Cryptography

I. INTRODUCTION

Data mining techniques are used to extract useful and meaningful information from huge database [1]. Prediction and description are the two fundamental goals of data mining. Data mining techniques such as classification, association rules, clustering are used to fulfill these goals. Among all these, association rule mining used in wide range of applications to determine interesting relationships among attributes in huge databases [1][2].

Mining unified frequent itemsets in distributed database with security is a challenging task. A number of techniques have been designed which suffer from security and efficiency. Proposed solution for finding unified frequent itemsets in distributed databases improves in terms of efficiency and simplicity. Simple cryptographic primitives are used.

IV. RELATED WORK

In [1], author explained about generating meaningful association rules from the huge transactional databases. Author describes the association between the itemsets in the huger databases. The author focus on algorithm for finding association rules and also managing buffer which is used to store huge transactional data of purchases from various categories of customers. In [2], authors presented improved algorithms named Apriori and AprioriTID are efficiently used for mining association rules which incorporates the meaningful association between the two items in the huge database. They have used the algorithm Apriori, which involves finding out large itemsets from database. Initially the candidate itemsets are generated and then the frequent itemset are found to generate rules. Author also explains a new algorithm AprioriHybrid by combining above two algorithms to get the benefits of both the algorithms. In [3], authors proposed the problem privacy of data while mining database for meaningful information. Solution involves constructing a decision tree classifier. The author explains usage of data which contains the values of each distinct records are disturbed. Hence it is difficult to guess real values in distinct data records the authors suggest a technique to precisely evaluate the scattered original data values. This is done by reconstructing the scattered real values. In [4], authors presented a solution for the privacy of data shared among multiple users problem by means of a secure multiparty computation method. The protocol describes the data which is split into horizontally. The main key concern is given for ID3 algorithm. In [5], authors proposed the problem of preserving privacy in data mining while discovering association rules when the data is divided horizontally. The proposed algorithm utilizes the some of techniques such as randomization, encryption of site results along with secure computation. The main vital part of the proposed algorithm is commutative encryption of itemsets from each player in circular fashion. In [6], authors recommends an enhanced Kantarcioglu and Clifton's protocol given by authors in [5], which is a two phase for preserving privacy while mining distributed data. This protocol enhance the security along with decreasing the transmission load for encrypted candidates generated in the first phase. They also focus on protecting the individual data shared rather thasn just final results. In [7], authors discuss a method of determining association rules for distributed databases of n sites which are horizontally partitioned. Here no site is considered to be a trusted party. Each site calculates global results. These results are computed from all sites database frequent item sets with specified support values. In [8], authors projected a new algorithm for an efficient association rule hiding that rises and drops the support specified for the left hand side and right hand side elements of the rule consistently in order to hide the rule. The recommended algorithm is profitable as it makes minimum change to the data entries to hide a collection of rules with reduced CPU time. In [9], authors discuss review of the up-to-date methods for privacy preservation in association rule mining. They also analyzes the techniques for privacy

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 5 Issue 01, January-2016

preserving while mining association rules and points out their advantages and disadvantages. End with the challenges and directions for future research. In [10], author proposes a protocol. This protocol involves secure calculation of the union of local private subsets. Partial databases are given as inputs. The desired output is the set of meaningful association rules. Rules must satisfy the user specified with given threshold support and confidence. These rules are discovered after merging the databases from various users termed as unified database. Author proposes a simple protocol to identify the inclusion of an item in the transaction by creating a binary vector, where one represents existence of an item and zero represents absence of the item.

II PRILIMINARY KNOWLEDGE

A. Association Rule Mining

To extract the meaningful information from the database we use association rule mining. Association rule mining is one of data mining technique. It is used to identify the predictabilities found in large volume of data. Such a technique can be used to identify and disclose unseen information that is reserved for an individual or organization [2][8].

An association rule mining plays a vital part in retail organizations. As we work with enormous amounts of sales data. There may be thousands of transactions occurring on daily basis. Every set of transaction can be referred as basket data. A record can be defined with a unique transaction number and set of items bought together in that particular transaction. Such databases are of key concern to the organizations for forecasting the future demand and analysis for further improvements in business [2]. This lead to development of data mining techniques such as association rules.

Example for an Association rule:

98% of customers who purchase milk will also purchase sugar. This rule may be very helpful for the retailer to keep the stock of these items up to date which are purchased together. It is very much necessary to find all such rules for customer segmentation based on buying patterns [1][2].

An association rule can be defined in the form of A➔B where A and B are items in the database. A is known as antecedent and B is known as consequent.

Support and confidence are the important keywords used as quality measures for finding the rule [2].

The support for the rule A➔B can be defined as the total number of transactions that includes A and B.

Support evaluates the frequency of the rule that is applicable for T, where T is transaction set. The support for a rule A➔B can be given by equation (1)

$$\text{Support}(A \rightarrow B) = \frac{|A \cap B|}{N} \quad \ldots\ldots\ldots\ldots \quad (1)$$

$A \cap B$ - All transactions which includes both A and B items.

N - Total number of transactions.

Confidence of a rule defines the percentage of transactions that includes A which also contains B. It is given by equation (2)

$$\text{Confidence}(A \rightarrow B) = \frac{|A \cap B|}{A} \quad \ldots\ldots\ldots \quad (2)$$

Confidence is a very much necessary to find the quality of rule. If we find the rule which will satisfy the specified confidence then we term such a rule as interesting rule. There are two important steps to mine association rules [1] [2].

a. Discovering all the itemsets contained in the data that are sufficient for mining association rules. These combinations have to demonstrate at least certain frequency and are hence called frequent itemsets.

b. Generating rules from the frequent itemsets which are discovered in the first step.

An itemset is said to be frequent if and only if the total count of transactions where itemset is appearing is more than or equal to the user specified minimum support. Otherwise it is termed as infrequent [2][7].

B. The need for privacy preserving

The regular data mining algorithms such as classification, association, and clustering deals with analyzing the stored raw data and mining meaningful knowledge discovery patterns from the database [9]. Privacy preserving data mining algorithms mainly concentrate the private or sensitive data of the user where privacy becomes an important factor [5][6][8].

The main goal of many distributed databases is mining the data along with preserving privacy. Allowing the user to perform useful calculations on the whole dataset. At the same time keeping the confidentiality of each individual site data [7]. Each site owner has to collaborate in getting combined results, but not completely trust other sites while distribution of their own private data sets [10].

In distributed data mining, preserving privacy is one of vital aspect. Secure multi party computation one of the useful approach to preserve the privacy in distributed databases [8][9].

III SYSTEM ARCHITECTURE

The aim is to safeguard the data from external data miner. The data owner intends at privacy of the data before it is released. The main methodology is to apply encryption. After encryption the data mining is performed with the data records of individual data owners with desired privacy without being disclosed to each other.

The database is distributed among many players (users). All these players are willing to share the data and find the global frequent itemsets.

The encrypted data can be used to find the interesting patterns in the data, without releasing original information about records.

Merge all candidate item sets which are generated after local pruning i.e. retain item sets which are locally ‘s’ frequent, where ‘s’ is user specified minimum support.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 5 Issue 01, January-2016

Final result is global information about association rules that are supported locally in each database.

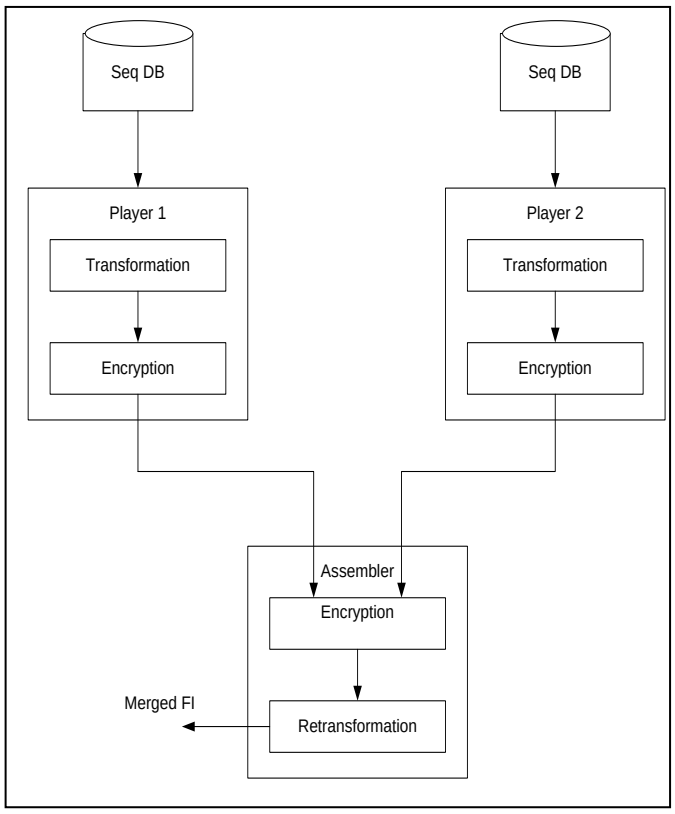

Figure 1: System Architecture

Figure 1 shows System architecture of mining frequent itemsets in distributed databases. There are many players (users) who share datasets in distributed databases, these databases are those that have the same schema but contain information about different entities. The input given is a sequence database, here each sequence is a list of transactions ordered by transaction-time, and each transaction is a set of items.

The application of some transformations is done to extract the item sets from given sequences and find frequent itemsets which are local to each player.

Our aim is to reduce the data revealed from the private sensitive database of individual players. Hence we apply encryption technique to protect the information for individual transactions in the different databases. Thus protecting the information which is global, such as which specific items are locally frequent in each of databases.

All encrypted itemsets will be sent to assembler for unification. We merge itemsets from all the players resulting a global frequent itemset.

The final results can be viewed by applying decryption technique.

ALGORITHM

Secure mining algorithm is divided into different phases. The given input is frequent itemset from individual site owner, termed as player 'P'. Frequent itemset is represented by 'F'. Candidate itemset is represented by 'C'. One of the frequent item is represented by 'X'. Each phase of the algorithm is explained below.

Algorithm for secure mining

Input : Every player Pm generates a input set Csk, m ⊆ Ap(Fsk-1), 1≤ m ≤ M

Output : $C_{sk} = U^{M}_{m=1}$ Csk, m

1: Initialization (Phase 0)
2: Player Pm , 1≤ m ≤ M, selects key Km.
3: Compute hash h(x) for all x ∈ Ap(Fsk-1).
4: Construct the lookup table
T = {(x, h(x)):x ∈ Ap (Fsk-1)}
5: Encryption of all itemsets (Phase 1)
6: For all Pm , 1≤ m ≤ M, do
7: set Xm=0.
8: For all x ∈ Csk, m do
9: Pm calculates EKm (h(x)) then adds it to Xm
10: end for
11: end for
12: Merging itemsets (Phase 2)
13: Every player broadcast encrypted set.
14: Assembler eliminates repeated set from the Unified list.
Denote the final list by ECsk.
15: Decryption (Phase 3)
16: For m=1 to M-1 do
17: Pm decrypts all itemsets in ECsk by means of key Km,
18: end for
19: Denote the resulting set by Csk.
20: Pm broadcast Csk.

Phase 0 : Each player choose the cryptographic primitives, a commutative cipher, and a private random key Km.

Phase 1 : Every player calculate a encryption of the hashed set Csk, m, 1≤ m ≤ M. each player Pm hashes every item sets in Csk, m Finally encrypts them by using the random key Km.

Phase 2 : all players will combine their lists of encrypted item sets. At the end of this stage the union set Csk = UMm=1 Csk, m is formed by eliminating the duplicate item sets and merging all the item sets from different players.

In Phase 3 : A set of decryptions are performed by using the key Km . At the end, the global results will be displayed in the original format.

TABLE I. SYMBOLS AND ITS DEFINITIONS

Symbols	Definition
$\{P_1, P_2 \ldots\ldots P_M\}$	No. of Players sharing datasets in the Distributed database
X	One of the frequent item
K	Key used for encryption and decryption
F	Frequent Itemset
C	Candidate Itemset
Ap	Apriori

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 5 Issue 01, January-2016

IV CONCLUSION

It is proposed to have a secure mining protocol for discovering frequent item sets in distributed databases. This algorithm is based on previous fast distributed mining algorithms which does not assure any security while mining the distributed databases. Secure mining algorithm is used reduce the itemset count by extracting the Frequent Itemsets from each site. These frequent Itemsets are encrypted so that no site owner discloses their local frequent itemsets. Proposed solution treats data confidentiality problem by using AES encryption scheme. Finally all the frequent itemsets from different sites are unified to get global itemsets. Encryption schemes are used in preserving privacy and avoiding sensitive information leakage. Efficient by using simple cryptographic primitives. Improvement in computational time compared to previous algorithms

REFERENCES

[1] Agrawal ,R., et al.: Mining association rules between sets of items in large database. In Proc. of ACM SIGMOD'93, D.C.pp.207-216, ACM Press, Washington, 1993.
[2] R. Agrawal and R. Srikant, "Fast Algorithms for Mining Association Rules in Large Databases," Proceedings 20th International Conference. Very Large Data Bases (VLDB), 1994.
[3] R. Agrawal and R. Srikant, "Privacy-Preserving Data Mining," Proc. ACM SIGMOD Conf., pp. 439-450, 2000.
[4] Y. Lindell and B. Pinkas, "Privacy Preserving Data Mining," Advances in Cryptology (CRYPTO 2000), pp. 36-54, 2000,
[5] M. Kantarcioglu and C. Clifton, "Privacy-Preserving Distributed Mining of Association Rules on Horizontally Partitioned Data" , IEEE Transactions Knowledge and Data Engineering., vol. 16, no. 9, pp. 1026-1037, 2004.
[6] Chin-Chen Chang, Jieh-Shan Yeh, and Yu-Chiang Li, "Privacy-Preserving Mining of Association Rules on Distributed Databases", IJCSNS International Journal of Computer Science and Network Security, Vol.6, No.11, 2006
[7] N V Muthu Lakshmi and Dr. K Sandhya Rani, "Privacy Preserving Association Rule Mining without Trusted Party for Horizontally Partitioned Databases", International Journal of Data Mining & Knowledge Management Process (IJDKP) Vol.2, No.2, March 2012
[8] Yogendra Kumar Jain, Vinod Kumar Yadav, Geetika S. Panday," An Efficient Association Rule Hiding Algorithm for Privacy Preserving Data Mining", International Journal on Computer Science and Engineering (IJCSE), Vol. 3 No.[8] .2011
[9] K. Sathiyapriya1 and Dr. G. Sudha Sadasivam " A Survey on Privacy Preserving of Association Rule Mining" article by, International Journal of Data Mining & Knowledge Management Process (IJDKP) Vol.3, No.2, 2013
[10] Tamir Tassa, "Secure Mining of Association Rules in Horizontally Distributed Databases", IEEE Transactions Knowledge and Data Engineering. Vol. 26 , NO. 4, 2014