www.ijecs.in
International Journal Of Engineering And Computer Science ISSN: 2319-7242
Volume 5 Issues 6 June 2016, Page No. 17102-17106

An Optimized and Secure Association Rule Mining Algorithm in Parallel and Distributed Environment

Sakharam.B.Kolpe[1] , N.V.Alone[2]
Department of Computer Engineering GES R.H.Sapat COE MSR,
Savitribai Phule Pune University, Maharashtra India-422005.

***Abstract*— Data Mining is a subfield of Artificial Intelligence which is mainly used for extraction of hidden predictive information from large available databases. To Handle Large Data set is very tedious and complex task. The solution to this problem is to apply distributed or parallel approaches. The field of distributed data mining has therefore gained increasing importance in the last decade. In Distributed data mining Security is the major problem with respect to association rule mining projects. Association rule mining is become one of the core data mining tasks and has attracted tremendous interest among data mining Researchers. The Apriori algorithm by Rakesh Agarwal has emerged as one of the best Association Rule mining algorithms. It also serves as the base algorithm for most parallel algorithms. The performance of data mining algorithm can be accelerated from O(N) to O(N/k) with parallelism, where N = number of data records and k =number of nodes in distributed system. By using various cryptographic techniques and Method interesting associations and patterns between variables of big database can be observed securely. This system addresses the problem of Secure association rule mining over the horizontally distributed database system. In Current technique for association rule mining huge processing time and also high threat from various attacks. The proposed system can provide security at the time of mining data from various data sources and also reduce computation time by using parallel programming environment like OpenMp. The goal of proposed approach is to find all association rule with maintain support s and confidence c and minimize the information disclosed about the private databases held by those player .This system is based on distributed mining algorithm, K&C and AES algorithm. Distributed mining algorithm used here is the distributed version of apriori algorithm. With proposed approach speed up is acquired using Parallel Computing while preserving the privacy of the data .**

Keywords- Apriori Algorithm, KC , AES , Association Rule Item Sets, Distributed Mining, Parallel Computing.

I. INTRODUCTION

The problem of securely mining association rule in distributed environment is studied here. In this system there are several sites that hold databases; these databases are distributed horizontally over different sites participating in transaction for experimentation. The goal is to mine these datasets for finding all association rules with support count at least s and confidence count at least c by pruning using minimal support count s and confidence size c, also hold for the unified database. Processing of dataset may become impossible due to limitations on processor and memory. The solution to the problem is to accelerate the mining process with the help of parallel or distributed approaches. Advances in computing and networking technologies have conclusion in distributed and dynamic sources of data. The proposed systems address the problem of securely mining association rule in distributed Environment. In this proposed system there are several sites or player that holds homogeneous partitioned databases [1], these databases are distributed using horizontally over different sites participating in transaction.

The important objective of the proposed system is to minimize the information reveal about the private database held by the multiple sites. The information that is protected here is not only individual transactions of every site but also information in the different database at every site, and also global information like association rules supported restrictedly by each one of those database at different sites [1].

Here in proposed work the design of an alternative protocol has been planned and implemented for securely computing the union of private subsets. The systems relay on and offer simplicity and efficiency as well as confidentiality. In addition this system does not depend on encryption like commutative encryption [4], [5]. The main objectives for implementing this system are as first is to handle an large size of big data sets, second is to acquire and obtain speed by utilizing resources available in distributed system and last objective is to provide more security by using cryptographic technique for preserving data secretly.

This paper is organized as follows. Section II reviews existing system used for association rule mining in distributed environment. In Section III, present proposed system. In Section IV, it provides the implementation details. Section V evaluates the dataset and results.

II.LITERATURE SURVEY

Data mining is the technique of extracting interesting patterns, knowledge from large data. It consists of set of activities which is used to find new, hidden and unexpected patterns from large available data warehouse. Almost in now days every application Data mining techniques and data warehousing are used. Most of the data mining tools operate by gather together all the data into a centralize site ,then applying data mining algorithm on this available data. However, security is main issue in the building a centralized warehouse. To obtaining and interested frequent pattern in this Centralized scenario having problem of security and commutation efficiency .So to resolve this issue data can be distributed in multiple site and the interested pattern can be be obtained efficiently in distributed environment but again problem in this scenario is secure multiparty computation. Here assume homogeneous databases in this All sites have the same schema, but each site has information on dissimilar entities.

The objective is to produce association rules that hold globally, while minimizing the information shared about each site. in the mining of association rule the research work is divided into two main settings . In first setting the data owner and the miner are two different entities, and second, which consist of data distribution among several sites or players, these players or sites jointly performs mining on the data held by those sites or players.
In the existing systems [1], [5] they proposed protocol for securely computing the union of private subsets at each one site in the transaction is suggested. Here a multi-party-computation is considered and in that implementation of cryptographic techniques is done like encryption, decryption, commutative encryption, and hash functions are used. In that systems it is more difficult to mine association rules through security assumptions in addition it disclose the data during the mining process. The use of such cryptographic techniques increase communication cost and computation cost [6]. In the existing system though these techniques are used but it causes some misuse of information, therefore it is not perfectly secure. Thus the union of private subsets is not perfectly calculated, so the system is proposed to defeat with this problem.

Kantarcioglu and Clifton [8] have introduce the protocol for securely computing union of each private subsets held by the different sites. The private subset of a prearranged site includes the item sets which are s-frequent in his own database. This implementation become more costly and time consuming and in this cryptographic techniques such as commutative encryption, oblivious transfer are used. Yao[9] proposed the protocol for securely computing the union of private sub-sets at each site. In the existing systems discussed so far these techniques causes some disclose and loss the security of private information hold by every site.
The proposed system overcomes this problem of information leakage and privacy of data. It is not possible to mine globally valid results from distributed data without losing security of private information. Secure mining of association rule in distributed environment is costly in terms of computational cost and communication.

III. PROPOSED WORK

Here in proposed work, the design of an alternative protocol has been proposed and implemented for securely computing the union of private subsets. The system relay on and offers simplicity and efficiency as well as privacy. In addition this system does not depend on encryption like commutative encryption. The main objectives for implementing this system are as first is to handle an large size of big data sets, second is to acquire and obtain speed by utilizing resources available in distributed system and ultimate and last objective is to provide more security by using cryptographic technique like AES for preserving data secretly. The algorithms used in proposed System are Distributed Mining Algorithm, AES, Kantarcioglu and Clifton Algorithm. In Proposed System the main three algorithm are used and this algorithm as follows.

1) Distributed Mining Algorithm With Parallel Environment: - The DM algorithm is the distributed version of apriori algorithm, this algorithm worked as follows:-

i. Initialization
ii. Player or Site ItemSets Generation - Each site will generate its local frequent itemset. Check weather frequent itemset is locally frequent in private database at own site and itemset is globally frequent.
iii. Local Pruning-Retains Locally frequent item sets.
iv. Identification of the candidate item sets each site broadcasts its itemset.
v. Computation of local supports - Compute local supports of all itemsets.
vi. Broadcast Mining Results - Here it is identified that each locally frequent item is subset of globally frequent itemset. Algorithm Proceeds until it finds no (k + 1) item are longest globally frequent itemsets. Here k is number of itemsets .

2) AES Algorithm: - Advanced Encryption Standard or AES is a symmetric block cipher which explained as

i. AES is a block cipher with a block length of 128 bits.
ii. AES allows for three different key lengths: 128, 192, or 256 bits.
iii. AES works by repeating the same defined steps multiple times.
iv. AES is a secret key encryption algorithm.
v. AES operates on a fixed number of bytes.
vi. AES as well as most encryption algorithms are reversible.

in AES Mostly the same steps are performed for completing both encryption and decryption in reverse order. The AES Encryption algorithm operates on bytes, which makes it simpler to implement and explain. This key is divided into individual sub keys, a sub key for each operation round. This process is called KEY EXPANSION. As mentioned before AES is an iterated block cipher, means that the same operations are performed many times on a fixed number of bytes. These operations can easily be divided in to the following functions:-
1] ADD ROUND KEY
2] BYTE SUB
3] SHIFT ROW
4] MIX COLUMN

3) Kantarcioglu and Clifton Algorithm(K&C): The K&C Algorithm is mainly used for implementing step number 4 of the DM Algorithm. This Algorithm is mainly used for obtaining locally frequent itemsets securely in distributed Environment. The algorithms work as follows:

i. Each site adds to his private subset fake itemsets, in order to hide its size.
ii. Sites jointly compute the encryption of their private subsets.
iii. Each site adds his own layer of encryption using his private secret key.
iv. Every itemset in each subset is encrypted by all of the sites. The usage of a commutative encryption scheme ensures that all itemsets are, eventually, encrypted in the same manner.

v. Sites compute the union of those subsets in their encrypted form.
vi. Sites decrypt the union set and remove fake itemsets from it.

FUNCTIONALITY

Input	Process	Output
User Details	Registration	DBCS'
UID, Password LDB1, LDB2,	Authentication	Valid User
Encoding Algorithm	Encoding	Encoded Data
Encoded Data [At Server]	Encryption	Encrypted Data
Encrypted Data	Decryption	Decrypted Data
Decrypted Data	Mining, Decoding	Mining Result

B. METHODOLOGY

The proposed approach distributes each transaction D into partitions and in each local partition frequent item sets are find out simultaneously using multiple core on single CPU at every site. After finding local frequent itemsets all local frequents item sets are combined together to find candidate item set. In last stage global frequent item sets are found. In experimentation a datasets from UCI repository are used namely mushroom, monks problem, vote, soybean, disease and item sets. The system is implemented as shown in figure1, In implementation of system the database is distributed horizontally among several sites in the transaction.
The data at each site is encrypted and decrypted using Advanced Encryption System algorithm. Encrypted data is processed by each site. and Apriori algorithm is applied to data stored at server side and attempts to find frequent itemsets. The K & C is applied at the intermediate stage of distributed mining algorithm for effectively mining global association rules. Round robin technique is used for Horizontal distribution of Data sets to simplify the data skew. The Join key is present at all the sites where the database is distributed. While implementation, one database at individual site in the transaction is rendered as primary and it is considered as ”Master” of the process or system. The main objective of this system is to obtain Association rules in very less amount of time by using parallel approach. To provide more security at the time mining of data in distributed environment AES Encryption Algorithms[10],[11] are used The Figure1 Shows the working of proposed System. In this System Database is Kept at centralized server and using horizontal data partitioning data is distributed at multiple site, then using distributed mining algorithm frequent itemsets are calculated at every site simultaneously in optimal of amount of time and AES Encryption used to provide security at the time mining of association rule.

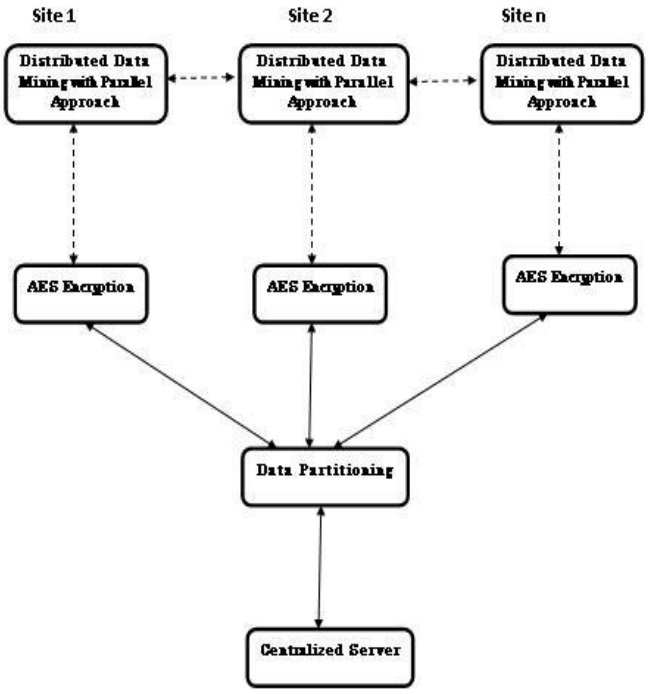

Fig.1.System Architecture

C. EXPERIMENTAL SETUP

In this proposed system the performance of secure implementations of the Distribute Mining algorithm is compared. In implementation, we have implemented Step 4 of the DM algorithm in the secure manner using AES Encryption Technique.
The experiment performed on multiple machines with core i3 processor 2.0 GHz and 4GB of main memory. The operating system is Ubuntu 14.04 and all algorithms are implemented in JDK 1.7. We have tested the two implementations with respect to some measures which are enlisted below:

i. Total computation time of the complete algorithms (DARM and K & C) over all sites. That measure includes the Apriori computing time, and the time to identify the globally s-frequent item sets.
ii. Total computation time of the unification over all sites.
iii. Total message size.

D. Experimental Outcome

- It can handle Big Data sets (KDD with10,000 records)
- Speed up is acquired in computation process by utilizing resources available in distributed system.
- Provides security in distributed computing environment.

This proposed System worked on three experiment sets, where each set is tested with abovementioned measures.
The result is compared with current Sequential System. It is observed that System performed the fast computation for computing frequent item sets in distributed and Parallel Environment and it requires small amount of Computation time and also provides more security.

D. RESULTS AND DISCUSSION

The following results show the performance of the proposed system. The proposed system provides security while doing mining task in the distributed environment. To verify the performance of the system, the encryption/decryption over-head and support overhead is evaluated. In experimentation datasets from UCI repository are used namely mushroom, monks problem, vote, soybean, disease and itemsets. The figure 3 shows the time required for mining association from data sets by sequential and the proposed approach. Datasets are tested for all the algorithms i.e. Apriori, AES and K and C algorithm. In this system distributed mining algorithm is used for data mining task. It also shows the increase in the Speed up and Computation time using association rule mining. The security provided by system gives better performance gain.

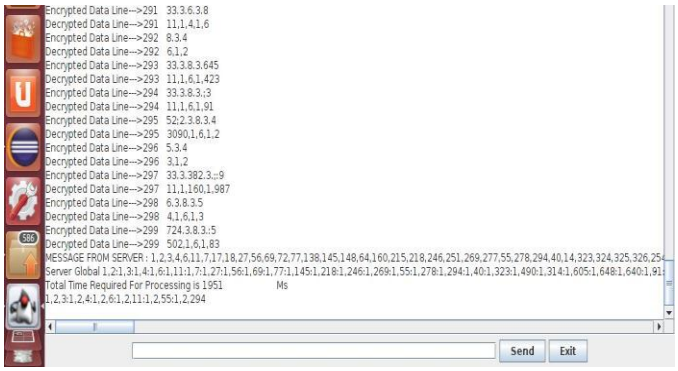

Fig.2.Total Computation Time and Applying Encryption on transaction site.

This system is totally independent of oblivious transfer and commutative encryption which makes it simple and it also contributes to the relatively less cost of computation and communication and raised in computational time. The graph shown in figure 2 gives the time required for generating frequent itemsets. The graph shown in figure 4 gives the computational time required for sequential and proposed system. In experimentation, datasets from KDD community, extended bakery dataset, frequent Itemset Mining Dataset Repository, Bioinformatics Data Set, IBM Almden Quest research group etc. are used. The datasets are namely chess, connect, algebra and test.

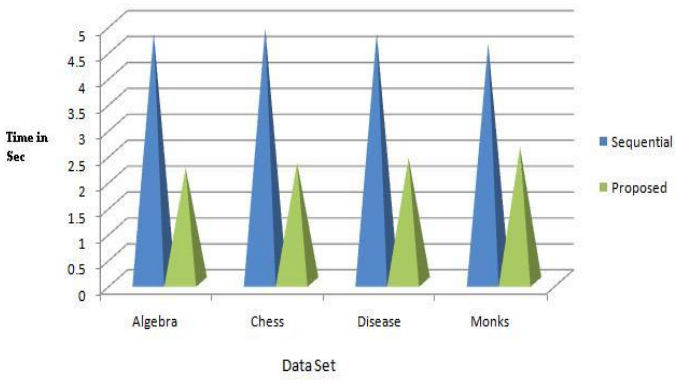

Fig.3. Time required for mining association rules from datasets by sequential and the proposed approach

Let *Ts* is a time required by sequential system for mining and *Tp* is a time required by proposed system for mining. The speed up is calculated by equation 1.

$$Speed\ up = \frac{Ts-Tp}{Ts} * 100 \qquad (1)$$

By evaluating the average speed up for algebra dataset is 50%.The average speed up for chess dataset is 55%, the average speed up for Disease dataset is 51%, the average speed up for monks dataset is 50% and the average speed up for disease dataset is 55%. Table 1 shows the speed up of sequential and proposed system.

Data Set	Ts	Tp	Speedup
Algebra	90	46	51
Chess	84	42	50
Disease	88	49	55
Monks	82	37	54

Table 1:- Speed Up Acquired during mining Process.

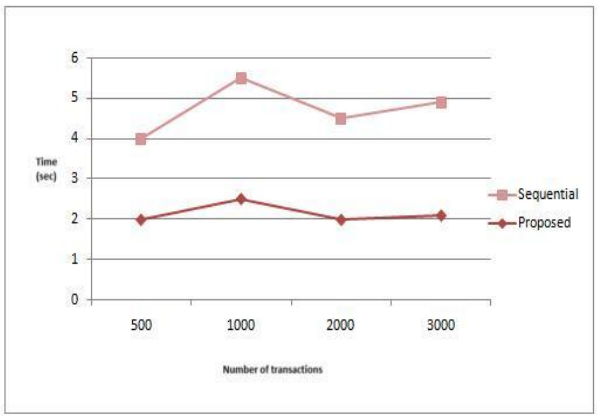

Fig.4.Time for Generation of Frequent ItemSets

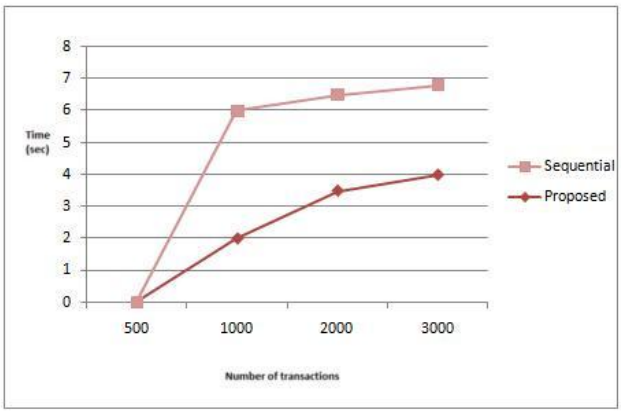

Fig.4.Computation Time

E.CONCLUSION

The problem of securely mining association rule in distributed environment is address here. In this proposed system there are several sites that hold homogeneous databases and these databases are distributed horizontally over different sites participating in transaction. The goal is to mine these data sets for finding all association rules with support count at least *s* and confidence count at least *c*. The given

minimal support count s and confidence size c, also hold for the unified database. The distinguished objective of the proposed system is to minimize the information disclosed about the private database held by the sites. The information that is protected here is individual transactions information in the different database for each one site, and also global information like association rules supported locally by each of those databases at different sites. Due to use of these techniques of secure distributed mining, performances enhanced from O(N) to lower bound O(N /k), where N = number of data instances and k = number of nodes(that is 4) secure distributed association rule mining is done with a reasonable cost, and time.

References

[1] T.Tassa "Secure Mining of association rules in horizontally distributed Database" proc, IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, VOL. 26, NO. 4, APRIL 2014.

[2] A.V. Evfimievski, R. Srikant, R. Agrawal, and J. Gehrke "Privacy preserving mining of association rules." In KDD, pages 217228, 2002.

[3] D.W.L. Cheung, J. Han, V.T.Y. Ng, A.W.C. Fu, and Y. Fu "fast distributedalgorithm for mining association rules." In PDIS, pages 3142, 1996.

[4] 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.

[5] 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.

[6] M. Kantarcioglu and C. Clifton, "Privacy-preserving distributed mining ofassociation rules on horizontally partitioned data" IEEE Transactions onKnowledge and Data Engineering, 16:10261037, 2004.

[7] G. Alex and A. Freides,"Scalable, high-performance data mining with parallel processing" ,in Principles and Practice of Knowledge Discovery in Databases,1998.

[8] R. Agrawal and R. Srikant, "Fast algorithms for mining association rules in large Database In VLDB, pages 487499, 1994.

[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 TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, VOL. 8, NO. 6,

[10] T. Tassa and E. Gudes. Secure distributed computation of anonym zed views of shared database, Transactions on Database Systems, 37, Article 11, 2012.

[11] R.L. Rivest, A. Shamir, and L.M. Adleman, A Method for Obtaining Digital Signatures and Public-Key no. 2, pp. 120-126, 1978.

[12] Priyanka khairnar "Secure Distributed Data Mining", IJCA Proceedings on Innovations and Trends in Computer and Communication Engineering(ITCCE), pp. 9-12, Dec 2014.

[13] Sakharam Kolpe "An Optimized and Secure Association Rule Mining Algorithm in Parallel and Distributed Environment" International Journal In Engineering Technology, Management and Applied Science (IJETMAS),Volume.2. DEC 2015 72-78.