Distributed Data Mining based on Random Projection with Optimal Communication

T.Revathi, P.Sumathi

Abstract— Distributed data mining discovers hidden useful information from data sources distributed among several sites. Privacy of participating sites becomes great concern and sensitive information pertaining to the individual sites needs high protection when data mining occurs among several sites. Different approaches for mining data securely in a distributed environment have been proposed but in the existing approaches, collusion among the participating sites may reveal sensitive information about other participating sites and they suffer from the intended purposes of maintaining privacy of the individual participating sites, reducing computational complexity and minimizing communication overhead. The proposed method finds global frequent itemsets in a distributed environment with minimal communication among sites and ensures higher degree of privacy with randomized site selection. The experimental analysis shows that proposed method generates global frequent itemsets among colluded sites without affecting mining performance and confirms optimal communication among sites.

Index Terms— Distributed data mining, privacy, secure multiparty computation, frequent itemsets.

I. INTRODUCTION

Data mining or has a goal to discover knowledge out of data and present it in a form that is easily comprehensible to humans. The term data mining refers to extracting or mining knowledge from a massive amount of data Knowledge detection in databases is precise process consisting of a number of distinct steps[2]. Data mining is the foundation step, which outcome in the discovery of unknown but helpful knowledge from huge databases. A formal definition of Knowledge discovery in databases is given as follows: "Data mining, or knowledge discovery, is the computer-assisted process of digging through and analyzing enormous sets of data and then extracting the meaning of the data. Data mining tools predict behaviors and future trends, allowing businesses to make proactive, knowledge-driven decisions [3].Data mining expertise provide a consumer-leaning approach to new and unknown patterns in the data. The exposed knowledge can be used by the healthcare administrators to progress the superiority of service.

Major technological developments and innovations in the field of information technology have made it easy for organizations to store a huge amount of data within its affordable limit. Data mining techniques come in handy to extract useful information for strategic decision making from voluminous data which is either centralized or distributed (Agrawal & Srikant, 1994; Han & Kamber, 2001).Data mining functionalities like association rule mining, cluster analysis, classification, prediction etc. specify the different kinds of patterns mined. Association Rule Mining (ARM) finds interesting association or correlation among a large set of data items. Finding association rules among huge amount of business transactions can help in making many business decisions such as catalog design cross marketing etc. A best example of ARM is market basket analysis. This is the process of analyzing the customer buying habits from the association between the different items which is available in the shopping baskets. This analysis can help retailers to develop marketing strategies. ARM involves two stages

(i) Finding frequent itemsets Privacy Preserving Distributed Data Mining using Randomized Site Selection.
(ii) Generating strong association rules.

II. RELATED WORK

Privacy preserving data mining is an active research area provides methods for finding patterns without revealing sensitive data. Numerous research works are underway to preserve privacy both in individual data source and multiple data sources. There are two broad approaches for privacy-preserving data mining (Wang, Lee, Billis, & Jafari, 2004). The first approach alters the data before it is delivered to the data miner so that real values are hidden. It is called data sanitization. The second approach assumes that the data is distributed between two or more sites, and that these sites cooperate to learn the global data mining results without revealing the data at their individual sites. The second approach was named by Goldreich as Secure Multiparty Computation (SMC) (Goldreich. 1998, Lindell & Pinkas, 2009).

A typical example of a privacy-preserving data mining problem occurs in the field of market-basket analysis. Consider the case of two or more companies have a huge data source where the customers' buying habits are stored as records. They wish to jointly mine their customers' data to make strategic decisions using the patterns mined. However, some of these companies may not want to share some sensitive strategic patterns hidden within their own data with other parties. In such cases, classical data mining solutions cannot be used. Hence, Secure Mutiparty Computational (SMC) solutions can be applied to maintain privacy in distributed association rule mining (Lindell & Pinkas, 2009*)*. The goal of SMC in distributed association rule mining is to find global frequent itemset without revealing the local support count of participating sites to each other.

In distributed privacy preserving data mining, the participating sites may be treated as honest, semi-honest or dishonest (Clifton, 2001). The semi-honest parties are honest but try to learn more from received information. The dishonest parties are malicious and they do not follow the defined protocols. When all the parties are honest the question of privacy will not arise. The real need for concealing the data of

Manuscript received on January, 2013.

T.Revathi, Research Scholar, Manonmaniam Sundaranar University & Assistant Professor, Dept. of Computer Science , PSG College of Arts & Science, Coimbatore,India.

P.Sumathi,Assistant Professor ,PG & Research Department, Dept. of Computer Science & Applications, Govt.College of Arts & Science, Coimbatore,India.

individual site arises when the parties are semi-honest or dishonest. Here, a new collusion free solution is proposed to find the global frequent itemsets among dishonest parties with minimum communication cost and time complexity.

The subsequent sections of the paper are organized as follows. Firstly, related existing works are reviewed. Secondly, the proposed approach and its performance evaluation are discussed. Thirdly discussion and limitations are presented. Lastly, a suitable conclusion and future work for maintaining privacy is attempted. mining on sanitized data results in loss of accuracy, while SMC protocols give accurate results with high computation or communication costs (Inan, Saygyn, Savas, Hintoglu & Levi, 2006). Representative works from each of the approaches is discussed in the following sections.

Clifton & Marks, (1996) have provided a well designed scenario which clearly reveals the importance of data sanitization. In this scenario, by providing the original unaltered database to an external party, some strategic association rules that are crucial to the data owner are disclosed with serious adverse effects. The sensitive association rule hiding problem is very common in a collaborative association rule mining project, in which one company may decide to disclose only part of knowledge contained in its data and hide strategic knowledge represented by sensitive rules. These sensitive rules must be protected before its data is shared. Also they suggest different measures to protect sensitive data such as limiting access, altering the data, eliminating unnecessary groupings, and augmenting the data in a single data source.

Atallah, Bertino, Elmagarmid, Ibrahim, & Verykios, (1999) proposed the concept of data sanitization to resolve the association rule hiding problem. Its main idea is to select some transactions from original database and to modify them through some heuristics. They also proved that the optimal sanitization is an NP-hard problem. Moreover, data sanitization can produce a lot of I/O operations, which greatly increases the time cost, especially when the original database includes a large number of transactions.

The sanitizing algorithm explained in the paper (Stanley, Oliveira & Zaıane, 2003) requires two scans of data source. The first scan is required to build the index for speeding up the sanitization process. The second scan is used to sanitize the original data source. This represents a more significant improvement compared to the other algorithms which require various scans depending on the number of association rules that are to be hidden. Lee, Chang & Chen (2004) define a sanitization matrix in their recent work. By multiplying the original transaction database and the sanitization matrix, a new database, which is sanitized for privacy concern is created. However, the construction of sanitization matrix for large data sources is a tedious process.

In order to hide sensitive rules, two fundamental approaches are presented in (Verykios, Elmagarmid, Bertino, Saygin & Dasseni, 2004). The first approach prevents rules from being generated by hiding the frequent sets from which they are derived. The second approach reduces the importance of the rules by setting their confidence below a user-specified threshold. The approaches used in the paper are moderately successful in hiding sensitive data, but they are computationally intensive and have side effects like generation of artifactual new rules and hiding the existing legitimate rules.

Wu, Chiang & Chen, (2007) have suggested a method for hiding sensitive rules with limited side effects. Templates are generated for sensitive rules to be hidden in order to minimize the side effects. But the cost involved in the template generation increases if the number of sensitive rules and the number of items in the individual sensitive rules are large.The data sanitization approach is suitable for privacy preserving in case of a single data source, whereas multiple data sources requires the cooperation of sites to learn the global data mining results, yet this approach may fail and yield incorrect global mining results. Thus, secure multiparty computational techniques were proposed.

The concept of Secure Multiparty Computation (SMC) was introduced in (Yao, 1986). In many applications the data is distributed between two or more sites, and for mutual benefit these sites cooperate to learn the global data mining results without revealing the data at their individual sites. The basic idea of SMC is that this computation is secure if at the end of the computation no party is unaware about the other participating sites except its input and the results.

Vaidya & Clifton (2002) proposed a method for privacy preserving association rule mining in vertically partitioned data. Each site holds some of the attributes of each transaction. An efficient scalar product protocol was proposed to preserve the privacy among two parties. This protocol did not consider the collusion among the parties and was also limited to boolean association rule mining.

Kantarcioglu & Clifton (2004) proposed a work to preserve privacy among semi-honest parties in a distributed environment. It works in two phases assuming no collusion among the parties

III. KNOWLEDGE DISCOVERY AND DATA MINING

Data mining is a form of knowledge discovery. Data are gathered and studied collectively for purposes other than those for which they were originally created. New knowledge may be obtained in the process while eliminating one of the largest costs, *viz.*, data collection. Medical data, for example, often exists in vast quantities in an unstructured format. The application of data mining can facilitate systematic analysis in such cases. Medical data, however, requires a large amount of preprocessing in order to be useful. **Here** numeric and textual information may be interspersed, different symbols can be used with the same meaning, redundancy often exists in data, misspelled medical terms are common, and the data is frequently rather sparse. A robust preprocessing system is required in order to extract any kind of knowledge from even medium-sized medical data sets. The data must not only be cleaned of errors and redundancy, but organized in a fashion that makes sense to the problem.

Soft computing is a consortium of methodologies that works synergistically and provides, in one form or another, flexible information processing capability for handling real-life ambiguous situations [8]. Its aim is to exploit the tolerance for imprecision, uncertainty, approximate reasoning, and partial truth in order to achieve tractability, robustness, and low-cost solutions. The guiding principle is to devise methods of computation that lead to an acceptable solution at low cost by seeking for an approximate solution to an imprecisely/precisely formulated problem [10].

Soft computing methodologies (involving fuzzy sets, neural networks, genetic algorithms, and rough sets) are most widely applied in the data mining. Fuzzy sets provide a natural framework for the process in dealing with uncertainty. Neural

networks and rough sets are widely used for classification and rule generation. Other approaches like case based reasoning [5] and decision trees [2], [3] are also widely used to solve data mining problems.

Clustering methodology is used to explore a data set where the goal is to separate the sample into groups or to provide understanding about the underlying structure or nature of the data. The results from clustering methods can be used to prototype supervised classifiers or to generate hypotheses. Clustering is called unsupervised classification because we typically do not know what groups there are in the data or the group membership of an individual observation. The two main methods for clustering are hierarchical clustering and *k*-means clustering.

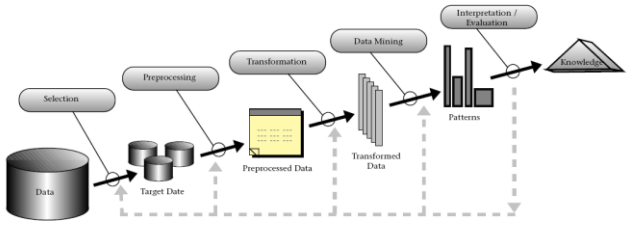


Fig.1. Steps involved in knowledge discovery

A. Definition

Data mining may be defined as "the exploration and analysis, by automatic or semiautomatic means, of large quantities of data in order to discover meaningful patterns and rules" [5]. Hence, it may be considered mining knowledge from large amounts of data since it involves knowledge extraction, as well as data/pattern analysis [6].

B. Tasks

Data mining techniques can be broadly classified based on what they can do, namely description and visualization; association and clustering; and classification and estimation, which is predictive modeling. Description and visualization can contribute greatly towards understanding a data set, especially a large one, and detecting hidden patterns in data, especially complicated data containing complex and non-linear Interactions.

In association, the aim is to decide which variables go jointly [7]. For example, market-basket analysis (the most popular form of association analysis) refers to a method that generates probabilistic statements such as, "If patients undergo treatment A, there is a 0.35 probability that they will exhibit symptom Z" [8]. With clustering, the objective is to group objects, such as patients, in such a way that objects belonging to the same cluster are similar and objects belonging to different clusters are dissimilar.

The most common and important applications in data mining probably involve predictive modeling. Classification refers to the prediction of a target variable that is categorical in nature, such have to be first transformed into information. The healthcare industry can benefit greatly from data mining applications [4]. The objective of this article is to explore relevant data mining applications by first examining data mining concepts; then, classifying potential data mining as predicting healthcare racket [10]. Estimation, on the other hand, refers to the prediction of a target variable that is metric (i.e., interval or ratio) in nature, such as predicting the length of stay or the amount of resource utilization. For predictive modeling, the data mining techniques commonly used include traditional statistics, such as multiple discriminate analysis and logistic regression analysis. They also include non-traditional methods developed in the areas of artificial intelligence and machine learning .The two for the most part significant models of these are neural networks and decision trees.

C. Distributed Data mining

Let $I = \{i_1,i_2 \ldots i_m\}$ be a set of m distinct items. Let D denote a database of transactions where each transaction T is a set of items such that $T \subseteq I$. Each transaction has a unique identifier, called TID. A set of item is referred to as an itemset. An itemset that contains k items is a k-itemset. Support of an itemset is defined as the ratio of the number of occurrences of the itemset in the data source to the total number of transactions in the data source. Support shows the frequency of occurrence of an itemset. The itemset X is said to have a support s if s% of transactions contain X. The support of an association rule $X \; Y$ is given by

Support = (Number of transactions ontaining X U Y) / (Total number of Transactions) where X is the antecedent and Y is the consequent

An itemset is said to be frequent when the number of occurrences of that particular itemset in the database is larger than a user-specified minimum support. Confidence shows the strength of the relation among the items. The confidence of an association rule is given by,

Confidence = (Number of transactions containing X U Y) / (Total number of Transactions containing X)

An association rule is said to be *strong* when its confidence is larger than a user-specified minimum confidence. Association rules with support and confidence above the minimum support and minimum confidence alone are mined. Many algorithms have been proposed for frequent itemsets generation. They are Apriori, Pincer search, Frequent pattern tree, etc. (Agrawal & Srikant, 1994; Lin & Kedem, 2002; Han, Pei, Yin & Mao, 2004).

In the present situation, information is the key factor which drives and decides the success of any organization and it is essential to share information pertaining to an individual data source for mutual benefit. Thus, Distributed Data Mining (DDM) is considered as the right solution for many applications, as it reduces some practical problems like voluminous data transfers, massive storage unit requirement, security issues etc. Distributed Association Rule Mining (DARM) is a sub-area of DDM. DARM is used to find global frequent itemsets from different data sources distributed among several sites.

In DARM, the local frequent itemsets for the given minimum support are generated at the individual sites by using data mining algorithms like Apriori, FP Growth tree, etc. (Agrawal et al. 1994; Han et al. 2001). Then, global frequent itemsets are generated by combining local frequent itemsets of all the participating sites with the help of distributed data mining algorithm (Cheung, D., Ng, Fu & Fu, 1996; Ashrafi, Taniar & Smith, 2004). The strong rules generated by distributed association rule mining algorithms satisfy both minimum global support and confidence threshold.

Let D_1, $D_2 \ldots$ D_n be the data sources which are geographically distributed. Let m be the number of items and $I = \{i_1,i_2 \ldots i_m\}$ be a set of items. Global Support of an itemset is defined as the ratio of the number of occurrences of the itemset in all the data sources to the total number of

transactions in all the data sources. An itemset is said to be globally frequent when the number of occurrences of that particular itemset in all the data sources is larger than a user-specified minimum support.

During global frequent itemset generation, local frequent itemsets of individual sites need to be shared. So, the participating sites learn the exact support count of all other participating sites. However, in many situations the participating sites are not interested to disclose the support counts of some of their itemsets which are considered as sensitive information. Thus, it is essential to share information pertaining to an individual data source without revealing sensitive information (Vaidya & Clifton, 2004). This type of information sharing or allowing access to sensitive information of a data source is likely to cause some serious privacy issues in real time situations.

Some real life situations collusion is inevitable. The first phase identifies the global candidate itemsets using commutative encryption which is computationally intensive and the second phase determines the global frequent itemsets. This work only determines the global frequent itemsets but not their exact support counts. Whereas Ashrafi, Taniar & Smith (2005) proposed privacy preserving algorithm for finding the exact support of global frequent itemsets among semi-honest parties. Here, each site generates local frequent one-length items. Randomization is applied to find the global one-length items in a secure manner. Then, higher length frequent itemsets are identified with the help of global one-length items. Through collusion and by closely watching the input and output of a particular site, details of local frequent itemsets of the site can be identified and the same can be used to estimate the supports of the higher length itemsets of the same site.

An algorithm using Clustering future (CF) tree and secure sum is proposed to preserve privacy of quantitative association rules over horizontally partitioned data (Luo, 2006) and fixing of proper threshold value for constructing the CF tree is not easy. The efficiency of the algorithm is unpredictable since it depends on the threshold value chosen to construct the CF tree. The present analysis proposes a novel, computationally simple and secure multiparty computation algorithm for dishonest parties. Since the Elliptic Curve Cryptography is used for encryption and decryption, the working principle of ECC is precisely given in the following section

IV. PROPOSED APPROACH

This paper proposes a new method to compute globally frequent itemsets from distributed data sources while preserving the privacy of the participating sites. Each participating site is assumed as dishonest and the mining process can be initiated by any one of them. The proposed method works in two phases to protect sensitive information of the individual participating sites in order to withstand any kind of collusion among the parties. In the first phase, global candidate itemsets are collected at the site who initiates the mining process using Elliptic Curve Cryptography (ECC) and in the second phase exact support count of the candidate itemsets are calculated by using randomized site selection, thus global frequent itemsets are collected in the initiator site (Rajalakshmi, Purusothaman & Pratheeba, 2010).

In *phase I* each site encrypts its local frequent itemsets in the first round and decrypts them in the second round using ECC algorithm. Initially, the site who wants to initiate the mining process encrypts all its local frequent itemsets and sends them to the next site. The next site encrypts all its local frequent itemsets along with its itemsets received from the initiator and pass it to its next site and so on. The procedure is continued for all the sites. Once the turn reached to the initiator, initiator start decrypts the encrypted message and passes it to the next site. The process continues till the initiator site has been reached. At this point, all global candidate itemsets are collected in the initiator site. Figure 2 illustrates the encryption processe.

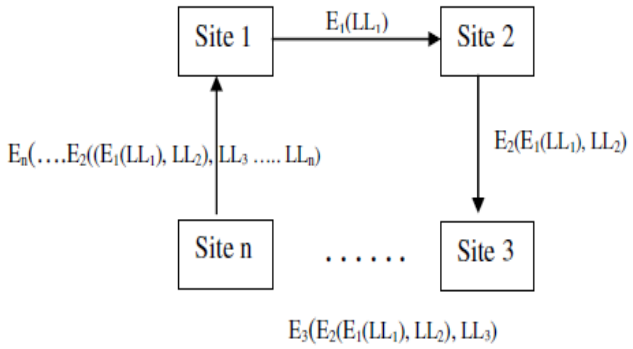


Fig. 2 Block diagram of encryption process

After collecting global candidate itemsets in *phase I*, initiator starts the second phase to count the support of candidate itemsets. In this process, the initiator first adds spurious support count to all the candidate sets in order to prevent the disclosure of exact support count of initiator itemsets to the next randomly selected site and splits the global candidate set into two sets randomly called the *retained set* and the *released set*. The *retained set* is the set which is retained in the current site itself and the *released set* is released to the randomly selected site. Thus, each participating site will send the sets in two rounds. Random site selection makes it complex for the colluders to find the exact itemsets of any particular site.

a. Problem Definition

Let $\{S_1, S_2 \ldots S_n\}$ be the set of participating sites where $n>2$. Let $D_1, D_2 \ldots D_n$ be the data sources of sites $S_1, S_2 \ldots S_n$ respectively which are geographically distributed and let $I = \{i_1, i_2 \ldots i_m\}$ be a set of items. Each transaction T in D_i such that $T \subseteq I$, where i=1 to n. LL_i be the local frequent itemset generated from a participating site S_i and G be the global frequent itemset. To generate the global frequent itemset G, each site sends its respective support counts of its local frequent itemsets to other participating sites. The intended goal of this proposed approach is to discover the global frequent itemsets without revealing the sensitive information of all the participating sites, where the sites are assumed to be colluding.

b. Elliptic Curve Cryptography

Data transmitted across a network is insecure and vulnerable to many types of attack nowadays. Asymmetric cryptography based communication is comparatively much secured than symmetric cryptography because asymmetric cryptographic algorithms have the property of using a pair of keys. One of the keys i.e. the public key is used for encryption and its corresponding private key must be used for decryption. The critical feature of asymmetric cryptography, which makes it useful, is the fact that one of the keys cannot be obtained from the other. This feature of asymmetric cryptosystems greatly simplifies key exchanges and key

management.

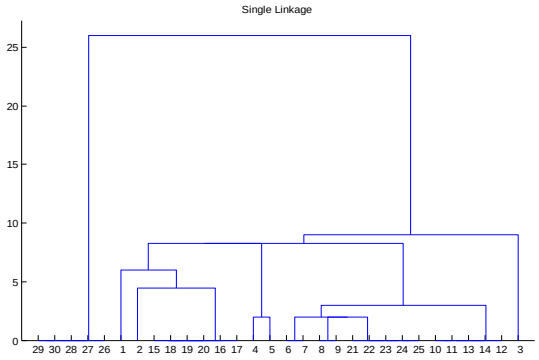
Diffie-Hellman, RSA and Elliptic Curve Cryptography (ECC) are the examples of asymmetric cryptography algorithms (Yao.A.C. 1986). Among those algorithms ECC is efficient due to shorter key size, faster key generation and faster decryption and its inverse operation gets harder, faster, against increasing key length than do the inverse operations in Diffie Hellman and RSA. The security strength of ECC algorithm is mainly determined by a mathematical principle called as the discrete logarithmic problem.

The ECC approach is a mathematically richer procedure than standard systems like RSA. The basic units for this cryptosystem are points (x,y) on an Elliptic curve, E(Fp), of the form, $y^2 = x^3+ax+b$, with x, y, a, b$^{\epsilon}$ Fp ={1, 2, 3, ... , p-2, p-1}, where Fp is a finite field of prime numbers. One basic condition for any cryptosystem is that the system is closed, i.e. any operation on an element of the system results in another element of the system. In order to satisfy this condition for elliptic curves it is necessary to construct nonstandard addition and multiplication operations.

V. RESULT AND DISCUSSION

To encrypt P, a user picks an integer, k, at random and sends the pair of points (k*BP, P+k*PUBKEY). The decryption is done by multiplying the first component with the secret key, s, and subtracting from the second component, (P+ k*PUBKEY) - s*(k*BP) = P+k*(s*(BP)) - s*(k*BP) = P. Then reverse the embedding process to produce the message, m, from the point P. This system requires a high level of mathematical abstraction to implement (Burnett, Winters & Dowling, 2002).

The database is preprocessed by removing duplicate records and supplying missing values. The clustering of data with single linkage and complete linkage using the Euclidean distances are shown in Fig.4(a) and 4(b). After clustering, we would like to measure the validity of the clustering. One way to do this would be to compare the distances between all observations with the links in the dendrogram. If the clustering is a valid one, then there should be a strong correlation between them. We can measure this using the cophenetic correlation coefficient.



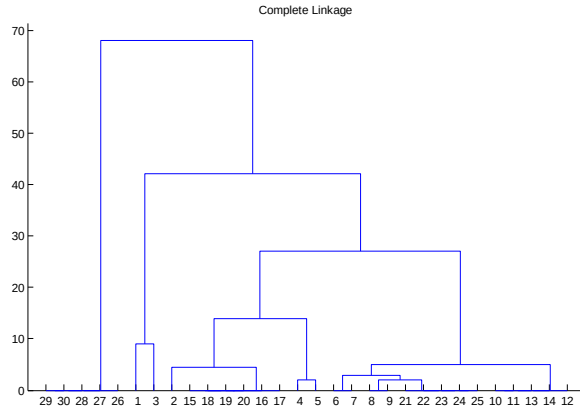
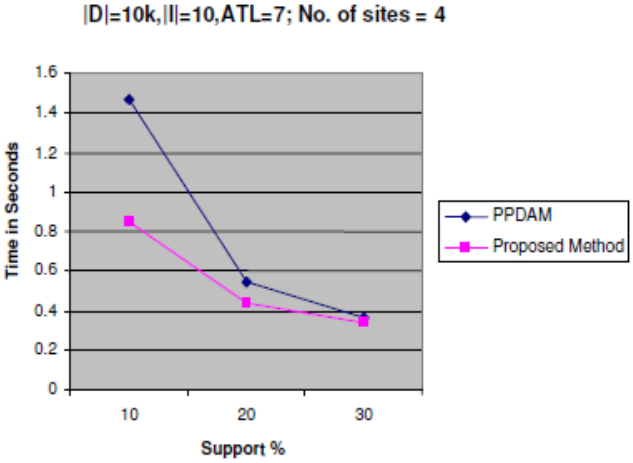


Fig.3(a) Single Linkage (b) Complete linkage

Time complexity is another major criterion considered for performance evaluation. Figure 4 compares the time complexity of the proposed method with the existing method. From the simulation results, the proposed method takes minimum time compared to the PPDAM. All these experiments are conducted for the supports varying from 30% to 10%. The time requirement to generate the global frequent itemsets on data source 100Ks by the existing method was from 0.45 seconds to 12.34 seconds and by the proposed method was from 0.36 seconds to 5.6 seconds. From this, it is evident that the proposed method performs better on dense data sets.



VI. CONCLUSION

Distributed data mining is useful for retrieving information from data sources distributed among several sites. Existing approaches suffer from the intended purposes of maintaining privacy of the individual participating sites, reducing computational complexity and minimizing communication overhead. The proposed method determines global frequent itemsets in a distributed environment with least communication among sites and ensures higher degree of privacy with randomized site selection. The experimental analysis shows that proposed method generates global frequent itemsets among colluded sites without affecting mining performance and confirms optimal communication among sites.

REFERENCES

[1] J.P. Bigus.(1996),"Data Mining with Neural Networks", New York: McGraw- Hill,

[2] A survey of Knowledge Discovery and Data Mining process models

The Review, Vol. 21:1- 2006, Cambridge University Press Printed in the United Kingdom.

[3] Sellappan, P., Chua, S.L.: "Model-based Healthcare Decision Support System", Proc. Of Int. Conf. on Information Technology in Asia CITA'05, 45-50, Kuching, Sarawak, Malaysia, 2005

[4] Atallah, M., Bertino, E., Elmagarmid, A., Ibrahim, M. & Verykios, V.S. (1999). Disclosure limitation of sensitive rules. In Proceedings of the IEEE Knowledge and Data Exchange Workshop (KDEX'99). IEEE Computer Society, 45-52.

[5] Burnett.A, Winters.K, and Dowling.T, (2002). A Java implementation of an elliptic curve Cryptosystem-Java programming and practices. In Proceedings of the inaugural conference on the Principles and Practice of programming.

[6] Cheung, D., Ng, V., Fu, A. & Fu, Y.(1996). Efficient Mining of Association Rules in Distributed Databases. IEEE Transactions on Knowledge and Data Engineering. 8(6), 911-922.

[7] Clifton, C. (2001) Secure Multiparty Computation Problems and Their Applications: A Review and Open Problems. In Proceedings of the Workshop on New Security Paradigms, Cloudcroft, New Mexico.

[8] Clifton, C., Kantarcioglu, M. & Vaidya, J.(2004). Defining privacy for data mining. Book Chapter Data Mining, Next generation challenges and future directions.

[9] Mehmed, K.: "Data mining: Concepts, Models, Methods and Algorithms", New Jersey: John Wiley, 2003.

[10] Mohd, H., Mohamed, S. H. S.: "Acceptance Model of Electronic Medical Record", Journal of Advancing Information and Management Studies. 2(1), 75-92, 2005m]. Volume(issue), paging if given. Available: http://www.(URL)

T.Revathi is B.Sc (Applied Science-Computer Technology), MCA,M.Phil., from PSG College of Arts & Science, Coimbatore. She is presently pursuing her Ph.D in Manonmaniam Sundaranar University, Tirunelveli. From 1997 to 2001 August 8 she was working in Sri Ramakrishna Polytechnic College, Coimbatore. Since 2001 August she has been working as Assistant Professor in Department of Computer Sicence,PSG College of Arts & Science, Coimbatore. Her area of research is Data Mining. She can be reached at **trevathi_psg@yahoo.co.in.**

Dr.P.Sumathi is MCA from Kongu Engineering College, Perundurai, M.Phil from Mother Teresa University, and Ph.D from PSG College of Arts & Science. She was working as Lecturer in Sengunthar Arts And Science College , Tiruchengode from June 1997 to May 1998. From June 1998 to February 2011 she was working as Assistant Professor & Head, Department of Computer Science, PSG College of Arts & Science, Coimbatore. Worked as Assistant Professor in Chikkanna Government Arts College, Tirupur and presently she is working as Assistant Professor in PG & Research Department, Department of Computer Science, Government Arts College, Coimbatore. Her area of research includes Data Mining and Grid Computing.