Privacy Preserving Association Rule Mining in Vertically Partitioned Databases

N. V. Muthu Lakshmi[1] & K. Sandhya Rani[2]
[1]Research Scholar, [2]Professor
Dept of Computer Science, S.P.M.V.V
Tirupati, Andhra Pradesh, INDIA

ABSTRACT
Data mining techniques are useful to discover hidden patterns from the large databases. Association rule mining is one of the important data mining techniques to discover relationships between items or item sets. In many organizations the database may exist in centralized or in distributed environment. In distributed environment, database may be partitioned in different ways such as horizontally partitioned, vertically partitioned and mixed mode which consists of both horizontal and vertical partitioning methods. The sites in the distributed environment interested to find association rules by participating themselves in the mining process without disclosing their individual private data/information. In this paper, a new model is proposed to find association rules by satisfying the privacy constraints for vertically partitioned databases at n number of sites along with data miner. This model adopts cryptography techniques such as encryption, decryption techniques and scalar product technique to find association rules efficiently and securely for vertically partitioned databases.

Keywords
Privacy Preserving Association Rule Mining, Distributed Databases, Cryptography, Scalar Product

1. INTRODUCTION
The main aim of data mining technology is to explore hidden information from large databases. Many data mining techniques are exist such as association rule mining, clustering, classification and so on are well known and have wide applications in the real world. In recent years, many organizations are showing interest to share the knowledge with other parties to get mutual benefits but at the same time no organization is willing to provide their private data. To achieve this, new area of research that is privacy preserving data mining has evolved. The main aim of privacy preserving data mining is to discover the uncovered information from large database while protecting the sensitive data/information of individuals.

The issue of privacy arises in two situations namely centralized and distributed environment. In centralized environment, database is available in single location and the multiple users are allowed to access the database .The main aim of privacy preserving data mining in this situation is to perform the mining process by hiding sensitive data/information from users. In distributed environment, the database is available across multiple sites and the main aim of privacy preserving data mining in this environment is to find the global mining results by preserving the individual sites private data/information. Every site can access the global results which are useful for analysis. In recent years, many researchers are focusing on privacy preserving data mining in distributed environment as it is having lot of applications in diverse fields.

In distributed database environment, the database among different sites can be partitioned as horizontally, vertically and mixed mode. Many privacy preserving data mining algorithms have been proposed for different partitioning methods in order to find the global mining results by satisfying the privacy constraints. In horizontally partitioned database, each site possess different set of tuples for the same set of attributes where as in the case of vertically partitioned databases each site possess the common set of transactions for distinct set of attributes. In mixed partitioning method, data is partitioned horizontally and then each horizontally partitioned database is further partitioned into vertical and vice versa. Among many data mining techniques, association rule mining is receiving more attention from the researchers to discover the associations between item sets. When many users are interested to know the global mining results without disclosing their private data, the issue of privacy arises in distributed environment. The issue of privacy also arises even in centralized environment where sensitive data/information exist and which has to be protected from the users. In this case sensitive rules are to be hidden.

In this paper, privacy preserving association rule mining for n number of vertically partitioned databases at n sites along with data mine where no site can be treated as trusted party is considered and is discussed in the next section.

2. PRIVACY PRESERVING ASSOCIATION RULE MINING FOR VERTICALLY PARTITIONED DATABASES
The association rule mining is getting more attention from researchers since its wide usage in many real life applications which helps the people to take right decisions to improve the performance of the business or service organization. But the main threat to the association rule mining is privacy when there is a requirement that knowledge is to be shared among many users who may be called as legitimate or partners. The association rule mining problem can be formally stated as follows:

Let $I = \{i_1, i_2, \ldots i_M\}$ be the set of items and value of M indicates the number of attributes in item set list I. Let $D = \{T_1, T_2, T_3, \ldots T_N\}$ be the set of transactions in database and Ti represents transaction identifier of the i^{th} transaction. To find the support of an item set, count the number of transactions where the values for all the attributes in the item set are 1. To find whether an item set is frequent or not, compare its support count with minimum

support threshold given by the user and if its support is greater than or equal to minimum support threshold then the item set is frequent otherwise it is infrequent. An association rule is an implication of the form $X \rightarrow Y$, where $x_i = 1$ and $X \cap Y = \phi$. The rule $X \rightarrow Y$ holds in the transaction set D with confidence c if c% of transactions in D that contains X also contains Y . In other words, to consider the rule to be a strong rule, confidence of an association to be computed which is $\sup(X \rightarrow Y) / \sup(X)$ and whose value must be greater than or equal to minimum confidence threshold value.

In distributed environment, the frequent item sets computed from all sites databases is called global frequent item sets where as individual site's frequent item sets computed from their database is called local frequent item sets. The process of finding global frequent item set for two parties can be specified as follows:

Let $Site_1$ and $Site_2$ are two sites possessing vertically partitioned databases DB_1 and DB_2.

$Site_1$ has L number of attributes and $Site_2$ has M number of attributes. Let MinSup be the support threshold specified by the user, and n be the total number of transactions. So the total number of attributes for two databases is $L + M$, where $Site_1$ has attributes A_1 through A_L and $Site_2$ has the remaining M attributes B_1 through B_M. Transactions are for the two databases consists of values of zero's or one's for $L + M$ attributes. Let $\vec{X}$ and $\vec{Y}$ are vectors represent columns in the database, that is $x_i = 1$ 1, if and only if row i has value 1 for attribute X. The scalar product of two cardinality n vectors $\vec{X}$ and $\vec{Y}$ is defined as

$$\vec{X}.\vec{Y} = \sum_{i=1}^{n} x_i * y_i$$

To find whether an item set (XY) is globally frequent or not by comparing value of $\vec{X}.\vec{Y}$ with MinSup. If the value is greater then or equal to MinSup then the item set is globally frequent otherwise globally infrequent.

The same definition can be extended to any number of sites and for any distinct set of attributes in sites.

Many researchers proposed various methods for privacy preserving association rule mining for both centralized and distributed databases. The various methodologies such as randomization, perturbation, heuristic and cryptography techniques are proposed by the authors to find privacy preserving association rule mining in horizontally and vertically partitioned databases. Among many techniques cryptography is the most popular and widely used technique to apply for horizontal, vertical and mixed mode partitioned databases since it gives exact solution which provides informational accuracy to users and at the same time privacy constraints are satisfied.

Earlier work in privacy preserving association rule mining in distributed environment is as follows:

The authors discussed the state of art in the area of privacy preserving data mining techniques [1]. They also discussed about classifications of privacy preserving techniques and privacy preserving algorithms such as heuristic-based technique, reconstruction based technique and cryptography-based technique. In [2], authors presented a framework for comparing different privacy preserving data mining algorithms and also discussed the results based on evaluation criteria for specific set of algorithms. Four efficient methods namely secure sum, secure set union, secure size of set intersection and scalar product for privacy preserving data mining in distributed environment are proposed in [3]. The author addressed in [4], the problem of finding association rules in case of horizontally partitioned databases, vertically partitioned and mixed partitioned databases and presented new algorithms for each case. Each algorithm is discussed with privacy preserving data mining evaluation metrics. The authors in [5] proposed private scalar product protocol based on homomorphic encryption for n number of vertically partitioned databases and can be applied to massive data sets. A new method is proposed in [6], to find association rules using secure scalar product for n number of vertically partitioned databases. Bit representation of item set inclusion in transactions is used to compute the frequency of the corresponding item sets in and this is a key step to find frequent item sets among n sites. The author in [7] proposed an efficient algorithm for finding privacy preserving association rule mining for vertically partitioned databases.

An enhanced kantarcioglu and Clifton scheme protocol is proposed in [8], which is a two phase privacy preserving distributed data mining for horizontally partitioned databases and this protocol is resilient to collision and can be applied to both cases that is with trusted party or without trusted party. Authors in [9] addressed cryptography role based access control for privacy preserving data mining. They proposed a new solution by integrating the advantages of the first approach which protects the privacy of the data by using an extended role based access control approach and the second approach which uses cryptographic technique to store sensitive data and providing access to the stored data based on an individual's role. In [10], the authors addresses the common errors when secure multi party computation techniques are applied to privacy preserving data mining. They also discussed the relationship between secure multi party computation and privacy preserving data mining. In [11][12] the authors presented algorithms for privacy preserving association rule mining for horizontally partitioned databases.

In [13], authors proposed a general protocol for privacy preserving multiparty scalar product computations which be used for obtaining trust values from private recommendations. They also proposed credential based trust model where the trustworthiness of the user is computed based on his/her affiliations and role assignments. The authors in [14] proposed a protocol for conducting secure regressions and similar analyses on vertically partitioned data. This protocol allows data owners to estimate coefficients and standard errors of linear regressions, and to examine regression model diagnostics, without disclosing the values of their attributes to each other and no third parties are involved. They also discussed the basis of an algorithm for secure matrix multiplication, which is used by pairs of owners to compute off-diagonal blocks of the full data covariance matrix. In [15], simple technique of transforming the categorical and numeric sensitive data using a mapping table and graded grouping technique respectively are discussed. The authors also discussed the proposed technique with data mining tasks such as classification, clustering and association rule mining and the results are analyzed. The authors presented

an extremely efficient and sufficiently secure protocol for computing the dot-product of two vectors using linear algebraic techniques [16]. Using analytical as well as experimental results, they demonstrated the performance in terms of computational overhead, numerical stability, and security.

An efficient two party scalar product protocol is proposed with an un trusted third party [17] and also analysis is discussed to demonstrate its effectiveness. A new paradigm, in which an acceptable security is used in the proposed model [18] that allows partial information disclosure. The basic idea in this model is lowering the restriction on the security can achieve a much better performance and also that people do accept a less secure but much more efficient solution. In [19],authors presented efficient protocols for 1-out-of-N Oblivious Transfer.

Many scalar product protocols have been proposed by the authors in [20]. The authors presented secure permutation algorithm which simultaneously computes a vector sum and permutes the order of the elements in the vector. The authors in [21] proposed a new architecture for finding privacy preserving data mining technique using two different entities Miner and Calculator without using secure computation or perturbation. They presented algorithms for vertically and horizontally partitioned databases. A novel hash function based sensitive information obfuscation technique is proposed for association rule mining on vertically partitioned database [22].

In this paper, new methodology is proposed to find privacy preserving association rule mining for vertically partitioned distributed database with data miner (DM). Cryptography techniques are used such as encryption & decryption technique and scalar product protocol are used to find global frequent item sets along with support values while protecting one's private data/information from others accessing, . In this method, a special site is designated and this site owner is called data miner (DM) who initiates the process of finding association rules without knowing any one's individual data/information even when he receives processed results from the last site. The function of the proposed model is discussed in the following section.

3. PROPOSED MODEL

The proposed model consists of n number of sites and a data miner (DM). Each site, $Site_i$ (i=1,..n) consists of a database DB_i and each DB_i consists of disjoint attributes for the same set of transactions that is the same transaction with different set of attributes at all sites. The role of DM is to initiate the process by sending MinSup threshold and public key to all sites. DM also participates in the encryption and decryption process for frequent item sets in order to protect individual sites attributes information that is names of attributes & number of attributes exists in a site and their support values. DM is having privileges to find the global frequent item sets and their support values. He also generates the association rules which are then broadcasted to all sites.

The main goal of the proposed model is to find the global association rules without revealing any individual sites data/information. The communication among three sites and DM is shown in the following diagram.

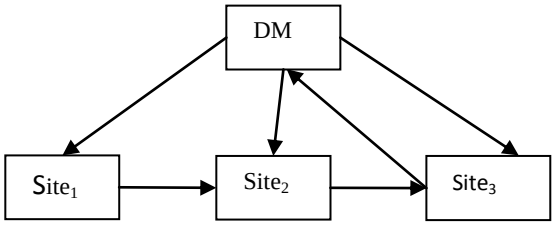

Figure1: Communication among three sites and DM

Every site communicates with its successor site only except the last site, $Site_n$ communicates with DM. DM is having communications with all sites, $Site_1$ to $Site_n$. Each site performs the computations by using scalar product concept with its own computed results and the computed results obtained from its predecessor site. The various steps in the proposed model are as follows:

Step1. DM initiates the process by broadcasting MinSup threshold and public key to all sites.

Step2. Each site converts its database into Transaction Identifier (TID) list approach.

Step3. Each site finds local frequent item sets for itsTID list based on the MinSup threshold which is received from miner.

Step4. For each $Site_k$, k ranges from 1 to n , prepares a matrix M_k in which each row represents a local frequent item set's transactions. In this matrix, if $M_k(i,j) = 1$ indicates that j^{th} transaction supports the i^{th} local frequent item set at $Site_k$.*Step5.* Each site, $Site_k$ prepares a vector V_k, (k ranges from 1 to n), which consists of local frequent item sets. It is very important to maintain a relationship between vector and the matrix that is i^{th} element in the vector corresponds to the transactions for the i^{th} row of the matrix.

Step6. Each site encrypts all frequent item sets in vector V_k by using public key, which is received from DM.

Step7. The first site sends matrix M_1 and the encrypted frequent item set list enV_1 to $Site_2$.

Step8. The second site($Site_2$) performs $M_1.M_2$ by using the concept of scalar product and prepares a matrix M_{12} which consists of only frequent item sets of $M_1.M_2$. $Site_2$ then prepares a matrix M_2' which consists of M_1, M_2 and M_{12}.

Step9. $Site_2$ prepares a vector enV_2' which consists of encrypted frequent item set list(s) enV_1, enV_2 and enV_{12} where enV_{12} represents the encrypted frequent item sets of M_{12} . $Site_2$ sends matrix M_2' along with vector enV'_2 to its successor site.

Step10. Each site, $Site_i$ in the remaining sequence of $Site_3,\ldots$ $Site_n$ performs step8 based on the received matrix and vector (M'_{i-1}, enV'_{i-1}) from its predecessor site and its own matrix (M_i) & vector (enV_i).

Step11. The last site, ($Site_n$) possess a matrix M'_n and enV'_n. $Site_n$ applies a sorting technique on enV'_n based on the length of encrypted form of frequent item sets in descending order. According to the position of frequent item set in the sorted list, the matrix M'_n is rearranged to preserve the order. This matrix M'_n along with enV'_n is sent to the DM.

Step12. The DM applies the decryption algorithm using private key for each element in the vector enV'_n to get the frequent item sets. The decrypted frequent item sets are nothing but global frequent item sets. The DM finds the support for each global frequent item set by counting the number of one's in the corresponding row of a matrix M'_n and prepares a list which consists of global frequent item sets with their support values.

Step13. Based on the list, DM generates association rules for each global frequent item set by using minimum confidence threshold specified by the user.

Step14. The generated rules are broadcasted to all sites.

4. IMPLEMENTATION OF PROPOSED MODEL WITH SAMPLE DATABASES

The proposed model is illustrated by taking three sites and each site possesses vertically partitioned databases. The three sites $Site_1$, Site2 and $Site_3$ have databases DB_1, DB_2 and DB_3 respectively. The sample databases consist of 6 transactions of different set of attributes at three sites and are shown in the following tables.

Table 1: Database DB_1 at $Site_1$

TID\Item	A_1	A_2	A_3
T1	1	1	1
T2	1	0	1
T3	1	0	1
T4	0	1	0
T5	1	0	1
T6	0	1	0

Table 2: DB_2 at $Site_2$

TID\Item	A_4	A_5
T1	1	0
T2	0	1
T3	1	1
T4	1	0
T5	1	0
T6	0	1

Table 3: DB_3 at $Site_3$

TID\Item	A_6	A_7	A_8	A_9
T1	1	0	0	1
T2	0	1	0	0
T3	0	1	1	1
T4	1	1	0	0
T5	0	0	1	0
T6	1	1	1	1

DM requests all three sites to participate in the mining process in order to find global frequent item sets by sending minimum support threshold value 40%.

Every site converts its database into Transaction Identifier (TID) list form and applies frequent item set generation algorithm to find set of locally frequent item sets based on user specified minimum support threshold 40%.

At site1:

Site 1 prepares a matrix M_1 and a vector V_1.

$$M_1 = \begin{bmatrix} 1 & 1 & 1 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 & 0 & 1 \\ 1 & 1 & 1 & 0 & 1 & 0 \\ 1 & 1 & 1 & 0 & 1 & 0 \end{bmatrix}$$

$$V_1 = \{A_1, A_2, A_3, (A_1, A_3)\}$$

$Site_1$ encrypts each element of V_1. The encrypted form of locally frequent item sets at $Site_1$ as

$$enV_1 = \{e(A_1), e(A_2), e(A_3), e(A_1, A_3)\}$$

$Site_1$ sends M_1 and enV_1 to $Site_2$ to compute frequent item sets between their individual frequent item sets.

At $Site_2$:

$Site_2$ has matrix M_2 and enV_2 (encrypted form of enV_2) as shown as

$$M_2 = \begin{bmatrix} 1 & 0 & 1 & 1 & 1 & 0 \\ 0 & 1 & 1 & 0 & 0 & 1 \end{bmatrix}$$

$$enV_2 = \{ e(A_4), e(A_5) \}$$

$Site_2$ finds matrix M_{12} and vector enV_{12} based on M_1, enV_1, M_2 and enV_2.

$$M_{12} = \begin{bmatrix} 1 & 1 & 1 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 & 0 & 1 \\ 1 & 1 & 1 & 0 & 1 & 0 \\ 1 & 1 & 1 & 0 & 1 & 0 \end{bmatrix}^{M_1} \cdot \begin{bmatrix} 1 & 0 & 1 & 1 & 1 & 0 \\ 0 & 1 & 1 & 0 & 0 & 1 \end{bmatrix}^{M_2}$$

$$\therefore M_{12} = \begin{bmatrix} 1 & 0 & 1 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 & 1 & 0 \end{bmatrix}$$

M'_2 can be computed by appending M_2, M_{12} to M_1 and enV'_2 is formed by appending enV_2, enV_{12} to enV_1 and is shown as

$$M_2' = \begin{bmatrix} 1 & 1 & 1 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 & 0 & 1 \\ 1 & 1 & 1 & 0 & 1 & 0 \\ 1 & 1 & 1 & 0 & 1 & 0 \\ 1 & 0 & 1 & 1 & 1 & 0 \\ 0 & 1 & 1 & 0 & 0 & 1 \\ 1 & 0 & 1 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 & 1 & 0 \end{bmatrix}$$

$$enV'_2 = \{e(A_1), e(A_2), e(A_3), e(A_1, A_3), e(A_4), e(A_5), e(A_1, A_4)), e(A_3, A_4)), e((A_1, A_3), A_4)\}$$

Now $Site_2$ sends M'_2 and enV'_2 to $Site_3$ to find frequent item sets between their frequent item sets.

At $Site_3$:
$Site_3$ has matrix M_3 and vector enV_3 as

$$M_3 = \begin{bmatrix} 1 & 0 & 0 & 1 & 0 & 1 \\ 0 & 1 & 1 & 1 & 0 & 1 \\ 0 & 0 & 1 & 0 & 1 & 1 \\ 1 & 0 & 1 & 0 & 0 & 1 \\ 1 & 0 & 0 & 1 & 0 & 1 \end{bmatrix}$$

Encrypted form of vector enV_3 is

$$\{e(A_6),\ e(A_7),\ e(A_8),\ e(A_9),\ (e(A_6,A_7))\}$$

$Site_3$ computes M'_{23} by doing scalar product with M'_2 with M_3 as specified below:

$$M'_{23} = \begin{bmatrix} 1 & 1 & 1 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 & 0 & 1 \\ 1 & 1 & 1 & 0 & 1 & 0 \\ 1 & 1 & 1 & 0 & 1 & 0 \\ 1 & 0 & 1 & 1 & 1 & 0 \\ 0 & 1 & 1 & 0 & 0 & 1 \\ 1 & 0 & 1 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 & 1 & 0 \end{bmatrix} \cdot \begin{bmatrix} 1 & 0 & 0 & 1 & 0 & 1 \\ 0 & 1 & 1 & 1 & 0 & 1 \\ 0 & 0 & 1 & 0 & 1 & 1 \\ 1 & 0 & 1 & 0 & 0 & 1 \\ 1 & 0 & 0 & 1 & 0 & 1 \end{bmatrix}$$

$$\therefore M'_{23} = \begin{bmatrix} 1 & 0 & 0 & 1 & 0 & 1 \\ 1 & 0 & 0 & 1 & 0 & 1 \\ 0 & 1 & 1 & 0 & 0 & 1 \end{bmatrix}$$

$$enV'_{23} = \{\ e(A_2,A_6),\ e(A_2,(A_6,A_7)),\ e(A_5,A_7)\}$$

$Site_3$ prepares a matrix M'_3 by appending M_3, M'_{23} to M'_2 . The encrypted vector is formed by appending enV_3, enV'_{23} to enV'_2.

$$\therefore enV'_3 = \begin{Bmatrix} \{e(A_1),\ e(A_2),\ e(A_3),\ e(A_1,\ A_3), e(A_4),\ e(A_5), \\ e(A_1,A_4)), e(A_3,A_4)),\ e((A_1,A_3),A_4),\ e(A_6), \\ e(A_7),\ e(A_8),\ e(A_9), e(A_6,A_7),\ e(A_2,A_6), \\ e(A_2,(A_6,A_7)),\ e(A_5,A_7)\} \end{Bmatrix}$$

Since $Site_3$ is the last site, sorts the frequent item sets in vector, enV'_3 in descending order based on the length of the item set list.
Therefore Sorted list of vector enV'_3 is as

$$\{e((A_1,A_3),A_4),\ e(A2,(A_6,A_7)),\ e(A_1,\ A_3), e(A_1,A_4)),$$
$$e(A_3,A_4)),\ e(A_6,A_7),\ e(A_2,A_6),\ e(A_5,A_7),\ e(A_1),$$
$$e(A_2),\ e(A_3),\ e(A_4), e(A_5), e(A_6),\ e(A_7),\ e(A_8),\ e(A_9)\}$$

According to the sorted frequent item set (in encrypted form) of vector enV'_3 matrix M'_3 is rearranged.

The matrix M'_3 and corresponding rearranged matrix ,

RM'_3 is specified in the following table.

Table 4: $Site_3$ Computed Matrix and Rearranged Matrix

Matrix, M'_3	Rearranged Matrix, RM'_3.
1 1 1 0 1 0	1 0 1 0 1 0
1 0 0 1 0 1	1 0 0 1 0 1
1 1 1 0 1 0	1 1 1 0 1 0
1 1 1 0 1 0	1 0 1 0 1 0
1 0 1 1 1 0	1 0 1 0 1 0
0 1 1 0 0 1	1 0 0 1 0 1
1 0 1 0 1 0	1 0 0 1 0 1
1 0 1 0 1 0	0 1 1 0 0 1
1 0 1 0 1 0	1 1 1 0 1 0
1 0 0 1 0 1	1 0 0 1 0 1
0 1 1 1 0 1	1 1 1 0 1 0
0 0 1 0 1 1	1 0 1 1 1 0
1 0 1 0 0 1	0 1 1 0 0 1
1 0 0 1 0 1	1 0 0 1 0 1
1 0 0 1 0 1	0 1 1 1 0 1
1 0 0 1 0 1	0 0 1 0 1 1
0 1 1 0 0 1	1 0 1 0 0 1

Now $Site_3$ sends matrix RM'_3 and enV'_3 to DM to find global frequent item sets.

At site DM:
DM receives the above results from last site ($Site_3$) and then applies decryption algorithm to find actual frequent item sets from the encrypted form of frequent item sets vector by using private key. He also finds global support for each frequent item set by counting number of one' in the corresponding row in the received matrix, RM'_3 and finally prepares a list consisting of only global frequent item sets (whose support value is greater than or equal to 40% of the database) along with their support. The following table shows global frequent item sets and their supports for three vertically partitioned databases DB_1, DB_2 and DB_3 at three sites $Site_1$, $Site_2$ and $Site_3$.

Table 5: Global frequent item sets and their supports

Item Sets	Sup	Item Sets	Sup	Item Sets	Sup
A_1	4	A_7	4	(A_2,A_6)	3
A_2	3	A_8	3	(A_5,A_7)	3
A_3	4	A_9	3	(A_6,A_7)	3
A_4	4	(A_1,A_3)	4	(A_1,A_3,A_4)	3
A_5	3	(A_1,A_4)	3	(A_2,A_6,A_7)	3
A_6	3	(A_3,A_4)	3		

The DM generates association rules for each global frequent item set based on user specified minimum confidence threshold value.

The following illustrates how to generate association rules for any global frequent item set based on minimum confidence threshold value 70%.

Let us consider a global frequent item set, (A_1, A_4).

The rules can be constructed as

$A1 \rightarrow A4$

$A4 \rightarrow A1$

By computing the confidence value for these rules we can find strong rules as

Confidence of a rule

$$\text{ConfidenceofaRule}\, A_4 \rightarrow A_1 = \text{Sup}(A_1, A_4) / \text{Sup}(A_4) = 3/4 = 75\%$$

$$\text{ConfidenceofaRule}\, A_4 \rightarrow A_1 = \text{Sup}(A_1, A_4) / \text{Sup}(A_4) = 3/4 = 75\%$$

As confidence values are greater than or equal to 70%, both are strong rules.

Hence the above rules are strong rules for item set (A_1, A_4) In the similar way association rules can be determined based on the user specified minimum confidence threshold for each global frequent item set. Finally the DM broadcast all the strong rules to all three sites.

5. PERFORMANCE OF THE PROPOSED MODEL

- In the proposed model, each site's database is represented in TID form which facilitates easy computations of local frequent item sets for its database by using scalar product technique. This TID form also helps to find the scalar product between the predecessor site´s computed results with its own results in order to obtain all the frequent item sets for all possible combinations of attributes related to all the sites databases which are processed so far (all predecessor sites and its own).
- By adopting encryption, decryption cryptography technique in the proposed model, no successor site can predict its predecessor site's data/information when it receives processed results from predecessor site.
- By adopting scalar product technique in the proposed model, every successor site can efficiently determines the frequent item sets between its own frequent item sets and all predecessors sites frequent item sets. The scalar product technique helps to explore all possible combination of predecessor site's frequent item sets with successor site's frequent item sets. This technique also helps to determine which frequent item sets are by counting the number of one's in the computed matrix and if the value of count is greater than or equal to MinSup then the item set is declared to be frequent for further processing.
- Although every site appends its computed results to the received results (consists of processed results of all predecessor sites) sent by its predecessor site in finding globally frequent item sets, no site can predict any predecessor site's private data/information such as attributes, local frequent item sets, support values as frequent item sets are in encrypted form in the received results.
- DM can not predict any site's private data/information even when DM is having certain privileges such as initiation of the mining process, decryption of frequent item sets, finding global frequent item sets and their supports, generation of association rules.
- The DM receives processed results from sites which consist of local frequent item sets of all possible combinations of attributes of all sites and related supporting transactions. These transactions are obtained after completion of process at all sites and based on these information, DM can not guess any individual site's private data/information.
- The data transfers between sites and last site to miner is performed as a bulk data transfer instead of single data transfer for each frequent item set. In the proposed model, only one data transfer is required for sending processed results from each predecessor site to its successor site. So only n number of data transfers are needed to obtain all sites processed results in order to find the global frequent item sets.
- As each site is having distinct set of attributes for the same set of transactions, the proposed model efficiently finds global frequent item sets by searching all possible combinations of attributes of all sites.

From the above discussion, the proposed model is easily, efficiently and with minimum number of data transfers, finds the global association rules for vertically partitioned databases without revealing any sites private data/information to any site and DM.

6. CONCLUSION

In this paper, a new model which utilizes the concept of scalar product is proposed to find global association rules when the database is partitioned vertically among n number of sites. In the proposed model, DM has privileges to initiate the mining process, finding global association rules. Secured computations for association rules is achieved with this model by preserving the privacy of the individual sites information. The functioning of the proposed model is illustrated with sample databases. With the proposed model, association rules can be generated easily, efficiently with minimum number of computations and communications by satisfying privacy constraints. The performance of this model is analyzed in terms of privacy and communications.

7. REFERENCES

[1] Verykios, V.S., Bertino, E., Nai Fovino, I., Parasiliti, L., Saygin, Y., and Theodoridis, Y. (2004), State-of-he-art in privacy preserving data mining, SIGMOD Record, 33(1):50–57.

[2] *Elisa Bertino , Igor Nai Fovino Loredana Parasiliti Provenza* ,A Framework for Evaluating Privacy Preserving Data Mining Algorithms, Data Mining and Knowledge Discovery, 2005, 11, 121–154.

[3] Chris Clifton, Murat Kantarcioglu, Jaideep Vaidya, Xiaodong Lin, and Michael Y. Zhu(2003),Tools for privacy preserving distributed data mining, SIGKDD Explorations, Vol. 4, No. 2 pp1-7.

[4] Mahmoud Hussein, Privcacy preserving in assocaiton rule mining using cryptography, Master thesis, Menofyia University,2009

[5] Bart Goethals, Sven Laur, Helger Lipmaa, and Taneli Mielik¨ainen, On Private Scalar Product Computation for Privacy-Preserving Data Mining,

[6] Vaidya, J. and Clifton, C. 2002. Privacy preserving association rule mining in vertically partitioned data,

8th ACM SIGKDD, International Conference on Knowledge Discovery and Data Mining, ACM Press,pp. 639–644.

[7] Pradeep Shenoy, Jayant R. Haritsa, S. Sundarshan, Gaurav Bhalotia, Mayank Bawa, and Devavrat Shah. Turbo-charging vertical mining of large databases, In Proceedings of the Nineteenth ACM SIGMOD International Conference on Management of Data, pages 22{33, Dallas, TX, 2000.

[8] *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, November 2006.

[9] Lalanthika Vasudevan , S.E. Deepa Sukanya, N.Aarthi ,Privacy Preserving Data Mining Using Cryptographic Role Based Access Control Approach, Proceedings of the International Multi Conference of Engineers and Computer Scientists 2008 Vol I, IMECS 2008.

[10] *Y. Lindell and B. Pinkas,* Secure Multiparty Computation for Privacy-Preserving Data Mining, The Journal of privacy and Confidentiality (2009) , 1, Number 1, pp. 59-98.

[11] A.C. Yao. Protocols for secure computations, In Proceedings of the 23rd Annual IEEE Symposium on Foundations of Computer Science, 1982.

[12] Ashraf El-Sisi , Fast Cryptographic Privacy Preserving Association Rules Mining on Distributed Homogenous Database, The International Arab journal of Information Technology, Vol. 7, No. 2, April 2010.

[13] Danfeng Yao, Roberto Tamassia ,Seth Proctor, Distributed Scalar Product Protocol with Application to Privacy Preserving Computation of Trust.

[14] Alan F. Karr, Xiaodong Lin, Ashish P. Sanil, JeromeP. Reiter, Privacy-Preserving Analysis of Vertically Partitioned Data Using Secure Matrix Products, Journal of Official Statistics, Vol.25, No.1, 2009. pp. 125–138.

[15] E. Poovammal, M. Ponnavaikko, Utility Independent Privacy Preserving Data Mining on Vertically Partitioned Data, Journal of Computer Science 5 (9): 666-673, 2009, Science Publications.

[16] Ioannidis I., Grama A., Atallah M., A secure protocol for computing dot products in clustered and distributed environments, Proceedings of International Conference on Parallel Processing, 2002. 379 – 384.

[17] Yiqun Huang, hengding Lu, HepingHu, Privacy preserving association rule mining with scalar product, Natural Language Processing and Knowledge Engineering, 2005. IEEE NLP-KE '05,Proceedings of 2005 IEEE, International Conference.

[18] Wenliang Du, Zhijun Zhan, A practical approach to solve Secure Multi-party Computation problems,Proceedings of NSPW'2002, workshop on New security paradigms, ACM publications.

[19] Moni Naor and Benny Pinkas. Efficient oblivious Transfer protocols. In Proceedings of SODA 2001 (SIAM Symposium on Discrete Algorithms), Washington, D.C., January 7-9 2001.

[20] Wenliang Du and Mikhail J. Atallah. Privacy preserving statistical analysis. In Proceeding of the Seventeenth Annual Computer Security Applications Conference, New Orleans, LA, December 10-14 2001.

[21] Alex Gurevich, Ehud Gudes, privacy Preserving data Mining Algorithms without the use of secure Computation or perturbation, 10th International Database Engineering and Applications Symposium (IDEAS'06),2006, IEEE.

[22] Syed Shams-ul-Haq, Csilla Farkas, Research on Contemporary Algorithms for Privacy Preserving Distributed Data Mining.