Privacy preserving association rule mining for n number of disjoint vertically partitioned databases with a Data Miner

Kasala Nageswar Rao[1],
[1]Lecturer, Department of Computer Science, S.R. & B.G.N.R Government Degree College.(Autonomous) Khammam. Telangana.

ABSTRACT

A novel technique is proposed to identify privacy preserving association rule mining for n sites with data miner for vertically partitioned databases. With no trusted party among n vertically partitioned sites, this method utilizes the services of data miner who is located at a special site. Cryptography techniques such as encryption, decryption techniques and scalar product technique are adopted by this model to identify association rule very efficient and secure for vertically partitioned databases. Encryption and decryption techniques helps in protecting one's local frequent item sets from other sites while computing global frequent item sets. Scalar product concept is applied to derive frequent item sets among the sites to identify global frequent item sets and their supports without knowing the attributes, support values and local frequent item sets. The data miner holds special privileges to start the mining to find global results.

1.INTRODUCTION

Privacy preserving association rule mining for n number of disjoint vertically partitioned databases is considered . The methodology proposed in this paper is to identify privacy preserving association rule mining for n number of disjoint vertically partitioned databases with data miner. The vertically partitioned distributed database model consists of two or more sites and each site possesses disjoint vertically partitioned database. Every vertically partitioned database consists of values of all transactions for only subset of attributes/items. Many authors proposed methodologies to identify privacy preserving association rule mining for vertically partitioned databases such as randomization, perturbation, heuristic and cryptography techniques. Cryptography is the most popular and also widely used technique when compared to many techniques. This technique is used to apply for vertical and that gives accurate results while each site's privacy constraints are satisfied.

Every site owner is keen to obtain global results like global association rules that satisfy privacy constraints based on user

specified confidence threshold values. In distributed partitioned model, the frequent item sets computed from all sites databases is known as global frequent item sets where as individual site's frequent item sets computed from their database is known as local frequent item sets. Global association rule is the association rule that is computed from global frequent item sets and their support values. This can be done only when individual site's frequent item/item sets are combined with one another while support values are considered.

The method of finding global frequent item set for two parties can be mentioned as follows:

Let Site1 and Site2 are two sites possessing vertically partitioned databases DB1 and DB2

Site1 has L number of attributes and Site2 has M number of attributes. Let MinSup be the support threshold specified by the user, and n be the total no.of transactions. So, total number of attributes for two databases is L + M, where Site1 has attributes A1 through AL and Site2 has the remaining M attributes B1 through BM. Transactions for the two databases consists of values of zero or one for L+M attributes. Let X and Y are vectors representing columns in the database, that is xi = 1, if and only if row i has value 1 for attribute X. The scalar product of two cardinality n vectors X and Y is defined as

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

To identify whether XY (item set) is globally frequent or not frequent by comparing value with MinSup. If this value is found to be greater than or equal to MinSup then it is said to be globally frequent else it is known as globally infrequent.

2. PROPOSED SYSTEM

The proposed model contains n number of sites and a data miner (DM). Each site, Sitei (i=1,.., n) contains database DBi and each DBi contains disjoint attributes for the same set of transactions with different set of attributes at all sites. 'The DMinitiate is the process of sending MinSup threshold and public key to all other sites. DM will participates in the encryption and decryption techniques to identify frequent item sets in-order to protect attributes information of individual sites like attributes name & number of attributes which are exists in a site and their corresponding support values. DM has the rights to find the global frequent item sets and their support values. It also produces the association rules that are to be broadcasted to all the sites'. The key objective of the proposed model is to identify the global association rules without revealing the individual sites data/information. The following diagram explains the communication between three sites and DM.

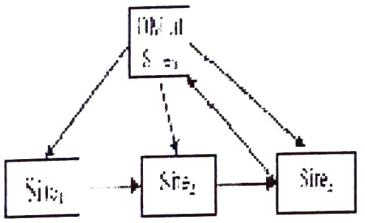


Figure 5.1: Communication Among Three Sites and DM

Each site communicates to its successor site except the last site, The last Site n will communicates with DM whereas DM communicates with all the sites, like Site1 till Site n. Each site executes the computations by applying scalar product perception with its own calculated results and the calculated results attained from its predecessor site. The steps in the proposed model are presented in the following section.

2.1 Algorithm

Step 1 DM initiates the process by broadcasting MinSup threshold and also public key to all sites. **Step 2** Every site converts its database into Transaction Identifier (TID) list approach.

Step 3 Every site finds local frequent item sets for its TID list based on the MinSup threshold which is received from miner.

Step 4 For each Sitek, k ranges from 1 to n , prepares a matrix Mk in which each row represents a local frequent item set's transactions. In this matrix, if Mk(i,j) =1 indicates that jth transaction supports the ith local frequent item set at Sitek.Step5. Every site, Sitek prepares a vector Vk, (k varies from 1 to n), that contains local frequent item sets. This is very essential to maintain a relationship between vector and the matrix that is ith element in the vector corresponds to the transactions for the ith row of the matrix.

Step 5 Each site encrypts all frequent item sets in vector Vk by using public key, which is received from DM.

Step 6 The first site directs matrix and the encrypted cipher text frequent item set list to Site2.

Step 7 The second site performs them by using the concept of scalar product and creates a matrix M12 which contains only frequent item sets of M1.M2. Site2 then prepares a matrix M2′ which consists of M1, M2 and M12. 149

Step 8 Site2 prepares a vector enV2′ which contains an encrypted frequent item set list(s) enV1, enV2 and enV12 where enV12 represents the encrypted frequent item sets of M12 . Site2 finally sends matrix M2′ and also vector enV′2 to its successor site.

Step 9 Each site, Site i in the left out sequence of Site3,...Siten executes step8 based on the obtained matrix and vector (M′i-1, enV′i-1) from its predecessor site and its own matrix (Mi) & vector (enVi).

Step 10 The last site, (Siten) possess a matrix M′n and enV′n..Siten applies a sorting technique on enV′n based on the length of encrypted form of frequent item sets in descending order. Based on 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.

Step 11 The DM enforces the decryption algorithm with the help of private key for every element in the vector enV′n to acquire the frequent item sets. The decrypted frequent item sets are the required global

frequent item sets. The DM computes the support for each global frequent item set by calculating the number of one's in the same row of a matrix M'n and organizes a list which contains global frequent item sets and their corresponding support values.

Step 12 Based on the list, DM generates association rules for every global frequent item set by using minimum confidence threshold mentioned by the user.

Step13 The generated rules are broadcasted to all sites.

2.2 Illustration of the Proposed Model:

The model proposed here is explained by considering three sites and each site possesses vertically partitioned databases. The three sites Site1, Site2 and Site3 have databases DB1, DB2 and DB3 respectively. The sample databases consist of 6 transactions of different set of attributes at three sites and are shown in the following tables.

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 1 Sample Database

T-ID\Item	A_4	A_5
T1	1	0
T2	0	1
T3	1	1
T4	1	0
T5	1	0
T6	0	1

Table 2 DB_2 at $Site_2$

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

Table 3 DB_3 at $Site_3$

DM requests all three sites to participate in the mining process to identify global frequent item sets with the help of 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 with the help of user specified minimum support threshold 40%.

At site1:

Site1 prepares a matrix M1 and a vector V1.

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

V1= {A1, A2, A3, (A1, A3)} Site1 encrypts each element of V1. The encrypted form of locally frequent item sets at Site1 as enV1 = {e(A1), e(A2), e(A3), e(A1, A3)} Site1 sends M1 and enV1 to Site2 to compute frequent item sets between their individual frequent item sets.

At Site2: Site2 has matrix M2 and enV2 (encrypted form of enV2) as shown as

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

enV2 = { e(A4), e(A5)} Site2 finds matrix M12 and vector enV12 based on M1, enV1, M2 and enV2.

$$M_{1.2} = \begin{bmatrix} 1 & 1 & 1 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 & 0 & 1 \\ 1 & 1 & 1 & 0 & 1 & 0 \\ 1 & 1 & 1 & 0 & 1 & 0 \end{bmatrix} \cdot \begin{bmatrix} 1 & 0 & 1 & 1 & 1 & 0 \\ 0 & 1 & 1 & 0 & 0 & 1 \end{bmatrix}$$

M′2 can be computed by appending M2, M1.2 to M1 and enV′2 is formed by appending enV2, enV12 to enV1 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(A1), e(A2), e(A3), e(A1, A3), e(A4), e(A5), e(A1,A4)), e(A3,A4)), e((A1,A3),A4) Now Site2 sends M′2 and enV′2 to Site3 to find frequent item sets between their frequent item sets.

At Site3:

Site3 has matrix M3 and vector enV3 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 enV3 is {e(A6), e(A7), e(A8), e(A9), (e(A6,A7))} Site3 computes M'23 by doing scalar product with M'2 with M3 as specified below:

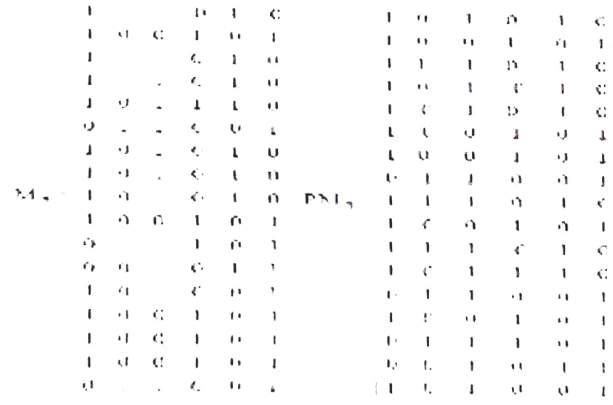
$$M'_{2.3} = \begin{bmatrix} 1 & 1 & 1 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 & 0 & 1 \\ 1 & 1 & 1 & 0 & 1 & 0 \\ 1 & 1 & 1 & 0 & 1 & 0 \\ 1 & 0 & 1 & 1 & 1 & 0 \\ 0 & 1 & 1 & 0 & 0 & 1 \\ 1 & 0 & 1 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 & 1 & 0 \end{bmatrix} \cdot M_3 = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 1 \\ 0 & 0 & 1 \\ 1 & 0 & 1 \\ 1 & 0 & 0 \end{bmatrix}$$

enV'23 = { e(A2,A6), e(A2,(A6,A7)), e(A5,A7)} Site3 prepares a matrix M'3 by appending M3, M'23 to M'2 . The encrypted vector is formed by appending enV3, enV'23 to enV'2.

enV'3 ={e(A1), e(A2), e(A3), e(A1, A3), e(A4), e(A5), e(A1,A4)), e(A3,A4)), e((A1,A3),A4), e(A6), e(A7), e(A8), e(A9), e(A6,A7), e(A2,A6), e(A2,(A6,A7)), e(A5,A7)} Since Site3 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((A1,A3),A4), e(A2,(A6,A7)), e(A1, A3), e(A1,A4)), e(A3,A4)), e(A6,A7), e(A2,A6), e(A5,A7), e(A1), e(A2), e(A3),

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 as follows:-



Now Site3 sends matrix RM'3 and enV'3 to DM to find global frequent item sets.

At Site0 (DM): DM receives the above results from last site (Site3) and then by applying decryption algorithm he finds 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 calculating number of one' in the corresponding row in the received matrix, RM'3 and finally prepares a list that holds only global frequent item along with their support. The following Table 5.4 shows global frequent item sets and their supports for three vertically partitioned databases DB1, DB2 and DB3 at three sites Site1, Site2 and Site3.

Table 4 Global Frequent Item Sets and Supports

The DM generates association rules for every global frequent item set with the help of user specified minimum confidence threshold value. The resulting demonstrates how to generate association rules for any global frequent item set with the help of minimum confidence threshold value 70%.

Let us consider a global frequent item set, (A1, A4). The two rules can be constructed based on this two item set (A1, A4) as A1→A4 and A4→A1

By computing the confidence value for these rules we can find strong rules as Confidence of a rule A1→A4 = Sup(A1,A4)/ Sup(A1) = ¾ = 75% Confidence of a rule A4→A1= Sup(A1,A4)/ Sup(A4) = ¾ = 75% As confidence values is always greater than or equal to 70%, both are strong rules. Hence the above rules are strong rules for item set (A1, A4). In the same manner association rules is determined with the help ofuser specified minimum confidence threshold for

Item Sets	Sup	Item Sets	Sup	Item Sets	Sup
[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
[illegible]	[illegible]	[illegible]	[illegible]		

each global frequent item set. Finally the DM broadcast all the strong rules to all three sites.

3.PERFORMANCE OF THE PROPOSED METHODOLOGY

In the proposed model, every site's database is represented in TID form that facilitates ease computations of local frequent item sets for its database by applying scalar product technique. This TID form also helps in finding the scalar product among the predecessor site´s computed results with its own results in order to acquire 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).

[redacted] ting encryption, decryption techniques in the proposed model, it is impossible for any successor site to predict its predecessor site's data/information when it receives processed results from predecessor site.

[redacted]

the proposed model, every successor site can efficiently determine the frequent item sets between its own frequent item sets and all predecessors sites frequent item sets. The scalar product technique helps in exploring all possible combination of predecessor site's frequent item sets with successor site's frequent item sets. This technique also aids to determine frequent item sets in-order to count 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.

Even though every site appends its computed results to the received results (consists of processed results of all predecessor sites) by its predecessor site in finding globally frequent item sets, it is impossible for any site to predict any predecessor site's private data/information like attributes, local frequent item sets, support values as the frequent item sets are encrypted in the obtained results.

[redacted] the site's private data/information even though it has certain privileges like decryption of

frequent item sets, initiation of the mining process, finding global frequent item sets and their supports, generation of association rules.

The DM receives processed results from sites that contain 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 this information, it is not feasible for DM to guess any individual site's private data/information.

to miner is performed as a bulk data transfer instead of single transfer for every frequent item set. In the proposed model, only one data transfer is needed for sending processed results from every predecessor site to its successor site. So only n number of data transfers are required to acquire all sites processed results in order to identify the global frequent item sets.

As every site has distinct set of attributes with the same set of transactions, the proposed model efficiently finds global frequent item sets by searching all possible combinations of attributes of all sites.

Several experiments are conducted by considering synthetic dataset which consists of boolean transactions possessing 10 attributes for three vertically partitioned databases. The first site contains three attributes, second one contains five attribute and third one two attributes. For comparison purpose, algorithm proposed in [109] is considered. Experiments are conducted with the existing algorithm and proposed algorithm. Both these algorithms adopted a cryptography technique, scalar product.

4. CONCLUSION

The proposed methodology outperforms the existing algorithm since it utilizes the T-ID representation which makes easy computation of scalar product in order to identify the global association rules. The computation time for existing algorithm is more since algebraic computations are to be performed to preserve the privacy. From the above discussion, the proposed model finds the global association rules for vertically partitioned databasesefficiently with minimum number of data transfers, without revealing any sites private data/ information to any site and DM.

REFERENCE

[1] 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.

[2] M. Hussein, A. El-Sisi, and N. Ismail, "Mining Association Rules on Mixed Database with Privacy Preserving", The 4thInternational Computer Engineering Conference Information Society Applications in the Next Decade, ICENCO2008, Giza, Egypt, 2008.

[3] Asha Khatri, SwathiKabra, Shamsher Singh, Durgeshkumar Mishra, "Architecture for Preserving Privacy During Data Mining by Hybridization of Partitioning on Medical Data", 2010, Fourth Asia International Conference on Mathematical/Analytical Modeling and Computer Simulation. pp 93-97,2010.