ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization) Vol.2, Special Issue 1, March 2014

Proceedings of International Conference On Global Innovations In Computing Technology (ICGICT'14)

Organized by

Department of CSE, JayShriram Group of Institutions, Tirupur, Tamilnadu, India on 6th & 7th March 2014

Privacy Preserving Outsourcing for Frequent Itemset Mining

M. Arunadevi[1], R. Anuradha[2]

PG Scholar, Department of Software Engineering, Sri Ramakrishna Engineering College, Coimbatore, India[1]

Assistant Professor (Sr. G), Department of CSE (UG), Sri Ramakrishna Engineering College, Coimbatore, India[2]

Abstract: Cloud computing uses the paradigm of data mining-as-a-service. A company/store lacking in mining expertise can outsource its mining needs to a service provider (server). The item-set of the outsourced database are the private property of the data owner. To protect this corporate privacy, the data owner encrypts the data and sends to the server. Based on the mining queries sent from client side, server conducts data mining and sends encrypted pattern to the client. To get true pattern client decrypts encrypted pattern. In this paper we have studied the problem of outsourcing the frequent item-set within corporate privacy preserving framework. We have proposed an attack model based on the basic assumption, attacker knows items and support of the item, he may know the details of encryption algorithm and some pairs of item and corresponding cipher values. Based on this basic assumption we have improved the security of the system, to reduce the item and item-set based attack, and to reduce the processing time.

Index Terms: Outsourcing, Frequent itemset, and mining pattern

I. INTRODUCTION

Outsourcing aims to provide a service in a corporate privacy preserving framework. Privacy protection is a main issue in data mining. Companies usually do not want to share their own private information with other companies. The idea is that data is published by Client for the benefit of allowing analysts to mine encrypted patterns from the encrypted database. As an example, the transactional database from various companies can be shipped to a third party (server) which provides mining services. The company management do not want to employ an in-house team of data mining experts. Besides, periodically data is sent to the service provider who is in charge of maintaining the encrypted data and conducting mining on it in response to requests from company analysts of the company management. The data owner is a client and the server is referred to as the service provider. One of the main issues with this paradigm is that the server has access to valuable data of the owner and may reveal sensitive information from the data. For example, by looking at the transaction database, the server (or an attacker who gains access to the server) can infer or reveal which products (items) are co-purchased, and in turn, the mined encrypted patterns that describe the company customers' details.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization) Vol.2, Special Issue 1, March 2014

Proceedings of International Conference On Global Innovations In Computing Technology (ICGICT'14)

Organized by

Department of CSE, JayShriram Group of Institutions, Tirupur, Tamilnadu, India on 6th & 7th March 2014

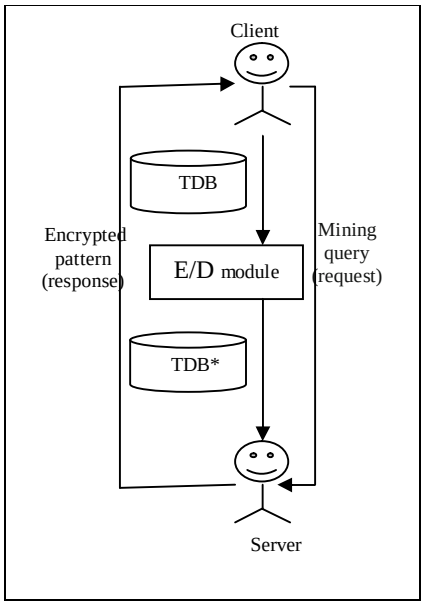

Figure 1: Architecture of mining-as-service paradigm

In this context, both the sale transaction database and the mined encrypted patterns and all the details of the company that can be extracted from the data are the property of the company management and should remain safe from the server and any other attacker. In reality the knowledge mined from the data can be used from the company management in important marketing decisions to improve their services. A client is a data owner or company owner of the company. A company wants their data to be secret but a company does not have sufficient mining expertise for data mining. In such a case, client outsources data for mining purpose. The data will be stored in a server side. Multiple companies have access the server. So, the data will be disclosed. For that purpose, client encrypts its data and stores it in a serer in some other format. Based on the mining queries server conducts mining and sends encrypted pattern to the client. Finally client decrypts the encrypted pattern and gets true support of the original transactions.

II. RELATED WORKS

The research of privacy-preserving data mining (PPDM) is an important issue recently. The main model here was that private details were gathered from a number of companies by a server for the purpose of combining the data and conducting mining. The server was not trusted with protecting the privacy, so data was subjected to a random perturbation as it was gathered. Techniques have been used for perturbing the data so as to protect privacy while ensuring the extracted patterns or other properties were sufficiently close to the patterns extracted from original database. This body of work was established by [2] and has been followed up by several papers since [7]. This idea was not suited for corporate privacy framework, in that some of the properties were revealed. Another issue was secure multiparty mining over distributed database (SMPM).

Mining of the data has been partitioned, horizontally or vertically, and distributed among several parties. The partitioned data were not shared and should protect as private but the results of mining on the combined data were shared among the clients, by means of multiparty secure protocols [4], [3], [5]. They were not considered as third parties. This idea partially implemented the corporate privacy framework, as local databases were kept private, but it was not sufficient for our outsourcing problem, as the resulting patterns were revealed to multiple parties. The particular problem addressed in our paper [1] is outsourcing of pattern mining within a corporate privacy-preserving framework.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization) Vol.2, Special Issue 1, March 2014

Proceedings of International Conference On Global Innovations In Computing Technology (ICGICT'14)

Organized by

Department of CSE, JayShriram Group of Institutions, Tirupur, Tamilnadu, India on 6th & 7th March 2014

A key difference between this problem and the above declared PPDM problem was that, in our setting, original data and the mined pattern results were not shared among multiple parties should protect as private property. In particular, when the server have possessed background knowledge of the transaction database and conducted attacks based on the background knowledge, it should not be able to guess the correct candidate item or set of items corresponding to a given cipher item or set of items with a probability above a given user specified threshold value. The works that were most related to ours were [9] and [8].

Similar to our work, they have assumed that the adversary possessed prior knowledge of the frequency of items or itemsets, details of encryption method and some pairs of cipher items for corresponding original items. The work [9] used a one-to-n item mapping together with non-deterministic addition of cipher items and itemsets to protect the individual items.

A recent paper [6] has proven that the encryption system in [9] can be broken without using specific information. The success of the attacks in [6] mainly focused on the existence of unique, common and fake items, described in [9]; in our scheme, the fake items were not created, and the attacks in [6] were not suitable to our scheme. Tai et al. [8] have assumed the attacker knew exact frequency of single items, more over similar to us. They utilized a similar privacy model as ours, which needs that each original item must have the same frequency as k – 1 other item in the outsourced database. They described that their outsourced database satisfies k-support anonymity.

III.RESEARCH METHODOLOGY

The approach followed in this paper for privacy preserving outsourcing for frequent item set mining was shown in the Fig 2,3,4.

A. Encryption

In this section we introduced the concept of encryption scheme called 1-1 substitution cipher method which transformed a transaction database D into its encrypted version D*. There was not sufficient security by using this encryption methodology.

For example, consider the transaction database in table 1(a) and its associated cipher items in table 1(b). The items were arranged in alphabetical order and each item was substituted by corresponding cipher items. The fake transaction added with encrypted database for improving security. The group partition method was used for constructing fake transaction to be combined with encrypted database.

TDB
Soda Nuts
Soda Milk
Milk Soda
Nuts Milk
Soda Dates
Nuts Soda
Soda Egg
Nuts Cake
Cake

TDB*
e_6 e_5
e_6 e_4
e_4 e_6
e_5 e_4
e_6 e_2
e_5 e_6
e_6 e_3
e_5 e_1
e_1

Table1 (a): TDB Table1 (b): TDB*

B. Fake transaction construction

The fake transaction was constructed based on the noise value by using group partition method which partitioned the data into frequent items and non frequent items based on user specified threshold value.

Step 1: Weighted support construction

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization) Vol.2, Special Issue 1, March 2014

Proceedings of International Conference On Global Innovations In Computing Technology (ICGICT'14)

Organized by

Department of CSE, JayShriram Group of Institutions, Tirupur, Tamilnadu, India on 6th & 7th March 2014

This approach was started with calculation of support of the items. Support count was defined as number of time the items occurred in the original transaction database. i.e. frequency of the item in the transaction database. The weight value was generated randomly.

The weights for each item were automated in random order. The weighted support was calculated by adding the support of the item and the weight of the corresponding item shown in table2 (a) and finally based on the weighted support the items has been arranged in descending order shown in table2 (b).

Step 2: Group partition method

Given the items weighted support table, several strategies were followed to cluster the items into groups of size k. We started from a simple grouping method called group partition. We assumed the item weighted support table was sorted in descending order of weighted support.

Item	Support	Weight	Weighted support
e_1	2	1	3
e_2	1	1	2
e_3	1	2	3
e_4	3	1	4
e_5	4	2	6
e_6	6	1	7

Table2 (a): Weighted support table

item	Weighted support
e_6	7
e_5	6
e_4	4
e_1	3
e_3	3
e_2	2

Table2 (b): Weighted support in descending order

Assume e1, e2. . . . en is the list of cipher items in descending order of weighted support (with respect to D), the groups created by group partition method are {e1, . . . , ek-1}, {ek+1, . . . , en}. The group can be created based on the user specified threshold value.

Item	Weighted support
e_6	7
e_5	6
e_4	4

Item	Weighted support
e_1	3
e_3	3
e_2	2

Table3 (a): Group Partition

Item	Weighted support	Noise value
e_6	7	0
e_5	6	1
e_4	4	3

e_1	3	0
e_3	3	0
e_2	2	1

Table 4 (a): Noise table construction

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization) Vol.2, Special Issue 1, March 2014

Proceedings of International Conference On Global Innovations In Computing Technology (ICGICT'14)

Organized by

Department of CSE, JayShriram Group of Institutions, Tirupur, Tamilnadu, India on 6th & 7th March 2014

For example, consider table3 (a) the user specified threshold value is 4. The weighted support of the items which are more than or equal to 4 are considered as frequent items and it can be grouped into one group. The weighted support of the items which is lesser than 4 are considered as non frequent items and it can be grouped into another group.

Step 3: Noise table construction

The output of group partitioning method can be represented as the noise table. It expands the item weighted support table with an extra column "Noise value" representing, for each cipher item e, the difference among the weighted support of the most frequent cipher item in e's group and the weighted support of e itself, as reported in the item weighted support table. We denote the noise value of a cipher item e as N (e). Continuing the example, the noise table obtained with group partition method is reported in Table4 (a). The noise value represents the tool for generating the fake transactions to be added with encrypted database.

Step 4: Fake transaction construction

Given a noise table mentioning the noise N (e) required for each cipher item e, we create the fake transactions as follows. First, we leave the rows with zero noise, associated with the most frequent items of each group or to other items with the weighted support equal to the maximum weighted support of a group. Second, we arrange the remaining rows in descending order of noise value. The following two fake transactions are generated: 1 instance of the transaction $\{e_4\}$, 1 instance of the transaction $\{e_4, e_3, e_1\}$, and 1 instance of the transaction $\{e_4\}$. Finally, the following 3 fake transactions are generated: 1 instances of the transaction $\{e_4\}$, 1 instance of the transaction $\{e_4, e_5\}$, and 1 instance of the transaction $\{e_4, e_2\}$. So, adding longer fake transactions technically does not form privacy protection. However, for added protection, we can decrease the lengths of the added fake transactions so that they are in line with the transaction lengths in transaction database D.

Step 5: Matrix formation

The observation produces a compact outline for the client of the constructed fake transactions. The purpose of using a compact outline is to decrease the storage overhead at the side of the data owner/client who may not be provided with sufficient computational resources, mining expertise, and storage, which is common in the outsourcing data model framework. In order to implement the outline efficiently, we use a matrix formation.
Shown in table5 (a).

items	e_2	e_4	e_5
e_2	0	1	1
e_4	1	1	1
e_5	0	1	0

Table5 (a): Matrix formation

C. Decryption

When the client has requested the execution of a data mining query to the server, mentioning a user specified threshold value σ, the server responded the encrypted frequent patterns from encrypted database. Clearly, for every itemset S and its associated cipher itemset E, we have that $wsupD(S) \leq wsupD^*(E)$. For each cipher pattern E returned by the server together with $wsupD^*(E)$, the E/D module obtained the corresponding plain pattern S. It required recreating the exact support of S in D. To attain this goal, the E/D module adjusted the weighted support of E by removing the effect of the fake transactions. $supD(S) = wsupD^*(E) - wsupD^*\backslash D(E)$. Note that after the data owner/client outsourced the encrypted database (including the fake transactions), there was not needed to preserve the fake transactions in its own storage. Instead the compact outline was maintained by the client side, which stored all the information required on the fake transactions, for recovery of real supports of item sets. The size of the outline was linear in the number of items and was much smaller than that of the fake transactions.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization) Vol.2, Special Issue 1, March 2014

Proceedings of International Conference On Global Innovations In Computing Technology (ICGICT'14)

Organized by

Department of CSE, JayShriram Group of Institutions, Tirupur, Tamilnadu, India on 6th & 7th March 2014

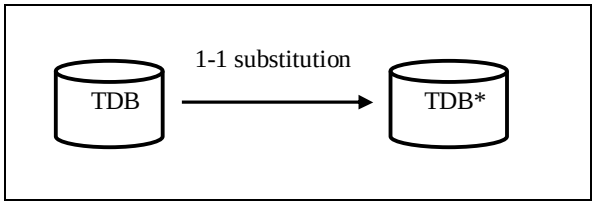

Figure2: Encryption

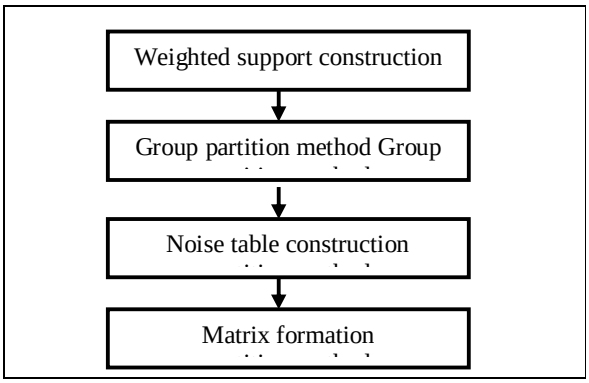

Figure3: Fake transaction construction

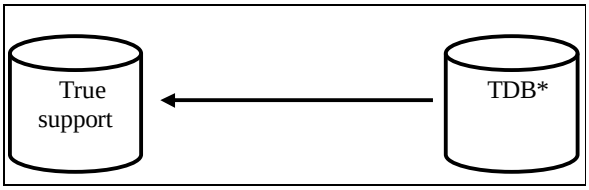

Figure4: Decryption

IV. PERFORMANCE EVALUATION

The random values have been generated by the following method.

$X_{n+1} = (aX_n + b) \bmod m$ (1)

Where a, b – user specified value

x_n -- previously calculated random value

m – Table size

The weighted support was calculated by the weights which are entirely based on random numbers. The group has been partitioned into two groups based on user specified threshold value (α).

The noise value is calculated by the following method.

$N(e) = WS(m(e)) - WS(e)$ (2)

Where N(e) – Noise value of e

WS(m(e)) – Weighted support of most frequent item of e in e's group

WS(e) – Weighted support of e

Decryption has been done by the following

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization) Vol.2, Special Issue 1, March 2014

Proceedings of International Conference On Global Innovations In Computing Technology (ICGICT'14)

Organized by

Department of CSE, JayShriram Group of Institutions, Tirupur, Tamilnadu, India on 6th & 7th March 2014

supD(S) = wsupD*(E) − wsupD*\D(E). (3)
where supD(S) –Real support of D
wsupD*(E) – Weighted support of D*
wsupD*\D(E) – Weighted support of fake transaction.
The comparison between rob frugal and group partition algorithm was shown in the figure.

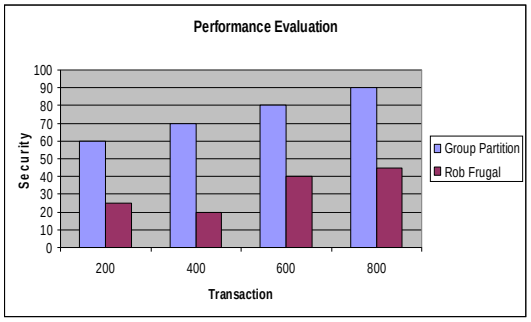

Figure5: security vs transaction.

V. CONCLUSION AND FUTURE WORK

In this paper, we studied the problem of (corporate) privacy preserving Mining framework of frequent patterns (from which association rules were easily computed) on an encrypted outsourced transaction database. We assumed a conservative model where the adversary knew the domain of items and their exact support, details of the encryption method and some pairs of cipher item corresponding to the plain items and has used this knowledge to identify cipher items and cipher itemsets. The proposed encryption scheme, called Group partition method, that is based on 1-1 substitution ciphers for items and adding together fake transactions to make each cipher item share the same frequency as $\geq k - 1$ others. It has been utilized the compact outline of the fake transactions from which the true support of mined patterns from the server were efficiently obtained. We proved that our method was robust against an adversarial attack based on the original items and their exact support. The fake transactions were not addressed in this paper. It could be interesting to reduce the fake transaction.

REFERENCES

[1] Fosca Giannotti, Laks V.S. Lakshmanan, Anna Monreale, Dino Pedreschi, and Hui (Wendy) Wang, "Privacy-preserving Mining of Association Rules from Outsourced Transaction Databases", In IEEE system journal,2012.
[2] Rakesh Agrawal and Ramakrishnan Srikant. Privacy preserving data mining. In SIGMOD, pages 439–450, 2000.
[3] Gilburd B, Schuste A, and Wolff R. k-ttp: A new privacy model for large scale distributed environments. In VLDB, pages 563 – 568, 2005.
[4] Murat Kantarcioglu and Chris Clifton. Privacy preserving distributed mining of association rules on horizontally partitioned data. TKDE, 16(9):1026–1037, 2004.
[5] P. Krishna Prasad and C. Pandu Rangan. Privacy preserving birch algorithm for clustering over arbitrarily partitioned databases. In Advanced Data Mining and Applications, pages 146–157, 2007.
[6] Ian Molloy, Ninghui Li, and Tiancheng Li. On the (in)security and (im)practicality of outsourcing precise association rule mining. In ICDM, pages 872–877, 2009.
[7] Shariq J. Rizvi and Jayant R. Haritsa. Maintaining data privacy in association rule mining. In VLDB, pages 682–693, 2002.
[8] C. Tai, P. S. Yu, and M. Chen. k-support anonymity based on pseudo taxonomy for outsourcing of frequent itemset mining. In KDD, pages 473–482, 2010.
[9] W. K. Wong, David W. Cheung, Edward Hung, Ben Kao, and Nikos Mamoulis. Security in outsourcing of association rule mining. In VLDB, pages 111–122, 2007.
[10] Rajkumar Buyya, Chee Shin Yeo, and Srikumar Venugopal. Marketoriented cloud computing: Vision, hype, and reality for delivering it services as computing utilities. In HPCC, 2008.