International Journal of Innovative Research in Information Security (IJIRIS)
Issue 2, Volume 5 (May 2015)
ISSN: 2349-7017(O)
ISSN: 2349-7009(P)
www.ijiris.com

Privacy-Preserving Mining of Association Rules On Cloud by improving Rob Frugal Algorithm

Mr.Vishal R. Redekar
Department of Computer Engineering
Smt.Kashibai Navale College of Engineering
Pune-411041, India

Dr.Kishor N. Honwadkar
Department of Computer Engineering
Smt.Kashibai Navale College of Engineering
Pune-411041, India

Abstract— Cloud computing uses the ideal model of information mining-as-a service, Utilizing these it seems to be an obvious choice for companies saving on the cost of contributing to secure, manage and keep up an IT infrastructure. An organization/store lacking in mining ability can outsource its mining needs to service provider on a cloud server. However, both the association rules and item-set of the outsourced database are viewed as private property of the organization. The data owner encrypts the data and sends to the server to preserve the corporate privacy. Client sends mining queries to server, and then server conducts data mining and sends encrypted pattern to the client. To get true pattern client decrypts encrypted pattern. In this paper, we consider the issue of outsourcing the association rule mining task within a corporate privacy-preserving system. Thus Privacy Preserving Data Mining is a examination area concerned with the security determined from personally identifiable information when considered for data mining. The Rob Frugal encryption strategy is introduced to overcome the security vulnerabilities of outsourced information, which is focused on one to one substitution ciphers for items and including fake patterns for database. However, it contains a various fake patterns which increase the capacity overhead. To overcome this issue, the proposed procedure includes addition of weighted support in original support of items so as to reduce the number of fake patterns and to enhance the security level for outsourced information with less complexity. The fake transaction table data is converted into matrix format to reduce the storage overhead. Also the guessing attack and man in the middle attack are possible on basic Rob frugal algorithm. To overcome these attack we utilize Elliptic Curve Diffie Hellman key exchange algorithm after Rob frugal encryption scheme in order to provide privacy preserving outsourced mining. In our proposed work we improved the security as item and item-set based attack are not possible on the system; also we reduce the processing time.

Keywords: - Cloud Computing, Association rule mining, Privacy-preserving outsourcing, Elliptic Curve Diffie Hellman, Rob Frugal.

I. INTRODUCTION

Cloud computing is computing in which large groups of remote servers are networked to allow centralized data storage and online access to computer services or resources. With the arrival of cloud computing and its model for IT services based on the web and big data centers, the outsourcing of data and computing services is acquiring a novel relevance, which is certainly required to skyrocket in the near future. In business, outsourcing involves the contracting out of a business process to another party. Outsourcing aims to provide a service in a corporate privacy preserving framework. Privacy protection is the main issue in data mining. Organizations, generally, do not want to share their own private data 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 illustration, the transactional database from different organizations can be transported to an outsider which gives mining services. The organization management would prefer not to utilize an in-house group of data mining specialists. Additionally, periodically data is sent to the server or 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 primary issues with this standard is that the server has entry to valuable information of the owner and may uncover sensitive information from the data. For example, by looking at the transaction database, the server can derive or uncover which products or items are co-purchased and thus, the mined encrypted patterns that describe the organization customers' details.

In this paper, we study the issue of outsourcing the association rule mining task inside a corporate security preserving structure [7]. Therefore, Privacy Preserving Data Mining is an exploration area concerned with the security determined from actually identifiable data when considered for data mining. 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 company or data owner wants their data to be secret but a company does not have sufficient mining expertise for data mining, for this we make the following

contributions. We develop an encryption scheme, called Improved RobFrugal in that the Encrypt/Decrypt module can employ to transform client data before it is shipped to the server. Second, to allow the E/D module to recover the true patterns and their correct support, we propose that it creates and keeps a compact structure, called synopsis. Third, we introduce addition of weighted support in original support of items and matrix formation of fake transaction to reduce the storage overhead. Fourth In order to provide privacy preserving outsourced mining, we utilize Elliptic Curve Diffie Hellman (ECDH) key exchange algorithm after Improve Rob Frugal encryption scheme. With use of ECDH algorithm guessing attack and man in the middle attack are not possible on our proposed system. Fifth, for better performance Enhance FP-Growth algorithm is used instead of Apriori algorithm [13] for association rule generation. At last, we direct test investigation of our pattern utilizing a large real dataset, our results demonstrate that our encryption schema is effective, scalable, and attain to the desired level of security.

Literature Survey is described in the next section. Section III presents the Implementation Details; it includes Encryption Decryption module details, ECDH algorithm and Enhance FP-Growth algorithm. In Section IV, Experimental analysis and Result are discussed. Section V concludes the paper.

II. LITERATURE SURVEY

A. Substitution cipher techniques:

W. K. Wong et al. [1] proposed substitution cipher techniques in the encryption of transactional data for outsourcing association rule mining. After recognizing the non-trivial dangers to the clear one-to-one item mapping substitution cipher, we propose a more secure encryption plan based on a one-to-n item mapping that transforms transactions non deterministically, yet guarantees correct decryption. They develop an effective and efficient encryption algorithm dependant on this method.

B. Data Perturbation:

There are two approaches that can protect sensitive information. Is to an encryption function that transforms the original data to a completely new format [4] [2]. The second could be to apply data perturbation, which modifies the original raw data randomly [3]. The perturbation approach is less attractive since it could only provide approximate results; nevertheless, the employment of encryption allows the same rules that they are recovered.

C. k-support anonymity

The background knowledge such as the supports of frequent item sets can be utilized to obtain privacy information in the outsourcing of frequent item set mining.
In this paper [5], C. Tai, P. Yu proposed k-support anonymity to provide protection against a knowledgeable attacker with exact support information. To achieve k-support anonymity, they introduce a pseudo taxonomy tree and have the third party mine the generalized frequent item sets instead. The construct of the pseudo taxonomy tree facilitates hiding of the original items and limits the fake items introduced in the encrypted database. The experimental results showed that the methods of k-support anonymity achieve very good privacy insurance with moderate storage overhead.

D. Corporate privacy preserving mining

In this paper [6], F. Giannotti, A. Monreale studied the problem of privacy preserving mining of frequent patterns on an encrypted outsourced transaction database. We accept a progressive model where the adversary knows the domain of items and their exact frequency and can utilize this learning to identify cipher items and cipher item sets. We proposed an encryption scheme, called Rob Frugal that is based on 1-1 substitution ciphers for items and adding fake transactions to make each cipher item share the same frequency. It uses the compact synopsis of the fake transactions from which the true support of mined patterns from the server can be efficiently recovered.

E. Association rule mining by Evmievski

Evmievski et al. [8] proposed an approach, for conducting the privacy preserving association rule mining. Kargupta et al. [9] proposed a method based on random matrix spectral filtering to recover original data from the perturbed data. Huang et al. [10] proposed further, the two data reconstruction methods, first PCA-DR and second, MLE-DR.

F. Randomized Response

The first person to propose the Randomized Response (RR) was Warner [14].The RR scheme was initially developed in the statistics community. It used to collect the information from individuals such that, the survey interviewers and the data processors do not know which of the two alternative questions are respondent have answered. In data mining, the method of randomization is a simple technique, can be very easily applied at data collection time. It was a useful technique for hiding individual data in privacy reserving data mining. The randomization method is more efficient [11]. Though, it results in high information loss. The literature on Privacy-Preserving Mining of Association rules can be classified into Pattern mining task, privacy model, and finally Encryption/Decryption scheme.

International Journal of Innovative Research in Information Security (IJIRIS)
Issue 2, Volume 5 (May 2015)
ISSN: 2349-7017(O)
ISSN: 2349-7009(P)
www.ijiris.com

III. IMPLEMENTATION DETAILS

In this paper, we propose improve security of Mining of Association Rules from Outsourced Transaction Databases using improved RobFrugal. We use Elliptic Curve Diffie Hellman (ECDH) after rob frugal encryption scheme. We further use Enhance FP-Growth algorithm, to generate association rules. The details of our improve schemes are presented, it includes following algorithms: Improved RobFrugal algorithm, ECDH algorithm, Enhance FP-Growth Algorithm. The System architecture behind our model is shown in Figure 1. The client/owner encrypts its data using encrypt / decrypt (E/D) module.

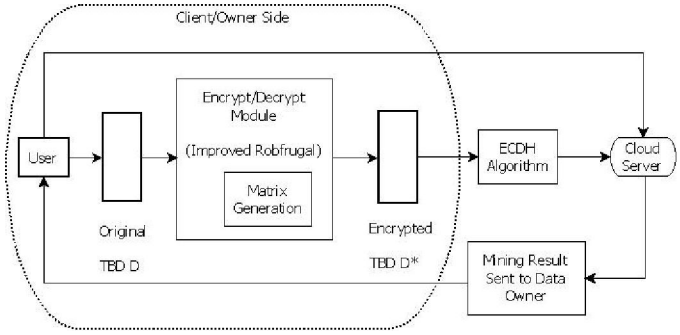

Figure 1.System Architecture

A. *Encryption:*

The section represents the concept of encryption scheme it uses 1-1 substitution cipher method which transformed original transaction database D into its encrypted version D^*. To improve the security fake transaction are added with encrypted database. The weight addition method was used for constructing fake transaction to be combined with encrypted database. Table 1(a) shows original transaction while Table 1(b) shows transaction after one to one substitution (encrypted)

Table1 (a): TDB

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

Table1 (b): TDB*

TDB*
e6 e5
e6 e4
e4 e6
e5 e4
e6 e2
e5 e6
e6 e3
e5 e1
e1

B. *Fake Transaction Construction:*

The fake transaction was constructed based on the adding noise value to the original transaction database.

Step 1: Weighted support construction-

This approach was started with calculation of support of the items. Support count is the number of time the items occurred in the original transaction database. The weight value was generated randomly. The weights for every item were automated in irregular order. The weighted support was figured by adding the support of the item and the weight of the corresponding item shown in Table 2(a) and finally based on the weighted support the items has been arranged in descending order according to frugal grouping shown in Table 2(b).

Step 2: k-Grouping method-

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

ISSN: 2349-7017(O)
ISSN: 2349-7009(P)
www.ijiris.com

Table 2(a): Addition of Weighted Support

Item	Support	Weight	Weighted support
e1	2	1	3
e2	1	1	2
e3	1	2	3
e4	3	1	4
e5	4	2	6
e6	6	1	7

Table 2(b): Descending order of items based on their weighted item support

Item	Weighted Support
e6	7
e4	6
e3	4
e1	3
e3	3
e2	2

Assume e1, e2… ,en is the list of cipher items as shown in table 2[b] in descending order of weighted support (with respect to D), the groups created are as {e1, . . . , ek-1}, {ek+1, . . . , en}. Based on the user specified threshold value the groups can be created.

Table 3: Noise table construction

Item	Weighted Support	Noise value
e6	7	0
e5	6	1
e4	4	3

Item	Weighted support	Noise value
e5	3	0
e3	3	0
e1	2	1

For example, consider Table3 the client specified threshold value is 4. The weighted support of the items which are more than or equivalent to 4 are considered as regular items and it can be assembled into one group. The weighted support of the items which is less than the specified threshold are considered as non frequent items and it can be grouped into alternate group.

Step 3: Noise table construction-

The yield of gathering apportioning technique can be spoken to as the clamor table. It extends the thing weighted help table with an additional segment "Commotion quality" speaking to, for each one figure thing e, the distinction among the weighted backing of the most successive figure thing in e's gathering and the weighted backing of e itself, as reported in the thing weighted help table. We indicate the commotion estimation of a figure thing e as N (e). Proceeding with the sample, the clamor table acquired with gathering allotment technique is accounted for in Table3. The commotion worth speaks to the apparatus for creating the fake exchanges to be included with encoded database.

The output of group partitioning technique can be represented as the noise table. It expands the item weighted support table with an additional column “Noise value” representing, for each one cipher item e, the difference among the weighted support of the most successive cipher item in e’s gathering and the weighted support of e itself, as reported in the item weighted support table. We indicate the noise value of a cipher item e as N (e). Proceeding with the example, the noise table obtained with k-grouping method is shown in Table3. 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 specifying the noise N (e) needed for every cipher item e, we make the fake transactions as follows. First, we leave the rows with zero noise, connected with the most frequent items of every group or to other items with the weighted support equal to the maximum weighted support of a group. Second, we organize the remaining rows in descending order of noise value. The accompanying two fake transactions are produced: 1 instance of the transaction {e4}, 1 instance of the transaction {e4, e3, e1}. Finally, the following 3 fake transactions are produced: 1 instances of the transaction {e4}, 1 instance of the transaction {e4, e5}, and 1 instance of the transaction {e4, e2}. 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.

ISSN: 2349-7017(O)
ISSN: 2349-7009(P)
www.ijiris.com

Step 5: Matrix formation-

The observation creates a compact diagram for the client of the constructed fake transactions. The reason for utilizing 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 capacity, which is common in the outsourcing data model framework. With a specific end goal to execute the outline efficiently, we use a matrix formation demonstrated in Table 4.

Table 4: Matrix format

Item	e2	e4	e5
e2	0	1	0
e4	1	1	1
e5	0	1	0

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) ≤ 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 accurate support of S in D. To accomplish this objective, the E/D module adjusted the weighted support of E by removing the effect of the fake transactions. supD(S) = wsupD*(E)−wsupD*\D(E). Note that after the data owner/client outsourced the encrypted database (including the fake transactions), there was not expected to preserve the fake transactions in its own storage. Instead the compact outline was maintained by the client side, which put away all the data needed on the fake transactions, for recovery of real support of item sets. The size of the outline was straight in the quantity of items and was much smaller than that of the fake transaction.

D. Algorithm

For securing shared secret between two devices A and B

(ECDH - Elliptic curve Diffie-Hellman)[14]:

1. Let dA and dB be the private key of device A and B respectively, Private keys are irregular number less than n, where n be the domain parameter.
2. Let QA = dA*G and QB = dB*G be the public key of device A and B individually, where G is a domain parameter
3. A and B exchanged their public keys
4. The end A computes K = (xK, yK) = dA*QB
5. The end B computes L = (xL, yL) = dB*QA
6. Since K=L, shared secret is chosen as xK

To demonstrate the agreed shared secret K and L at both devices A and B are the same From 2, 4 and 5
K = dA*QB = dA*(dB*G) = (dB*dA)*G = dB*(dA*G) = dB*QA = L Hence K = L, therefore xK = xL Since it is practically impossible to find the private key dA or dB from the public key QA or QB, it is not possible to obtain the shared secret for a third party.

E. Algorithm (Enhance FP-Growth):

FP-Growth methodology is based on divide and conquers strategy for creating the frequent item sets. FP-growth is basically utilized for mining frequent item sets without candidate generation. Significant steps in FP -growth is-

Step1- It firstly compresses the database indicating frequent item set into FP-tree. FP-tree is built utilizing 2 passes over the dataset.

Step2: It partitions the FP-tree into a set of conditional database and mines every database independently, thus extract frequent item sets from FP-tree straightforwardly.

It comprise of one root labeled as invalid , a set of item prefix sub trees as the offspring of the root, and a frequent item header table. Every node in the item prefix sub tree comprises of three fields: item-name, count and node link where the item-name registers which item the node represents; count registers the number of transactions represented by the portion of path reaching this node, node link links to the next node in the FP- tree.

Every item in the header table comprises of two fields---item name and head of node connection, which indicates the first node in the FP-tree carrying the item name.

ISSN: 2349-7017(O)
ISSN: 2349-7009(P)
www.ijiris.com

Input: Built FP-tree
Output: complete set of frequent patterns
Method: Call FP-growth (FP-tree, null).
Procedure FP-growth (Tree, α)
{

1) **If the event that Tree contains a single path P then**
2) **For every combination do generate pattern β U α with support = minimum support of nodes in β.**
3) **Else For every header ai in the header of Tree do {**
4) **Generate pattern β = ai U α with support = ai.support;**
5) **Construct β.s restrictive example base and after that β.s conditional FP-tree Tree β**
6) **If Tree β = null**
7) **Then call FP-growth (Tree β, β)}**

}

F. Mathematical Model

Mathematical Model(I)

Let S be the system such that,
S=fIS,FS,X,Y,IT,TD,,DD,ND,ST,SMEM,CPUCT,ESg
Where,
Let IT = fi1, i2, i3....ing be the set of items.
Let TD = ft1, t2, t3... tmg be the set of transaction database.
Each transaction (TID) in TD has a unique transaction ID
and contains a subset of the items in IT.
TID 1 = fSoda Nutsg
TID 2 = fSoda Milkg
TID 3 = fMilk Sodag
TID 4 = fNuts Milkg
TID 5 = fSoda Datesg
TID 6 = fNuts Sodag
TID 7 = fSoda Eggg
TID 8 = fNuts Cakeg
TID 9 = fCakeg
Where, I = fSoda, Nuts, Milk, Dates, Egg, Cakeg
IS=Be the Initial State which provides list of transactions.
FS=Be the Final State Which gives the Rules.
X=Set Of Inputs (Transactions)
Y=Set of Output (Association Rules)
DD=Deterministic Data, it helps identifying the load-store function or assignment function.
ND=Non Deterministic Data of the system to be solved.
ST= Set Of Transactions.
SMEM= Memory required to process all these operations,
CPUCT= more the number of count double the speed and performance.
ES = Null value.
The support supp(X) of an Itemset X is defined as the proportion of transactions in the data set which contain the Itemset.
For ex, the Itemset = Soda Nuts has a support of 2/9= 0.2
The confidence of a rule is defined conf(X) Y) = supp(X U Y) / supp(X)
Supp(X U Y) =0.2, where X and Y are Soda and Nuts respectively.
Supp(X) =6/9=0.6
Conf(X) Y) =supp(X U Y) / supp(X) =0.2/0.6 = 0.33= 33%
Mathematical Model(II)

ISSN: 2349-7017(O)
ISSN: 2349-7009(P)
www.ijiris.com

IV. RESULT AND DISCUSSION

In this section, we compare the performance between Enhance FP-Growth, and Apriori Algorithm. Graphs show the execution time of implementations over the various instances.Figure 2 shows the comparison graph between Apriori algorithm and Enhance FP-Growth algorithm based on Table 5.

Table 5. Time required to rules for different instances in the data set.

Number of Instances	Apriori (time in ms)	Enhance FP growth (time in ms)
1000	29	3
2000	36	5
3000	48	7

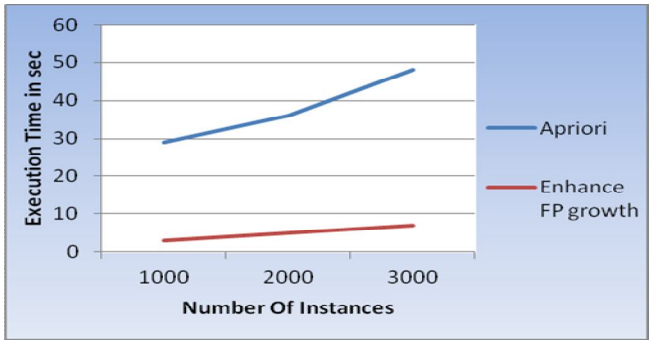

Figure 2. Comparison between Enhance FP-Growth and Apriori

The comparison between improved robs frugal and Rob frugal algorithm was shown in the figure 3.

Table 6.Average Accuracy of a Query Result

Number of transaction	Improved RobFrugal	Rob frugal
300	60	30
600	70	34
900	80	46
1200	90	49

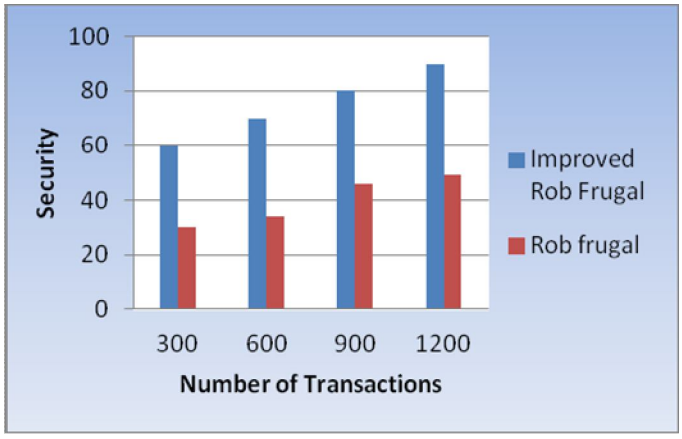

Figure 3. Comparison between Enhance FP-Growth and Apriori

ISSN: 2349-7017(O)
ISSN: 2349-7009(P)
www.ijiris.com

V. CONCLUSION AND FUTURE WORK

We studied the problem of privacy-preserving mining of frequent patterns from which association rules can easily be computed on an encrypted outsourced Transitional database. We assumed that a conservative model where the adversary knows the domain of items and their exact frequency and can use this knowledge to identify cipher items and cipher item sets. We proposed an encryption plan; called improved Rob Frugal That is based on adding weighted support in original item support transactions to reduce the fake transaction table data also the matrix is generated to reduce the storage overhead. We also proposed Elliptic Curve Diffie Hellman key exchange algorithm after Rob frugal encryption scheme to overcome the guessing attack and man in the middle attack. Unlike previous works, we formally proved that our method is robust against an adversarial attack based on the original items and their exact support.

In future the Privacy-preserving tools for individuals and incorporating privacy protection in engineering process can be generated.

1) Privacy-preserving tools for individuals: The privacy preserving techniques in research is proposed only for information holders; however individual record owners should additionally have the rights and obligations to ensure their own particular private information.

2) Incorporating security assurance in engineering process: The privacy issue should be considered as a essential necessity in the engineering process of creating new technology. This includes formal detail of protection prerequisites and formal confirmation devices to demonstrate the rightness of a privacy-preserving framework.

REFERENCES

[1] W. K. Wong, D. W. Cheung, E. Hung, B. Kao, and N. Mamoulis, “Security in outsourcing of association rule mining,” in Proc. Int. Conf. Very Large Data Bases, 2007, pp. 111-122.
[2] G. I. Davida, D. L. Wells, and J. B. Kam. “ A database encryption system with sub keys.” ACM TODS, 6(2):312-328, 1981.
[3] J. He and M. Wang. Cryptography and relational database management systems. In IDEAS, 2001.
[4] B. Iyer, S. Mehrotra, E. Mykletun, G. Tsudik, and Y. Wu. A framework for efficient storage security in RDBMS. In EDBT, 2004.
[5] C. Tai, P. S. Yu, and M. Chen, “K-support anonymity based on pseudo taxonomy for outsourcing of frequent item set mining,” in Proc. Int. Knowledge Discovery Data Mining, 2010, pp. 473-482.
[6] F. Giannotti, L. V. Lakshmanan, A. Monreale, D. Pedreschi, and H.Wang, “Privacy preserving data mining from outsourced databases,” in Proc. SPCC2010 Conjunction with CPDP, 2010, pp. 411-426.
[7] M. Kantarcioglu and C. Clifton, “Privacy-preserving distributed mining of association rules on horizontally partitioned data,” IEEE Trans. Knowledge Data Eng., vol. 16, no. 9, pp. 1026-1037, Sep. 2004.
[8] S. J. Rizvi and J. R. Haritsa, “Maintaining data privacy in association rule mining”, in Proc. Int. Conf. Very Large Data Bases, 2002.
[9] A. Evfimievski, R. Srikant, R. Agrawal, J. Gehrke, “Privacy Preserving Mining of Association Rules”, Information System, 2004.
[10] H. Kargupta, S. Datta, Q. Wang, K. Sivakumar, “On the Privacy Preserving Properties of Random Data Perturbation Techniques”, In Proceedings of the 3rd International Conference on Data Mining, 2003.
[11] Z. Huang, W. Du, B. Chen, “Deriving Private Information from Randomized Data”, In Proceedings of the ACM SIGMOD Conference on Management of Data, 2005.
[12] S. L. Warner, “Randomized Response: A Survey Technique for Eliminating 999999999999Evasive Answer Bias”, J. Am. Stat. Assoc., 1965.
[13] A. Evfimievski, R. Srikant, R. Agrawal, J. Gehrk, “Privacy Preserving Mining of Association Rules”, In Proceedings the 8th ACM SIGKDD International Conference on Knowledge Discovery in Databases and Data Mining, 2002.
[14] Ram Ratan Ahirwal, Manoj Ahke Samrat Ashok, “Elliptic Curve Diffie- Hellman Key Exchange Algorithm for Securing Hypertext Information on Wide Area Network” et al, / (IJCSIT) International Journal of Computer Science and Information Technologies, Vol. 4 (2) , 2013, 363 368

