An Approach to Mining Association Rules in Horizontally Distributed Databases with Anonymous ID Assignment

Ms. Manali Rajeev Raut
Mtech CSE (IV sem)
Dept. of Computer Science and Engineering
G.H. Raisoni Institute of Engineering and Technology for Women, Nagpur
RTMNU, Nagpur
manaliraut2@gmail.com

Ms. Hemlata Dakhore
Assistant Professor
Dept. of Computer Science and Engineering
G.H. Raisoni Institute of Engineering and Technology for Women, Nagpur
RTMNU, Nagpur
hemlata.dakhore@raisoni.net

***Abstract*—Data Mining is the technique of automated extraction of interesting data patterns used to represent knowledge, from the large data sets but sometimes these datasets are divided among various parties. Association rule mining is a popular mining technique that identifies interesting correlations between database attributes. In this paper, proposed a protocol Privacy Preserving Fast Distributed Mining (PPFDM) for association rules mining in horizontally distributed databases which is based on the Fast Distributed Mining (FDM) algorithm. FDM is an unsecured distributed version of the Apriori algorithm devoted to generate a small number of candidate sets and considerably cut down the number of messages to be passed at mining association rules. PPFDM adopts two major ideas: one that computes the union of private subsets that each of the interacting player holds and another that evaluate the inclusion of an element held by one player in a subset held by another. An implementation of a PPDM algorithm is developed in Java framework and performance results are presented for synthetic data generation and association rules as well as indexing is provided to the user. It is simpler and significantly more efficient in the matter of communication rounds, communication cost and computational cost.**

Index Terms— association rules , Privacy preserving data mining, distributed database, frequent item sets, anonymous ID assignment.

I. INTRODUCTION

Data mining is inclined to extract important knowledge from large datasets. It is a powerful new technology with extreme potential to assist organizations to concentrate on the most important information in their data warehouses. Now a day, many organizations manage huge amount of data which is mined to gain valuable knowledge by using several available data mining techniques. Data mining is useful within an organization as well as can provide more benefits with the combined data of multiple organizations. But sharing data within multiple organizations creates probable privacy problems. The motivation to keep one's own delicate data secret is an important challenge of preserving privacy while mining.

The firmly related field of Privacy Preserving Data Mining (PPDM) appends the dimension of privacy to the problem, trying to asset the ways that organizations can collaborate to mine their databases collectively, while at the same time preserving the privacy of their records. Association rule mining is one of the data mining techniques applied in distributed databases, discloses some interesting relationships between locally large as well as globally large item sets[2]. The Distributed databases are a sole logical databases that is spread across more than one node or locations that are all associated via some communication link. Horizontal Partitioning is defined as 'the different sites may have different sets of records consisting the same attributes'. The idea is to use association rules for prediction purposes: if bread, butter and milk often appear in the same transactions, then the presence of butter and milk in a shopping cart commend that the customer may also buy bread. More generally, knowing which items a shopping cart may contains, shop-keeper can assume other items that the customer is likely to add before proceeding to the checkout counter.

The paper defines a problem of secure multi-party computation. In the corresponding problems, there are N players that clasp private inputs, x=(x1 . . . xN), and they desire to calculate y =h(x1, . . . , xN) for some mutual function h. If a trusted third party is available, the players could hand over to the their party, their particular inputs and third party would execute the function evaluation and deliver them back the resulting output. But when there is no appearance a trusted third party, it is essential to create a protocol that the players can evaluate by their own for the purpose to gain the required output. Then that protocol is treated perfectly immune if there is no player able to study from his perspective of the protocol more than what he

978-1-4799-8553-1/15/$31.00 © 2015 IEEE

would have studied in the setting where the computation is done by a trusted third party.

This paper plotted the above problem of association rules mining in horizontally distributed databases and contemplated synthetic database generation, after that generate association rules. The purpose is to find all association rules with support s and confidence c to minimize the information disclosed about the private databases held by those players [1]. This paper also designed an algorithm, PPFDM, privacy preserving fast distributed mining algorithm for horizontally distributed data sets and assets interesting association or correlation relationships among a large set of data items and to integrate cryptographic techniques to reduce the information which is going to shared with residual, while appending lmited overhead to the mining task [1].

II. LITERATURE REVIEW

Tamir Tassa in the paper [1] proposed the problem of secure mining of association rules in horizontally distributed databases. The main thought of using the protocol Fast Distributed Mining algorithm (FDM) is that after finding the locally s-frequent item sets the player should check each item to find out globally s-frequent item set. The main aim of the paper is to offer better privacy and reduced communicational and computational cost while the solution is still leaks some excess information [1].The paper [3] proposed two methods, Elliptic Curve based Digital Signature Algorithm (ECDSA) and Elliptic Curve Integrated Encryption Scheme (ECIES). These algorithms used to mine the association rules with the minimum iterations as well as consume less time[3].

In the paper [4], the combination of Apriori Algorithm and Extended Distributed Rk Secure Sum protocol is proposed where firstly apriori algorithm mine frequent items from all the individual parties then global result is obtained by Extended Distributed Rk- secure sum protocol. Basically, the RK-secure sum protocol is secure multi party computation protocol which is used to hold global items without affecting privacy of the individual parties. All the parties are arranged in the bus networks, where first one is protocol initiator called as p1.

Like paper [4], the paper [5] also uses the combination of Apriori Algorithm and playfair cipher technique. This paper described the two parts of association rule; Antecedent and consequent. The item found in the database is called as Antecedent and the item found in the combination is consequent. Here, firstly the apriori algorithm for generates association rules then the playfair cipher encrypts some pair of letters. This technique requires 5 by 5 table which contains keyword.

The paper [6] proposed FP tree which is used to generate global frequent item sets with the help of association rules. FP tree is defined as the compact data structure because it finds the global frequent item set without generating candidate item sets. Data Encryption Standard (DES) is used here to provide privacy to the resultant item set. Data Encryption Standard (DES) is also called as Double Encryption because it provides higher security to database by the help of two keys which double encryption and double decryption. This paper shows the zero percent data leakage with minimum time complexity.The paper [7] deals with the two important problems of association rule mining i.e. hiding of data and hiding of knowledge. Data hiding is described as the method of removing confidential information from the database before its disclosure.

Apriori algorithm is devoted to generate candidate itemsets. It scan the all database first for pruning and then gives result in the form of frequent itemset. But apriori algorithm is not as efficient as over large database hence the paper [8] improved the apriori and put forward a new method called Sampling. It is used to sample the data from the large database to find frequent itemsets. The paper[8] also described about the SamplingHT algorithm, is the combination of hash table and sampling algorithm. In the SamplingHT algorithm, the sample size of the data and negative border is calculated. After that hash table generate the frequent itemset. The negative border is used to reduce the running time of the algorithm.

The paper [9] is the survey paper which introduce the 5 data mining algorithm i.e. MSApriori, Apriori Algorithm, Aprori with Systematic rules, MCISI Algorithm and HMT. Apriori algorithm is inclined to gain candidate itemsets. It scan all the database first for pruning and then delivers resulting frequent itemset and discard those candidate item set value which is lower than defined support. Minimum support apriori (MSApriori) used to provide the distinct minimum item support values for different items in the database. In the MCISI algorithm, search many imperfectly sporadic rule and also sporadic item sets. Systematic rules are also described in [9] paper where user is not allow to specify minimum support value to gain frequent item sets and timing algorithm is also devoted to save time with scanning of the whole transactional database. The Hash Mapping Table (HMT) is availed to abbreviate the given data sets.

Now a day, privacy preserving for the data and the owner is becoming a problem in case of distributed server sharing. The solutions are exists but for central server model which is computationally expensive and because of low data security and

huge bandwidth tradeoff it is not useful for distributed server model. Hence the paper [11] focuses on distributed model assigning IDs to nodes (user) which are anonymous. Each node chooses random values with the help of Anonymous ID assignment (AIDA) algorithm. These IDs can used for sharing communication bandwidth as it uses network setup where number of clients can register and shares data and also for data storage. The advantage of this algorithm is at the transaction, no id being visible to any group member or person. AIDA is not a cryptographic algorithm hence it saves memory space. This paper shows that the privacy preserving with the help of anonymous ID assignment is successful. Like paper [12], the paper [12] addresses an algorithm to share the simple integer data, it allows the secure sum to be collected with the guarantees of anonymity. This paper addresses the complexities of the secure multiparty computation. The Anonymous IDs are used in sensor networks to secure the individual nodes.

III. PROPOSED METHODOLOGY

The proposed methodology section describes the rationale for the application of specific techniques used to identify, select, and analyze information applied to understanding the problem. This paper proposed a novel algorithm for optimization of association rule mining, Privacy Preserving Fast Distributed Mining (PPFDM). It takes partial databases as input and gives result as list of association rules which consist of unified databases with support s and confidence c[1]. The information that would like to protect in this paper is not only personal transaction in the different-different databases, yet also further global or public data such as what association rules are supported locally in each of those databases[1][2].

In particular, PPFDM protocol independent on commutative encryption and oblivious transfer of the data. The proposed protocol improves upon that in [2] in terms of simplicity and efficiency as well as privacy and also resolves the problem of leakage of information. The Privacy preserving fast distributed mining (PPFDM) algorithm is a combination of Fast Distributed mining algorithm (FDM)[1] and anonymous ID assignment (AIDA)[10]. FDM is an unsecured distributed version of the Apriori algorithm and AIDA is used for security of the databases. The Privacy preserving fast distributed mining (PPFDM) protocol involves following steps.

4.1 Synthetic database generation

Data generation has always been an important phase in Data Mining. There are so many techniques like random sequences, normal distribution, etc to generate data of certain distribution. Despite the keen interest, the research is still going on for the generation of synthetic data for numerous applications. In this method, the input data is generated to test the correctness of the given data mining algorithm. For instance: The synthetic data is generated here to evaluate the performance of the PPFDM algorithms over a large range of dataset. In the real world, it considered that the people tend to buy sets of items together and each set has potentially large itemset, a transaction in the database contains a transaction ID and an itemset. Acquisition of itemsets practiced for transaction databases and bunch of association rules can be characterized as binary matrices with columns corresponding to the items and rows corresponding to the itemsets. An example of a binary matrix containing itemsets for the database shown in the following table.

TABLE I: Example of a collection of itemsets.

Itemsets	Items			
	i1	i2	i3	i4
X1	1	1	0	0
X2	0	1	0	1
X3	1	1	1	0
X4	0	0	1	0

The generation of synthetic data is an implicate process of data anonymization means synthetic data is a subdivision of anonymized data. Researchers or software developers availed to evaluate against a safe data set without disturbing or even approaching the original data, to protect them for privacy and security and also let them create enormous data sets. For different ambitions, it's devotion for synthetic data sets to display more or less selected properties of the original data sets. To generate a dataset, the synthetic data generation takes the parameters shown in table below.

TABLE II: Parameters for Generating the Synthetic Database

Parameters	Interpretations
N	Number of transactions in the whole database
L	Number of items
At	Transaction average size
Af	Average size of maximal potentially large itemsets
Nf	Number of maximal potentially large itemsets
CS	Clustering size
PS	Pool size
Cor	Correlation level
MF	Multiplying factor

4.2 Apriori Algorithm

Apriori is an algorithm, implemented by Agrawal and Srikant in1994, determine the frequent individual items in the database and enhance them to the larger item sets. The term of algorithm, Apriori is planted on the fact that the this algorithm causes a prior knowledge of frequent itemset properties. It finds candidate item sets having length L from item sets of length L-1. Then it prunes the candidates to check the infrequent pattern and assure that all the subsets of the candidate sets are earlier accepted to be frequent itemsets. After that, scanning is done to evaluate frequent item sets among the candidates. The frequent item sets determined by Apriori can be purposed to determine association rules, this application is called as 'market basket analysis'. The output of Apriori is a bunch of rules that express how often items are accommodate in sets of data.

4.3 Association Rules

Association rule mining is one of the data mining technique inclined to asset the association relationship amongst the large set of data with the help of minimum support and minimum confidence. The problem is usually dissolved into two sub problems. One is to gain those itemsets whose occurrences beat a already defined threshold in the database; those item sets are known as frequent or large itemsets with the compulsion of minimal confidence. For instance, in supermarket, The set of items is I = {eggs, bread, butter, milk}. Rule of Supermarket could be {eggs, bread} → {butter}, indicates that if user buy milk and bread then he may buy butter too.

Table III: An example supermarket database.

Transaction ID	Items
1	eggs, bread
2	bread, butter
3	milk
4	juice, bread, butter
5	bread, butter

Let I={item1, item 2,…. item m } be a group of items. Let D is the task relevant data and a group of dataset transaction where individual transaction T is a group of items such that $T \subseteq I$. Each transaction is correlated with an identifier, called TID. Let A be the group of items. A transaction T is encompasses A if and only if $A \subseteq T$. An association rule is an assumption of the form $A \Rightarrow B$, where $A \subset I, B \subset I$ and $A \cap B = \emptyset$. The rule $A \Rightarrow B$ clasps the transaction set D with the aid of support s, where s is assumed as the percentage of transaction in D that consist of $A \cup B$. This is assumed to be the probability, $P(A \cup B)$. The rule $A \Rightarrow B$ has confidence c in the transaction D, where c is known as the percent of transaction in D consisting A that also consist B. This is assumed to be the conditional probability, p (B|A). Hence,

$$\text{Support } (A \Rightarrow B) = P(A \cup B)$$
$$\text{Confidence } (A \Rightarrow B) = p(B|A)$$

4.4 Privacy Preserving Data Mining

Many researchers designed many techniques for privacy preserving association rule mining for databases. This paper also described the Anonymous ID assignment algorithm (AIDA) to preserve privacy for association rules in the area of distributed database. It is assign ID numbers to the nodes ranging from 1 to N so that the more complex data can be shared. In the secure multiparty computation, AIDA allows multiple players on a network to execute data globally; hence, the data hold by each player remains unknown to other players. This paper considers that the players (data owners) are semi-honest. This technique is called as anonymous because IDs are unknown to the other members of the group as well as required evaluations are distributed without using a trusted central authority. For instance, assume there is a group of hospitals with individual datasets and desire to evaluate and share only the some of a data item, like the number of hospital having infection of flu, without revealing the value of this data item to others. Thus, the data items have the IDs and desire to evaluate and share the value. The dynamic unique IDs are useful for storage and sharing of data. In AIDA, random integers between 1 and S are chosen by each node such that $S \geq N$ [10].

Algorithm: Given nodes, n1,n2....nM uses distributed computation to search an anonymous indexing permutation.

$$s: \{1,...,.....,M\} \rightarrow \{1,...,.....,M\}.$$

1) Firm the number of assigned nodes, X=0.
2)Each unassigned node ni chooses a random number ri in the range 1 to S. A node assigned in a previous round chooses ri=0 .
3) The arbitrary numbers are divided anonymously. Denote the shared values by A1,qA2,..... qAM .
4) Let qA1....qAk denote a improved list of shared values with duplicated and zero values fully removed where k is the number of unique arbitrary values. The node ni which tied the unique random numbers then determine their index si from the position of their random number in the improved list as it would appear after being sorted:
$si=X+ \text{Card } \{qj : qj \leq ri\}$
5) Update the number of nodes assigned: X=X+k.
6) If X<M then return to step (2).
Example: Suppose that four user participate to seek for an AIDA. Random numbers are 6, 10, 6 and 2. The resultant indexing is then 3,2,4,1 as shown in fig 5.

4.5 The Fast Distributed Mining Algorithm

This paper is planted on the Privacy Preserving Fast Distributed Mining algorithm (PPFDM) which is a

combo of privacy preserving and Fast Distributed mining algorithm which is an sloppy dispersed version of the Apriori algorithm. Its main belief behind is that any s-frequent item set have to be also locally s-frequent in at least one of the datasites. Hence, in order to search all globally s-frequent item sets, individual player shows his locally s-frequent item sets and then the players evaluate each of them to check if they are s-frequent also globally. The FDM algorithm continues as follows:

1. Candidate Sets Generation: Individual player p_m evaluates the group of all (k-1) item sets, L_{K-1} that are locally frequent and also globally frequent. The idea behind the candidate set generation is that if an itemset X has minimum support, so achieve all subsets of X. Hence the player then employees group the Apriori algorithm on L_{K-1} in order to develop the group of candidate k-item sets.

2. Local Pruning: The pruning step excludes the expansion of (K-1) itemsets which are not gain to be frequent. Here, player p_m evaluates suppm(X). He then returns only those item sets that are locally s-frequent. The paper describe this collection of item set by $C_s^{k,m}$.

3. Computing local supports: All players evaluates the local supports of all item sets in $C_s^{k,m}$.

4. Broadcast mining results: Particular player discloses the local supports that he gain earlier. From that, anyone can evaluate the global support of each item set in $C_s^{k,m}$. At last, F_s^k is the subset of $C_s^{k,m}$ that contains all globally s-frequent k-item sets.

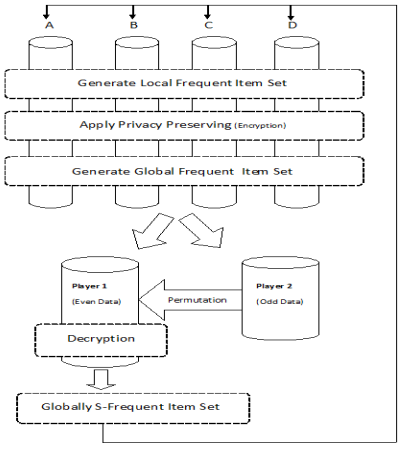

Fig 1: Architecture Of PPFDM (Privacy Preserving Fast Distributed Data Mining)

The paper designed the architecture of privacy preserving Fast distributed data mining (PPFDM) with help of Fast distributed Mining algorithm and Anonymous ID assignment. The evaluation time of this protocol is expected to reduce the cost of secure computation of the individually frequent itemsets. Thus, the security offered by this protocol is customized by increased implementation cost.

IV.EXPERIMENTAL SETUP

This paper present the results obtained by the implementation of the algorithm. Firstly the synthetic data is generated by the same techniques that were implemented in [1]. After that the association rule is derived so that the new set of rules is generated. This algorithm is implemented in Java Language because it contains the Java Data Mining (JDM) API to develop the applications in data mining and tools.

The snapshots of Synthetic data, association rules and data sharing are as follows which are acquired during implementation.

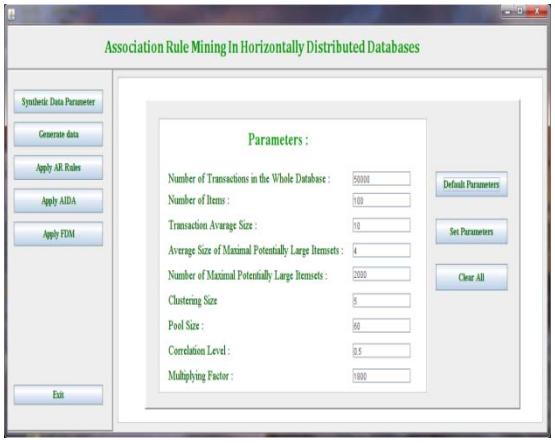

Fig 2:The screenshot of Synthetic Data Generation default parameters

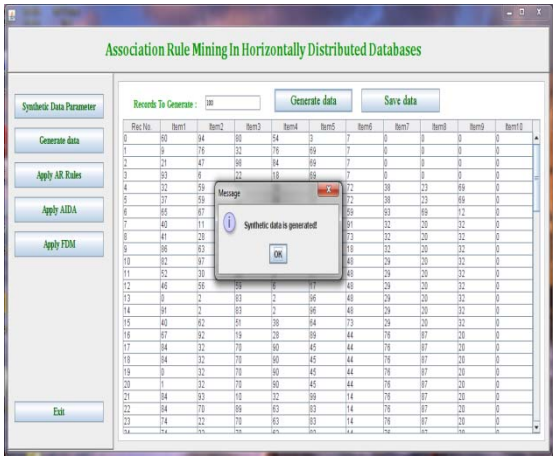

Fig 3 : The screenshot of Synthetic Data is generated. The message dialog box acknowledges that data is generated.

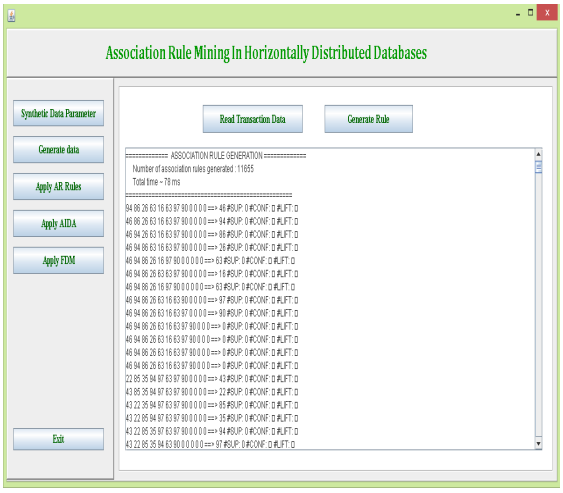

Fig 4: The screenshot of association rules generated.

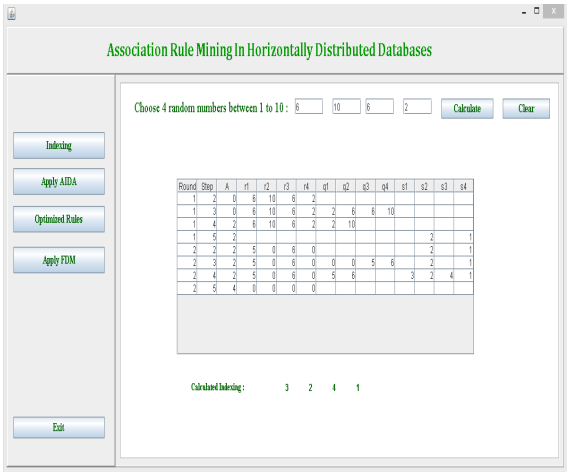

Fig 5: The indexing is calculated by Anonymous ID assignemnt.

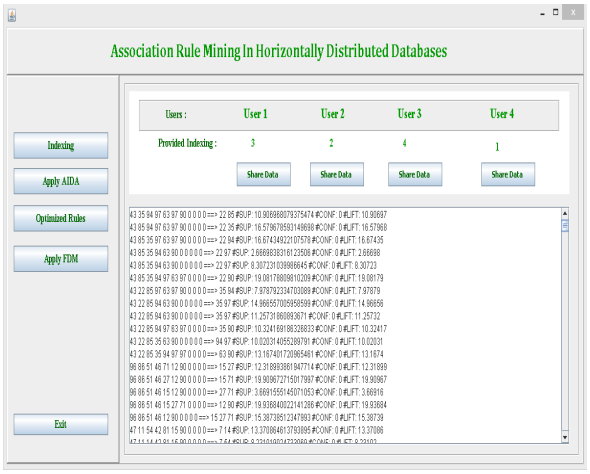

Fig 6: Data is shared according to the indexing provided

V. CONCLUSION

The paper proposed a novel protocol for optimization of mining of association rules for horizontally distributed databases. The PPFDM algorithms have the potential to advances the production of vital knowledge for individuals as well as organizations. The paper shown that the distributed association rule mining can be implemented efficiently with the reasonable security assumption. The current protocol improves in terms of privacy and security. As well, the objective of the paper is achieved by successfully generating the synthetic data and association rules and indexing is provided. This paper studied and suggest research problem is that the implementation of technique shown here to the problem of association rule mining in the vertical setting.

VI. REFERENCES

[1] Tamir Tassa,"Secure mining of association rule in horizontally distributed databases" ,IEEE trans. Konwledege and Data Engg. ,Vol. 26, no.2, April 2014

[2] M. Kantarcioglu and C. Clifton, "Privacy-Preserving Distributed Mining of Association Rules on Horizontally Partitioned Data," IEEE Trans. Knowledge and Data Eng., vol. 16, no. 9, pp. 1026-1037, Sept. 2004.

[3] Krishna Pratap Rao, Adesh chaudhary, Prashant johri "Elliptic Curve Cryptography Based Algorithm for Privacy Preserving in Data Mining", International Journal for research in Applied Science and Engineering Technology (IJRASET) ,Vol. 2 Issue V, May 2014

[4] Meera Treesa Mathews, Manju E.V," Extended Distributed RK- Secure Sum Protocol in Apriori Algorithm for Privacy Preserving", International Journal of Engineering and Advanced Technology (IJEAT), Volume-3, Issue-4, April 2014

[5] P. Jagannadha Varma, Amruthaseshadri,.M. Priyanka, M.Ajay Kumar, B.L.Bharadwaj Varma, " Association Rule Mining with Security Based on Playfair Cipher Technique" (IJCSIT) International Journal of Computer Science and Information Technologies, Vol. 5 (1) , 2014

[6] Jyotirmayee Rautaray, Raghvendra Kumar, "Privacy Preserving In Distributed Database Using Data Encryption Standard (DES) ", International Journal of Innovative Research in Science, Engineering and Technology Vol. 2, Issue 3, March 2013

[7] Prof. Geetika. Narang, Anjum Shaikh, Arti Sonawane, Kanchan Shegar, Madhuri Andhale," Preservation Of Privacy In Mining Using Association Rule Technique", International Journal of Scientific & Technology Research, Volume 2, Issue 3, March 2013

[8] Zhi Liu,Tianhong Sunand Guoming Sang," An Algorithm of Association Rules Mining in Large Databases Based on Sampling ", International Journal of Database Theory and Application Vol.6, No.6 , 2013

[9] Priyanka Asthana, Anju Singh , Diwakar Singh," A Survey on Association Rule Mining Using Apriori Based Algorithm and Hash Based Methods ", International Journal of Advanced Research in Computer Science and Software Engineering. Volume 3, Issue 7, July 2013

[10] Larry A. Dunning, Member, IEEE, and Ray Kresman,"Privacy Preserving Data Sharing With Anonymous ID Assignment", IEEE Transaction On Information Forensics and security, VOL. 8, NO. 2, FEBRUARY 2013

[11] Ms.R.Kalaivani, Ms.R.Kiruthika,"Automated Anonymous ID Assignment For Maintaining Data privacy", Proceedings of 2nd International Conference on Science,Engineering and Management,Srinivasan Engineering college,TamilNadu,India,March 28-29,2014

[12] Shiny. I.S , S. Gayathri,"Secure Multiparty Computation and Privacy Preserving Data Sharing with Anonymous ID Assignment", International Journal of Engineering Research and Applications (IJERA) ISSN: 2248-9622 International Conference on Humming Bird ,01st March 2014