European Journal of Academic Essays 1(2): 35-39, 2014
ISSN: 2183-1904
www.euroessays.org

Privacy Preservation in Distributed Database

Pattnaik Dr. Prasant Kumar[1], Kumar Raghvendra[2], Sharma Dr. Yogesh[3]

[1] School of Computer Engineering, KIIT University,Bhubaneswar, Odisha, India
[2] Ph.D Scholor, Jodhpur National University, Jodhpur, Rajsthan, India
[3]Faculty of Engineering and Technology, Jodhpur National University, Jodhpur, Rajsthan, India
E-mail: [1]*patnaikprasant@gmail.com*, [2]*raghvendraagrawal7@gmail.com*, [3]*dryogeshsharma121@rediffmail.com*

Abstract: Data mining is a new era of technology of identifying patterns and trends from large quantities of large business data sets. Mining association rule is important data mining problem for finding the association between the different attributes in the given business data sets .To solve mining association rule problem have many different types of technique are available for example Apriori, FP tree. The privacy concept arises when the data is distributed in the distributed database environment, concept of security and privacy arises here so that no any unauthorized person not able to decrypt the data and see the result of the middle or unauthorized party. For providing the security to the distributed database we used hash based secure sum cryptography technique without trusted party and another concept is with trusted party when there are one of the middle party that will control all the details of the all the party presents in the distributed database environment because when the Party to find the global result may be that result is frequent or infrequent. In this paper we mainly compare the result privacy preserving technique with and without trusted party for horizontal partitioned data. And provide high security to data Parties with percentage of data leakage is zero percent by using the Apriori algorithm.

Keywords: Data mining, Distributed Database, Privacy Preserving Association Rule Mining, Cryptography Technique, Horizontal Partitioned Database, Secure Sum.

1. Introduction

Data mining techniques [1] [2] [3] are available to retrieve useful information from large database. Prediction and description are the two fundamental goals of data mining. To full fill these goals many data mining techniques exists such as association rules, classification, clustering and so on. Among these, association rule has wide applications to discover interesting relationship among attributes in large databases. Association rule mining is used to find the rules which satisfy the user specified minimum support and minimum confidence. In the process of finding association rules, the set of frequent item sets are computed as the first step and then association rules are generated based on these frequent item sets. Two types of database environments exist namely centralized and distributed [4] [5]. Many associations struggle with their management information systems and their membership databases. One of their primary struggles is the lack of centralized information. Too often, assns and non-profits keep separate databases for membership, events, sales, and other processes. When at all feasible, these databases should be combined into a single, centralized database. A centralized system means that for member Joe Smith, there is only one place a user has to go to find his name, primary address, and activities within the association. There are several benefits to moving your data to a centralized system. Benefit in centralized database .data integrity, Valuable broad marketing information or history, Ease of training, Support. Distributed database is defined as collection of logically distributed database which are connected with each other through a network. A distributed database management system is used for managing distributed database. Each side has its own database and operating system. Charge in view of the motivation to have as a feature of privacy in data mining techniques to save from harm the confidential data of the user, there evolved an innovative stream in data mining period that is privacy preserving in data mining. There exists a key difference among regular data mining algorithms Under a variety of data mining techniques similar to classification, association, clustering and privacy Preserving data mining algorithms that is the recognized algorithms deals with how to evaluate the Stored raw data and how to take out useful knowledge discovery patterns from the database Whereas in the afterward, it essentially deals with the sensitive information of the user records where privacy factor is the main concern and it is measured to be vital issue. The main aim in many scattered methods for privacy preserving data mining is to agree to useful aggregate computations on the complete data set through preserving the privacy of the individual Parties data or information. Each Party owner is interested to work together in obtaining combined results, but not fully trust other Parties in conditions of the distribution of their own data sets. Several data mining system should satisfy the important property that is privacy preserving of data or information. Particularly in distributed data mining, privacy preserving is individual crucial feature. Secure multi party computation is a useful approach to save the privacy in distributed data mining. Privacy preserving data mining

Corresponding Author: Kumar Raghvendra, Email: raghvendraagrawal7@gmail.com

utilizes a mining algorithm to obtain mutually beneficial global data mining objectives without helpful private data. Therefore, in many data mining applications privacy preserving has become a significant subject.

2 Privacy in Rule Mining using Cryptography based Technique

In this paper we describe the Privacy Preserving association rule mining [5] [6] [7] technique for a horizontally partitioned or vertical partitioned or mixed partitioned data set across multiple Parties in wireless or wired medium its depend on the network designing. Let I = {i1, i2, .in} be a set of items and T = {T1, T2,… Tn} be a set of transactions where each $T\ I \cup$. A transaction Ti contains an item set $X \cup I$ only if $i\ X \cup T$. An association rule implication is of the form $X \cup Y(X \cup Y \cup 0)$ with support s and confidence C if S% of the transactions in T contains $X \cup Y$ and C% of transactions that contain X also contain Y. In a horizontally partitioned database, the transactions are distributed among n Parties.
Support ($X \cup Y$) = probe ($X \cup Y$) /Total number of transaction the global support count of an item set is the sum of all local support counts.

Support g (X) = Support1(x) +Support2(x) +……+Support n(x).

Confidence of rule ($X \cup Y$) = Support ($X \cup Y$) / Support(X)

The global confidence of a rule can be expressed in terms of the global support.

Confidence g ($X \cup Y$) = Support g ($X \cup Y$) / Support g(X)

The aim of the privacy preserving association rule mining is to find all rules with
Global support and global confidence higher than the user specified minimum support and confidence. The following steps, utilizing the secure sum and secure set union methods described earlier are used. The basis of the algorithm is the Apriori algorithm which use the (k-1) sized frequent item sets to generate the k sized frequent item sets. The problem of generating size 1 item sets can be easily done with secure computation on the multiple Parties.

1. *Candidate Set Generation*: Overlap the globally frequent item [5] [6] set of size (k-1) with locally frequent (k-1) item set to get candidates. From these, use the Apriori algorithm to get the candidate k item sets.

2. *Local Pruning:* For each X in the local candidate set, scan the local database to compute the support of X. If X is locally frequent, it's included in the locally frequent item set.

3. *Item set Exchange*: Calculate a Secure union of the large item sets over all Parties.

4. ***Support Count*:** Compute a Secure Sum of the local supports to get the global support.

Privacy preserving association rule mining is very necessary because when the data is distributed among different partitioned like horizontal partition, vertical partition and mixed partition but in this we will describe only privacy preserving in horizontal partition. In case of horizontal partition the data is distributed in among different Party so to find the global support, global confidence and life. Then privacy is play very important role to find global result, there are mainly three important method to provide security in horizontal partition are cryptography technique, heuristic based technique and reconstruction based technique but in this we mainly focus only on cryptography technique to provide security to horizontal partitioned data. And why the cryptography technique [7] [8] [9] is more useful because two main reasons behind that. It has a well recognized and well amorphous model meant for privacy which can essentially provide good number of methodologies for verifying and validating intention. Cryptography branch has a broad mixture of tool set to incorporate privacy in data mining.

3Process of Privacy Preserving Rule Mining for Horizontal Partitioned Data without Party

In horizontal partition data is distributed in among Party in the wired or wireless medium the number of Party will be grater then 2 (n>2). And no Party is consider as a trusted party all the party have their individual private data and no other party will able to know other party data .in this method basically we are using hash based secure sum technique .in secure sum each Party will calculate their own data value and send to next Party that near to original Party and this will going till the original Party will collect all the value of data after that the parent Party will calculate the global support and global confidence and it also not necessary that the result that found is globally frequent or infrequent [7] [8] [9] its depend on value which will found after collect all the value may be that globally frequent and may not be locally frequent to say that item is globally frequent its consider that item may or may not be locally frequent. Figure 1 shows that how the data is distributed in among different Parties.

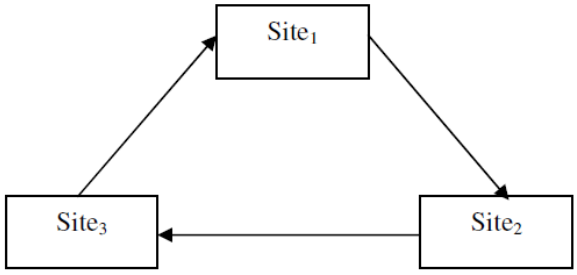

Figure1:-Communication among three sites

For this we are using one of method to find the global result of confidence and support.
There are mainly number of steps to find the global confidence and global confidence.
Step1: Each Party will calculate their frequent item sets and infrequent item sets and store the data value on memory.
Step2: Each Party will generate their own random number because we are using hash based secure sum protocol so that

each Party have two random number one of its own and other is received by previous Party.
Step3: Now the Party 1 will calculate the partial support value by using the following formula.
Ps j= Xj .support – Min support*|DB|-RN1-RNn where RN is random number.

After that Party1 calculate the mask value
PSj = PSj + mask value
Step4: Party 2 compute the PSj for each item received the list using the formula
PSj= PSj + xj. Sup - min*support |DB| +Rn1-Rn (i-1)
Step5: After that the value of Psj calculated by Party 2 send to next coming Party and after that all the value is send to the original Party and that original Party will calculate all global support .
Step6: Party 1 will find whether that global support is grater then zero or not if the value is grater then zero then it will be global frequent otherwise is infrequent.
Step7: Like that the entire Party will calculate the will calculate the global support Party 3 Party 4 Party 5…….Party n.
Step8: Finally the Party 1 calculates the actual support by using the formula in the wired or wireless medium
ASj = PSj + mask value
Step9: At last the Party 1 will send the calculated value of actual support and global frequent item set to all other Party in the horizontal partition.
Step10: Each Party will generate the association rule by using their confidence value.
Party 1 calculate the mask value by using the following formula
Mask value is calculated by using two different hash functions
Key1=Hash (key) = key mod N
And after that
Mask key = Hash2 (Key1) = Key+Mkey1
Double hash function is used to make the association rule more secure

Table1:- Database for Site1

TID/site	A1	A2	A3	A4	A5	A6
T1	1	0	1	1	1	0
T2	0	0	1	1	1	0
T3	1	0	1	1	0	1
T4	0	1	1	0	1	0
T5	1	1	0	1	0	1
T6	0	0	1	1	0	1
T7	1	1	1	0	1	0

Table2:- Database for Site2

TID/site	A1	A2	A3	A4	A5	A6
T1	1	0	1	1	1	0
T2	0	0	1	1	1	1
T3	1	1	0	0	1	1
T4	1	1	1	0	1	0
T5	1	1	0	1	0	1
T6	0	0	1	1	0	1
T7	1	1	1	0	1	0
T8	0	0	0	1	0	1
T9	1	0	1	0	0	0

Table3:- Database for Site3

TID/site	A1	A2	A3	A4	A5	A6
T1	0	0	0	0	1	0
T2	0	0	1	1	1	0
T3	1	0	1	1	0	1
T4	0	1	1	0	1	0
T5	1	1	0	1	0	1
T6	0	0	1	1	0	1
T7	1	0	1	1	1	0

4. Execution of Privacy of Preserving Rule Mining without Trusted Party

At Party 1: The list of frequent item at Party 1 {A1, A3, A4, A5, (A3, A4), (A3, A5)}
At Party 2: The list of frequent item at Party 2 {A1, A3, A4, A5, A6}
At Party 3: The list of frequent item at Party 3 {A3, A4, (A3, A4)}
Consider the item set {A2, (A3, A5)}
Select the random number RN1=10, RN2=20, RN3=10
Key =110, M=2
Hash key=key mod M
Mask key=hash key- M^{key}
Hash key=110 mod 2=0
Mask key=110-2^0=109
FOR FIRST ITEM SET I= {A2}
PS=I1Support- minimum support*DB+(RNi-RNi-1)+ Mask key
STEP1:-
PS11= 3-.5*7+ (10-20) +109=98.5
PS12=4-.5*9+ (20-10) +98.5=108
ps13=2-.5*7+ (10-20)+108=96.5
Global encrypt support (GES) = partial support-mask key
GES= 96.5-110=-13.5
Actual support=global support+ database*minimum support
AS=-13.5+23*.5=-2

5. Process of Privacy Preserving Rule Mining in Horizontal Partition Data with Trusted Party

In this method all the task will done by help of trusted party and trusted party play important role in this method as well as trusted party is helpful of providing security to the database [10] [11] [12] [13] in among Party and also data leakage is zero percent in both the wired and wireless medium. The following steps to calculate the association rule in horizontal partition with the help of trusted party is in shows fgure2.

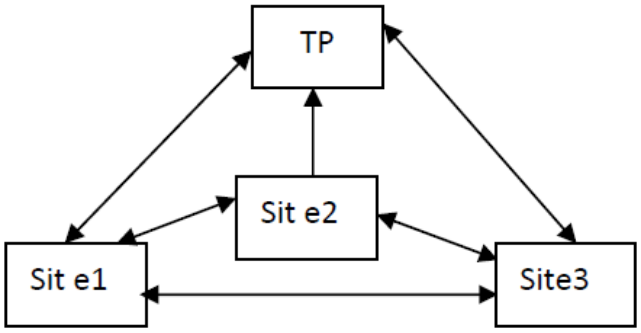

Figure2:- Communication between Sites and Trusted Party

Step1: Trusted party will send a request to calculate the locally frequent item by the help of public key and hash function as well as minimum threshold support value.
Step2: After that each Party will calculate the frequent item set and send to the trusted party without interfering of other party.
Step3: Trusted party will see all the frequent item set that coming from different Party by help of private key and merged all frequent items set and remove all the duplicate item sets. For each Party TP will generate the random number and sign (+ or -) and send to all the Party that present to the database and that Party is indicate weather that random number will added or subtracted its depend on the sign value.
Step4: Every Party computes partial support for each item set in the merged list which is received from TP by using the formula
PS ij= Xj.sup – Min Sup × | D Bi | + (Sign i) RN i
Where i indicate the ith Party, ranges from 1 to n and j indicates jth item set in the merged list, ranges from 1 to k. Each Party then broadcast its computed PSij values for all the item sets in the merged list to all other Parties.
Step5: Every Party computes Total PS ij for each item set Xj by using the formula.
Total PS ij =∑for each j = 1 to k and sends to the Trusted party.
Step6: Trusted party will send the value of total partial support to all Parties that present in database if any duplication will occur then the step 5 will follow again to calculate the partial support.
Step7: Trusted party computes Global Excess Support for each item set Xj by using the formula
Global support j = TotalPS1j – Sign Sum RN
Where Sign Sum RN is computed by adding all the random numbers with their signs by trusted party. If the computed value of GES j ≥ 0 then the item set Xj is globally frequent otherwise it is globally infrequent.
Step8: For each global frequent item set Xj, Trusted party finds Actual Support as
Actual support j = Global support j + Minimum Support * |DB|
Where |DB| =DB1+DB2……DBN
Step9: Trusted party will send the list of frequent item sets to all other Parties present in distributed database.
Step10: Every Party will generate association rule by using the confidence to every Parties and the minimum support that received by trusted party.

6 Execution of Privacy Preserving Rule Mining with Trusted Party

For implementation we will take above database that contain of three different tables.
At Party 1: The list of frequent item at Party 1 {A1, A3, A4, A5, (A3, A4), (A3, A5)}
At Party 2: The list of frequent item at Party 2 {A1, A3, A4, A5, A6}
At Party 3: The list of frequent item at Party 3 {A3, A4, (A3, A4)}
Consider the item set {A2, (A3, A5)}
Select the random number RN1=10, RN2=20, RN3=10
So that the size of database that contain |DB1|=7, |DB2|=9, |DB3|=7 so the size of global database is 23.
Trusted party will compute the signsumRN BY adding three random numbers
SignsumRN= (+) 10+ (+20) + (+) 10=40
Partial supports for A2 at different Parties are computed as follows.
At Party1
PS11=A2.Sup-50%of DB1+ (Sign1) RN1
PS11=3-.5*7+ (+) 10=9.5
At Party2
PS21=A2.Sup-50%of DB2+ (Sign2) RN2
PS21=4-.5*9+ (+) 20=19.5
At Party3
PS31=A2.Sup-50%of DB3+ (Sign3) RN3
PS31=1-.5*7+ (+) 10=7.5
Party1 broadcast 9.5 to all other Party Party2 and Party3, Party2 broadcast 19.5 to Party3 and Party1, Party3 broadcast 7.5 to Party 1 and Party2.
Total PS11=PS11+ (PS21+PS31) =9.5+19.5+7.5=36.5
Total PS11=PS21+ (PS11+PS31) =19.5+9.5+7.5=36.5
Total PS11=PS31+ (PS21+PS11) = 7.5+19.5+9.5=36.5
Trusted party receives 36.5 as total support of an item set A2 from three Parties which ensures the computations performed by all others Parties is correct. Trusted party when calculates the Global Excess Support by subtracting the Sign sum RN from the TotalPS11
Global Excess support=TotalPS11-SignsumRN=36.5-40=-3.5
The value of global support is -3.5 then it means that the item sets are globally infrequent.
Actual support of A2is computed by adding minimum support of the total database of global excess support
AS11=Global Excess Support+ minimum support*|DB|=-3.5+.5*23=8
Hence the Global frequent item A2 Support is 8.

7. Conclusion

The complexity of preserving privacy in rule mining when the database is distributed horizontally in the environment when the number of Parties in greater than two in wireless or wired medium when no trusted party is considered. A replica which adopts a hash based secure sum cryptography technique to find the global association rules is propose in this paper by preserving the privacy constraints. Double hashing function is adopted to enhance the privacy further. The proposed replica capably finds global frequent item sets even when no Party can be treated as trusted. And next in this paper we compare with trusted party. The trusted party

initiates the process and prepares the merged list. All the Parties computes the partial supports and total supports for all the item sets in the merged list using the sign based secure sum cryptography technique and based on these results finally trusted party finds global frequent item sets. And after comparing the result of these we find output that data leakage with trusted party is more as compare to without trusted party so privacy also without trusted party is more as compare to with trusted party.

References

[1]. Agrawal, R., et al "Mining association rules between sets of items in large database". In: Proc. of ACM SIGMOD'93, D.C, ACM Press, Washington, pp.207-216, 1993.

[2]. Agarwal, R., Imielinski, T., Swamy, A. "Mining Association Rules between Sets of Items in Large Databases", Proceedings of the 1993 ACM SIGMOD International Conference on Management of Data, pp. 207-210, 1993.

[3]. Srikant, R., Agrawal, R "Mining generalized association rules", In: VLDB'95, pp.479-488, 1994.

[4]. Agrawal, R., Srikant, R, "Privacy-Preserving Data Mining", In: proceedings of the 2000 ACM SIGMOD on management of data, pp. 439-450, 2000.

[5]. Lindell, Y., Pinkas, B, "Privacy preserving Data Mining", In: Proceedings of 20th Annual International Cryptology Conference (CRYPTO), 2000.

[6]. Kantarcioglu, M., Clifton, C, "Privacy-Preserving distributed mining of association rules on horizontally partitioned data", In IEEE Transactions on Knowledge and Data Engineering Journal, IEEE Press, Vol 16(9), pp.1026-1037, 2004.

[7]. Han, J. Kamber, M, "Data Mining Concepts and Techniques". Morgan Kaufmann, San Francisco, 2006.

[8]. Sheikh, R., Kumar, B., Mishra, D, K, "A Distributed k- Secure Sum Protocol for Secure Multi-Party Computations". Journal of Computing, Vol 2, pp.239-243, 2010.

[9]. Sugumar, Jayakumar, R., Rengarajan, C "Design a Secure Multi Party Computation System for Privacy Preserving Data Mining". International Journal of Computer Science and Telecommunications, Vol 3, pp.101-105. 2012.

[10]. N V Muthu Lakshmi, Dr. K Sandhya Rani ,"Privacy Preserving Association Rule Mining without Trusted Party for Horizontal Partitioned database", International Journal of Data Mining & Knowledge Management Process (IJDKP) Vol.2, pp.17-29, 2012.

[11]. N V Muthu lakshmi, Dr. K Sandhya Rani, "Privacy Preserving Association Rule Mining in Horizontally Partitioned Databases Using Cryptography Techniques", International Journal of Computer Science and Information Technologies(IJCSIT), Vol. 3 (1) , PP. 3176 – 3182, 2012.

[12]. Goldreich, O., Micali, S. & Wigerson, A. ,"How to play any mental game", In: Proceedings of the 19th Annual ACM Symposium on Theory of Computing, pp.218-229, 1987.

[13]. Franklin, M., Galil, Z. & Yung, M.,"An overview of Secured Distributed Computing". Technical Report CUCS- 00892, Department of Computer Science, Columbia University.

1. Dr. Prasant Kumar Pattnaik
Associate Professor, KIIT University, India
Senior member of IACSIT.
Research Area: MANET, Wireless Sensor Network, Cloud Computing.
Email id: patnaikprasant@gmail.com

2. **Raghvendra Kumar** received B. Tech. in Computer Science and Engineering from SRM University Chennai (Tamil Nadu), India, in 2011, M. Tech. in Computer Science and Engineering from KIIT University, Bhubaneswar, (Odisha) India in 2013, and pursuing Ph.D. in Computer Science and Engineering from Jodhpur National University, Jodhpur (Rajasthan), India. Currently he is working as Assistant Professor in Computer Science and Engineering Department at L.N.C.T Group of College Jabalpur, M.P. India. He has published many research papers in international journal including IEEE and ACM. He attends many national and international conferences and also He Received best paper award in IEEE 2013 for his research work in the field of distributed database in Tamil Nadu. His researches areas are Computer Networks, Data Mining, cloud computing and Secure Multiparty Computations, Theory of Computer Science and Design of Algorithms

3. Dr. Yogesh Sharma
Head of Department, Mathematics
Jodhpur National University, Jodhpur, Rajsthan, India