ISSN (Online) 2278-1021
ISSN (Print) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 4, Issue 9, September 2015

Distributed Databases Rules of Mining in the field of Association

K.Rajeshwar Rao[1], A.Prakash[2], B.Venkat Ramudu[3]

Asst.Professor, Department of CSE, MREC(A), Hyderabad , India[1,3]

Asso.Professor, Department of CSE, St. Martyn's Engineering College, Hyderabad, India[2]

Abstract: Association Rule mining is one among the data mining technique employed in distributed databases. This technique disclose some attention grabbing relationship between locally massive associate degreed globally massive itemsets and proposed an algorithm, Fast distributed mining of association(FDM) , unsecured distributed version of Apriori algorithm and generates a low range of candidate sets and well reduces the amount of message to be passed at mining association rules. Two secure multi party algorithms are main ingredients in our protocol. One that computes the union of private subsets that every of the interacting players hold, and another that tests the inclusion of element by one player in a subset held by another. This protocol offers increased privacy with relation to the protocol and additionally, it is less complicated and considerably a lot of efficient in terms of Communication rounds, communication cost and computation cost.

Index terms: Privacy Preserving Data Mining, Association Rules, Distributed Databases.

I. INTRODUCTION

Aim of data mining is to extract vital information from massive datasets, however typically these datasets are split among varied parties. Data mining is defined as the technique for extracting hidden, predictive and knowledge data from large distributed databases. The technology that has emerged as method of identifying patterns and trends from large quantities of knowledge. This paper studies the matter of association rule mining in horizontally distributed databases. In the distributed databases, there are many players thathold same databases that share same schema however hold data totally on different entities. The goal is to search out all association rules with support s and confidence c to attenuate the data disclosed regarding the personal databases command by those players [1].

Kantarcioglu and Clifton studied the problem whenever a lot of suitable security definitions that permit parties to settle on their desired level of security are required, effective solutions that maintain the required security [2]. So they devised a protocol for its solution. The main part of that protocol is the sub protocol for secure computation of union of private subsets that are controlled by the various players. It makes the protocol expensive and its implementation depends upon cryptographic primitive's strategies, oblivious transfer and hash function also the leakage of data renders the protocol not secured perfectly.

This paper proposed an algorithm privacy preserving distributed data mining algorithm for horizontally distributed data sets and to get most interesting association or correlation relationships among an outsized or large set of data items and to incorporate encryption security techniques to reduce the information that goes to shared with others, while adding little overhead to the mining task [1]. In the proposed association rule that hold within the unified database with support and confidence no smaller than the given threshold s and c, accordingly.

The data that would like to secure during this paper is not only individual transaction within the completely different databases, however more public or global information like what association rules are supported locally in every of these databases. The proposed protocol improves upon that in Kantarcioglu and Clifton [2] in terms of simplicity efficiency such as privacy and security.

II. LITERATURE REVIEW

Data mining in centralized model assumes that required data is either available at or can sent to a central site and data mining algorithm applied at central site. A simple approach not to share data is, perform data mining tool at each site independently and combine the results. However, it fails in globally valid results because of data skewness and disparity between local and global results. Individual sites are unable to detect cross-site correlations in data mining. Because the same item may be duplicated at different sites and there will be an overweight in the results.

FP tree is a compact data structure and without candidate set generation it finds the frequent itemset by traversing itemsets through FP tree. Paper [10] uses the FP tree algorithm for association rule mining in distributed environment to find the frequent itemsets. Data encryption standards are used to generate the privacy to database by using two keys. With the key1, first party encrypts the dataset. This encrypted dataset again encrypted with key2. At the receiver side at first, it decrypt the dataset with key2 and then after with key1. It provides high security to the databases compared with other cryptographic techniques and it is also known as Double encryption. With data leakage of zero percentage it provides higher privacy to the database; however it applicable to homogeneous databases only.

ISSN (Online) 2278-1021
ISSN (Print) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 4, Issue 9, September 2015

The paper [11] uses apriori algorithm with extended Distributed RK sum secure protocol for privacy preserving data mining. At first, on given dataset Apriori algorithm is applied then to find the frequent itemsets of individual parties and then to get global results Extended distributed RK sum secure algorithm is applied. Distributed RK sum secure protocol used for secure multi party protocol but drawback is more than two parties join together, they can know the data of some party. The drawback is reduced in extended distributed RK sum protocol with more privacy to the data, the communication and computation complexity O (n).

The paper [13] studied five different algorithms such as Apriori algorithm, MCISI algorithm, MS algorithm, Apriori with systematic rules and HMT (Hash Mapping Table). Apriori algorithms combined with hashing techniques to reduce the time and space complexity. MSApriori (minimum support) method provides different minimum support itemset values for different items. Apriori systematic rules used to save the time by scanning the database only once rather than multiple scans.

Privacy Preserving Data Mining

In [4] studied the problem in privacy preserving association rule mining in horizontally distributed databases among n (n>2) sites, with no sites can be treated as trusted. The proposed protocol uses hash based cryptographic techniques to find the global frequents itemsets. Privacy preserving rule mining categorized into three types based on privacy preserving/protection techniques.

1. Heuristic based techniques
2. Reconstruction based techniques
3. Cryptographic based techniques

Heuristic based techniques uses adaptive modification and it only modifies the selected values, utility loss is minimized. Cryptographic based techniques are used for both horizontal distributed databases and vertically distributed databases and it is based on the secure multi party computation. Here no database knows anything except his input database and the result, so that computation is secured. Reconstruction based techniques are used for different data such as numerical, binary and categorical data, its work on the problem of privacy preservation by perturbing the data. Original distribution of data is constructed from randomized data by Reconstruction based techniques.

Association Rule:

Association rule mining is well known and popular method finding interesting relationships between from large databases. Today association rules are used in many applications including intrusion detection, biometrics, web usage mining and continuous production. The concept was popularized by Agrawal et al. article in 1993. It is a two step process i.e. Minimum support and confidence, minimum support used to find all frequent itemsets from given database and the constraint confidence used to form the rules. The association rule mining problem is stated/defined as Let I = {i1, i2, i3…..in} are the set of items, database D consists set of transaction T. Each transaction T is a set of items from I such that T⊆I, with a unique transaction id TID. Each transaction T conatins X, and a set of some items in I, if X⊆T. An association rule is an implication of the form , X⇒Y, where X ⊂ I, Y ⊂ I and X ∩ Y= φ. The association rule X⇒Y holds in the transaction set D, with confidance c if c% of transaction in D that contain X also contain Y.The rule X⇒ Y has support s in the transaction set D if S % of transactions in D contain X U Y. In shopping centers association rules are used to place the items side by side , so that more items will be sold. In Amazon , association rules mining to recommended you to buy the products based on present item buying or browsing. Google uses assocaition rules for mining, when you type the words for search it will give frequent associated words.

Apriori Algorithm

In transactional databases, apriori algorithm is used for mining of frequent itemset and association rule learning. Apriori algorithm proceeds by identifying the frequent itemsets and extending them to larger and larger itemsets in database. It identifies the itemset at least C transaction in the database. Apriori algorithm uses the bottom up approach and proposed by Agarwal and Srikanth in 1994. It usage is more in market basket analysis. To count candidate itemset efficiently Apriori algorithm uses breadth –first search and hash tree structure. Apriori algorithm works based on Apriori principle and to generate candidate k-item sets from (k-1) frequent itemsets, pruning technique used to avoid the measuring certain itemsets. Hash based methods with Apriori algorithms are used to reduce the time and space complexity.

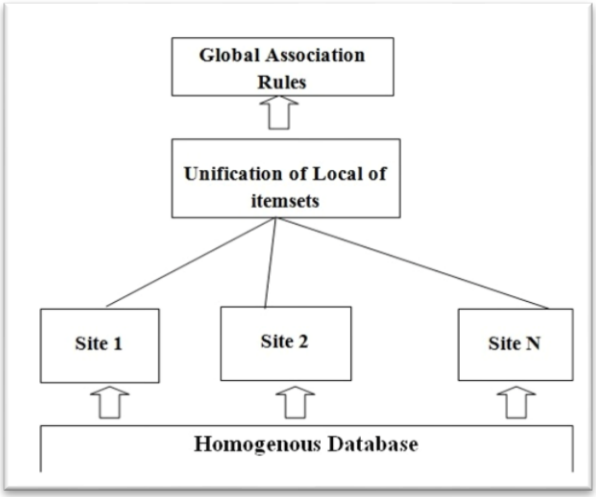

Fig1 Distributed Databases Architecture of the system

Synthetic Database Generation

To calculate the performance of algorithms over a very large range of data characteristics the synthetic transactions are used and the process involved here is data anonymization. It is used in data mining process. Synthetic data are generated to get certain conditions and

ISSN (Online) 2278-1021
ISSN (Print) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 4, Issue 9, September 2015

to meet specific needs that may not be found in real, original data. Following are the parameters used in synthetic data generation and these values are similar to used in [2], [5], [12].

Parameter	Interception
N	No. of transactions in database
L	No. of items
A_t	Average size of transaction
A_f	Maximal potentially large itemsets average size
N_f	No. of maximal potentially large itemsets
CS	Clustering size
PS	Pool size
C or M F	Correlation level Multiplying factor

FDM Algorithm

Following are the steps for FDM algorithm
1. Initialization
2. Candidate set generation
3. Local Pruning
4. Unifying the candidate itemset
5. Computing local support
6. Broadcast the mining results

Cheung [4] proposed Fast distributed mining for association rules, for distributed databases, Apriori algorithm for association rule mining for individual databases. In [2] proposed an insecured version of FDM algorithm but algorithm violates privacy in two stages, in step four and step six. In step 4 whenever the players broadcast the itemsets that are locally frequent in their private databases. In step 6 where they broadcast the sizes of the local supports of candidate itemsets. Kantarcioglu & Clifton [2] proposed secure implementations of these steps and also in [1] describe the various implementations and proceed to analyze implementations in terms of privacy, efficiency and compare them. Finally it show that our protocol offers better/enhanced privacy which is less complicated and is considerably additional efficient in terms of communication rounds, communication cost and computation cost.

III. DISCUSSIONS

Following are the steps to secure computation itemsets using FDM K&C protocol.

1. Selection of Cryptographic primitive cipher.

- Players choose the required commutative cipher and its corresponding private key and selected hash function applied on all itemsets for encryption.
- A lookup table with hash values is generated to compute pre image of hash values.
- Encrypt all the itemsets.

2. Merging of itemsets

- Odd players send its encrypted set to player1.
- Even players send its encrypted set to player2.
- Player1 unifies all itemsets were sent by odd players and duplicates are removed.
- Players 2 unifies all itemsets were sent by odd players and duplicates are removed.
- Player 2 sends his permuted list of itemsets to player 1.
- Player 1 unifies his list of itemsets and the list received from player 2 and from the unified list duplicates are removed.
- Denote the final list as EC_S^k.

3. Decryption.

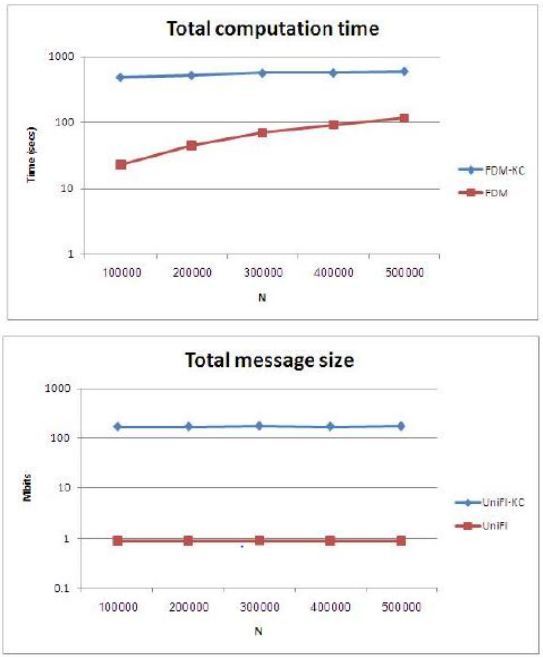

Fig2 Computation and communication cost versus number of transactions N

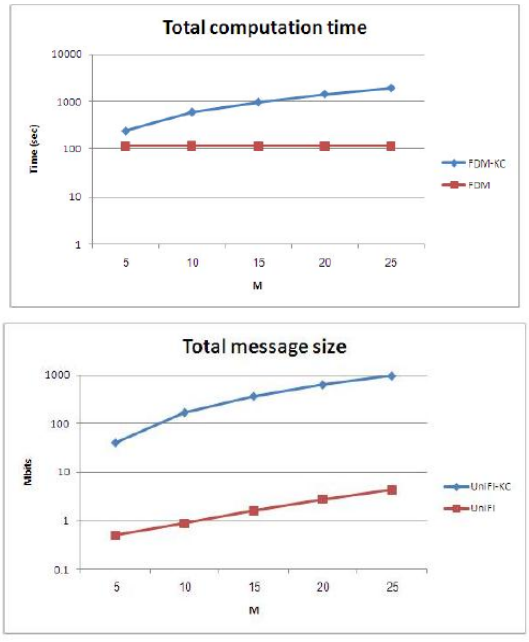

Fig3 Computation and communication costs versus the number of players M

DOI 10.17148/IJARCCE.2015.49105

ISSN (Online) 2278-1021
ISSN (Print) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 4, Issue 9, September 2015

The fig. 3 illustrates the leading protocol K and C. In this protocol main ingredients are two novel secure multiparty algorithms in which union and intersection operations. This protocol works if the number of players is more than two.

IV. CONCLUSION

The proposed protocol secure mining of association rules in horizontally distributed databases offers enhanced privacy and security than the leading protocol [2]. Two secure multi party algorithms are main ingredients in our protocol.

One that computes the union of private subsets that every of the interacting players hold, and another that tests the inclusion of element by one player in a subset held by another. Proposed protocol is more significant in communication cost, communication rounds and computation cost

REFERENCES

[1] Tamir Tassa, "secure Mining of Association rules in Horizontally Distributed Databases" IEEE trans. Knowledge and Data Engg. Vol. 26, no. 2, April 2014.

[2] M. Kantarcioglu & C. Clifton," Privacy preserving distributed mining of association rules on horizontally partitioned data. IEEE Transactions on Knowledge and Data Engineering, 16:1026-1037, 2004.

[3] Agarwal and R. Srikanth "Fast algorithms for mining association rules in large databases. In VLDB, pages 487-499, 1994.

[4] D.W.L Cheung, V.T.Y. Ng, A.W.C. Fu, and Y. Fu. Efficient mining of association rules in distributed databases. *IEEE Trans. Knowl. Data Eng.*, 8(6):911–922, 1996Tavel, P. 2007 Modeling and Simulation Design. AK Peters Ltd.

[5] D.W.L. Cheung, J. Han, V.T.Y. Ng, A.W.C. Fu, and Y. Fu. A fast distributed algorithm for mining association rules. In *PDIS*, pages 31–42, 1996.

[6] A.V. Evfimievski, R. Srikant, R. Agrawal, and J. Gehrke. Privacy preserving mining of association rules. In *KDD*, pages 217–228, 2002.

[7] 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.

[8] J. Vaidya and C. Clifton, "Privacy preserving association rule mining in vertically partitioned data," in *The Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, Edmonton, Alberta, Canada, July 23-26 2002, pp. 639–644.

[9] H. Grosskreutz, B. Lemmen, and S. R¨uping. Secure distributed subgroup discovery in horizontally partitioned data. *Transactions on Data Privacy*, 4:147–165, 2011.

[10]] Jyotirmayee Rautaray, Raghavendra Kumar, "Privacy Preserving Databases Using Data Encryption (DES) " International Journal of Innovative Research in Science , Engineering and Technology Vol. 2 , Issue 3, March 2013.

[11] Meers 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.

[12] J.S. Park, M.S. Chen, and P.S. Yu. An effective hash based algorithm for mining association rules. In *SIGMOD Conference*, pages 175–186, 1995.

[13] Priyanka Asthana, Anju Singh , Diwakar Singh," A Survey onAssociation 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.

BIOGRAPHIES

K .RAJESHWAR RAO ,Working as Asst.professor, in Department of CSE in Malla Reddy Engineering College , Hyderabad . He has vast experiene in teaching. His research includes data mining , Web technologies, Artificial intelligence , Software Engineering and Computer networks.

A.PRAKASH, currently working as Head of the Department, CSE, St.Martin's Engineering College. He completed his M.Tech in Software Engineering from JNTUH. He had an experience of more than 12 years in teaching and his areas of interest are Education Technologies, Computer Networks, Cryptography and security.

B VENKATA RAMUDU, is working as Asst.professor Malla Reddy Engineering College (Auto nous) Hyderabad Education Technologies, Computer Networks, Cryptography and security.