IJARCCE

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

International Journal of Advanced Research in Computer and Communication Engineering
ISO 3297:2007 Certified
Vol. 5, Issue 11, November 2016

Subgroup Identification in Vertically Partitioned Data Using UNIFI Protocol

Dr. Madhavi Karanam[1]

Professor, CSE Department, Gokaraju Rangaraju Institute of Engineering and Technology, Hyderabad, India[1]

Abstract: Association rule learning is a popular and well researched method for discovering interesting relations between variables in large databases. It is intended to identify strong rules discovered in databases using different measures of interestingness. Based on the concept of strong rules for discovering regularities between products in large-scale transaction like data recorded by point-of-sale (POS) systems in supermarkets. In data mining, association rules are useful for analysing and predicting customer behaviour. They play an important role in shopping basket data analysis, product clustering, catalogue design and store layout. The proposed work is aimed to implement the protocol used in the existing system to the problem of distributed association rule mining in the vertical data. The proposed work uses unifying lists of locally frequent item sets (UNIFI) protocol to find out the subgroup in vertically partitioned data.

Keywords: Association Rule; Data mining; Horizontally Distributed Databases., Privacy-Preserving.

I. INTRODUCTION

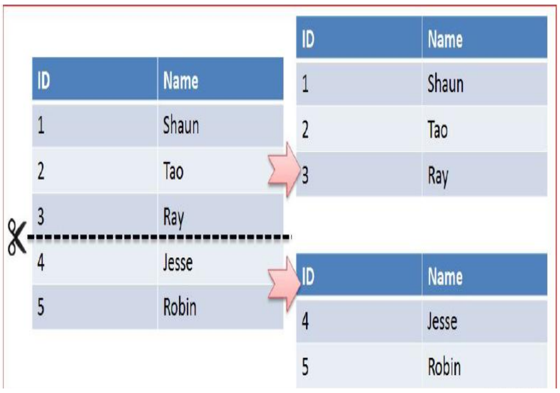
The amount of data kept in computer files is growing at a phenomenal rate. The data mining field offers to discover unknown information. Data mining is often defined as the process of discovering meaningful, new correlation patterns and trends through non-trivial extraction of implicit, previously unknown information from the large amount of data stored in repositories, using pattern recognition as well as statistical and mathematical techniques [13]. An SQL query is usually stated or written to retrieve specific data, while data miners might not even be exactly sure of what they require. Whether data is personal or corporate data, data mining offers the potential to reveal what others regard as sensitive (private). In some cases, it may be of mutual benefit for two parties (even competitors) to share their data for an analysis task. However, they would like to ensure their own data remains private. In other words, there is a need to protect sensitive knowledge during a data mining process. This problem is called Privacy-Preserving Data Mining (PPDM). Most organizations may be very clear about what constitutes examples of sensitive knowledge. Challenge is to identify what is non-sensitive knowledge because there are many inference channels available to adversaries. It may be possible that making some knowledge public (because perceived as not sensitive), allows an adversary to infer sensitive knowledge. In fact, part of the challenge is to identify the largest set of non-sensitive knowledge that can be disclosed under all inference channels.

While considering data, data may be distributed among the various systems. Most of the businesses share their information along with their personal information for getting equal benefits. Sharing of this type of personal information arise the privacy issue. Though businesses share their private information but still they focus on to the data remains as a private only. This is known as secure mining. For the user distributed database is like a single compartment it is not in scattered format. As data is increasing day by day we need to store it on different computer and whenever user want to access it, it works like a single unit though we are storing the data on different machines. The data on several computers can be simultaneously accessed and modified using a network. In a network each server is linked by its local database management system (DBMS), and each cooperates to maintain the consistency of global database. To maintain privacy of the data many scientist put their efforts so that we get data without losing the privacy of that related data. Whenever we are concerning with data mining, Security is major issue while extracting data. Privacy Preserving Data Mining concerns with the security of data and provide the data on demand as well as amount of data that is required. Sometimes it may happen that we get the information but not complete. Privacy preserving algorithm is concerns on the basis of its performance, data utility, and level of uncertainty or resistance. There are various techniques and tools for security are used. For mining the data many protocols were proposed by various scientists keeping common goal as to protect sensitive data. While studying about this problem of privacy preserving most of the scientist goes through by searching frequent item set and accordingly related association rule. Association rule indicates association between various entity while fetching data or getting the result. Suppose anyone who wanted to buy bread at that time there is maximum possibility of buying the milk. This is association rule, where the things are connected with each other. Many business areas use this association rule for getting the benefits.

In horizontal distributed databases system, the data is stored on different machines. Other generic solution discussed in [14], and this information belongs to one

DOI 10.17148/IJARCCE.2016.511126

particular related subject. Consider one example for proper understanding of this concept. Let there is one table which have lots of records in rows and columns format. Some system may store some records which belong to the same column and other columns along with their information may store on the next machines. So, data is distributed along the various machines. If the data stored on different machines and is divided by rows means while storing the data on different machines some rows are stored on one machine and others are stored on different machines, i.e. partitioning the data according to rows is called horizontally partitioned (Distributed Databases) and if that data is stored and partitioning according to the column wise then it is called as a vertical partitioning (Distributed databases). Some may prefer to store or partition of data according to vertical and some may prefer to use horizontal partition. Most of the scientist uses semi-honest model where node may follow or not all protocol for accessing the data among distributed system.
Following figure illustrate the horizontal partition.



ID	Name
1	Shaun
2	Tao
3	Ray
4	Jesse
5	Robin

ID	Name
1	Shaun
2	Tao
3	Ray

ID	Name
4	Jesse
5	Robin

Fig. 1. Illustration of Horizontal partition

Whenever we consider databases, data are distributed among various parties and whatever result i.e. collective result of these all data either of the vertical partitioned or horizontal partitioned. As name specifies in distributed system data is not store in a single computer and as data is increasing day by day we need to store it on multiple storing devices. That’s why we are using to store it on multiple machines. In proposed method though we are storing this data in scattered format, it appears like a single system for the user. User need not to know anything about which data is stored on which machine. User need to know only about the information which has to be return as a result. As data is stored on different machines we need to access it simultaneously and also need to update the data or need to modify the data if any user changes that data within the network.

II. RELATED WORK

There are various methods that are used by different authors for developing secure protocol for mining of data. Most of the time data is distributed for sharing purposes but sharing or collective analysis of this data is not possible for the security purpose. And everyone wants their data to be private always. If we got the data which does not use to get only private information or which does not effect on anyone’s private data then that data mining will not all have the privacy issue. While privacy issue is concerned researchers focuses on two settings. Privacy preserving can be divided into following two categories
1. Perturbation and randomized based approach
2. Secure multiparty computation based approach

Second approach is based on cryptographic tools for mining the data. But as it concerns with the multi-party, i.e. multiple user that’s why computation cost and communication cost is higher than first one. As number of user increases the cost required to handle is also large. That’s the main thing which is reduced by modifying the data mining algorithm for perturbation technique which will build classifier directly. Businesses like hospital or bank need to preserve personal information and they need to share the person specific record. There are some generalized techniques that were used at the cost of loss of information. And these techniques were used to solve the external linkage problem. There are mainly two things that come along with the result: Quasi identifier and sensitive attribute to protect our data we can do it by simply do not display the result together by quasi identifier and sensitive attribute. The main help of these, quasi identifier and sensitive attribute, is to support data mining tasks that consider both type of attribute. To improve the method for privacy we are transforming a part of quasi-identifier and personalizing the sensitive attribute values. Existing works implements clustering. Clustering is nothing but grouping. That means while considering data, it is the proper grouping of inter related data. While sharing this data in a group, privacy is the major issue. To preserve data privacy we may go through the use of synthetic data generation. So whenever we are applying the synthetic data generation technique IPSO families of methods are used. It uses to generate accurate result in cluster.

Yao [2] was the first to propose a generic solution for this problem in the case of two players. Other generic solutions, for the multi-party case, were later proposed in [3], [4]. Existing work considered, partial databases as inputs and the list of association rules is the required output that hold in unified database with no smaller than some defined support and confidence [5], [6], [7], [8], [9]. Kantarcioglu and Clifton studied and developed a protocol. This protocol computes union of private subsets that are possessed by different players. In this main part of the protocol is a sub-protocol and hence it increases its cost which is implemented by hash function, obvious transfer and encryption [10].

III.PROPOSED WORK

Main objective of this paper is to implement the protocol to the problem of distributed association rule mining in the vertical data. This paper uses the unifying lists of locally frequent item sets (UNIFI) [1] protocol to find out the subgroup in vertically partitioned data. The proposed work also uses two secure multi party algorithms, one for computing the union of private subsets and other for testing the inclusion of an element held by one player in a subset held by another player. The proposed implementation of protocol tries to improve privacy and efficiency to greater extent.

A. The Fast Distributed Algorithm

Protocol used in proposed work is based on the Fast Distributed Mining (FDM) [1] algorithm which is an unsecured distributed version of Apriori algorithm. Its main idea is that any s-frequent item set must be also locally s-frequent in at least one of the sites.

The Fast Distributed Mining algorithm

- Step 1: Initialization It is assumed that the players have already jointly calculated Fs $^{k-1}$. The goal is to proceed and calculate Fs ^{k}.
- Step 2: Candidate Sets Generation Pm computes the set Fs $^{k-1}$,m ∩ Fs $^{k-1}$. Then apply on that set the Apriori algorithm in order to generate the set of Bs ^{k,m} candidate k-itemsets.
- Step 3: Local Pruning For each X € Bs ^{k,m}, Pm computes suppm(X). Then retain only those itemsets that are locally s-frequent denoted as Cs ^{k,m}.
- Step 4: Unifying the candidate itemsets. Each player broadcasts his Cs ^{k,m} and then all players compute Cs ^{k} := Um=1 ^{M} Cs ^{k,m}.
- Step 5: Computing local supports all players compute the local supports of all itemsets in Cs ^{k}.
- Step 6: Broadcast Mining Results each player broadcasts the local supports that he computed. From that, everyone can compute the global support of every itemset in Cs ^{k} . Finally, Fs ^{k} is the subset of Cs ^{k} that consists of all globally s-frequent k-itemsets.

B. Modified UNIFI-KC

UNIFI-KC Protocol UNIFI-KC (Unifying lists of locally Frequent Item sets- Kantarcioglu and Clifton). Works as follows:

- First, each player adds to his private subset C^{k,m} $_{s}$ fake item sets, in order to hide its size.
- Then, the players jointly compute the encryption of their private subsets by applying on those subsets a commutative encryption, where each player adds, in his turn, his own layer of encryption using his private secret key.

C. Phases of Modifies UNIFI-KC

Protocol UNIFI-KC securely computes the union of private subsets of some publicly known ground set (Ap(F^{k-1} $_{s}$)). Such a problem is equivalent to the problem of computing OR of private vectors.

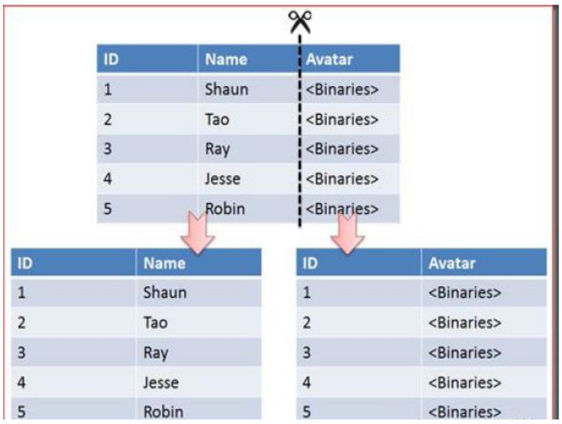
Indeed, if the ground set is Ω = {w1, . . ., €wn}, then any subset B of Ω may be described by the characteristic binary vector b =(b1,. . . , bn) € Z^{n}_{2} where bi = 1 if and only if w_i € B. Let bm be the binary vector that characterizes the private subset held by player Pm, 1<= m <= M. Then the union of the private subsets is described by the OR of those private vectors, b = W^{M} $_{m=1}$ bm. The UNIFI function has three major phases:

- In Phase 0 the players select the needed cryptographic primitives: They jointly select a commutative cipher, and each player selects a corresponding private random key. In addition, they select a hash function h to apply on all item sets prior to encryption.
- In Phase 1, all players compute a composite encryption of the hashed sets Cs $^{k, m}$, 1 < = m < = M. First, each player Pm hashes all item sets in Cs $^{k, m}$ and then encrypts them using the key Km. (Hashing is needed in order to prevent leakage of algebraic relations between item sets.
- In Phase 2, the players merge the lists of encrypted item sets. At the completion of this stage P1 holds the union set Cs ^{k} = U^{M} $_{m=1}$ Cs $^{k, m}$ hashed and then encrypted by all encryption keys, together with some fake item sets that were used for the sake of hiding the sizes of the sets C s $^{k, m}$; those fake item sets are not needed anymore and will be removed after decryption in the next phase.
- In Phase 3, a similar round of decryptions is initiated. At end, the last player who performs the last decryption uses lookup table T that was constructed in Step 4 in order to identify and remove the fake item sets and then to recover Cs ^{k} . Finally, that player broadcasts C s ^{k} to all his peers.

D. Vertically Partitioned Data

The horizontal data is quite opposite to vertical data. In order to handle vertical data the proposed system tries to implement frequent item sets which uses the top down and bottom up searching.

The horizontal data is traversed from left to right, where modified UNIFI Protocol traverses the data from top to bottom using index search technique. Whenever searching any data base starts with index search technique stores the location of particular file in an index and for the next time without searching entire database the searching technique just utilizes same index to search directly.



ID	Name	Avatar
1	Shaun	<Binaries>
2	Tao	<Binaries>
3	Ray	<Binaries>
4	Jesse	<Binaries>
5	Robin	<Binaries>

ID	Name
1	Shaun
2	Tao
3	Ray
4	Jesse
5	Robin

ID	Avatar
1	<Binaries>
2	<Binaries>
3	<Binaries>
4	<Binaries>
5	<Binaries>

Fig. 2. Process of Vertical Partition

IV. EXPERIMENTAL RESULTS

Analysis phase considers three metrics to evaluate the proposed technique. For comparisons UNIFI & UNIFI-KC are considered.

Considered metrics are:

- Time taken to UNIFI
- Total computation cost
- Total message size

First and foremost metric is to calculate time taken to UNIFI the item sets where x-axis defines the number of item sets and y-axis defines the time taken in seconds. Results are clearly depicting that the proposed technique takes constant time irrespective of the item set size.

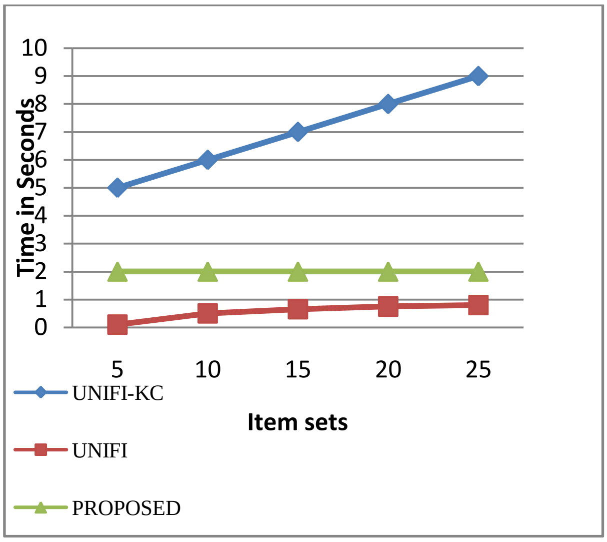


Fig. 3. Time to Unifi Candidate Item sets

Next metric is to calculate time taken for computation of the complete association rule mining. This time depicts performance of the proposed and existing systems.

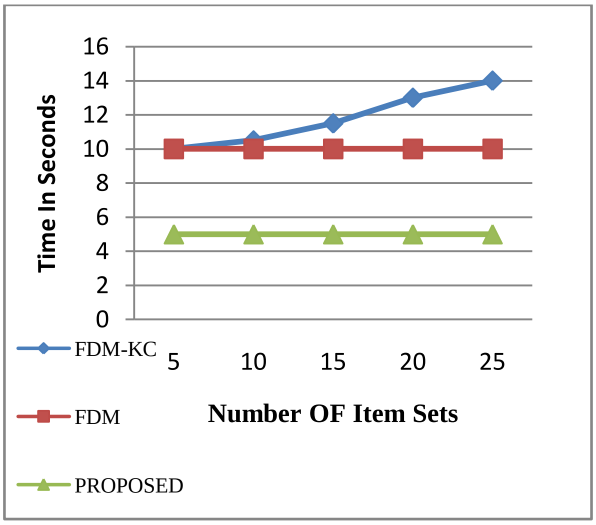


Fig. 4. Total Computation Time

Figure 4 shows that the total computation time taken is minimum for the proposed technique compared to both existing schemes. Final metric is the total message size taken to transmit which is measured in Mbits. The lesser the message size then performance will be higher. Figure 5 depicts the results of total message size. Figure 5 clearly depicts that UNIFI and the proposed schemes takes constant message size irrespective of the item sets. Proposed technique performs well compared to UNIFI because it is tested on vertical data.

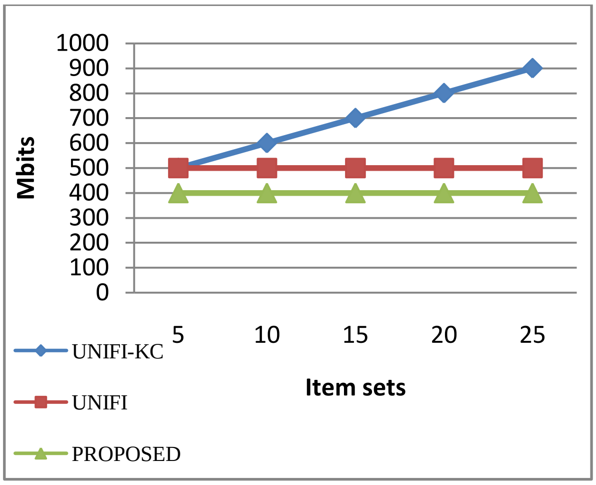


Fig. 5. Total Message size

V. CONCLUSION

Security of data is defined in this paper which is a major problem in data mining and then proposed a secure mining technique for association of data for vertically distributed data. Horizontal data is traversed from left to right. The modified UNIFI Protocol traverses the data from top to bottom using index search technique. The proposed technique is tested under three basic metrics and in all these three aspects the results are good compared to traditional techniques.

REFERENCES

[1] TamirTassa,‖Secure Mining of AssociationRules in Horizontally Distributed Databases,‖ IEEE trans. Knowledge and Data Eng., vol.26, no.4, pp.970-98, 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] 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,‖ Proc. Fourth Int'l Conf. Parallel and Distributed Information Systems (PDIS), pp. 31-42, 1996.

[4] J. Vaidya and C. Clifton, ‖Privacy Preserving Association Rule Mining in Vertically Partitioned Data,‖ Proc. Eighth ACM SIGKDD Int'l Conf. Knowledge discovery and Data Mining (KDD), pp. 639-644, 2002.

[5] Y. Lindell and B. Pinkas, ―Privacy Preserving Data Mining,‖ Proc. Crypto, pp. 175-186, 2005.

[6] H. Grosskreutz, B. Lemmen, and S. Ruping, ―Secure Distributed Subgroup Discovery in Horizontally Partitioned Data,‖ Trans. Data Privacy, vol. 4,no. 3,pp. 147-165,2011.

DOI 10.17148/IJARCCE.2016.511126

[7] R. Chen, K. Sivakumar, and H. Kargupta. Distributed web mining using bayesian networks from multiple data streams. In The 2001 IEEE International Conference on Data Mining. IEEE, Nov. 29 - Dec. 2 2001.

[8] D. W.-L. Cheung, V. Ng, A. W.-C. Fu, and Y. Fu. Efficient mining of association rules in distributed databases. Transactions on Knowledge and Data Engineering, 8(6):911–922, Dec. 1996.

[9] W. Du and M. J. Atallah. Secure multi-party computation problems and their applications: A review and open problems. In Proceedings of the 2001 New Security Paradigms Workshop, Cloudcroft, New Mexico, Sept. 11-13 2001.

[10] W. Du and M. J. Atallah. Secure multi-party computational geometry. In Proceedings of the Seventh International Workshop on Algorithms and Data Structures, Providence, Rhode Island, Aug. 8-10 2001.

[11] Ford Motor Corporation. Corporate citizenship report.http://www.ford.com/en/ourCompany/communityAndCulture/buildingRelationships/strategicIssues/firestoneTireRecall.htm, May 2001.

[12] O. Goldreich, S. Micali, and A. Wigderson. How to play any mental game - a completeness theorem for protocols with honest majority. In 19th ACM Symposium on the Theory of Computing, pages 218–229, 1987.

[13] U.Fayyad et al.”Advances in knowledge discovery and datamining”,AAAI press,1996.

[14] D. Beaver, S. Micali, and P. Rogaway. The round complexity of secure protocols. In STOC, pages 503–513, 1990.

BIOGRAPHY

Dr. K. Madhavi, working as a Professor in Computer Scince and Engineering Department, Gokaraju Rangaraju Instittute of Engineering and Technology. She has completed her B.E in 1997, M.Tech from JNTUA in 2003 and awarded Ph.D from JNTUA in 2013. She has 19 years of teaching experience. She has published several papers in reputed international journals and international conference. Her research interest include sofware engineering , Model Driven Engineering, Data Mining, and other areas.

DOI 10.17148/IJARCCE.2016.511126