INTERNATIONAL JOURNAL OF RESEARCH
In Advanced Engineering Technologies

Volume 2, Issue 3 OCT 2014

IMPLEMENTATION OF UNIFY ALGORITHM FOR MINING OF ASSOCIATION RULES IN PARTITIONED DATABASES

KATARU MANI[1], N. JAYA KRISHNA[2]

[1]Mtech Student, Department of CSE.

EMAIL: manicsit@gmail.com

[2]Assistant Professor, Department of CSE.

EMAIL: jaya1238@gmail.com

ABSTRACT:

Many algorithms have been proposed to provide privacy preserving in data mining. These protocols are based on two main approaches named as: the Randomization approach and the Cryptographic approach. Sometimes the mining data is spilt into various parties, which can share the data. For example, insurance companies share the data from medical hospitals. Privacy concerns may prevent the parties from directly sharing the data. Here this project addresses secure mining of association rules over horizontally partitioned data. The drawbacks in the existing work are inaccuracy, inefficiency and lacking of security In this project we propose a protocol for secure mining of association rules in horizontally partitioned databases. The current integral protocol is that of Kantarcioglu and Clifton prominent as K&C protocol. This protocol is predicated on an unsecured distributed version of the Apriori algorithm designated as Fast Distributed Mining (FDM) algorithm of Cheung et al. The key ingredients in our protocol are two novel secure multi-party algorithms one that computes the coalescence of private subsets that each of the interacting players hold and another that tests the whether an element held by one player is included in a subset held by another. This protocol offers better privacy with deference to the protocol. In integration it is simpler and is significantly more efficient in terms of communication cost, communications rounds and computational cost.

Index Terms: Security, Privacy, Data Mining, Frequent Item sets, Association Rules, multi-party

1. INTRODUCTION

While considering data, data may be distributed among the sundry systems. Most of the businesses share their information along with their personal information for getting equipollent benefits. Sharing of this type of personal information arise the privacy issue. We study here the quandary of secure mining of association rules in horizontally distributed databases. In that setting, there are several sites that hold homogeneous databases, i.e., databases that distribute the same schema but hold information on different entities. The main aim is to find all association rules with support at least s and confidence at least c, for some given minimum support size s and confidence level c, that hold in the integrated database, while minimizing the information disclosed about the

private databases held by those players. The information that we would relish to forefend in this context is not only individual transactions in the different databases, but withal more global information such as what association rules are fortified locally in each of those databases. That goal defines a quandary of secure multi-party computation. In such quandaries, there are M players that hold private inputs, x1, . . . , xM, and they optate to securely compute y = f(x1, . . . , xM) for some public function f. If there subsisted a trusted third party (TP), the players could submit to him their inputs and he would perform the function evaluation and send to them the result. In the absence of such a trusted third party (TP), it is needed to devise a protocol that the players can run on their own in order to arrive at the required output y. Such a protocol is considered impeccably secure if no player can learn from his view of the protocol more than what he would have learnt in the idealized setting where the computation is carried out by a trusted third party. Yao was the first to propose a generic solution for this quandary in the case of two players.

2. ASSOCIATION RULE MINING

Association rule mining discovers the frequent patterns among the item sets. It aims to extract fascinating associations, frequent patterns, and correlations among sets of items in the data repositories . For Example, In a Laptop store in India, 80% of the customers who are buying Laptop computers additionally buy Data card for internet and pen drive for data portability. The formal verbal expression of Association rule mining quandary was initially designated by Agrawal .

Let I = I1, I2, … , Im be a set of m different attributes, T be the transaction that comprises a set of items such that T □ I, D be a database with different transactions Ts. An association rule is an insinuation in the form of X □ Y, where X, Y □ I are sets of items termed item sets, and X □Y = □. X is named antecedent. Y is called consequent

The rule means X implies Y. The two significant basic measures of association rules are support(s) and confidence(c). Since the database is enormous in size, users concern about only the frequently bought items. The users can pre-define thresholds of support and confidence to drop the rules which are not so useful. The two thresholds are named minimal support and minimal confidence [20]. Support(s) is defined as the proportion of records that contain X □ Y to the overall records in the database. The amount for each item is augmented by one, whenever the item is crossed over in different transaction in database during the course of the scanning.

$$\text{Support (XY)} = \frac{\text{Support sum of XY}}{\text{Overall records in the database D}}$$

Confidence(c) is defined as the proportion of the number of transactions that contain X □ Y to the overall records that contain X, where, if the ratio outperforms the threshold of confidence, an association rule X □ Y can be generated.

$$Confidence(X/Y) = \frac{Support\ (XY)}{Support\ (X)}$$

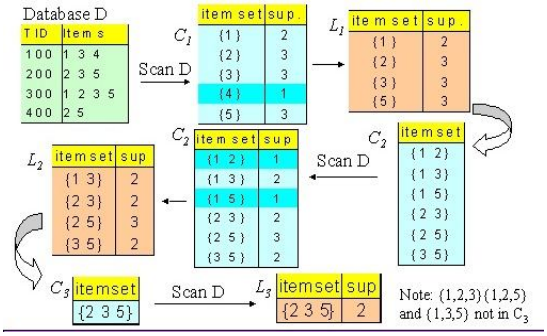
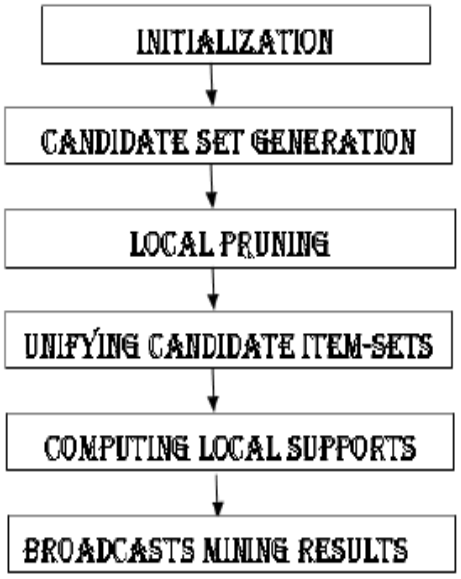


Fig1. Frequent Item sets

2.1 FDM Algorithm:

Fast Distributed Mining (FDM) algorithm is an unsecured distributed version of the 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. Hence, in order to find all globally -frequent item sets, each player reveals his locally

s-frequent item sets and then the players check each of them to see if they are s-frequent also globally. In the first iteration of FDM algorithm, when k=1, Cs 1,m the set that the m^{th} player computes (Steps 2-3) is just Fs 1,m , namely, the set of single items that are s-frequent in Dm. The complete FDM algorithm starts by finding all single items that are globally s-frequent. It then proceeds to find all 2-itemsets that are globally s-frequent, and so forth, until it finds the longest globally s-frequent item sets. If the length of such item sets is k, then in the $(k+1)^{th}$ iteration of the FDM it will find no (k+1)-item sets that are globally s-frequent, in that case it terminates.FDM algorithm steps are as follows:



.

3.IMPLEMENTATION OF THE PROPOSED MODEL:

A incipient model is proposed in this paper to find efficiently privacy preserving association rule mining in horizontally partitioned databases. The proposed model can be applied to any number of sites and for any number of transactions in the databases of the sites. Many tasks such as findings of locally frequent item sets, partial fortifies and total fortifies for each item set in the merged list are performed independently at different sites. Hence the computation time of the proposed model is less. The efficiency of the proposed method in terms of privacy and communication is discussed as follows

- Privacy is ascertained by utilizing encryption and decryption techniques at the time of transferring the frequent item sets from different sites to trusted party. From this, trusted party can ken only local frequent item sets of each site but he does not ken the fortifies of any item and cannot soothsay anything cognate to sites database.
- At the time of calculation of Partial Fortifies of an item set at each Sitei , MinSup * DBi is subtracted and the value of denotement * arbitrary number is integrated to the fortifies of the item at that site. So Partial Fortifies are in dissimulated form and broadcast to the sites securely. Each site is not having any conception about the denotement, desultory number which are assigned by trusted party to other sites and the database size of other sites is withal not kenned. So from the Partial Fortifies, no site can soothsay other sites data/information. In this way, partial fortifies of item sets can be broadcast to all other sites by preserving privacy of individual data. Hence, the denotement predicated secure sum concept which is utilized in the computation of partial fortifies enhances the privacy.
- Trusted party receives total partial support of each item set from all sites in order to find the global frequent item sets. By having these total fortifies, trusted party cannot find sites data/information since the database size of any site and local fortifies of any item at any site is not kenned by trusted party. Although trusted party assigned arbitrary numbers, signs to all sites and total database size is kenned, he cannot prognosticate any site's private data.
- Finally results that are global frequent item sets and their fortifies are broadcasted by trusted party to all sites. With these results, no site owner can soothsay local support of any global frequent item sets, as global frequent item sets may not be frequent in all sites and any site owner can not

prognosticate the contribution of other sites database which makes the item set globally frequent. In distributed environment, the cost of communication is quantified in terms of the number of communications for data transfer among all the sites and trusted party which are involved in the process of finding global association rules.

• The efficiency of an algorithm is assessed in terms of the communication costs incurred during information exchange. The proposed model minimizes the number of data transfers by sanctioning the transfer of bulk of data at a time from one site to another site and trusted party to sites. For example each site sends local frequent item sets of their database in a single data transfer to trusted party and even the sites sends its partial support for each item to other sites in a single transfer in lieu of sending one item set‘s partial support in one transfer to other sites. Hence the proposed model needs less communications.

• Trusted party additionally broadcast all the global frequent item sets for all sites in a single transfer. Hence the proposed model is more economy in terms of communication cost as it utilizes bulk data transfers. The above discussion pellucidly designates that the proposed model is efficient for finding global association rules by slaking privacy constraints.

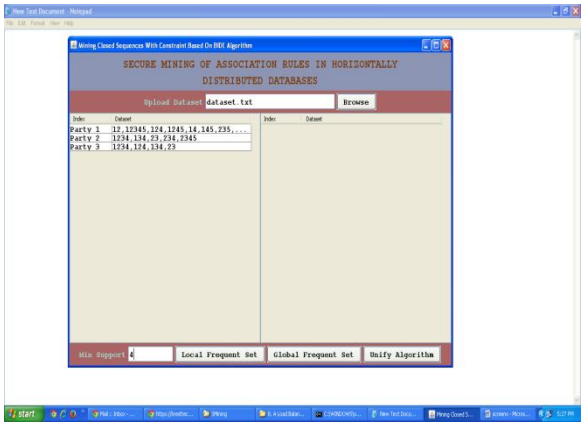


Fig 2. Frequent Item sets (local and global)

4. CONCLUSIONS AND FUTURE WORK

We proposed a protocol for secure mining of association rules in horizontally distributed databases that ameliorates significantly upon the current leading protocol in terms of privacy and efficiency. One of the main ingredients in our proposed protocol is a novel secure multi-party protocol for computing the coalescence (or intersection) of private subsets that each of the interacting players holds. Another ingredient is a protocol that tests the inclusion of an element held by one player in a subset held by another. Those protocols exploit the fact that the underlying quandary is of interest only when the number of players is more preponderant than two. One research quandary that this study suggests was described in Section 3; namely, to devise an efficient protocol for inequality verifications that utilizes the esse of a semi honest third party. Such a protocol might enable to further ameliorate upon the communication and computational costs. The second and third stages of the protocol of, as described. Other research quandaries that this study suggests is the implementation of the techniques presented here to the quandary of distributed association rule mining in the vertical setting , the quandary of mining generalized association rules , and the quandary of subgroup revelation in horizontally partitioned data.

REFERENCES

[1] J.Brickell and V. Shmatikov. Privacy-preserving graph algorithms in the semi-honest model. In ASIACRYPT, pages 236– 252, 2005.

[2] 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.

[3]. M. Kantarcioglu and C. Clifton. Privacy-preserving distributed mining of association rules on horizontally

partitioned data. IEEE Transactions on Knowledge and Data Engineering,, 16(9):10261037, September 2004.

[4]. X. Lin, C. Clifton, and M.Y. Zhu. Privacy-preserving clustering with distributed em mixture modeling. Knowl. Inf. Syst., 8:68–81, 2005.

[5]. Y. Lindell and B. Pinkas. Privacy preserving data mining. In CRYPTO, pages 36–54, 2000

[6] M. Kantarcioglu, R. Nix, and J. Vaidya,"An efficient approximate protocol for privacy-preserving association rule mining", In PAKDD, pages 515–524, 2009.

[7] M. Freedman, Y. Ishai, B. Pinkas, and O. Reingold. "Keyword search and oblivious pseudorandom functions", In TCC, pages 303–324, 2005.

[8].J.C. Benaloh. Secret sharing homomorphisms: Keeping shares of a secretsecret. In Crypto, pages 251–260, 1986.

[9]. J. Brickell and V. Shmatikov. Privacy-preserving graph algorithms inthe semi-honest model. In ASIACRYPT, pages 236–252, 2005.

[10] C. Clifton, M. Kantarcioglu, and J. Vaidya, "Defining privacy for data mining," in National Science Foundation Workshop on Next Generation Data Mining, H. Kargupta, A. Joshi, and K. Sivakumar, Eds., Baltimore, MD, Nov. 1-3 2002, pp. 126–133.

[11] B. A. Huberman, M. Franklin, and T. Hogg, "Enhancing privacy and trust in electronic communities," in Proceedings of the First ACM Conference on Electronic Commerce (EC99). Denver, Colorado, USA: ACM Press, Nov. 3–5 1999, pp. 78–86.

[12] Evfimievski, A., Srikant, R., Agrawal, R. and Gehrke, J. (2002) "Privacy preserving mining of association rules," Proceedings of the 8th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Alberta, Canada, July, pp.217-228