International Journal on Recent and Innovation Trends in Computing and Communication ISSN: 2321-8169
Volume: 2 Issue: 12 4091 - 4094

Literature Review on Secure Mining of Association Rules in Horizontally Distributed Databases

Ms. Priyanka B. Korde.
Department of Computer Engineering
KJ College Of Engineering & Management Research
Pune, India
kordepriyanka311@gmail.com

Prof. Mr. N R. Bogiri.
Department Of Computer Engineering
KJ College Of Engineering & Management Research
Pune, India
mail2nagaraju@gmail.com

***Abstract*—** Data and knowledge Engineering is one of the area under data mining. Which can extract important knowledge from large database, but sometimes these database are divided among various parties. This paper addresses a fast distributed mining of association rules over horizontally distributed data. This paper presents different methods for secure mining of association rules in horizontally distributed databases. The main aim of this paper is protocol for secure mining of association rules in horizontally distributed databases. The current main protocol is that of Kantarcioglu and Clifton. This protocol, like theirs, is based on the Fast Distributed Mining (FDM) algorithm of Cheung et al., which is an unsecured distributed version of the Apriori algorithm. The main components in this protocol are two novel secure multi-party algorithms — one that computes the union of private subsets that each of the interacting players hold, and another that tests the inclusion of an element held by one player in a subset held by another. This protocol offers improved privacy with respect to the protocol in. In addition, it is simpler and is significantly more efficient in terms of communication rounds, communication cost and computational cost.

***Keywords*—** *Privacy Preserving Data Mining; Distributed Computation; Frequent Item sets; Association Rules.*

*****

I. INTRODUCTION

Data mining methodology has emerged as a means of identifying patterns and trends from large quantities of data. Data mining go hand in hand: most tools operate by gathering all data into a central site, then running an algorithm against that data. This paper addresses the problem of computing association rules within such a scenario.

Here there is problem of secure mining of association rules in distributed databases. In that there are various sites that holds the homogeneous databases where databases contain same schema holds the information of different entities.

Where the inputs are partial databases and the output is the list of association rules that hold in unified database with exceeding level of support and confidence. The aim is to find out association rules with predefined level of support and confidence and also protect the content of the information which is not only local but also more global. So that to overcome the problem of security another protocol is proposed for secure computation of union of private subsets. The proposed protocol improves upon that in terms of simplicity and efficiency as well as privacy. In particular, our protocol does not depend on commutative encryption and oblivious transfer(what simplifies it significantly and contributes towards much reduced communication and computational costs). While our solution is still not perfectly secure, it leaks excess information only to a small number (three) of possible coalitions, unlike the protocol of that discloses information also to some single players. In addition, we claim that the excess information that our protocol may leak is less sensitive than the excess information leaked by the protocol.

II. LITERATURE SURVEY

We have to exploration of the Data Mining Literature Survey: Data mining derive its name from the similarities between searching for valuable business information in a large database —for example, finding linked products in terabytes of store scanner data —and mining a mountain for a vein of valuable ore. Both processes require either sifting through an immense amount of material, or intelligently probing it to find exactly where the value resides. Given databases of sufficient size and quality, data mining technology can generate new business opportunities by providing these capabilities:

Automatic prediction of trends and behaviors:

Data mining automates the process of finding predictive information in large databases. Questions that traditionally required extensive hands-on analysis can now be answered directly from the data quickly [2]. A typical example of a predictive problem is targeted marketing. Data mining uses data on past promotional mailings to identify the targets most likely to maximize return on investment in future mailings[7]. Other predictive problems include forecasting impoverishment and other forms of default, and identifying segments of a population likely to respond similarly to given events.

Automatic discovery of previously unknown patterns:

Data mining tools sweep through databases and identify previously hidden patterns in one step. An example of pattern discovery is the analysis of retail sales data to identify seemingly unrelated products that are often purchased together [4]. Other pattern discovery problems include detecting fraudulent credit card transactions and identifying anomalous data that could represent data entry keying errors.

SQL extensions are defining aggregate functions for association rule mining. Their optimizations have the purpose of avoiding the joins to convey unit (cell) formulas, but are not optimized to perform partial Transposition for each group of result rows [3]. Conor Cunningalam proposed an optimization

and Execution strategies in an RDBMS which uses two operators i.e ., PIVOT operator on tabular data that exchange rows and columns enabled data transformations are useful in data modeling, data analysis, and data presentation. They can quite easily be implemented inside a query processor system, much like select, project, and join operator. Such design provides the opportunities for better performance, both during query optimization and query execution. Pivot is an extension of Group By with an unique restrictions and optimization opportunities, and this makes it is very simple to introduce incrementally on top of existing grouping implementations. H Wang.C.Zaniolo proposed a s mall but Complete SQL Extension for Data Streams Data Mining and. This technique is a powerful database language and system that enables users develop complete data-intensive applications in SQL by writing new aggregates and table functions in SQL, rather than in procedural languages as in current Object -Relational systems .

III. RELATED WORK

Association Rule mining is one of the most important data mining tools used in many real life applications. It is used to reveal unexpected relationships in the data. In this paper, we will discuss the problem of computing association rules within a horizontally partitioned database. We assume homogeneous databases. All sites have the same schema, but each site has information on different entities. The goal is to produce associate ion rules that hold globally, while limiting the information shared about each site to preserve the privacy of data in each site.

Association Rule Mining:

Association rule mining is used to find interesting associations and/or correlation relationships among large sets of data items [1]. Association rules show attributes value conditions that occur frequently together in a given dataset.

Apriori Algorithm :

The Apriori Algorithm proposed to finds frequent items in a large amount of database[14][20]. Apriori is an influential algorithm in market basket analysis for mining frequent item sets for Boolean association rules. The name of Apriori is based on the fact that the algorithm uses a prior knowledge of frequent itemset properties. Apriori employs an iterative approach known as a level wise search, where k item sets are used to explore (k+1) itemsets .Apriori algorithm is an in fluential algorithm for mining frequent itemsets for Boolean association rules. This algorithm contains a number of passes over the database. During pass k, the algorithm finds the set of frequent itemsets Lk of length k that satisfy the minimum support requirement. Apriori is designed to operate on databases containing transactions. The purpose of the Apriori Algorithm is to find associations between different sets of data. It is sometimes referred to as "Market Basket Analysis". Each set of data has a number of items and is called a transaction [16]. The output of Apriori is sets of frequent itemsets tell us how often items are contained in sets of data. Authentication is provided by data integrity, the auditor erases the local data.

IV. METHODOLOGY

a. Process Design:

Let consider D be a transaction database.
The database is partitioned horizontally between P1, P2 . . . , Pm players, denoted 1 M. Player Pm holds the partial database Dm that contains Nm = |Dm | of the transactions in D, $1 \leq m \leq M$. The unified database is D =D1 U⋯U DM. An itemset X is a subset of A. Its global support, supp(X), is the number of transactions in D that contain it. Its local support, sup (X), is the number of transactions in Dm that contain it.

Support

The rule $X \Rightarrow Y$ holds with support s if s% of transactions in D contain $X \cup Y$. Rules that have a s greater than a user-specified support is said to have minimum support or threshold support. The support of rule is defined as,

sup(X) = no of transactions that contain X / total no of Transactions.

Confidence

The rule $X \Rightarrow Y$ holds with confidence c if c% of the transactions in D that contain X also contain Y. Rules that have a c greater than a user-specified confidence is said to have minimum confidence or threshold Confidence. The confidence of a rule is defined as,

conf(X =>Y) = sup(X U Y)/ supp(X).

FDM Algorithm:

Fast Distributed Mining (FDM) algorithm is an unsecured distributed version of the Apriori algorithm. Its main idea is that any s-frequent itemset must be also locally s-frequent in at least one of the sites. Hence, in order to find all globally -frequent itemsets, each player reveals his locally s-frequent itemsets and then the players check each of them to see if they are s-frequent also globally[20].
In the first iteration of FDM algorithm, when k=1, Cs1,m the set that the mth player computes (Steps 2-3) is just Fs1,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 itemsets. If the length of such itemsets is k, then in the (k+1)th iteration of the FDM it will find no (k+1)-itemsets that are globally s-frequent, in that case it terminates.

FDM algorithm steps are as follows:

1) Initialization
2) Candidate Sets Generation

3) Local Pruning
4) Unifying the candidate itemsets
5) Computing local supports
6) Broadcast Mining Results

Unifi-KC(FDM-KC):

The input that each player Pm has at the beginning of Protocol UNIFI-KC (Unifying lists of locally Frequent Itemsets — Kantarcioglu and Clifton) [14]. is the collection $Cs^{k,m}$, as defined in Steps 2-3 of the FDM algorithm. Let $Ap(F^{k-1})$ denotes the set of all candidate k-itemsets that the Apriori algorithm generates from F_s^{k-1} s . The output of the protocol is the union $C_s^k = \bigcup_{m=1}^{M} C_s^{k,m}$ .In the first iteration of this computation, and the players compute all -frequent 1- itemsets (here F_S^0 = s {∅ }). In the next iteration they compute all s-frequent 2-itemsets, and so forth, until the first ≤ in which they find no s-frequent k-itemsets. After computing that union, the players proceed to extract from C_s^k the subset F_s^k that consists of all k-itemsets that are globally s-frequent; Finally, by applying the above described procedure from k=1 until the first value of k≤L for which the resulting set F_s^k is empty, the Players may recover the full set of $F_s := \bigcup_{k=1}^{L} F_s^k$ all globally –frequent item sets. Protocol UNIFI-KC works as follows: First, each player adds to his private subset $C_s^{k,m}$ 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. At the end of that stage, every item set in each subset is encrypted by all of the players; the usage of a commutative encryption scheme ensures that all item sets are, eventually, encrypted in the same manner. Then, they compute the union of those subsets in their encrypted form. Finally, they decrypt the union set and remove from it item sets which are identified as fake.
Steps For secure computations of all item sets (by K&C):

1. Cryptographic Primitive Selection

- Player selects needed commutative cipher and corresponding private key
- Player selects hash function to apply on all itemsets prior to encryption
- Player compute lookup table with hash values to find preimage of given hash values.

2. All itemsets Encryption

3. Itemset Merging
- Each odd player sends his encrypted set to player 1.
- Each even player sends his encrypted set to player 2.
- Player 1 unifies all sets that were sent by the odd players and removes duplicates.
- Player 2 unifies all sets that were sent by the even players and removes duplicates.
- 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 then remove duplicates from the unified list. Denote the final list by EC_S^K.

4. Decryption

V. CONCLUSION

In this we present various techniques for secure mining of association rules in horizontally partitioned distributed databases. In this, presented protocol is more efficient than current leading K and C protocol. The main components of this protocol are two novel secure multiparty algorithms in which these two main steps are union and intersection. The protocol utilizes the fact that the underlying problem is of interest only if the number of player is more than two. The direction to future work is to devise an efficient protocol for inequality verifications that uses the existence of semi-honest third party and another in Implementation of the techniques to the problem of distributed association rule mining in vertical setting.

ACKNOWLEDGMENT

The authors would like to thanks the Dept of Computer Engineering ,KJ College Of Engineering and Management Research Pune, Maharashtra, India, And cooperation of all teaching staff.

REFERENCES

[1] Tamir tassa, “Secure Mining of Association Rules in Horizontally Distributed Databases”, IEEE transactions on knowledge and data engineering, 2013.
[2] 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.
[3] M.Kantarcioglu and C. Clifton., “Privacy-preserving distributed mining of association rules on horizontally partitioned data”, *IEEE Transactions on Knowledge and Data Engineering*, 16:1026–1037, 2004.
[4] R.Agrawal and R. Srikant.,“Privacy-preserving data mining”, *SIGMOD Conference*, pages 439–450, 2000.
[5] A.V. Evfimievski, R. Srikant, R. Agrawal, and J. Gehrke, “Privacy preserving mining of association rules”, In *KDD*, pages 217–228, 2002.
[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] Chang, C.-C. and Lin, C.-Y. (2005) “Perfect hashing schemes for mining association Rules,” *The Computer Journal*, Vol. 48, No. 2, pp.168-179.
[9] Chen, M.-S., Han, J. and Yu, P.S. (1996) “Data mining: an overview from a database perspective,” *IEEE Transactions on Knowledge Data Engineering*, Vol. 8. No. 6, pp.866-883.
[10] Cheung, D.W.-L., Han, J., Ng, V.T.Y., Fu, A.W.-C. and Fu, Y. (1996) “A fast distributed algorithm for mining association rules,” *Proceedings of the 1996 International*

Conference on Parallel and Distributed Information Systems, Miami Beach, Florida, December, pp.21-42.

[11] Dwork, C. and Nissim, K. (2004) "Privacy-preserving data mining on vertically partitioned databases," in Franklin, M.K. (Ed.): *Lecture Notes in Computer Science*, Vol. 3152, Springer-Verlag, pp.528-544.

[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.

[13] Grahne, G. and Zhu, J. (2005) "Fast algorithms for frequent itemset mining using FP-trees," *IEEE Transactions on Knowledge and Data Engineering*, Vol. 17, No. 10, pp.1347-1362.

[14] Han, J., Pei, J., Yin, Y. and Mao, R. (2004) "Mining frequent pattern without candidate generation: a frequent pattern tree approach," *Data Mining and Knowledge Discovery*, Vol. 8, No. 1, pp.53-87.

[15] Kantarcioglu, M. and Clifton, C. (2004) "Privacypreserving distributed mining of association rules on horizontally partition data," *IEEE Transactions on Knowledge and Data Engineering*, Vol. 16, No. 9, pp.1026- 1037.

[16] Kantardzic, M. (2002) "Data mining: concepts, models, methods, and algorithms," *John Wiley & Sons, Inc.*, New York.

[17] Li, Y.-C. and Chang, C.-C. (2004) "A new FP-tree algorithm for mining frequent itemsets," in Chi, C.-H. and Lam, K.-Y. (Eds.): *Lecture Notes in Computer Science*, Vol. 3309, Springer-Verlag, pp.266-277.

[18] Lindell, Y. and Pinkas, B. (2002) "Privacy preserving data mining," *Journal of Cryptology*, Vol. 15, No. 3, pp.177-206.

[19] Oliveira, S.R.M. and Zaïane, O.R. (2002) "Privacy preserving frequent itemset mining," *Proceedings of the* IJCSNS International Journal of Computer Science and Network Security, VOL.6 No.11, November 2006 266 *2002 IEEE ICDM Workshop on Privacy, Security and Data Mining*, Maebashi City, Japan, December, pp.43-54.

[20] Oliveira, S.R.M. and Zaïane, O.R. (2003) "Algorithms for balancing privacy and knowledge discovery in association rule mining," *Proceedings of 7th International Database Engineering and Applications Symposium*, Hong Kong, China, July, pp.54-63.

[21] Pohlig, S.C. and Hellman, M.E. (1978) "An improved algorithm for computing logarithms over GF(P) and its cryptographic significance," *IEEE Transaction on Information Theory*, Vol. 24, No. 1, pp.106-110.