A COMPARATIVE STUDY ON PRIVACY OF ASSOCIATION RULE MINING AMONG HYBRID PARTITIONED DATABASES

GONUGUNTLA SUJATHA,
M.Tech (CSE), Priyadarshini Institute of Technology & Management

ISMAIL KHAN,
Assistant Professor (Dept. of CSE),
Priyadarshini Institute of Technology & Management

ABSTRACT:

Data mining is the drawing out of obscure examples from walloping information. Association rule mining distinguishes the connections among the attributes of the information. Mining from taken information bases is tough because the data ought to not be unveiled among the databases. The taken databases is partitioned off horizontally and vertically or hybrid that is combination of horizontal and vertical .This paper augment the association rule mining in vertically distributed databases. The protocol for this association mining is taking under consideration the Fast Distributed Mining (FDM) algorithmic program. For satisfying the privacy constraints in vertically partitioned off databases, algorithmic program supported cryptography techniques, Homomorphic coding, Shamir's secret sharing technique area unit used For horizontal partitioned databases that uses Paillier cryptosystem to process global supports are used. This paper contemplates the in depth ways used for mining association rules over distributed where as maintain privacy.

Keywords: *Fast Distributed Mining algorithm, Paillier cryptosystem, Association rules*

1. INTRODUCTION

Dealing with massive databases is one in all the key challenges in data processing analysis and development. Some databases are just too massive (e.g., with terabytes of data) to be processed at just once. Efficiency and space reasons, partitioning them into subsets for process is critical. However, since the amount of itemsets in every partitioned off information set is a combinatorial amount and every of them could also be as original itemsets, In data mining, a strategy for assessing the quality of model generalization is to partition the data source. A portion of the data, called the training data set, is used for preliminary model fitting. The rest is reserved for empirical validation and is often split into two parts: validation data and test data. The validation data set is used to prevent a modeling node from over fitting the training data and to compare models. The test data set is used for a final assessment of the model. Data mining results from these subsets can be very large in size. Therefore, the key to data partitioning is how to aggregate the results from these subsets. It is not realistic to keep all results from each subset, because the rules from one subset need to be verified for usefulness in other subsets. Association analysis in large databases has received much attention recently The problem of mining association rules is to generate all rules A → B that have both support and confidence greater than or equal to some user specified thresholds, called minimum support and minimum confidence, respectively[5]. To implement association analysis, a wide range of problems have been investigated over such diverse topics as models for discovering generalized associated rules (Srikant & Agrawal, 1997), efficient algorithms for

computing the support and confidence of an association rule measurements of interestingness mining negative association rules (Brin et al., 1997), and computing large itemsets online (Hidber, 1999). The main limitation of these approaches, however, is that they require multiple passes over the database. For a very large database that is typically disk resident, this requires reading the database completely for each pass resulting in a large number of disk I/Os. Consequently, the larger the size of a given database, the greater the number of disk I/Os. This means that existing models cannot work well when resources are bounded. Therefore, faster mining models have to be explored.

Horizontal and vertical partitioning are necessary important aspects of physical info framing that have the important impact on performance and traceableness. Horizontal partitioning permits access ways that like tables, indexes and materialized views to be partitioned into disjoint sets of rows that unit physically keep and accessed singly. Here 2 common sorts of horizontal partitioning are vary and hash partitioning. On the other hand, vertical partitioning permits a table to be partitioned into disjoint sets of columns. Like indexes and materialized views, every form of partitioning can significantly impact the performance of the use i.e., queries and updates that execute against the knowledge system, by reducing the worth of accessing and method info.

2. LITERATURE REVIEW

In [1], the author has presented a system for secure mining of association rules in horizontally distributed databases using Fast Distributed Mining (FDM) Algorithm and Secure Multiparty Algorithm. The Protocol facilitates enhanced privacy with respect to the protocol in [4]. In addition to that, it is simpler and is significantly more efficient in terms of communication cost, computational cost, and communication rounds.

In [2], the authors have presented a protocol for discovering association rules between items in Large Databases. Experiment results have shown that Apriori Hybrid Algorithm is faster than AIS [7] and SETM [8] Algorithm [2].

In [3], the authors have offered FairplayMP, a generic system for Secure Multiparty Computation which is an extension of the Fairplay system that supported secure computation by two parties.

In [4], the authors have presented a protocol for Privacy-Preserving Distributed Mining of Association Rules on Horizontally Partitioned Data, which addresses the problem of computing association rules where the data may be distributed among various custodians, none of which are permitted to transfer their data to another site. Based on Fast Distributed Algorithm (FDM) and Secure Multiparty computation, Association Rules have been computed [4].

In [5], the authors have proposed a system for privacy preserving Association Rule Mining in Vertically Partitioned Data. The system is demonstrated through a two-party algorithm for efficiently discovering frequent item sets with minimum support levels, without either site transmitting individual transaction values.

In [6], the authors have done a comparative study for different vertical partitioning

algorithm and shown how graph-based vertical partitioning algorithm has contributed towards the optimization of data fragmentation problem by providing an efficient way of improving performance of applications.

Fast Distributed Mining Algorithm: The protocol of our current algorithm is based on the Fast Distributed Mining (FDM) algorithm of author Cheung et al., which is an unsecured distributed version of Apriori. Hence, in order to find all globally s-frequent itemsets, each database reveals its locally s-frequent itemsets and then the databases check each of them to see if they are s-frequent also globally.

3. ASSOCIATION RULE MINING for DATA BASES

Association rule mining is one amongst the numerous data processing techniques that predicts associations between things or itemsets from large database. Database could incorporates highlevel of transactions that are extracted from single source of knowledge or from several sources. Counting on the necessities of applications, information is maintained at single location known as centralized information or the information is also distributed at multiple sites known as distributed information. In centralized surroundings, information is accessible in single location and therefore the multiple user's are allowed to access the information. The main aim of privacy conserving association rule mining during this scenario is to perform the mining method by concealment sensitive data/information from users. In distributed environment, the database is available across multiple sites and the distributed applications, databases are partitioned basically in two ways such as horizontally and vertically partitioned database, where each partitioned database is placed in one site or many sites. The site which owns the database has local autonomy over its database and no site can have access to any data/information belongs to any other site. Depending on the hierarchy of the distributed application, any site's partitioned database can be further partitioned into two or more and each partitioning may follow horizontal or vertical and this process of partitioning is called mixed/hybrid. In some distributed applications databases are partitioned into disjoint segments so every database is placed in a single location/site only.

Association rule concealment methodologies: aim at sanitizing the first information so as to attain the subsequent goals[14]:

1. No rule that's thought of as sensitive from the owner's perspective and might be deep-mined from the first information at pre-specified thresholds of confidence and support, are often conjointly disclosed from the modify information, once this information is deep-mined at constant or at higher thresholds

2. All the non sensitive rules that seem once mining the first information at prespecified thresholds of confidence and support are often with success deep-mined from the modify information at constant thresholds or higher.

3. No rule that wasn't resulting from the first information once the information was deep-mined at pre-specified entry of

confidence and support are often derived from its disinfected counterpart once it's deep-mined at constant or at higher thresholds. the primary goal needs that each one the sensitive rules disappear from the modify information, once the information is deep-mined underneath constant or higher levels of support and confidence because the original information. The second goal states that there ought to be no lost rules within the modify database[12]. That is, all the non sensitive rules that were deep-mined from the first information ought to even be deep-mined from its modify counterpart at constant or higher levels of confidence and support. The third goal states that no false rules conjointly referred to as ghost rules ought to be made once the modify information is deep-mined at constant or higher levels of confidence and support. A false (ghost) rule is Associate in Nursing association rule that wasn't among the principles deep-mined from the first information. an answer that addresses of these 3 goals is named actual. actual concealment solutions that cause the smallest amount doable modification to the first information ar known as ideal or best. Non-exact however possible solutions ar known as approximate. The privacy conserving association rule mining algorithms ought to

1. Stop the invention of sensitive information.

2. Not compromise the access and therefore the use of non sensitive knowledge.

3. be usable on giant amounts of Data.

4. Not have Associate in attention exponential procedure complexness.

4. ILLUSTRATION OF VERTICAL DATABASES:

To demonstrate the problem of data mining of frequent occurring patterns, consider the vertical databases shown in the below tables. The sample shows shopping carts of customers in 2 different databases with different entities. Given s=1/3 and N= 7, Min-sup=1/3 * 7= 2. Each database finds support count of each item and mining proceeds with the items satisfying min-support. Item 'c' is not taken into consideration as it doesn't satisfy min-sup. An authenticated third party does the join of items in the databases to form unified databases without disclosing their local frequent items

Item Key	A	b
10	1	0
11	1	1
12	0	0
13	1	1
14	0	0
15	1	1
16	0	1

Item Key	A	b	c
10	1	0	0
11	1	1	1
12	0	0	0
13	1	1	1
14	0	0	0
15	1	1	1
16	0	1	0

Table1 & 2 . Vertical Frequent Itemset

4.1 Data Representation

In real time the transactional database can be viewed as a two-dimensional matrix: rows represent individual transactions and the columns represent items. Data can be represented either in a horizontal view or a vertical view: Horizontal view this view represents each row with a unique transaction identifier and a bitmap to represent the items involved in the transaction. Considering 8 items involved in a transaction, and then the bit-string 1100010 means items 1, 2, and 7 are involved. Vertical view this view employs assigning a unique identifier to each column and a bitmap that represents the transactions in which that particular item is involved. For example, if there are 8 transactions that involve a particular item, then the bit string 11100000 means that transactions 1, 2, and 3 are involved. Lin suggests that a vertical representation is the best choice for mining frequent itemsets. Vertical representation is beneficial as it allows operating only on those item sets that are frequent. Hence bitmap representation can be discarded for those itemsets that are not frequent, which leads to a reduced memory storage.

5. THE FDM ALGORITHM

The FDM (Fast Distributed Algorithm for Data Mining) algorithm, proposed in (Cheung et al. 1996) has the following distinguishing characteristiques:

1. Candidate set generation is Apriori-like. However, some interesting properties of locally and globally frequent itemsets are used to generate a reduced set of candidates at each iteration, this resulting in a reduction in the number of messages interchanged between sites.

2. After the candidate sets were generated, two types of reduction techniques are applied, namely a local reduction and a global reduction, to eliminate some candidate sets from each site.

3. To be able to determine if a candidate set is frequent, the algorithm needs only O(n) messages for the exchange of support counts, where n is the number of sites from the distributed system. This number is much less than a disrect adaptation of Apriori, which would need O(n2) messages for calculating the support counts.

6.CONCLUSSION:

In this paper we focused on enhanced privacy among the association rule mining in vertically distributed databases. The protocol for this association mining is taking into account the Fast Distributed Mining (FDM) algorithm. For satisfying the privacy constraints in vertically and horizontally partitioned databases, uses Paillier cryptosystem to compute global supports in order to protect the privacy and also contemplates the extensive methods used for mining association rules over distributed while maintain privacy.

REFERENCES:

[1].Tamir Tassa, "Secure Mining of Association Rules in Horizontally Distributed Databases" IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, VOL. 26, NO. 4, APRIL 2014

[2].R. Agrawal and R. Srikant, "Fast Algorithms for Mining Association Rules in Large Databases," Proc 20th Int'l Conference. Very Large Data Bases (VLDB), pp. 487-499, 1994.
[3].A. Ben-David, N. Nisan and B. Pinkas, "FairplayMP - A System for Secure Multi-Party Computation," Proc. 15th ACM Conference. Computer and Comm. Security (CCS), pp. 257-266, 2008.
[4].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,September 2004
[5].J. Vaidya and C. Clifton, "Privacy Preserving Association Rule Mining in Vertically Partitioned Data," Proc. Eighth ACM SIGKDD Int'l Conference Knowledge Discovery and Data Mining (KDD), pp. 639-644, 2002.
[6].Vertical Partitioning Impact on Performance and Manageability of Distributed Database systems (A Comparative study of some vertical partitioning algorithms) (2006) by Hassan I. Abdalla, F. Marir 18th National computer conference 2006.
[7].R. Agrawal, T. Imielinski, and A. Swami. Mining Association Rules between Sets of Items in Large Databases. In Proc. of the ACM SIGMOD Conference on Management of Data, Washington, D.C., May 1993
[8].M. Houtsma and A. Swami. Set-Oriented Mining of Association Rules. Research Report RJ 9567, IBM Almaden Research Center, San Jose, California, October 1993.
[9].D.Kerana Hanirex, Dr.K.P.Kaliyamurthie," Mining Frequent Item sets Using Genetic Algorithm", Middle-East Journal of Scientific Research, 19 (6): 807-810, 2014.
[10].Kerana Hanirex.D, Dr.K.P.Kaliyamurthie, "Finding the Dominating Amino Acids in Dengue Virus Type1 Study on mining frequent item sets", 4(3): (B);880 – 89, Int. Journal of Pharma and Bio Sciences, July, 2013.
[11].D.Kerana Hanirex ,"Association Rule Mining in Distributed Database System", International Journal of Computer Science and Mobile Computing(IJCSMC), Vol3,Iss 4,pg 727-732,2014.
[12] Assaf Schuster, Ran Wolff, Bobi Gilburd," Privacy-Preserving Association Rule Mining in LargeScale Distributed Systems", fourth IEEE symposium on Cluster Computing and Grid, 2004.
[13] Tirumala prasad B, Dr. MHM Krishna Prasad, "Distributed Count Association Rule Mining Algorithm", International Journal of Computer Trends and Technology, July to Aug Issue 2011, pp.280-284.
[14] Gkoulalas-Divanis, Aris, Verykios, Vassilios S. "Association Rule Hiding for Data Mining", Springer Series: Advances in Database Systems, Vol. 41, 1st Edition., 2010, p.13