ISSN:2320-0790

An International Journal of Advanced Computer Technology

An Overview of Secure Mining of Association Rules in Horizontally Distributed Databases

Ms. Sonal Patil[1], **Mr. Harshad S. Patil**[2]

[1]Asst.Prof, GHRIEM, Jalgaon
[2]GHRIEM, Jalgaon

Abstract: In this paper, propose a protocol for secure mining of association rules in horizontally distributed databases. Now a day the current leading protocol is Kantarcioglu and Clifton. This protocol is based on the Fast Distributed Mining (FDM) algorithm which is an unsecured distributed version of the Apriori algorithm. The main ingredients in this protocol are two novel secure multi-party algorithms 1. That computes the union of private subsets that each of the interacting players hold, and 2. Tests the inclusion of an element held by one player in a subset held by another. In this protocol offers enhanced privacy with respect to the other one. Differences in this protocol, it is simpler and is significantly more efficient in terms of communication rounds, communication cost and computational cost [1].

Keywords: Privacy Preserving Data Mining; Distributed Computation; Frequent Item sets; Association Rules

I. INTRODUCTION

We are study here the problem of secure mining of association rules in horizontally partitioned databases. In that there are several places, several parties and several player that hold homogeneous databases, i.e., databases that share the same schema but hold information on different entities. The goal is to minimizing the information disclosed about the private databases held by those players. The information that we would like to protect in this context is not only individual transactions in the different databases, but also more global information such as what association rules are supported locally in each of those databases. In our problem, the inputs are the partial databases, and the required output is the list of association rules that hold in the unified database with support and confidence no smaller than the given thresholds *s* and *c*, respectively. As the above mentioned generic solutions rely upon a description of the function *f* as a Boolean circuit, they can be applied only to small inputs and functions which are realizable by simple circuits. In more complex settings, such as ours, other methods are required for carrying out this computation. In such cases, some relaxations of the notion of perfect security might be inevitable when looking for practical protocols, provided that the excess information is deemed benign (see examples of such protocols in e.g. [3], [4], [5], [6], [7]).

II. DATA MINING

Data mining is the process of extracting hidden patterns from data. As more data is gathered, with the amount of data doubling every three years, data mining is becoming an increasingly important tool to transform this data into knowledge. It is commonly used in a wide range of applications, such as marketing, fraud detection and scientific discovery. Data mining can be applied to data sets of any size, and while it can be used to uncover hidden patterns, it cannot uncover patterns which are not already present in the data set. Data mining extracts novel and useful knowledge from data and has become an effective analysis and decision means in corporation. Data sharing can bring a lot of advantages for research and business collaboration. However, large repositories of data contain private data and sensitive rules that must be preserved before published. Motivated by the multiple conflicting requirements of data sharing, privacy preserving and knowledge discovery, privacy preserving data mining (PPDM) has become a research hotspot in data mining and database security fields. Two problems are addressed in PPDM: one is the protection of private data; another is the protection of sensitive rules (knowledge) contained in the data.

The former settles how to get normal mining results when private data cannot be accessed accurately; the latter settles how to protect sensitive rules contained in the data from being discovered, while non-sensitive rules can still be mined normally. The latter problem is called knowledge hiding in database in (KHD) which is opposite to knowledge discovery in database (KDD). And association rule hiding problem we focus is one of problems in KHD[9].

III. THE FAST DISTRIBUTED MINING ALGORITHM

The protocol of [3], as well as ours, are based on the Fast Distributed Mining (FDM) algorithm of Cheung et al.[8], which 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 *s*-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.

The FDM algorithm proceeds as follows:

(1) Initialization:
(2) Candidate sets generation
(3) Local Pruning
(4) Unifying the candidate item sets
(5) Computing local supports
(6) Broadcast Mining Results

With the existence of many large transaction databases, the huge amounts of data, the high scal-ability of distributed systems, and the easy partition and distribution of a centralized database, it is im-portant to inuestzgate eficient methods for distributed mining of association rules. This study discloses some interesting relationships between locally large and glob-ally large itemsets and proposes an interesting distributed association rule mining algorithm, FDM (Fast Distributed Mining of association rules), which generates a small number of candidate sets and substantially reduces the number of messages to be passed at mining association rules. Our performance study shows that FDM has a superior performance over the direct application of a typical sequential algorithm. Further performance enhancement leads to a few variations of the algorithm.

IV. DISTRIBUTED DATABASE

A distributed database is database in which storage devices are not all attached to a common processing unit such as the CPU, controlled by a distributed database management system (together sometimes called a distributed database system). It may be stored in multiple computers, located in the same physical location; or may be dispersed over a network of interconnected computers. Unlike parallel systems, in which the processors are tightly coupled and constitute a single database system, a distributed database system consists of loosely-coupled sites that share no physical components. System administrators can distribute collections of data (e.g. in a database) across multiple physical locations. A distributed database can reside on network servers on the Internet, on corporate intranets or extranets, or on other company networks. Because they store data across multiple computers, distributed databases can improve performance at end-user worksites by allowing transactions to be processed on many machines, instead of being limited to one [2].

Two processes ensure that the distributed databases remain up-to-date and current: replication and duplication.

1. Replication involves using specialized software that looks for changes in the distributive database. Once the changes have been identified, the replication process makes all the databases look the same. The replication process can be complex and time- consuming depending on the size and number of the distributed database. This process also requires lot of time and computer resources.

2. Duplication, on the other hand, has less complexity. It basically identifies one database as a master and then duplicates that database. The duplication process is normally done at a set time after hours. This is to ensure that each distributed location has the same data. In the duplication process, users may change only the master database. This ensures that local data will not be overwritten.

Both replication and duplication can keep the data current in all distributive locations [2].

V. ASSOCIATION RULE

In Data mining, association rule is a popular and well researched method for discovering interesting relations between variables in large databases. Piatetsky-shapiro describes analyzing & presenting strong rules discovered in databases using different measures of interestingness. Based on the concept of strong rules, Agrawal et al introduced association rules for discovering regularities between products in large scale transaction data recorded by point-of-sale (POS) systems in supermarkets For example, the rule Found in the sales data of a supermarket would indicate that if a customer buys onions and potatoes together, he or she is likely to also buy beef. Such information can be used as the basis for decisions about marketing activities such as, e.g., promotional pricing or product placements. In addition to the above example from market basket analysis association rules are employed today in many application areas including Web usage mining, intrusion detection and bioinformatics. [9]

VI. REFERENCES

[1]. Tamir Tassa, "Secure Mining of Association Rules in Horizontally Distributed Databases."201
[2]. Jump up to: a b O'Brien, J. & Marakas, G.M. (2008) Management Information Systems (pp. 35-39). New York, NY: McGraw-Hill Irwin.
[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]. T. Tassa and D. Cohen. Anonymization of centralized and distributed social networks by sequential clustering. IEEE Transactions on Knowledge and Data Engineering, 2012.
[5]. T. Tassa and D. Cohen. Anonymization of

centralized and distributed social networks by sequential clustering. IEEE Transactions on Knowledge and Data Engineering, 2012.

[6]. J. Vaidya and C. Clifton. Privacy preserving association rule mining in vertically partitioned data. In KDD, pages 639-644, 2002.

[7]. S. Zhong, Z. Yang, and R.N. Wright. Privacy-enhancing kanonymization of customer data. In PODS, pages 139-147, 2005.

[8]. David W. Cheungt Jiawei Hans Vincent T. Ngtt Ada W. Fuss Yongjian FuI. "A Fast Distributed Algorithm for Mining Association Rules".

[9]. Shikha Sharma, Prof. Pooja Jain, " A Novel Data Mining Approach for Information Hiding".