"An Analysis on Effective, Precise and Privacy Preserving Data Mining Association Rules with Partitioning On Distributed Databases"

Shoban Babu Sriramoju[1] Dr. Atul Kumar[2]

[1]Research Scholar, CMJ University, Shillong, Meghalaya

[2]Prof. CMJ University, Shillong, Meghalaya

Abstract – Data mining techniques are used to discover hidden information from large databases. Among many data mining techniques, association rule mining is receiving more attention to the researchers to find correlations between items or items sets efficiently. In distributed database environment, the way the data is distributed plays an important role in the problem definition. The data may be distributed horizontally or vertically or in hybrid mode among different sites. There is an increasing demand for computing global association rules for the databases belongs to different sites in a way that private data is not revealed and site owner knows the global findings and their individual data only. In this paper a model is proposed which adopts a sign based secure sum cryptography technique to find global association rules with trusted party by preserving the privacy of the individual's data when the data is distributed horizontally among different sites.

Mining distributed databases is emerging as a fundamental computational problem. A common approach for mining distributed databases is to move all of the data from each database to a central site and a single model is built. This approach is accurate, but too expensive in terms of time required. For this reason, several approaches were developed to efficiently mine distributed databases, but they still ignore a key issue privacy. Privacy is the right of individuals or organizations to keep their own information secret. Privacy concerns can prevent data movement data may be distributed among several custodians, none of which is allowed to transfer its data to another site.

----------♦----------

INTRODUCTION

The goal of recent advances in data mining techniques is to efficiently discover valuable and non-obvious knowledge from large databases. The mining of association rules plays an important role in various data mining fields, such as financial analysis, the retail industry and business decision-making.

Modern organizations have their own databases, located in different places. Most mining techniques assume that the data is centralized or the distributed amounts of data can efficiently move to a central site to become a single model.

However, organizations may be willing to share only their mining models, not their data. These centralized techniques have a high risk of unexpected information leaks when data is released. Organizations urgently require evaluation to decrease the risk of disclosing information. Privacy-Preserving Data Mining (PPDM) can run a data mining algorithm to obtain mutually beneficial global mining objectives without exposing private data. Therefore, PPDM has become an important issue in many data mining applications.

A simple method of PPDM in distributed databases is to perturb the original data. The procedure of transforming the original database into a new one that hides some sensitive association rules is called the *sanitization process*. Performing a mining process on the sanitized database can reduce the risk of revealing the sensitive information. However, the mining result on the sanitized database is less precise than that of the original database.

Data mining has been viewed as a threat to privacy because of the widespread proliferation of electronic data maintained by corporations. This has led to increased

concerns about the privacy of the underlying data. Data mining techniques find hidden information from large database while secret data is preserved safely when data is allowed to access by single person. Now a days many people want to access data or hidden information using data mining technique even they are not fully authorized to access. For getting mutual benefits, many organizations wish to share their data to many legitimate people but without revealing their secret data.

In large applications the whole data may be in single place called centralized or multiple sites called distributed database. Methodologies are proposed by many authors for both centralized as well as distributed database to protect private data. This paper deals with privacy preserving in distributed database environment while sharing discovered knowledge/hidden information to many legitimate people.

In distributed environment, database is a collection of multiple, logically interrelated databases distributed over a computer network and are distributed among number of sites. As the database is distributed, different users can access it without interfering with one another.

In distributed environment, database is partitioned into disjoint fragments and each site consists of only one fragment. Data can be partitioned in different ways such as horizontal, vertical and mixed. In horizontal partitioning of data, each fragment consists of a subset of the records of a relation R whereas vertical partitioning of data, each fragment consists of a subset of attributes of a relation R. The another partitioning method is mixed fragmentation where data is partitioned horizontally and then each partitioned fragment is further partitioned into vertical fragments and vice versa. Figure 1 shows a method for mixed partitioned in which data is first partitioned vertically and then horizontally. Figure 2 shows another mixed method in which data is partitioned horizontally and then vertically partitioned.

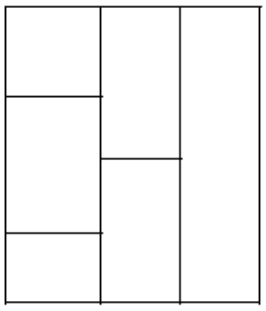

Figure1: Vertically partitioned database is further partitioned into horizontal

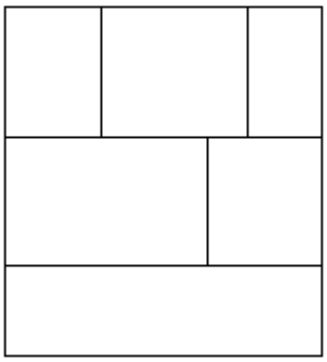

Figure 2: Horizontally partitioned database is further partitioned into vertical.

In data mining, association rule mining is a popular and well researched method for discovering interesting relations between variables in large databases. When data is distributed among different sites, finding the global association rules is a challenging task as the privacy of the individual site's data is to be preserved.

The process of preserving privacy in case of association rule mining can be termed as privacy preserving association rule mining. Database may consist of enormous amount of transactions which are extracted from a single source of data or from many sources. Depending on the requirements of applications, database is maintained at single location called centralized database or the database may be distributed at multiple sites called distributed database. The main aim of privacy preserving association rule mining in centralized database is, mining process can be done by hiding sensitive data/information from users other than database owner. In distributed environment, aim is finding the global mining results by preserving the individual sites private data/information from one another.

Global results are determined only when the necessary results/information is captured based on all sites' database individually like local frequent item sets and their support values of all sites are required to determine whether an item set is globally frequent or infrequent. As the individual database may possess some private data/information and in case of leakage of private data to anyone causes damage to database owners.

In distributed applications, databases are partitioned basically in two ways such as horizontal and vertical partitioned databases, 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. In this paper, privacy preserving association rule mining for two commonly used mixed partitioning (disjoint) methods in distributed database environment is considered.

We present an efficient approach for mining frequent itemsets in *horizontally-distributed* databases. Instead of sharing possibly privacy sensitive data to perform the distributed mining task, we chose to share just a small portion of each local model, which are used to construct the global model of frequent itemsets. This choice also makes our approach extremely efficient in terms of communication overhead, enabling it to be used for mining even geographically distributed databases. We also developed a communication mechanism to ensure privacy among the sites involved in the mining operation. We prove that the global model generated by our approach is totally accurate, we present bounds for the total amount of communication, and we demonstrate the performance of our approach through a set of experiments under a variety of conditions.

BACKGROUND

Association rule mining - The process of association rule mining includes two main sub-problems: the first is to discover all frequent itemsets; the second is to use these discovered frequent itemsets to generate association rules. Since each association rule can easily be derived from the corresponding frequent itemsets, the overall performance of the association rule mining is determined by the first sub-problem. Therefore, researchers usually focus on efficiently discovering frequent itemsets. Agrawal et al. presented the Apriori algorithm to efficiently identify frequent itemsets.

Apriori is a level-by-level algorithm including multiple passes. In each pass, Apriori generates a candidate set of frequent *k*-itemsets (frequent itemsets with length *k*). Each frequent k-itemset is combined from two arbitrary frequent (*k*-1)-itemsets, in which the first *k*-2 items are identical.

Then, Apriori scans the entire transaction database to determine the frequent *k*-itemsets. The process is repeated for the next pass until no candidate can be generated. Apriori employs the downward closure property to efficiently generate candidates in each pass. The property indicates that no subset of a frequent itemset is infrequent; otherwise the itemset is infrequent. The property can be used to eliminate useless candidates to speed up the mining process. Other methods have been proposed to efficiently discover frequent itemsets, such as level-wise algorithms and pattern-growth methods.

Secure Multiparty Computation - A Secure Multiparty Computation (SMC) problem is defined a situation in which some information can be exchanged by ideal functions without a leak of knowledge other than the final result among multiparties. The generic techniques for SMC have high polynomial-time complexity, resulting in it sometimes being impractical. Several studies focus on finding efficient privacypreserving algorithms for specific problems, such as privacy-preserving computation of decision trees, and mining of vertically partitioned databases.

Privacy-preserving mining on horizontally partitioned databases -Distributed association rule mining techniques can discover association rules among multiple sites. They do not require that each site discloses the individual database, but each site is required to exchange all global candidate itemsets and the corresponding support counts with each other. If the support count for each global candidate itemsets in each individual site is sensitive, the above approach reveals such sensitive information to other competition companies. Therefore, to enhance the security of distributed mining and reduce the computation complexity of SMC, Kantarcioglu and Clifton proposed a secure scheme for privacy-preserving association rule mining on horizontally partitioned databases. For the simplicity, we refer Kantarcioglu and Clifton's Scheme as KCS for the rest of the paper. For this issue, Veloso et al. also proposed an efficient method to speed up the global candidate generation and concern the privacy-preserving for discovering frequent itemsets on distributed databases.

DISTRIBUTED ASSOCIATION RULE MINING

Association rule mining technique is receiving more attention among data mining techniques to the researchers to explore correlations between items or item sets. These rules can be analyzed to make strategic decisions to improve the performance of the business or quality of the organization service and so on.

Association rule generation has two steps. Computation of frequent item sets from the database based on user specified minimum support threshold is the first step and this process is difficult since it involves searching all combinations of item sets. In the second step, the association rules can be easily generated based on user specified minimum confidence threshold for the frequent item sets which are generated in the first step.

Now a days, more people who prefer to provide mutual benefit to their partners want to get access over

association rules which are derived from large database even though they are neither owners nor possessing privileges to access. The database owners also wish to share their derived results that is association rules to get some benefits from them but they do not want to provide their secret data and also leakage of secret data may cause damage or loss. Sharing of knowledge is the main concern in some application for the mutual benefits in knowledge discovery system while preserving privacy of individual is another concern.

In distributed environment, the challenging task is how efficiently one can provide accurate knowledge to their partners to have goodwill while no single secret data is revealed to them. This issue makes the researchers to study further to propose methods for privacy preserving association rule mining.

PRIVACY PRESERVING ASSOCIATION RULE MINING PERTAINING TO COMBINED PARTITIONED STYLE

Privacy-preserving data mining in a distributed environment is a multidisciplinary field and requires close cooperation between researchers and practitioners from the fields of cryptography, data mining, public policy and law. Now, the question is how to compute the results without pooling the data in a way that reveals nothing but the final results of the data mining computation. This question of privacy-preserving data mining is actually a special case of a long-studied problem in cryptography called secure multiparty computation. This problem deals with a setting where a set of parties with private inputs wishes to jointly compute some function of their inputs. This joint computation should have the property that the parties learn the correct output and nothing else, even if some of the parties maliciously collude to obtain more information.

Clearly, a protocol is needed to solve privacy-preserving data mining problems. Earlier work in privacy preserving association rule mining is as follows. In 1996, Clifton et al. discussed and presented ideas related to the issue of protecting privacy of individuals in the database. The state of the art in the area of privacy preserving data mining techniques is discussed by the authors in. This paper also describes the different dimensions of preserving data mining techniques such as data distribution, data modification technique, data mining algorithms, data or rule hiding and approaches for privacy preserving data mining techniques. In, the authors proposed a framework for evaluating privacy preserving data mining algorithms and based on their frame work one can assess the different features of privacy preserving algorithms according to different evaluation criteria.

Evfimievski et al. presented a new framework for preserving privacy association rule mining. In order to find privacy preserving association rule mining in centralized database, a new algorithm is presented in which balancs privacy preserving and knowledge discovery in association rule mining. Gkoulalas Divanis, et al. addressed many issues related to privacy preserving data mining, association rule hiding, classes of association rule hiding methodologies and also rule hiding in classification technique, privacy preserving clustering & sequence hiding.

The problem of knowing who is richer without disclosing their wealth is addressed in two milliner's problem and which belongs to secure multi party computation. The authors proposed protocols for two milliner's problem and also proposed for multi-party case. Clifton proposed a toolkit consisting of Secure sum, Secure set union, Secure size of set intersection and Scalar product are the protocols that can be combined for specific privacy preserving data mining applications. The algorithms for privacy preserving association rules mining over horizontally, vertically and mixed partitioned database are presented in this study work. Secure mining of association rules over horizontally partitioned database using cryptographic technique to minimize the information shared by adding overhead to the mining process is presented in. In, authors addressed the problem of association rule mining in vertically partitioned database by using cryptography based approach. In, several private scalar product protocols for two party scalar product protocols is proposed with a un trusted third party using algebraic computations.

The authors in proposed architecture for privacy preservation in classification technique for mixed partitioned distributed database model which is a combination of vertical and horizontal for Breast cancer dataset. In, algorithm is presented for finding privacy preserving association rule mining in mixed partitioned database model.

In most of the real life applications, mixed partitioning models are used and partitioning follows the organization structure. Consider the two mixed models which are commonly used and its partitioning are shown in the Figure 3 and Figure 4. The first type partitioning method is one in which all or some horizontally partitioned databases are further partitioned into two or more vertically partitioned databases and second type is one in which all or some vertically partitioned database are further partitioned into two or more horizontally partitioned databases. Mixed Model-1 is shown in following figure.

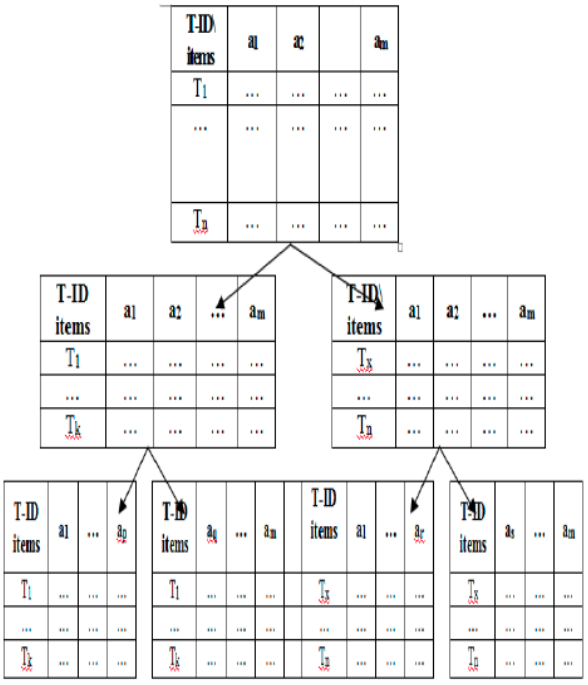

Fig.3. Two Horizontally Partitioned Databases are Further Partitioned into Two Vertical (Model-1)

Initially the database is partitioned into two horizontal partitioned databases. As partitioning is based on horizontal, all two partitioned sites possess same set of attributes but possessing different set of disjoint transactions. Each horizontally partitioned database is further partitioned into two vertical databases. The other Mixed Model-II is as shown below.

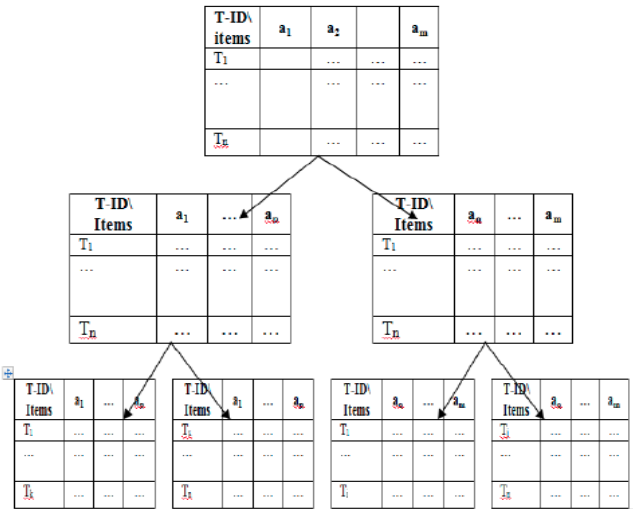

Fig. 4. Two Vertically Partitioned Databases are Further Partitioned into Two Horizontal (Model - 2)

In the above diagram, the database is partitioned into two vertical partitioned databases. As partitioning is based on vertical, all two partitioned sites possess disjoint subset set of attributes but possessing same set of transactions. In addition to the above common types of mixed partitioned databases, many other models exist like combinations of horizontal, horizontal and then vertical or vice versa. In this paper, the above two types of mixed model partitioning strategies are considered to find the global association rules.

CONCLUSION

Data mining techniques are very useful in extracting interesting information from databases. In this competitive but also cooperative business environment, companies need to share information with others, but not sharing the data. The research of privacy-preserving data mining on distributed databases has become an important issue. This study proposes an Enhanced Kantarcioglu and Clifton Scheme (EKCS) based on the two-phase method of the Kantarcioglu and Clifton Scheme (KCS). In the first phase, EKCS reduces the number of itemsets to be encrypted and transmitted without increasing the security risk.

Furthermore, in the second phase, this study introduces two protocols for enhancing security against collusion. Now, we are investigating the development of superior privacy-preserving algorithms to further reduce computation complexity and increase the security without sharing the data in distributed database environments.

The main threat in finding association rule mining in distributed database environment is privacy that is no site owner wish to provide database or local frequent item sets or support value to anyone. However every owner wishes to access mined result by participating indirectly in the mining process by providing partial results in disguised form.

Privacy is becoming a great research topic in the process of applying data mining techniques to various real applications. As the necessity makes the people to share the knowledge to the legitimate people in order to gain mutual benefits and this issue made to study privacy preserving data mining. Among many data mining techniques, privacy preserving association rule mining is a popular technique. But finding an efficient solution satisfying both privacy constraints as well as accuracy is a challenging task to researchers. A database in distributed environment can be partitioned in different ways like horizontal, vertical or mixed mode. In this paper, two new methodologies are presented to find global association rules in distributed environment by satisfying privacy constraints for two common mixed partitioned models.

Algorithms are also presented for each mixed model and implementation is discussed with suitable database. The efficiency of the proposed model is discussed in terms of privacy and communication.

REFERENCES

- A. Evfimievski, R. Srikant, R. Agrawal, and J. Gehrke. Privacy preserving mining of association rules. In *Proc. of 8__ACM SIGKDD Int'l Conf. on Knowledge Discovery and Data Mining*, July 2002.

- Agrawal, R. and Shafer, J.C. (1996) "Parallel mining of association rules," *IEEE Transactions on Knowledge and Data Engineering*, Vol. 8, No. 6, pp.929-969.

- 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.

- Charu C. Aggarwal and Philip S. Yu (July 2008), "A general survey of privacy - preserving data mining models and algorithms", Privacy-preserving data mining: models and algorithms.

- 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.

- *Elisa Bertino , Igor Nai Fovino Loredana Parasiliti Provenza ,*A Framework for Evaluating Privacy Preserving Data Mining Algorithms, Data Mining and Knowledge Discovery, 2005, 11, 121–154.

- Gkoulalas Divanis, V.S. Verykios(2010), "Association Rule Hiding for Data Mining", Advances in Database Systems, Vol. 41.

- J. Lin and M. Dunham. Mining association rules: Anti-skew algorithms. In *Proc. of 14 IEEE Int'l Conf. on Data Engineering*, October 1998.

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

- M tamer Ozsu Patrick Valduriez, *Principles of Distributed Database Systems,* 3rd Edition.

- Mohamed Hussein (2009), "Privacy Preserving in Association Rule Mining using cryptography", Master's Thesis, Menofia University, Egypt.

- *Y. Lindell and B. Pinkas,* Secure Multiparty Computation for Privacy-Preserving Data Mining, The Journal of Privacy and Confidentiality (2009) , 1, Number 1, pp. 59-98.

- Y. Lindell and B. Pinkas. Privacy preserving data mining. *Advances in Cryptology*, 1880:36–54, 2000.