FULLY HOMOMORPHIC ENCRYPTION BASED MULTIPARTY ASSOCIATION RULE MINING

Priyanka Dubey[1], Roshani Dubey[2]
[1]Department of Computer Sc. & Engineering, SRIT, RGPV University, Jabalpur, India
[2]Asst Prof, Department of Computer Sc. & Engineering,
SRIT, RGPV University, Jabalpur, India

ABSTRACT

Distributed data mining explores unknown information from data sources which are distributed among several parties. The Distributed Data Mining assumes that the data sources are distributed across multiple sites. Algorithms developed within this field address the problem of efficiently getting the mining results from all the data across these distributed sources. Since the primary (if not only) focus is on efficiency, most of the algorithms developed to date do not take security consideration into account. Privacy of participating parties becomes great concern and sensitive information pertaining to individual parties and needs high protection when data mining occurs among several parties. The proposed method finds global frequent item sets using association rule mining in a distributed multiparty environment with minimal communication among parties and ensures higher degree of privacy with fully homomorphic encryption.

KEYWORDS: *Distributed Data Mining, Multiparty Association Rule Mining, Homomorphic Encryption.*

1. INTRODUCTION

Data mining, or knowledge discovery, is the computer-assisted process of digging through and analyzing enormous sets of data and then extracting the meaning of the data. Data mining tools predict behaviors and future trends, allowing businesses to make proactive, knowledge-driven decisions. Data mining tools can answer business questions that traditionally were too time consuming to resolve. They scour databases for hidden patterns, finding predictive information that experts may miss because it lies outside their expectations.

Data mining algorithms discover patterns and interesting trends from large amount of data[6], which are critical to business success. Usually, the database is distributed across different sites and owned by different organizations. For example, two hospitals, each having patient records, want to discover common disease patterns from their joined data. In this case the privacy issue is raised, because privacy law does not permit the patient data to be disclosed. It is a challenge for them to perform data mining algorithms on their union data without disclosing the patient data.

The Distributed Data Mining assumes that the data sources are distributed across multiple sites. Algorithms developed within this field address the problem of efficiently getting the mining results from all the data across these distributed sources. Since the primary (if not only) focus is on efficiency, most of the algorithms developed to date do not take security consideration into account. Whether data is personal or corporate data, data mining offers the potential to reveal what others regard as sensitive (private). In some cases, it may be of mutual benefit for two parties (even

competitors) to share their data for an analysis task. However, they would like to ensure their own data remains private. In other words, there is a need to protect sensitive knowledge during a data mining process. This problem is called Privacy-Preserving Data Mining (PPDM). In the literature, the problem of parties who want to perform a computation on the union of their private data but do not trust each other, with each party wanting to hide its data from the other parties, is referred to as secure multi- party computation (SMC).

The aim of homomorphic cryptography is to ensure privacy of data in communication and storage processes, such as the ability to delegate computations to untrusted parties. If a user could take a problem defined in one algebraic system and encode it into a problem in a different algebraic system in a way that decoding back to the original algebraic system is hard, then the user could encode expensive computations and send them to the untrusted party. This untrusted party then performs the corresponding computation in the second algebraic system, returning the result to the user. Upon receiving the result, the user can decode it into a solution in the original algebraic system, while the untrusted party learns nothing of which computation was actually performed.

2. ASSOCIATION RULE MINING

Association rule mining (ARM) algorithms are typically a two stage process. The first stage consists of generating a list of frequent itemsets from the set of all known items. To avoid generating a list of all possible itemsets a threshold value is chosen, which is known as the support. This support value filters out most itemsets that would lead to uninteresting results. The second stage generates association rules from the list of frequent itemsets. The association rules are chosen, based on their support value. The two values, support and confidence, define how well we should trust an association rule generated from this process.

More formally, any combination of items are known as an itemset. That is, an itemset Is=(I1 U I2 U ….Ik) where, Ii I. An itemset with length k is known as a k-itemset. The general form of an association rule is X=>Y, where X & Y is subset of I and X∩Y=Φ.

Support and confidence are calculated as follows:[1][5]

$$Support_{AB \to C} = s = \frac{\sum_{i-1}^{sites} SupportCount_{ABC_i}}{\sum_{i-1}^{sites} DatabaseSize_i}$$

$$Support_{AB} = \frac{\sum_{i-1}^{sites} SupportCount_{AB_i}}{\sum_{i-1}^{sites} DatabaseSize_i}$$

3. MULTI PARTY MODEL

In Multiparty model data sources may be exists on more than one site which can be physically located anywhere. Here computation for getting result may also be on multiple points. There are two approaches: parallel & distributed. Parallel data mining (PDM)[7] deals with tightly-coupled systems including shared-memory systems (SMP), distributed-memory machines (DMM), or clussters of SMP workstations (CLUMPS) with a fast interconnect. Distributed data mining (DDM)[8], on the other hand, deals with loosely-coupled systems such as a cluster over a slow Ethernet local-area network. It also includes geographically distributed sites over a wide-area network like the Internet. The main differences between PDM to DDM are best understood if view DDM as a gradual transition from tightly-coupled, one-grained parallel machines to loosely-coupled medium- grained LAN of workstations, and finally very coarse-grained WANs. There is in fact a significant overlap between the two areas, especially at the medium- grained level where is it hard to draw a line between them. Secure Multi Party Computation solutions can be applied to maintain privacy in distributed rule mining. The main goal of Secure Multi Party Computation in distributed rule mining is to find global frequent item set without disclosing the local support count of participating sites to each other. In distributed privacy preserving data mining, the participating sites may be treated as honest, semi-honest. The semi-honest parties are honest but try to learn more from received information.

4. HOMOMORPHIC ENCRYPTION

Homomorphic encryption is a special form of encryption where one can perform a specific algebraic operation on the plain-text by applying the same or different operation on the cipher-text. If X and Y are two numbers and E and D denotes encryption and decryption function respectively, then homomorphic encryption holds following condition for an algebraic operation such as '+' [2]:

$$D[E(X) + E(Y)] = D[E(X + Y)]$$

Most homomorphic encryption system such as RSA (Rivest et al. 1978), ElGamal (El Gamal 1985), Benaloh (Clarkson 1994), Paillier (Paillier 1999) etc are capable to perform only one operation. But fully homomorphic encryption system can be used for many operations (such as, addition, multiplication, division etc.) at the same time.

5. PROPOSED WORK

Our proposed model extends existing system[1] for multiparty computation. Our goal is to perform fully homomorphic encryption based association rule mining for multiparty system in distributed environment. For this it is required first to partition the datasets according to partition scheme like horizontal, vertical or mixed[3].

Let there be k (> 2) parties P1, P2, . . . , Pk. The database is vertically partitioned between the k parties. The goal is to find association rules over the attributes across all of the parties. Steps for multiparty association rule mining at master and slave are as follows:

The master process

Step 1: spawn n slave process.
Step 2: divide database into n partitions.
Step 3: distribute partitions to each slave process.
Step 4: send global candidate (k-1)- itemsets to each slave process to count supports.
Step 5: wait and receive local supports from each slave process, and then compute global supports for the global candidate (k-1)- itemsets.
Step 6: make the global large (k-1)- itemsets with minimum support.
Step 7: send the global large (k-1)- itemsets to each slave process to generate candidate k-itemsets.
Step 8: wait and receive local candidate k-itemsets from each slave process.
Step 9: unionize local candidate kitemsets to the global candidate k-itemsets

Slave process

Step 1: receive the global candidate (k-1)-itemsets from the master process.
Step 2: count local supports for the global candidate (k-1)-itemsets.
Step 3: send local supports to master process.
Step 4: receive the global large (k-1)- itemsets from master process.

For privacy preservation we will uses fully homomorphic encryption scheme, that works when two party communicates or transfer support or frequent item information so that no other will know about this information.

6. CONCLUSIONS

This paper proposed a model for secure multiparty association rule mining over distributed dataset. Privacy/Security concerns have become an enduring part of society and commerce. It is increasingly necessary to ensure that useful computation does not violate legal/commercial norms for the safety of personal data. They are definitely applicable even beyond the scope of data mining.

REFERENCES

[1] Mohammed Golam Kaosar , Russell Paulet, Xun Yi "Fully homomorphic encryption based two- party association rule mining", Data & Knowledge Engineering 76–78 (2012) 1–15

[2] Md. Golam Kaosar Russell Paulet Xun Yi, "Secure Two-Party Association Rule Mining",9th Australasian Information Security Conference (AISC 2011), Perth, Australia, January 2011.

[3] Jayanti Danasana, Raghvendra Kumar and Debadutta Dey , "Mining Association Rule For Horizontally

Partitioned Databases Using Ck Secure Sum Technique", International Journal of Distributed and Parallel Systems (IJDPS) Vol.3, No.6, November 2012

[4] Caroline Fontaine and Fabien Galand, "A Survey of Homomorphic Encryption for Nonspecialists", Journal of Information Security, 2009, 1, 41-50

[5] Murat Kantarcıoˇglu and Chris Clifton," Privacy-preserving Distributed Mining of Association Rules on Horizontally Partitioned Data", 2003 IEEE

[6] Qiankun Zhao and Sourav S. Bhowmick, "Association Rule Mining: A Survey" Technical Report, CAIS, Nanyang Technological University, Singapore, No. 2003116 , 2003.

[7] Mohammed J. Zaki, "Parallel and Distributed Data Mining: An Introduction", Springer-Verlag Berlin Heidelberg 2000

[8] Neha Saxena, Rakhi Arora, Ranjana Sikarwar and Ashika Gupta, "An Efficient Approach of Association Rule Mining on Distributed Database Algorithm", International Journal of Information and Computation Technology. ISSN 0974-2239 Volume 3, Number 4 (2013), pp. 225-234

AUTHORS

Priyanka Dubey received his B. E. degree in Computer Science and Engineering from Department of Computer Science and Engineering, from Takshshila Institute of Engineering &Technology, Jabalpur under RGPV University Bhopal (M.P.). She is currently a student of Master of Engg (M.E.) at Shri Ram Institute of Technology Jabalpur. She has also good teaching experience of under graduate engineering students.