ISSN 2348 – 7968

Mining of Data Association in Distributed Databases

Shubhangi Ramteke[1]
Dept. of Computer Technology

Prof.R.K.Krishna[2]
Dept. of Electronics Engineering

[1,2]Rajiv Gandhi College of Engineering, Research & Technology
Chandrapur, Maharashtra State- 442 403 , India

Abstract

Data mining is the process of extracting information from data set and transform it into an understandable structure for further processing. The goal is to extract pattern and knowledge from large amount of data. Data mining can be used to obtain more accurate prediction result. Among many data mining techniques association rules mining is most widely used. This paper introduces a model for performing mining of data & association between them in distributed databases without revealing private data of data owner and to maintain data confidentiality. This model perform data mining by using the concept of cryptography technique in addition with some data mining algorithms when data is distributed on different sites.

Keywords: Association rules, Data Mining, Cryptography, Distributed Databases

1. Introduction

Data Mining has been viewed as a threat to privacy because of the rapid increase in the electronic data maintained by corporations. This results in the increased concerns about the privacy of the data. Data mining techniques find hidden information while confidential data is preserved safely when single person is allowed to access the data. But every person wants to access the confidential data by using data mining technique even though they are not allowed to do so. Most of the organizations wants to share their data with user but without providing access to their secret data. In most of the applications whole data is stored in a single central place called as a centralized database or number of places/sites called as distributed databases. This paper deals with the preserving or maintaining the privacy of confidential data while sharing hidden or secret data with many people. A distributed database consists of two or more data files located at different sites on a computer network. Because the database is distributed, different users can access it without interfering with one another. However, the DBMS must periodically synchronize the scattered databases to make sure that they all have consistent data. In this, 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 data mining, association rule mining is most popular method for finding interesting pattern of data in large distributed databases. In this paper, a model is proposed to find association between data when data is distributed among number of sites.

2. Association rules mining in distributed databases

Among many data mining techniques association rules mining is most widely used technique. These association rules describes association or correlations between items or items set. These rules then can be used for analysis purpose to gain profit or to improve the performance of the business or quality of the organization service by most of the organization. Association rule mining concept was introduced by Agarwal. An association rule can be defined as follows- Let I={i1,i2,i3,.....,im} be the set of attributes called items. The item set X consisting of one or more items. Let DB= {T1,T2,T3,....,Tn} be the database consisting of n number of boolean transactions, and each transaction Ti consisting of items supported by ith transaction. An item set X is said to be frequent when number of transactions supporting this item set is greater than or equal to the user specified minimum support threshold else it is considered as a infrequent. An association rule is an implication of the form X→Y where X and Y are disjoint subsets of I, X is called the antecedent and Y is called consequent. An association rule X→Y is said to be strong association rule only when its confidence is greater than or equal to user specified minimum confidence. It is very difficult task to provide their partners the accurate knowledge without revealing a single secret to them. This issue makes the researchers to study further to propose methods for privacy preserving data association mining in distributed databases without loss of confidentiality of data.

3. Mining of data association in horizontally distributed databases in a secure way

Many researchers provides different methods to maintain data confidentiality while finding association between data in distributed databases. For performing this different methods such as data distribution, some secure data mining algorithms, cryptographic techniques are used. In this paper, new methodology is proposed to find privacy preserving association rule mining for horizontally database with trusted third party. The proposed method perform this by using a cryptography technique. This technique protects data from unauthorized access to distributed data. In this paper to protect one's local frequent item sets from others unauthorized access, public private key algorithm is used in addition with data mining algorithm. In this method, a special site is designated and this site owner is called Trusted Party or simply the OWNER initiates the process of finding association rules without knowing any one's individual data/information but by taking processed results from all the sites in secure manner.

The following figure shows the communication between Trusted third party and sites in the proposed model.

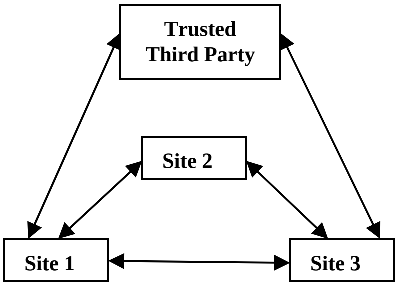

Figure 1. communication between different sites and Trusted Third Party

In the above figure, site 1, site 2 and site 3 are the three different sites from where trusted third party is accessing data for further processing. Every site needs global frequent item sets and supports to generate global association rules. So the goal is to determine global frequent item sets with supports based on the databases at all sites. Any item set is said to be globally frequent only when sum of support value of item sets at all sites is greater than or equal to minimum number of transactions required to support this item globally. An item set can be globally frequent only when it exists in at least one or more sites as frequent. Similarly an item set can be globally infrequent only when the item set is infrequent in at least one or more sites.

4. Implementation of the proposed model with sample data

It is very clear that no one is willing to reveal their local frequent item sets, supports , confidence and database size to any site owner as well as to trusted third party. To solve this problem the method provides special rights to trusted third party to capture local frequent item sets without taking the value of supports from all sites to determine all sites frequent item sets. Every site owner accepts to provide local frequent item sets in encrypted form to trusted third party to whom they trusts to generate merged or combined frequent item set result. The proposed model is illustrated by using horizontally partitioned distributed database for finding data association mining to preserve privacy of secret data. Trusted third party request to site to send encrypted form of local frequent item sets by sending two values such as minimum support, threshold and decryption key. After receiving the encrypted form of local frequent item sets from the site, trusted third party prepares a merged frequent item list after eliminating duplicates with the help of decryption key. Each site can generate global association rules for each global frequent item set based on the specified minimum confidence threshold. Each site computes partial support and broadcast to all other sites in order to find the total partial supports.

5. Maintaining confidentiality in proposed model

In this paper, a new model is proposed to find association between data by maintaining data privacy. This model can be used with any number of sites and with any number of transactions in a distributed databases. The proposed model is efficient in many ways such as Privacy is ensured by using encryption and decryption techniques at the time of transferring the frequent item sets from different sites to trusted third party. This ensures that the trusted third party can only access requested data and can't access the confidential data. Results that are global frequent item sets and their supports are broadcasted by trusted party to all sites. Because of this result no site owner can predict local support of any global frequent item sets, as global frequent item sets may not be frequent in all sites and any site owner can not predict the contribution of other sites database which makes the item set globally frequent.

5.1 Fast Distributed Mining Algorithm

In this model, Fast Distributed Mining (FDM) algorithm is also used which is an unsecured distributed version of the Apriori algorithm. In this it uses 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 site in a subset held by another. Our protocol offers enhanced privacy. It is simpler and is significantly more efficient in terms of communication rounds, communication cost and computational cost. the Fast Distributed Mining (FDM) algorithm of Cheung et al ,which is an unsecured distributed version of the Apriori algorithm. Its main idea is that any s-frequent item set 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 works 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

5.2 Apriori Algorithm

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 called a transaction. The output of Apriori is sets of rules that tell us how often items are contained in sets of data. It proceeds by identifying the frequent individual items in the database and extending them to larger and larger item sets as long as those item sets appear sufficiently often in the database. The frequent item sets determined by Apriori can be used to determine association rules which highlight general trends in the database. Apriori uses a "bottom up" approach, where frequent subsets are extended one item at a time and groups of candidates are tested against the data. The algorithm terminates when no further successful extensions are found.

It basically consists of two steps-
1. Self-Join
2. Pruning

Repeating these steps k times, where k is the number of items, in the last iteration you get frequent item sets containing k items.

Acknowledgments

It is a pleasure to acknowledge the assistance of several people and institutions in this effort. First and foremost, I feel indebted to my guide, Professor R.K. Krishna, Department of Electronics, Rajiv Gandhi College of Engineering, Research & Technology , Chandrapur for his valuable guidance, continuous support ,advice and constant encouragement throughout my work. A special word of thanks goes to Prof. Rahila Sheikh, Assistant professor, Department of Computer Technology and Prof. P.S.Kulkarni, Head, Department of Information Technology, R.C.E.R.T., Chandrapur for their encouragement to accomplish my work on time. I am also grateful to Prof. Nitin J. Janwe, Head, Department of Computer Technology, R.C.E.R.T., Chandrapur for his last minute instruction which helped me to focus my work in the right direction.

I would also like to extend my thanks to my guide, my family members without whose inspiration and support my efforts would not have come to true.

Above all, I would like to thank God for making all my efforts success.

References

[1] R Agarwal, T Imielinski and A Swamy, "Mining Association Rules between Sets of Items in Large Databases", Proceedings of the 1993 ACM SIGMOD International Conference on Management of Data, pp.207-210, 1993.

[2] 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.

[3] M. Kantarcioglu and C. Clifto. "Privacy-preserving distributed mining of association rules on horizontally partitioned data". In IEEE Transactions on Knowledge and Data Engineering Journal, volume 16(9), pp.1026–1037.

[4] Chin-Chen Chang, Jieh-Shan Yeh, and Yu-Chiang Li, "Privacy- Preserving Mining of Association Rules On Distributed Databases", IJCSNS International Journal of Computer Science and Network Security, VOL.6 No.11, November 2006.

[5] Agrawal, R., et al “Mining association rules between sets of items in large database”. In: Proc. of ACM SIGMOD’93, D.C, ACM Press,Washington, pp.207-216, 1993.

[6] A.V. Evfimievski, R. Srikant, R. Agrawal, and J.

Gehrke.“Privacy preserving mining of association rules.”In KDD, pp. 217–228, 2002.

[7] Tamir Tassa, "Secure Mining of Association Rules in Horizontally Distributed Databases", IEEE transaction on knowledge and data engineering VOL.pp.99 , 2013

[8] Vaidya, J. and Clifton,"Privacy preserving association rule mining in vertically partitioned data", 8th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, ACM Press, pp. 639–644,2002

[9] Srikant, R. Agrawal, “Mining generalized association rules”, In: VLDB’95, pp.479-488, 1994.

[10] J. Zhan, S. Matwin, and L. Chang."Privacy preserving collaborative association rule mining." In Data and Applications Security, pp.153–165, 2005.

Authors

Shubhangi Ramteke Received the Batchlor’s degree in Computer Science & Engineering from Swami Ramanand Tirth Marathwda University, Nanded. She is pursuing MTech in Computer Science & Engineering from Gondwana University. Her research interest includes Data mining, Information security, data warehousing , database security.

Prof. R. K. Krishna Professor - Department of Electronics Engineering, RCERT, Chandrapur(M.S.)

B.E., M. Tech, MBA, 23 years experience of teaching, More than 50 publications. His research interest includes Wireless sensor networks, communication engineering and image processing.