ISSN: 2277-9655
(I2OR), Publication Impact Factor: 3.785

IJESRT

INTERNATIONAL JOURNAL OF ENGINEERING SCIENCES & RESEARCH TECHNOLOGY

SECURE MINING OF ASSOCIATION RULES IN DISTRIBUTED DATABASE USING SEMI HONEST THIRD PARTY

Miss. Amruta Jadhav*,Prof. Vipul Bag

* Department of Computer Engineering, N.K.Orchid College of Engineering, Maharashtra, India

ABSTRACT

Data mining is used to discovering useful patterns hidden in a database from large datasets, but sometimes these datasets are split among various sites and none of the sites is allowed to expose its database to another site. Association Rule mining in distributed database is one of the important and well researched techniques of data mining. This technique discloses some interesting relationship between local as well as global item sets.Mining of association rules from distributed databasesare essential in different area such as market basket analysis.But sometimes there are problem to determine a useful pattern in distributed databases.Also the protection of information from illegal access has been a long term goal for businesses and government organizations.So that it requires enhanced privacy. In this paper, we have shown the Association rule mining algorithm over horizontal distributed databases. Using our approach is to generate strong association rules from different data sets spread over various geographical sites and also preserving privacy of data.

KEYWORDS: Secure Data Mining, Distributed Database, Frequent Item sets, Association Rules

INTRODUCTION

The data mining techniques is to efficiently discover valuable and non-obvious knowledge from large databases. Association rule mining is one of the mainly essential and fine researched methods of data mining. It aims to extort exciting correlations, common patterns, associations or informal structures amongst sets of objects in the transaction databases or additional data repositories. The mining of association rules is essential in various data mining fields, such as financial analysis, the retail industry and business decision-making. The market basket analysis used association rule mining in distributed environment.

Today modern organizations have their own databases, located in different places; 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. Therefore, secure mining has become an important issue in many data mining applications.The data mining may need to be processed among databases in some business environments. Nevertheless, data may be distributed among several sites, but none of the sites is allowed to expose its database to another site.

Association rules show attributes value conditions that occur frequently together in a given dataset. The process of association rule mining includes two mainsub-problems: the first is to discover all frequent itemsetsi.eto discover those item sets whose occurrences go above a predefined threshold in the database; those item sets are known as frequent or large item sets. the second is to use these discovered frequent itemsets to generate association rules by calculating amount support and confidence from a database using formula . Since each association rule can easily be derived from the corresponding frequent itemsets, the overall performance of the association rule mining isdetermined by the first sub-problem. Therefore, researchers usually focus on efficiently discoveringfrequent itemsets.

RELATED WORK

In the earlier research work [6],[7],[8],[9]the author uses various techniques such as transaction reduction ,clustering and algorithms such as Apriori and fpgrowth for mining frequent item sets using different datasets were analysed and compared.

Distributed association rule mining techniques can discover association rules among multiple sites [1, 5]. 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 itemset in each individual site is sensitive, the above approach reveals such sensitive information to other competition companies. Therefore, it is necessary 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 [3].

In Existing System, the problem of mining of association rules in distributed databases. [1]In that setting, there are several sites (or players) that hold homogeneous databases. The inputs are the partial databases, and the required output is the list of association rules which are less strong that hold in the unified database with support and confidence.There is less protecting the data records of each of the data owners from the other data owners. While such leakage of information renders the protocol not perfectly secure, the perimeter of the excess information is explicitly bounded and it leaks data that discloses information also to some sites. Also problem of inefficient association rules generated and communication and computation overhead.

Agrawal et al. presented the Apriorialgorithm to identify frequent itemsets [4].Apriori is a level-by-level algorithm including multiple passes. In each pass, Apriori generates a candidate set offrequent *k*-itemsets (frequent itemsets with length *k*). Eachfrequent 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 todetermine the frequent *k*-itemsets.The process is repeatedfor the next pass until no candidate can be generated. There are needs to eliminate useless candidates to speed up themining process. Agrawal et al. first introduced the problem of association rule mining over a market-basket transaction database in [4]. An example of a rule is as follows: 50% of transactions that purchase an 21" LCD monitor also purchase a video game. Such rules can provide valuable information on the customer buying behavior.

PROPOSED APPROACH

The goal of proposedsytemis to find all frequent item sets and association rules from distributed database and also to minimize the information disclosed about the private databases held by those sites.Privacy concerns are the major important scenario in recent approaches because each party may not want to reveal in their own partitioned database that exists relative data efficiency.

In this distributed databases, there are mining patterns and trends from large amount of data.So that we have develope algorithm for mining interesting association rules or correlation relationships from horizontally distributed database.For that here partial databases of different sites is given as input and the output is the list of association rules with support and confidence. The process of association rule mining will calculate minimum support 's' and minimum confidence 'c'and also find all global frequent item set as well as local frequent item set from dataset of each site in distributed database by using semi honest third party. It is called semi-honest party because it don't have any knowledge about original transaction in distributed database.Finally will get strong association rules in the form of original data to only that data owner.

Strong Association Rules is association rules are generated from the frequent items sets and whose confidence is greater than the minimum threshold confidence. Finally strong associations rules are grouped in this way are displayed to the data owner and user. Also incorporate cryptographic techniques to minimize the information reveal which is going to shared with others.So that this system protect the information is not only individual transaction in the different distributed databases, but also more global information such as what association rules are supported locally in each of those databases.

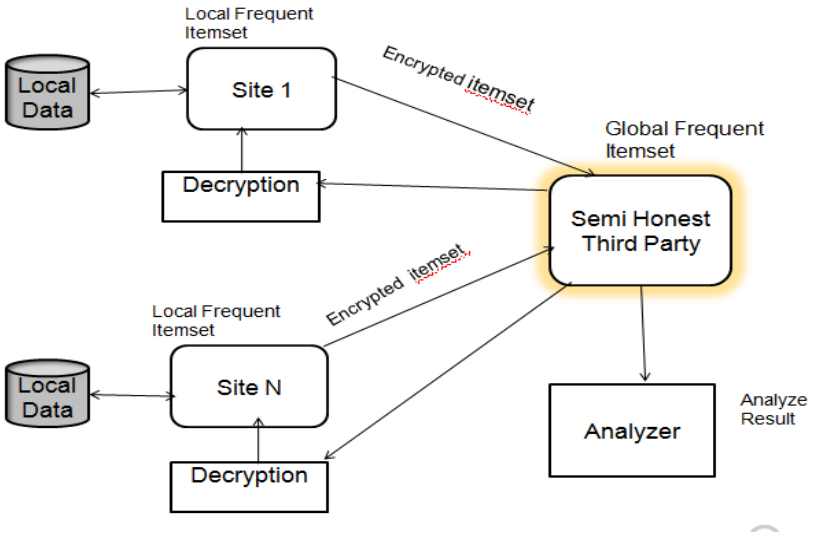

Fig1. System Overview

The proposed system consists the following modules.

1) Secure Mining of details

Secure mining of details this module displays the transaction details maintained in the horizontal databases without showing private attributes and item sets details of transaction.

2) Frequent Item sets

Thefrequent item sets are sets of items that have minimum support.In this module frequent item sets are displayedfrom the datasets by using formula.

Support is an important measure because a rule that has very low support may occur simply by chance. A low support rule is also likely to be uninteresting from a business perspective because it may not be profitable to promote items that customers infrequently buy together. For these reasons, support is often used to eliminate uninteresting rules.

Here local frequent item sets calculated at each sites in distributed database by using formula.

$$Support = \frac{(\text{containing the item combination})}{(\text{total number of record})}$$

Let the rule is "If a customer purchases **Cola**, then they will purchase **Frozen Pizza**"The support for this ……

= 2 (number of transaction that include both Cola and Frozen Pizza is) / 5(total records)
= 40%.

3) Association Rule Mining

An association rule is an implication expression of the form X → Y, where X and Y are disjoint item sets, i.e., X ∩ Y= Ø. The strength of an association rule can be measured in terms of its support and confidence.

$$\text{Confidence of a rule} = \frac{(\text{Support for the combination})}{(\text{Support for the condition})}$$

For the rule "If a customer purchases **Milk**, then they will purchase **Potato Chips**"

Confidence = support for the combination (Potato Chips + Milk) is 20%/ support for the condition (Milk) is 60%
=33%

Confidence, on the other hand, measures the reliability of the inference made by a rule. For a given rule X →Y, the higher the confidence, the more likely it is for Y to be present in transactions that contain X. It also provides an estimate of the conditional probability of Y given X.

Algorithm :

Input: Datasets from different sites,a set of n records within datasets.

Output: List of association rules with support and confidence.

Step1:Scan Dataset at each site.

Step2:Compute Local Frequent Itemsetat each sites by calculating support.

$$\text{Support, s}\,(X \rightarrow Y) = \frac{(X \cup Y)}{N}$$

Step 3:GetinitialFIS from all sites and calculate FIS.

Step3:Prune:Remove those candidatesitemsets(subset) that can not be frequent.

Step4:Compute Global Frequent Itemset.

Step5:Generate Association rules with support and confidence.

$$\text{Confidence, c}\,(X \rightarrow Y) = \frac{(X \cup Y)}{(X)}$$

Step6:Broadcast Mining Results to all sites.

we could derive the association rules:

- **{Cheese, Milk} =>Bread** [sup=5%, conf=80%]
- **{Bread, Jam} =>Milk** [sup=40%, conf=75%]
- **{ Milk, Jam} => Bread** [sup=40%, conf=75%]
- **Bread => { Milk, Jam}[**sup=40%, conf=75%]
- **Frozen Pizza => Cola**[sup=60%, conf=90%]
- **Cola => Chips**[sup=40%, conf=75%]
- **Laptop=>Antivirus**[sup=50%, conf=75%]
- **Laptop =>S/W** [sup=40%, conf=95%]

RESULTS AND DISCUSSION

The experiment is done with the below design considerations.

System is designed with synthetic databases those are horizontally distributed across the different databases. In this horizontal partitioning approach, some of the rows of a relation R are apportioned at one of the database,and other rows are assigned at another database.

This proposed system's developed algorithm is secure association rule mining (SAM) Algorithm which is implemented for finding the frequent Item Sets from horizontally distributed databases and the Association Rules are generated from frequent Item Sets found and grouped the strong association rules whose confidence is greater than minimum threshold confidence.

We have checked the total computation time for finding the Association Rules using this Algorithm. The computation time was measured based on the parameter "No of transactions". Table 1 shows total execution time taken for finding association rules using secure association rule mining (SAM) Algorithm against number of transactions.

Table 1

Execution Time for different No of Transactions

Number of Transactions	Execution Time (Seconds)	
	Apriori	**SAM**
1000	0.6	0.4
2000	0.8	0.44
3000	0.9	0.5
4000	1.1	0.6
5000	1.3	0.7

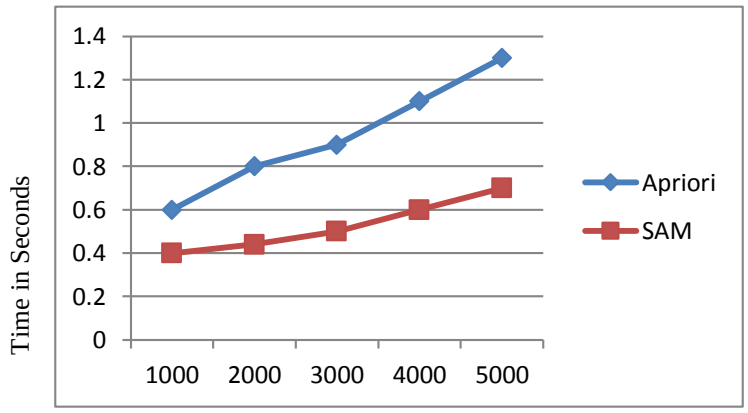

Fig. 2. Total Execution Time

CONCLUSION

From the above results have been demonstrated that the results of the SAM algorithm over horizontally distributed databases. The algorithm performance is analyzed based on the execution time and different no of transactions. From this results we see that our approach is better than of the previous works and generated association rules are efficient. In this system there is security of data i.e.it protects the information from each of data owner to some other data owners. By using this system we can shown how Market basket analysis using association rules works in determining the customer buying patterns.

REFERENCES

[1] TamirTassa, "Secure Mining of Association Rules in Horizontally Distributed Databases" IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, VOL. 26, NO. 4, APRIL 2014
[2] 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
[3] 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 2008
[4] 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 1993P.
[5] D.W.L Cheung, V.T.Y. Ng, A.W.C. Fu, and Y. Fu, "Efficient Mining of Association Rules in Distributed Databases," IEEE Trans.Knowledge and Data Eng., vol. 8, no. 6, Dec. 1996.
[6] H. Grosskreutz, B. Lemmen, and S. R¨uping. Secure distributed subgroup discovery in horizontally partitioned data. *Transactions on Data Privacy*,4:147–165, 2011.
[7] A. Schuster, R. Wolff, and B. Gilburd. Privacy-preserving associationrule mining in large-scale distributed systems. In *CCGRID*, pages 411–418, 2004.
[8] J. Zhan, S. Matwin, and L. Chang. Privacy preserving collaborative association rule mining. In *Data and Applications Security*, pages 153–165, 2005.
[9] D.KeranaHanirex ,"Association Rule Mining in Distributed Database System", International Journal of Computer Science and MobileComputing(IJCSMC), Vol3,Iss 4,pg 727-732,2014.

AUTHOR BIBLIOGRAPHY

Miss. AmrutaJadhav

She received B.E degree in Computer Science and Engineering from University of Solapur, Maharashtra, India and pursuing the M.E. degree in Computer Science and Engineering in NageshKarajagi Orchid College of Engg. & Technology, Solapur, India. She is doing her dissertation work under the guidance ofMr. VipulBag.,AssociateProfessorat NageshKarajagi Orchid College of Engg. & Technology, Solapur, Maharashtra, India..

Mr. Vipul Bag

Mr. Vipul Bag, is working as Assosiate Professor in Department of Computer Science and Engineering in NK Orchid College of Engineering and Technology, Solapur, Maharashtra, India. He has 16 years of teaching experience. He has co-authored over 20 International Journal Publications. He is pursuing PhD from SGGSIET, Nanded, Maharashtra, India. The current research interests are Recommendation systems, Data Mining and Machine Learning.