IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319 5940

International Journal of Advanced Research in Computer and Communication Engineering
ISO 3297:2007 Certified
Vol. 5, Issue 10, October 2016

Privacy Preserving Association Rule Mining in Partitioned Databases

Pankaj P. Joshi[1], Prof. R.M. Goudar[2]

Computer Department, MIT Academy of Engineering, Pune, India[1]

Asst. Prof, Computer Department, MIT Academy of Engineering, Pune, India[2]

Abstract: Fast improvement in the field of information technology rises the serious issue about huge information storage. To establish the correlation among set of data items, also becomes difficult challenge. To overcome these problems in huge data, association rule mining plays a very important role. In recent years, most of the researcher's point of interest is to enhance the effectiveness of association rule mining based algorithm and thereby increasing the speed of mining. In existing system, association rules are applied on horizontal partitioned database and such technique is depending on IDMA, EMADS communication. In this paper, we consider both horizontal as well as vertical partitioned database, while applying association rule mining. The proposed system also minimizes the communications overhead. Further to enhance the security of system, RC4 algorithm is used to secure the horizontal and vertical partitions of database. Finally, system utilizes the protocol for recognizing the fake and duplicate frequent rules.

Keywords: Association Rule Mining, RC4, horizontal partitioned, vertical partitioned.

I. INTRODUCTION

Generating the knowledge from the data, which is embedded in the database, is one of the main task of data mining technology. Data mining procedure extract the information from the frequently growing the quantity of data and the transferring this data into useful knowledge. Therefore, this becomes core point of attraction of many researchers. Numbers of technique have been developed by the researchers, for analysing the increasing volume of huge data. One of the most important types of data mining technique named as Association rule mining was introduced in 1993 by Agrawal. The term association rule mining refers to find out the correlation among large set of data items. Association rule mining plays a very important role in decision making algorithms. One of the characteristic instances of the association rule mining is the market base analysis.

Association rule mining performs the two major operations, which are listed here:

- Frequent item generation: This is the item set which satisfy the minimum threshold support.
- Rule generation: This is also called as the strong rules, since these rules have high confidence. These rules were found in the initial step of the association rule.

Database can be partitioned either horizontally or vertically. Security becomes major concern issue during data mining or partitioning. Here we focus on the issue occurs during the secure mining of association rules in the horizontal and vertical partitioned database. There are numerous websites are available, who handles the uniform database, which share the same information but hold the data on the distinct schemas. The main focus is to find out the association rules with the value of support and confidence level. Our main focus is not only to secure the information from the entity transaction in distinct databases, but also provide more universal information like the association rules which are supported in each of those databases.

Here we discuss the issues regarding to privacy of multiparty computation. In this issue, there are number of clients who hold the private inputs and they want to securely evaluate the public functions. If the existence of the TTP occurred, clients should provide his inputs and performing the function evaluation and sending them the resulting output. If there is no TTP exist, then there is need to devise the protocol, in which client can run on their own in order to arrive at the required output.

Here in this paper, we proposed the method for providing the secured computation of the union of private database. Also, we proposed the method which provides the security to both horizontal and vertical partitioning database and trying to reduce the communication overhead of the multiparty distributed system.

By using this algorithm, the generated fake and duplicate association rules are identified and deleted. We also proposed the method for detecting the fake user.

This paper is composed further as: Section II talks about related work studied till now. Section III presents implementation details, algorithm used and mathematical model. It also includes experimental setup. Section IV ends with the conclusions and presents future work.

IJARCCE
ISSN (Online) 2278-1021
ISSN (Print) 2319 5940
International Journal of Advanced Research in Computer and Communication Engineering
ISO 3297:2007 Certified
Vol. 5, Issue 10, October 2016

II. LITERATURE REVIEW

This section describes recent approaches for mobile based distributed secured mining.

- In [1], proposed an optimized distributed association rule mining algorithm for the biologically dispersed data which is used in parallel and dispersed environment for reducing the communication cost. In this approach the multiple nesting problems in XML data is handled appropriately for assuring the correctness of the results. The algorithm is used for the mining process in a parallel and distributed environment.

- In [2], authors tracing out the recently trends in parallel and distributed Apriori algorithm. In this paper author review distinct parallel and dispersed association rule mining which are developed based on the Apriori algorithm. In this paper discussed 10 distinct algorithm of association rule mining which are dynamic hashing and pruning, Éclat-frequent item set mining, hybrid distribution, intelligent data distribution, partition and non-partitioned, simply partitioned, and hashed partitioned Apriori, DMM and FDM.

- In [3], identify and solve the issues of mining association rules on shared nothing multiprocessor. Author presents three algorithms which discover a spectrum of trade-offs among computation, communication, memory usage, synchronization and the use of problem specific information.

- In [4], detailed use of association rule mining for extracting the patterns is explained, which occurred regularly within a dataset. Also shows the execution of the Apriori algorithm for mining association's rules from the dataset which contains the crime data concerning women.

- The paper [5], represent a well-organized algorithm which produce all important association rules among items in the database. The algorithm incorporates buffer management and novel estimation and pruning techniques.

- In [6], Matthias Klusch, Stefano Lodi, Gianluca Moro discussed one key fact about developing the huge amount of independent and varied data sources in the Internet is not only how to recover, gather and put together applicable information but to determine previously unknown, implied and precious knowledge. In recent years' numerous approaches to dispersed data mining and knowledge discovery have been developed, but only a few of them make use of intelligent agents.

- In [7], Stolfo, Prodromidis, et al illustrates the JAM system, a dispersed, scalable and moveable agent based data mining system which employs a universal approach for scaling data mining applications that called meta learning. JAM offers a set of learning programs, executed either as JAVA applets or applications that calculate models over data stored locally at a site. JAM also offered a set of meta-learning agents for merge multiple models that were learned at distinct sites.

- In [8], E.I. Aria, M. B. Senousy and M. M. Medhat proposed the model defeat the communication blockage by using mobile agent's example. This model split the DDM procedure into numerous phases that can be done in similar on distinct data sources: Preparation phase, data mining phase and knowledge integration phase. Author also includes a particular section on how present e-business models can use our model to strengthen the decision support in the association. A cost examination in terms of time consumed by each minor process is given to demonstrate the overheads of this model and the other models.

- In [9], Yun-Lan Wang, Zeng-Zhi Li and Hai-Ping Zhu discussed about the over rising dimension of data being accumulate in today's information schemes, unavoidably escorts to the dispersed database framework. Furthermore, many databases are dispersed in nature. It is essential to device resourceful methods for dispersed data mining. They proposed a mobile agent based dispersed knowledge discovery framework for data mining in the distributed, heterogeneous database systems. Based on this framework a supple and resourceful mobile-agent-based distributed algorithm for association rules is represented that can mine the global and local large item sets at the same time.

- In [10], authors proposed application that works with large datasets which cannot manage most feature of the data's divider and preparations. So far, concentration in data mining process has always alert on extracting information from data physically situated at one central site and they often do not consider the reserve restraint of dispersed and mobile environments. Few attempts were also made in similar data mining.

- In [11], WalidAdlyAtteya, KeshavDahal, M. Alamgir Hossain represent a dispersed Multi Agent based algorithm for mining association rules in dispersed environments. The dispersed MAS algorithm uses Bit vector data structure which was proved to have improved presentation in centralized environments. The algorithm is executed in the background of Multi-Agent systems and obeys with global communication standard Foundation for Intelligent Physical Agents (FIPA).

- In [12], Rakesh Agrawal and Ramakrishnan Srikant represent two novel algorithms for resolving the issues

IJARCCE
ISSN (Online) 2278-1021
ISSN (Print) 2319 5940
International Journal of Advanced Research in Computer and Communication Engineering
ISO 3297:2007 Certified
Vol. 5, Issue 10, October 2016

which are essentially distinct from the known algorithms. Experiential assessment shows that these algorithms better than known algorithms by factors ranging from three for small problems to more than an order of extent for big problems. Author also shows how the best features of the two proposed algorithms can be joint into a hybrid algorithm, called Apriori Hybrid.

- In [13], FerencBodon examines a trie based APRIORI algorithm for mining frequent item sequence in a transactional database. Author inspects the data structure, completion and algorithmic features mainly focusing on those that also occur in frequent item set mining.

III.IMPLEMENTATIONS DETAILS

A. System Overview

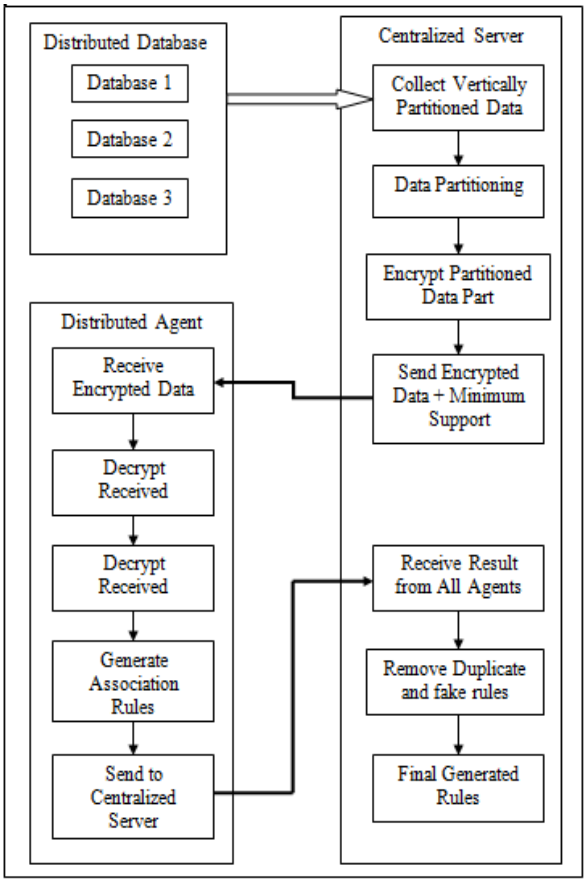

Fig 1. Propose System

Initially the distributed database i.e. the database from the distinct sites is send to the centralized server. We have denoted the dispersed database as database 1, database 2 and database 3.

- After this centralized server collects all the vertically partitioned dataset given by the dispersed database.
- Database is partitioned in two ways horizontally and vertically partitioned data.
- After partitioning the database, encryption algorithm is used to encrypt each partition. The partitioned data are encrypted for providing the security over data.
- After providing the security to the partitioned data, encrypted data and minimum support value send to the distributed agent.
- Distributed agent receives the encrypted data and decrypts it.
- After decrypting the encrypted data, associated rules are generated by using the Apriori algorithm. After generating the rules, distributed agent sends these rules to the centralized server.
- Centralized server receives result from all the agents, after receiving rules from all the agents, centralized server removes duplicate and fake rules from the server.
- Finally, the global list is generated by the centralized server.

B. Algorithm

Unifying lists of locally frequent item set

Input: Vertically partitioned database
Output: Global List generated association rules.
Step 1: Vertically partitioning of databases.
Step 2: For centralized Server Gather partitioned data from database.
Step 3: after combine data partitioned into no of parts.
Step 4: for each agent
Step 5: if (distributed agent is free)
Step 6: encrypt each part of data to send distributed processor agent.
Step 7: end If;
Step 8: end For
Step 9: send encrypt data to agent and minimum support value.
Step 10; at each agent
Step 11: decrypt received data
Step 12: generate association rules base on support and confidence.
Step 13: send generated rules above the minimum support to the centralized server.
Step 14: Merge all generated rules from different distributed agents.
Step 15: remove Duplicate and fake rules
Step 16: Generate Final association rules

C. Mathematical Model

Encryption method for RC4

for each message byte Mi
i = (i + 1) (mod 256)
j = (j + S[i]) (mod 256)
swap(S[i], S[j])
t = (S[i] + S[j]) (mod 256)
Ci = Mi XOR S[t]

Equation for generating the Confidence value

$$Confidence=\frac{\text{Support } (x \cup y)}{\text{Supp } (x)}$$

IJARCCE
ISSN (Online) 2278-1021
ISSN (Print) 2319 5940
International Journal of Advanced Research in Computer and Communication Engineering
ISO 3297:2007 Certified
Vol. 5, Issue 10, October 2016

Where, Where x and y item sets

Equation for generating the support value

$$Support = \frac{\text{Number of occurence of x}}{\text{Total number of transactions}}$$

Where
X is the item sets

Set theory

1. Identify the set of clients

C= {c1, c2, c3....,}
Where C is main set of Clients like c1, c2, c3....

2. Identify the set of severs

S= {s1, s2, s3...,}
Where S is main set of sever like s1, s2, s3....

3. Identify the set of association rule generated AR= {ar1, ar2, ar3}

Where AR is main set of association rules generated ar1, ar2, ar3

4. Identify the set of Encryption key generated. E= {e1, e2, e3...}

Where E is main set of Encryption Key Generated e1, e2, e3, ...

5. Identify the set of GFIL list generated.

G= {g1, g2, g3,}
Where G is main set of GFIL list generated g1, g2, g3....

6. Identify the processes as P.

P= {Set of processes}
P = {P1, P2, P3, P4,}
P1 = {e1, e2, e3}
Where
{e1= Making Client-Server Connection}
{e2= Association Rule Generation}
{e3= GFIL List Generated}

7. Identify failure cases as FL

Failure occurs when
FL= {F1, F2, F3...}
F1= = {f—f if the Client server connection is not established}

8. Identify success case SS:- Success is defined as-

SS= {S1, S2, S3, S4}
(a) S1={sj s if Client Server Connection is established g
(b) S2= {sj s if Association Rules is generated}
(c) S3={s j s if GFIL list Is generated}

9. Initial conditions as I0

(a) User wants to set client server connection

IV. CONCLUSION

Here proposed the method for securing the mining of associations rule in both horizontal and vertical partitioned databases. We explain about the basic overview of dispersed association rule mining and also we discuss about the framework of agent based dispersed data mining. We also discuss about the framework of the existing method of the data mining. Here, the architecture for mobile agent based dispersed association rule mining is designed. The architecture is used for dropping the communication overhead. This architecture ensures the security for the mobile agent. We also provide security to the proposed system by using the RC4 algorithm. By using this algorithm, the fake rules and duplicate rules generated in the framework have removed.

REFERENCES

[1] Dr (Mrs).Sujni Paul, "An Optimized Distributed Association Rule Mining Algorithm in Parallel and Distributed Data Mining with XML Data For Improved Response Time. ", International Journal of Computer Science and Information Technology, Volume 2, Number 2, April 2010.
[2] T. Anu Radha, P. Lavanya, "Recent Trends in Parallel and Distributed Apriori Algorithm", International Journal of Engineering Research and Applications, Vol. 1, Issue 4, pp.1820-1822
[3] R. Agrawal and J. Shefer, "Parallel Mining of Association Rules", IEEE transaction 6 December 1996
[4] Divya Bansal, LekhaBhambhu, "Execution of APRIORI Algorithm of Data Mining Directed Towards Tumultuous Crimes Concerning Women ", International Journal of Advanced Research in Computer Science and Software Engineering.
[5] R. Agrawal, T. Imielinski, and A. Swami, "Mining Associations between Sets of Items in Massive Databases,"Proceedings of the ACM SIGMOD, Washington DC, 1993.
[6] Matthias Klusch , Stefano Lodi , Gianluca Moro, "Agent-Based Distributed Data Mining: The KDEC Scheme "
[7] Stolfo, Prodromidis, et al "JAM: Java agents for meta-learning over distributed databases"in 1997.
[8] E.I. Aria, M. B. Senousy and M. M. Medhat, "Information and E-business model application for distributed data mining using mobile agents", Proceedings of the international conference WWW/ Internet, USA, 2003.
[9] Yun-Lan Wang, Zeng-Zhi Li and Hai-Ping Zhu, "Mobile Agent Based Distributed and Incremental Techniques for Association Rules". In Proceeding of the Second International Conference on Machine Learning and Cybernetics, 2003.
[10] U.P.Kulkarni, P.D.Desai, Tanveer Ahmed, J.V.Vadavi and A.R.Yardi, "Mobile Agent Based Distributed Data Mining", ICCIMA, 2007.
[11] WalidAdlyAtteya, KeshavDahal, M. Alamgir Hossain, "Distributed BitTable Multi-Agent Association Rules Mining Algorithm", Knowledge-Based and Intelligent Information and Engineering Systems Lecture Notes in Computer Science Volume 6881, 2011, pp 151-160.
[12] Rakesh Agrawal and Rama Krishnan Srikant, "Fast Algorithms for Mining Association Rules"
[13] FerencBodon, "A Trie-based APRIORI Implementation for Mining Frequent Item sequences", OSDM'05, August 21, 2005, Chicago, Illinois, USA.