ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 12, December 2015

Secured Association Rule Mining In Partitioned Database

Gurpreet Kaur Bhatti, Prof Ravi P. Patki

Department of Computer, D. Y. Patil College of Engineering, Talegaon, Pune, India

ABSTRACT: From last few years Data mining becomes a vital area of research for researcher. Data mining aims to find hidden information from large database although secret data is kept safely when data is allowed to access by single person. Association rule mining (ARM) is the vital technique in the field of data mining used in the distributed database and have received much attention from the database community. Association rule mining (ARM) is applicable in the various fields such as Banking, department stores etc. Main objective of ARM is to extract interesting correlation, frequent pattern, association or casual structure among set of item in the transaction database or other data repositories. In proposed work we design secured association rules from the distributed vertically partitioned database with minimum memory and communication overhead by eliminating duplicate and fake rules to increase performance and security by using RC4 algorithm.

KEYWORDS: Privacy, Association rule mining, data hiding.

I. INTRODUCTION

Tremendous growth in the IT field and the problems addressed during the storage of the huge data is the major problem. For discovering association between the large set of data items the data mining plays a very vital role. Data mining techniques are the result of a long process of research and product development. The major types of task performed by data mining techniques are Classification, Dependence Modeling, Clustering, Regression, Prediction and Association. Data mining aims to discover secret information from large database although secret data is kept safely when data is allowed to access by single person. It is new technology which has emerged as a means of identifying patterns and trends from large quantities of data. Generating the learning from the information inserted in the database is one of the basic needs of information mining innovation. Association rule mining (ARM) is the vital technique in the field of data mining used in the distributed database and have received much attention from the database area. Association rule mining (ARM) is applicable in the various fields such as Banking, department stores etc.

A frequent item sets is the basic component of many algorithms for mining association rules for large datasets subroutine. Main disadvantage of frequent item sets are very heavy as transactions data increasing. The existing approach Fast Distributed Mining (FDM) algorithm is an unsecured distributed version of the Apriori algorithm.

The traditional algorithm of association rules discovery performs in two steps. In first step, all frequent item sets are searches. The frequent item set is the item set that is incorporated in at least minimum support transactions. In second step, the association rules with the confidence at least minimum confident are generated.

However in this paper addresses the few problems associated with the existing work, for example, correspondence overhead, adaptability, execution and security amid the safe mining of the affiliation guideline in evenly apportioned database.

The existing framework is not ready to evacuate fake and copy rules. The proposed framework works with on a level plane divided database as well as with the vertically apportioned database. The proposed framework won't just be adaptable yet additionally be secured due to the utilization of RC4 (Ron Rivest Cipher) calculation.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 12, December 2015

II. LITERATURE SURVEY

In paper [1] author proposed a novel and a flexible framework called mobile-agent-based distributed algorithm for association rules (IDMA) with aim of to mine the global and local large itemsets at the same time. To mine the local large itemsets the author employed the incremental algorithm (IAA) which is in turn based on the heuristic selective scan technique to reduce the number of database scans required and to keep the size of the candidate itemset sets from increasing exponential. Evaluation illustrates that the algorithm IDMA is valid and has superior performance.

In paper [2] author explained displays a productive continuous information base structural engineering for multi-specialists based patient demonstrative framework for unending malady administration, fundamentally, the early recognition of Inflammation of urinary bladder and Nephritis of renal pelvis cause sicknesses. The model incorporates data put away heterogeneous and topographically appropriated social insurance focuses. The paper presents two fundamental commitments. Initial, proposed multi-operators based framework for mining continuous item-sets in dispersed databases. Second, the usage of this model on circulated restorative databases with a specific end goal to create shrouded therapeutic tenets. The proposed model can assemble data from every office or from distinctive healing facilities, and utilizing the helpful operators it breaks down the information utilizing affiliation rules as an information mining procedure. The proposed model enhances the demonstrative information and finds the infections in light of the base number of powerful tests, in this manner, giving precise medicinal choices in view of financially savvy medications. It can likewise anticipate the presence or the nonattendance of the illnesses, in this manner enhancing the medicinal administration for the patients. The proposed multi-specialists framework constitute an exertion toward the configuration of clever, adaptable, and incorporated extensive scale disseminated information mining framework.

In paper [3] author explained Many current data mining tasks can be accomplished successfully only in a distributed setting. The field of distributed data mining has therefore gained increasing importance in the last decade. The Apriori algorithm by Rakesh Agarwal has emerged as one of the best Association Rule mining algorithms. Ii also serves as the base algorithm for most parallel algorithms. The enormity and high dimensionality of datasets typically available as input to problem of association rule discovery, makes it an ideal problem for solving on multiple processors in parallel. The primary reasons are the memory and CPU speed limitations faced by single processors. In this paper an Optimized Distributed Association Rule mining algorithm for geographically distributed data is used in parallel and distributed environment so that it reduces communication costs. The response time is calculated in this environment using XML data.

In paper [4] author explained Distributed Data Mining (DDM) is worried with use of the traditional Data Mining (DM) approach in a Distributed Computing (DC) situations so that the accessible asset including correspondence systems, registering units and disseminated information archives, human elements and so forth can be used betterly and on-line, ongoing choice bolster based conveyed applications can be outlined. A Mobile Agent (MA) is a self-governing transportable system that can move under its own or host control starting with one hub then onto the next in a heterogeneous system. This paper highlights the specialists based methodology for mining the affiliation rules from the disseminated information sources and proposed an another system called Agent enhanced Mining of Strong Association Rules (AeMSAR) from Distributed Data Sources. As specialists innovation worldview of the DC has picked up loads of exploration in the late years, along these lines, making a collusion of operators and Association Rules Mining(ARM) will offer mining the Association some assistance with ruling in a Distributed domain betterly.

In paper [5] author explained Association Rule Mining is a popular and well researched method for discovering interesting relations between variables in large databases. It is intended to identify strong rules discovered in databases using different measures of interestingness. Most ARM algorithms focus on a sequential or centralized environment where no external communication is required. Distributed ARM algorithms, aim to generate rules from different data sets spread over various geographical sites; hence, they require external communications throughout the entire process. Distributed ARM is one of the major research fields of Data Mining (DM). DARM algorithm efficiency is highly dependent on data distribution. The paper reviews different algorithms developed for DARM and also discusses the different ways in which data is distributed. Agents are software entities developed to make distributed computing more

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 12, December 2015

efficient. They have also been used in Data Mining. The paper discusses the role of agents in DARM.

In paper [6] author explained the investigation of huge datasets has turned into an imperative apparatus in comprehension complex frameworks in territories, for example, financial aspects, business, science and designing. Such datasets are regularly gathered topographically conveyed way and can't practically speaking be assembled into a single store. Applications that work with such datasets can't control most parts of the information's dividing and courses of action. As such, consideration in information mining procedure has dependably concentrated on extricating data from information physically situated at one focal site and they frequently don't consider the asset limitations of disseminated and versatile situations. Few endeavors were additionally made in parallel information mining. However most genuine applications depend on information appropriated in a few areas. As a result both new architectures and new calculations are required. In this paper creator proposes a strategy that investigates the abilities of portable specialists to assemble a fitting edge work and a calculation that better suits the Appropriated Data Mining applications. It additionally makes the execution investigation and correlation with the current such strategy.

In paper [7] author explained an efficient real-time knowledge base architecture for multi-agent based patient diagnostic system for chronic disease management, basically, the early detection of Inflammation of urinary bladder and Nephritis of renal pelvis origin diseases. The model integrates information stored heterogeneous and geographically distributed healthcare centers. The paper presents two main contributions. First, a proposed multi-agent based system for mining frequent itemsets in distributed databases. Second, the implementation of this model on distributed medical databases in order to generate hidden medical rules. The proposed model can gather information from each department or from different hospitals, and using the cooperative agents it analyzes the data using association rules as a data mining technique. The proposed model improves the diagnostic knowledge and discovers the diseases based on the minimum number of effective tests, thus, providing accurate medical decisions based on cost effective treatments. It can also predict the existence or the absence of the diseases, thus improving the medical service for the patients. The proposed multi-agent system constitutes an effort toward the design of intelligent, flexible, and integrated large-scale distributed data mining system.

In paper [8] author explained that Very often data relevant to one search is not located at a single site, it may be widely-distributed and in many different forms. Similarly there may be a number of algorithms that may be applied to a single Knowledge Discovery in Databases (KDD) task with no obvious “best” algorithm. There is a clear advantage to be gained from a software organization that can locate, evaluate, consolidate and mine data from diverse sources and/or apply a diverse number of algorithms. Multi-agent systems (MAS) often deal with complex applications that require distributed problem solving. Since MAS are often distributed and agents have proactive and reactive features, combining Data Mining (DM) with MAS for Data Mining (DM) intensive applications is therefore appealing. This thesis discusses a number of research issues concerned with the viability of Multi-Agent systems for Data Mining (MADM). The problem addressed by this thesis is that of investigating the usefulness of MAS in the context of DM. This thesis also examines the issues affecting the design and implementation of a generic and extendible agent-based data mining framework. The principal research issues associated with MADM are those of experience and resource sharing, flexibility and extendibility, and protection of privacy and intellectual property rights. To investigate and evaluate proposed solutions to MADM issues, an Extendible Multi-Agent Data mining System (EMADS) was developed. This framework promotes the ideas of high-availability and high performance without compromising data or DM algorithm integrity. The proposed framework provides a highly flexible and extendible data-mining platform. The resulting system allows users to build collaborative DM approaches. The proposed framework has been applied to a number of DM scenarios. Experimental tests on real data have confirmed its effectiveness.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 12, December 2015

III. IMPLEMENTATION DETAILS

A.System Architecture

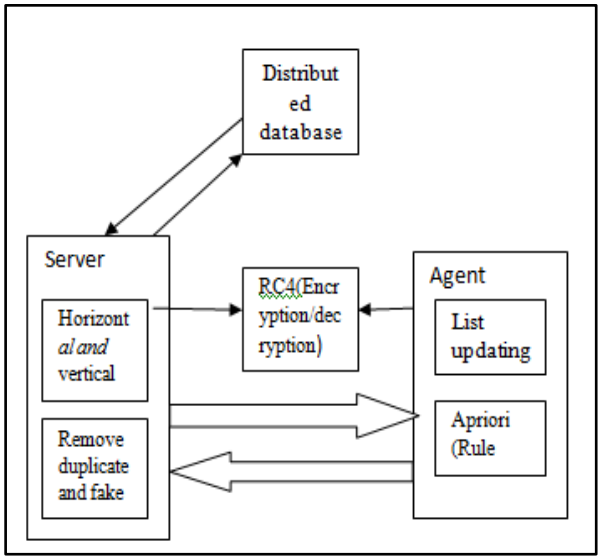


Figure1: System architecture

The section explains the overall architecture of proposed system. It explains how association rules get generated using partitioning technique. To do so, system has to traverse into number of modules. Explanation of each module is given below,

- *Distributed Database:* This module consists of the database from various shops, for which, association rules will get generated. This database will be given as input to proposed system, which acts as server in this architecture.

- *Server:* The server is responsible for horizontal and vertical partitioning. The server, fetch the data from database. While fetching it does horizontal partitioning on it. After it gets the partitioned data, it performs vertical partitioning on that data. Server also passes the generated association rules to the Distributed database from where it fetches the data.

- *Horizontal and Vertical Partitioning:* This module is within server. It performs horizontal and vertical partitioning of the database.

- *Fake and Duplicate Rules:* Server is also responsible for removing fake and duplicate rules.

- *RC4:* RC4 module is nothing but the encryption and decryption algorithm which is applied on the data that is passed between server and client.

- *Agent:* Agents are responsible for generating association rules. It consists of Apriori module which will be used to generate association rule. Agents also do list updating. After generating association rules, agents send those rules to the server.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 12, December 2015

B. Algorithm

Frequent Itemset Generation

Input: Vertically partitioned database
Output: Global List generated association rules.
DB_Data = DataBase_Partitioning (Vertical).
All_Transactions_Data=Collect_All (DB_Data);
Subparts [ ] =Divide_into_Subparts.
For (Each Agent Aj)
if (Aj is free)
encrypted_subparts= Encrypt (Subparts[i]).
End If and End For
For each agent Aj Send (encrypted_subparts)
Decrypted_data=Decrypt (Received_encrypted_subparts).
Rules = Generate_Rules (Decrypted _data).
Send _to _centralized server ((Rules)).
Merge_all_Rules _Aj (Rules)
Remove_Duplicate _Fake (Rules);
Generate_Final_Association_Rules

C. Mathematical Model:

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].

1. Identify the set of clients
C= {c1, c2, c3 ...cn},
Where C is main set of Clients like c1, c2, c3.
2. Identify the set of severs
S= {s1, s2, s3... sn}
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 ...en}.
Where E is main set of Encryption Key Generated e1, e2, e3.
5. Identify the set of GFIL list generated.
G= {g1, g2, g3 ...gn}.
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}

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 12, December 2015

{e3= GFIL List Generated}
7. Identify failure cases as FL
Failure occurs when
FL={FL=F1, F2, F3,...}
F1= = is 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}
(b) S2 = {sj s if Association Rules is generated} (c) S3 = {sj s if GFIL list is generated}
9. Initial conditions as I0
(a) User wants to set client server connection

IV. RESULT & DISCUSSION

Below graph shows the Total Computation Time. This graph represents Number of transaction in whole dataset vs Time. Proposed system improves the time efficiency as compared to existing system as shown in the below graph.

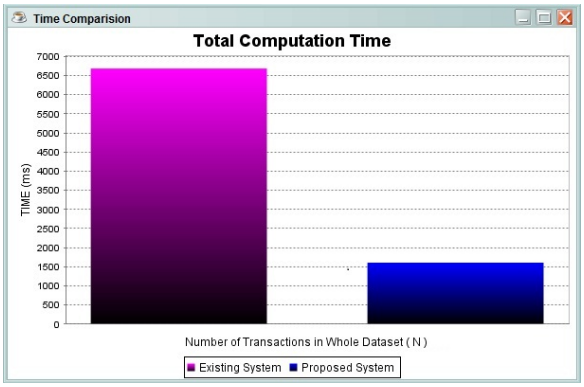


Figure2: Total Computation Time

	Existing System	Proposed System
N	*Time(ms)*	*Time(ms)*
100	887	125
300	1314	289
500	1704	480

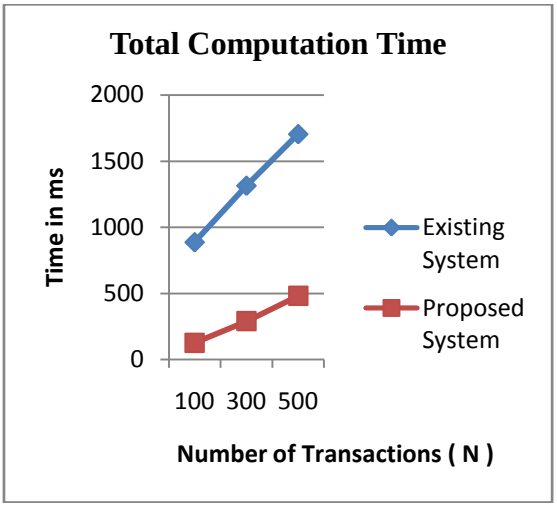
Table No. 1 Table show comparison between Proposed System and Existing system (N varies).

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 12, December 2015

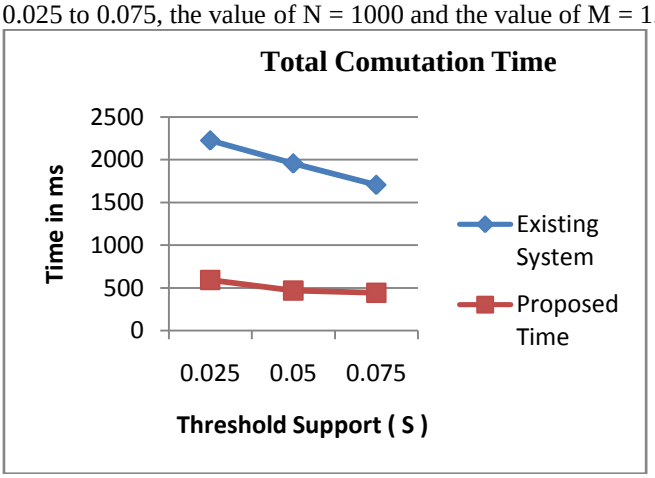


Graph No. 2 Time computation graph of Proposed System and Existing system.

In the above graph the value of threshold support (S) and number of agents (M) are constant where the value of transactions (N) are changed from 100 to 5000, the value of S = 0.1 and the value of M = 1 ,Here X-axis show the Number of Transactions (N) and Y-axis shows Time (ms).

		Existing System	Proposed System
S	N	Time (ms)	Time (ms)
0.025	1000	2226	593
0.05	1000	1956	468
0.075	1000	1705	443

In the below graph the value of transactions (N) and number of agents (M) are constant where the value of threshold support (S) are changed from 0.025 to 0.075, the value of N = 1000 and the value of M = 1.



Graph No. 3 Time computation graph of proposed system and existing system.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 12, December 2015

V. CONCLUSION

Tremendous growth in the IT field and the problems addressed during the storage of the huge data is the major problem. Data mining aims to discover secret information from large database although secret data is kept safely when data is allowed to access by single person. In this paper, we surveyed methods various papers advantages and disadvantages by identifying some open challenges that will be useful to research community in this area. This paper presents a novel association rule mining algorithm based on vertical partitioning.In proposed work we design secured association rules from the distributed vertically partitioned database with minimum memory and communication overhead by eliminating duplicate and fake rules to increase performance and security by using RC4 algorithm.

In future we try to use better association rule generation algorithm to improve the system performance in terms of time and memory.

REFERENCES

[1] 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.
[2] Divya Bansal, Lekha Bhambhu, "Execution of APRIORI Algorithm of Data Mining Directed Towards Tumultuous Crimes Concerning Women", International Journal of Advanced Research in Computer Science and Software Engineering.
[3] 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.
[4] G.S.Bhamra, "Agent Enriched Distributed Association Rule Mining: A Review"
[5] Walid Adly Atteya, Keshav Dahal, M. Alamgir Hossain, "Distributed Bit Table 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.
[6] U.P.Kulkarni, P.D.Desai, Tanveer Ahmed, J.V.Vadavi and A.R.Yardi, "Mobile Agent Based Distributed Data Mining", ICCIMA, 2007.
[7] Atteya, W.A., "Multi-agent system for early prediction of urinary bladder inflammation disease."
[8] Kamal Ali Albashiri, FransCoenen, and Paul Leng, "An investigation into the issues of Multi- Agent Data Mining" Ph.D-Thesis 2010.
[9] A. Ben-David, N. Nisan, and B. Pinkas, "FairplayMP - A System for Secure Multi-Party Computation," Proc. 15th ACM Conf. Computer and Comm. Security (CCS), pp. 257-266, 2008.
[10] H. Grosskreutz, B. Lemmen, and S. Ruping, "Secure Distributed Subgroup Discovery in Horizontally Partitioned Data," Trans. Data Privacy, vol. 4, no. 3, pp. 147-165, 2011.
[11] M. Kantarcioglu, R. Nix, and J. Vaidya, "An Efficient Approximate Protocol for Privacy-Preserving Association Rule Mining," Proc. 13th Pacific-Asia Conf. Advances in Knowledge Discovery and Data Mining (PAKDD), pp. 515-524, 2009.
[12] L. Kissner and D.X. Song, "Privacy-Preserving Set Operations," Proc. 25th Ann. Int'l Cryptology Conf. (CRYPTO), pp. 241-257, 2005.
[13] X. Lin, C. Clifton, and M.Y. Zhu, "Privacy-Preserving Clustering with Distributed EM Mixture Modeling," Knowledge and Information Systems, vol. 8, pp. 68-81, 2005.
[14] Y. Lindell and B. Pinkas, "Privacy Preserving Data Mining," Proc. Crypto, pp. 36-54, 2000.
[15] T. Tassa and D. Cohen, "Anonymization of Centralized and Distributed Social Networks by Sequential Clustering," IEEE Trans. Knowledge and Data Eng., vol. 25, no. 2, pp. 311-324, Feb. 2013.
[16] T. Tassa and E. Gudes, "Secure Distributed Computation of Anonymized Views of Shared Databases," Trans. Database Systems, vol. 37, article 11, 2012.
[17] T. Tassa, A. Jarrous, and J. Be Ya'akov, "Oblivious Evaluation of Multivariate Polynomials," J. Mathematical Cryptology, vol. 7, pp. 1-29, 2013.
[18]J. Zhan, S. Matwin, and L. Chang, "Privacy Preserving Collaborative Association Rule Mining," Proc. 19th Ann. IFIP WG 11.3 Working Conf. Data and Applications Security, pp. 153-165, 2005.
[19] S. Zhong, Z. Yang, and R.N. Wright, "Privacy-Enhancing k Anonymization of Customer Data," Proc. ACM SIGMOD-SIGACTSIGART Symp.Principles of Database Systems (PODS), pp. 139-147, 2005.
[20] J. Brickell and V. Shmatikov, "Privacy-Preserving Graph Algorithms in the Semi-Honest Model," Proc. 11th Int'l Conf. Theory and Application of Cryptology and Information Security (ASIACRYPT), pp. 236-252, 2005.