ISSN:2249-5789
Bhagyashri Waghamare et al , International Journal of Computer Science & Communication Networks,Vol 5(2),58-61

Review:Association Rule for Distributed Data

Bhagyashri Waghamare
Department of Computer Engg
Flora Institute of Technology, Pune
Maharashtra, India
bshri1992@gmail.com

Bharat Tidke
Department of Computer Engg
Flora Institute of Technology, Pune
Maharashtra, India
batidke@gmail.com

***Abstract*— we consider the problem of association rules for distributed database. There have been many research studies done on association rules in distributed databases. However, it is important to maintain such discovered rules in distributed databases because a database may allow frequent or timely updates for generating strong association rules. In this study, a many algorithms is proposed and implemented for generating and mining efficient association rules from distributed data for finding frequent item sets.**

Keywords—Data Mining; Association Rule; Disributed Data; Frequent Iemsets Mining

I. INTRODUCTION

A. Association Rules

Association Rule Mining (ARM) is the most important and researched techniques of data mining. ARM was first introduced by Agrawal et al. 1993 [1].It is association tools for analyzing customer purchasing habit, such as market-basket analysis. ARM aims to extract interesting frequent patterns, association among set of items or database. Association Rules are if/then statements that help to discover relationships among unrelated data in a data repository. Many algorithms are proposed for finding frequent itemsets for large datasets Association rule uses two criteria support and confidence to identify the relationships and rules are generated by analyzing data for frequent if/then pattern. Association rules are generally needs to satisfy a minimum support and a minimum confidence at the same time.

B. Distributed Association Rule Mining

In a distributed environment, data comes from multiple remote sources. Such an environment there is high communication overhead and wastes of resources when data is dynamic. In this situation, how to minimize the communication cost, how to combine frequency counts from multiple nodes, and how to mine data in parallel and update the associated information incrementally are additional issues we need to consider.

Association rule mining usually split into two separate steps:

1. Apply minimum support value to find all frequent item sets in a database. This steps required more attention.
2. Form rules by using frequent item sets and the minimum confidence value.

Support(S)-It is the percentage of records that holds union of X and Y to the total number of records in the database.

Confidence(C)-It is percentage of the number of transactions that contain union of X and Y to the total number of records that include X. Confidence is a measure of strength of the association rule.

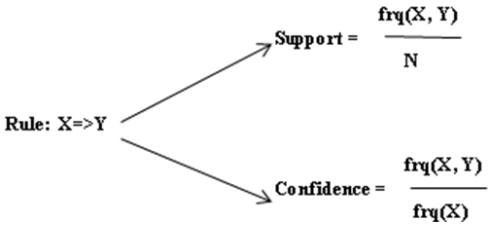

Fig. 1.Association Rule.

Association Rule Mining is to find the association rules that satisfy the user related minimum support and confidence from a given database. ARM is normally divided into two sub problems. First is to locate those item sets whose occurrences more than user related threshold in the database, these item sets are defined as frequent item sets. The Second problem is to generate association rules from those large item sets with a limitation of minimum confidence value.

C. Steps of Generating Association Rule

An association rules generation has the following steps:

1. The set of candidate k-item sets is generated by 1-extensions of the large (k-1) item sets generated in the previous slot.
2. Supports for the candidate k-items sets are generated by a scan over the database.
3. Item sets that do not have the minimum support are deleted and the remaining itemsets are called large k-item sets.

D. Generation Model of Association Rule

Association Rules Generation contains many process, they can easily understand by the following related model. The model of Association Rule Generation is in Fig. 1.2.

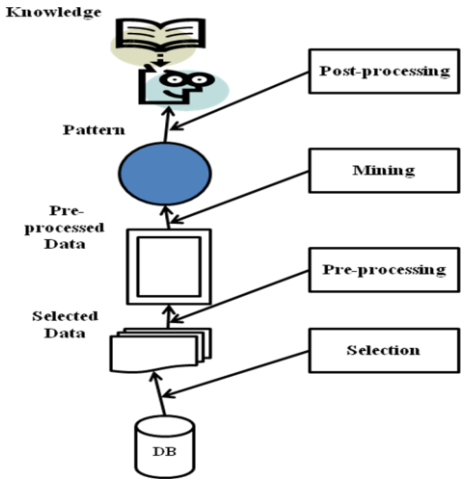

Fig. 2.Model of Association Rule Generation.

Association Rule Generation contain of processes as selection of database from the large repository, then preprocessing on selected data, after that mine the candidate frequent itemsets from the preprocessed data, then prune the frequent itemsets according to a given threshold. Such a way rules are generated then association rules are mined according to given support and confidence.

ARM is mainly used for mining the frequent and infrequent itemsets from the large databases. It is based on two principles - support and confidence. Association rules are the if/then sentences to show the relationship among the item sets. One of the most important ARM algorithms is Apriori algorithm. Section 2, it gives the survey of papers on Apriori algorithms with their merits and demerits.

II. Literature Survey

Mining Association Rules is one of the most important in data mining. Association rules are of interested in database researchers and data mining users. Since 90s, different approaches of data mining have been proposed for discovering useful knowledge from very large datasets [10]. A survey of previous research in the area is provided below.

- In 1994 R. Agrawal and R. Srikant was proposed an Apriori algorithm. The Apriori algorithm first finds all frequent itemsets and then generates association rules from the frequent itemsets. The process is repeated until no more frequent itemsets can be found. However, the performance of the Apriori algorithm is a major problem in finding frequent itemsets. More time is spent in dealing with the creations of candidate itemsets. The algorithm has linear dependence on the size of the dataset but the exponential growth on the size of the itemset. Thus, the time spent for the algorithm is considerable.[3]
- In 2001 Schuster and Wolff proposed a distributed algorithm called The Distributed Decision Miner (DDM), this algorithm belongs to the group of Apriori-based algorithms assuming a shared-nothing architecture as well. Here, after local frequency counts are computed on each node, the nodes perform a distributed decision protocol in each round in order to determine the set of globally frequent itemsets [8].
- In 2001 Zaiane et al. proposed a parallel algorithm that is based on frequent pattern –growth algorithm (fp-growth).The algorithm is called MLFPT (Multiple Local Frequent Pattern Tree). It assumes shared-memory architecture. Just like the centralized fp-growth algorithm, MLFPT does not generate candidates for frequent itemsets but instead builds multiple frequent pattern trees (FP-trees) [9].
- In 2003 Otey et al. proposed an algorithm named ZigZag. This algorithm assumes a shared-nothing architecture and a setting where the data is initially distributed on different sites (like network data for intrusion detection) [12].
- In 2003 Schuster, Wolff, et al. proposed a distributed sampling algorithm called D-Sampling. This algorithm is a combination of a centralized sampling algorithm and the DDM algorithm Schuster and Wolff presented in 2001. It assumes a shared-nothing architecture. D-Sampling assumes a centralized dataset and distributes it during runtime. Each node gets the "responsibility" for a set of items. The algorithm loads a sample of the dataset into memory. This sample is then distributed according to the responsibility of the different nodes, fragmenting the dataset vertically [13].
- In 2005 Emad Kadum Jabbar proposed algorithm called Association Rule with logical AND operation, which aims to produce association rules depending on logical AND operation by convert the database transaction into binary representation and neglecting any sum (column) less than threshold to find identical column in (k-1)-itemset table with column in k-itemset table which represents the association rules [18].
- In 2005 Claudio Silvestri, Salvatore Orlando proposed algorithm called Distributed Approximate Mining of Frequent Patterns. The proposed algorithm consists in the distributed exact computation of locally frequent itemsets and an effective method for inferring the local support of locally unfrequented itemsets [19].
- In 2007 Rawia Tahrir Salih proposed a new algorithm for distributed association rules called Extracting Association Rules for Distributed Association Rules (EAR4DAR) Algorithm; which aims to extract association rules for distributed association rules instead of extracting association rules from huge quantity of distributed data at several sites, and that is through collecting the local association these Local Association Rules turn in series of operations to produce global association rules over distributed systems [23].
- In 2007 Lamine M. Aouad, Nhien-An Le-Khac and Tahar M. Kechadi, proposed a distributed algorithm for frequent itemsets generation on heterogeneous clusters and grid

environments called Distributed Frequent Itemsets Mining in Heterogeneous Platforms. The proposed approach uses a dynamic workload management through a block-based partitioning, and takes into account inherent characteristics of the Apriori algorithm related to the candidate sets generation [24].

- In 1996 Cheung proposed the Fast Distributed algorithm (FDM) to mine rules from distributed data sets partitioned among different sites .FDM finds the local support counts and prunes all infrequent local candidate sets. After completing local pruning, each site broadcasts messages containing all the remaining candidate sets to all other sites to request their support counts. It decides whether large itemsets are globally frequent and generates the candidate itemsets from the globally frequent itemsets.[4]

TABLE I. COMPARATIVE ANALYSIS OF DIFFERENT FIM TECHNIQUES

Author	Works on Distributed Data	Algorithm	Characteristics	Benefits	Limitations
Agrawal et al. [3]	No	Apriori	Level wise search, Monotonicity property and Easy to implement	Generates frequent itemsets, candidate itemsets and association rules	Cannot handle large datasets and requires up to n number of scans, n – number of items in itemset
Han et al. [8]	No	FP-Growth	Recursive approach, Employs divide- and conquer method and FP-tree data structure	No candidate generation and repeated scans, Focused search of smaller databases	Recursive construction of the FP-tree affects the algorithm's performance
Cheung et al.[4]	Yes	FDM	Fat Distributed mining, Portioning approach	Generation among different sites	Local support count and pruning
Schuster et al.[8]	Yes	DDM	Distributed decision mining, Apriori based algorithm, Shared nothing architecture, Local frequency count	Generate set of global frequent itemsets	Local data sets
Zaiane et al.[9]	Yes	MLFPT	Multiple local frequent pattern tree, Parallel algorithm, Shared nothing architecture	Generate multiple frequent pattern trees	Not generate frequent item sets
Octey et al.[12]	Yes	Zigzag	Shared nothing architecture	Generate data to near on different sites	Local data sets
Schuster et al.[13]	Yes	D-Sampling	Centralized sampling algorithm, Distributed decision mining, Shared nothing architecture	Fragmenting dataset vertically	Apply on centralized dataset
Emad et al.[18]	Yes	ARM	Association rule mining, Logical AND operation, Convert DB transaction into binary representation	Need less space than threshold to find identical column	Depend on logical AND operation
Claudio et al.[19]	Yes	DAM	Distributed Approximate Mining	Generate locally frequented itemset	Local data sets
Rawla et al.[23]	Yes	EAR4DAR	Extracting association rules for distributed association Rules algorithm, Extracting, Distributing at several sites	Generate local association rule from each sites and storing them	Not generate global association rule
Lamine et al.[24]	Yes	DFISM	Distributed frequent item set mining, Heterogeneous users, Grid environment	Dynamically workload management, Block-based portioning	Local data sets
Yan Zhao et al.[22]	Yes	ED-ARM	Efficient distributed association	Finding large item sets, Greater performance	Local data sets

III. CONCLUSION

In this paper, authors have studied the various variations of association rule algorithm for distributed data. This review is focused on how to solve which is efficient algorithm for association rule generation for distributed data that further use for the reduction of association rule in distributed data.

Acknowledgment

Authors would like to express gratitude to Prof. Bharat Tidke, Head of the Computer Science and Engineering Department, Flora Institute of Technology, Pune Maharashtra, India for his guidance and support in this work. The authors are also thankful to the Principal, Flora Institute of Technology, Pune Maharashtra, India for being a constant source of inspiration.

References

[1] R. Agrawal, T. Imielinski, and A. Swami, “Mining association rules between sets of items in large databases," Proceedings of ACM SIGMOD Record, vol. 22, no. 2. ACM, 1993, pp. 207-216.

[2] Matthias Klusch, Stefano Lodi and Gianluca Moro,“Agent based distributed data mining: The KDEC Scheme”.R. Agrawal, T. Imielinski, and A. Swami, “MiningAssociations between Sets of Items in Massive Databases,” Proceedings of the ACM SIGMOD,Washington DC, 1993.

[3] R. Agrawal and R. Srikant, “Fast algorithms for mining association rules,” Proceedings of the 20th International Conference on VeryLarge Data Bases, pp. 487-499, 1994.

[4] D. W. Cheung, J. Han, V. Ng, A. Fu, and Y. Fu. “A fast distributed algorithm for mining association rules”. In Proc. 1996 Int. Conf. Parallel and Distributed Information Systems,pages 31–44, Miami Beach, FL, Dec. 1996.

[5] D. W. Cheung, V. T. NG, A. W. Fu, and Y.J. Fu. “Efficient mining of association rules in distributed databases. Special Issue in Data Mining”. IEEE Transactions on Knowledge and Data Engineering, 8(6):911–022, 1996.

[6] Naghibzadeh, M. (1998). “Modeling and performance evaluation of distributed system with coordinator”. Iranian Journal of Science and Technology, Transaction B: Engineering, Vol. 22, No. B3, pp 317-328.

[7] M. J. Zaki. “Parallel and distributed association mining: A survey”. IEEE Concurrency, pages 14–25, 1999.

[8] Yijun, Xuemin Lin, and C. Tsang, “An Efficient Distributed Algorithm for Computing Association Rules”. Springer-Verlag Berlin Heidelberg 2000.

[9] Assaf Schuster and Ran Wolff. “Communication-Efficient Distributed Mining of Association Rules”. In SIGMOD ’01: Proceedings of the 2001 ACM Sigmod International Conference on Management of Data, pages 473–484, New York, NY, USA, ACM Press.May 2001.

[10] Osmar R. Zaiane, Mohammad El-Hajj, and Paul Lu. “Fast Parallel Association Rule Mining without Candidacy Generation”. In ICDM, pages 665–668, 2001.

[11] Rawia Tahrir Salih Kadoori,“Extracting Association Rules From Distributed Association Rules”.MSc. Thesis Computer Science, University of Technology, 2002.

[12] B.-H. Park and H. Kargupta. “Distributed data mining: Algorithms, systems, and applications”. 2002.

[13] M.E. Otey, C. Wang, S. Parthasarathy, A. Veloso, and Jr. W. Meira. “Mining Frequent Itemsets in Distributed and Dynamic Databases”. In ICDM 2003: Third IEEE International Conference on Data Mining, pages 617– 620, Nov. 2003.

[14] Ran Wolff Assaf Schuster**, “**A High-Performance Distributed Algorithm for Mining Association Rules”. IEEE Conference on Data Mining (ICDM), Florida, 2003.

[15] E. I. Ariwa, 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.

[16] 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.

[17] Assaf Schuster, Ran Wolff, and Dan Trock. “A high-Performance Distributed Algorithm for Mining Association Rules”. In ICDM ’03: Proceedings of the Third IEEE International Conference on Data Mining, page 291, Washington, DC, USA, 2003.

[18] Ashrafi, M. Z., Taniar, D. & Smith, K. (2004). “ODAM: an optimized distributed association rule mining algorithm”. IEEE distributed systems online, Vol. 05, No. 3.

[19] Emad Kadum Jabbar, “New Algorithms for Discovering Association Rules”. PHD. thesis, Department of Computer Sciences of the University of Technology, 2005.

[20] Claudio Silvestri, Salvatore Orlando, “Distributed Approximate Mining of Frequent Patterns”. ACM Symposium on Applied Computing,Italy,2005.

[21] Schuster, A., Wolf, R. & Trock, D. (2005). “A high-performance distributed algorithm for mining association rules”. Knowledge And Information Systems (KAIS) Journal, Vol. 7, No. 4.

[22] Yan Zhao, Yong Yao“An Efficient Distributed Algorithm for Mining Association Rules” International Conference on Fuzzy Systems and Knowledge Discovery,2007

[23] U. P. Kulkarni, P. D. Desai, Tanveer Ahmed, J. V. Vadavi and A. R. Yardi, “Mobile Agent Based Distributed Data Mining”, ICCIMA, 2007.

[24] Nhien An Le Khac, Lamine M. Aouad and M-Tahar Kechadi, “Distributed Knowledge Map for Mining Data on Grid Platforms”. IJCSNS International Journal of Computer Science and Network 98 Security, VOL.7 No.10, October 2007.

[25] Nhien An Le Khac, Lamine M. Aouad and M-Tahar Kechadi,” Distributed Knowledge Map for Mining Data on Grid Platforms”. IJCSNS International Journal of Computer Science and Network 98 Security, VOL.7 No.10, October 2007.

[26] Lamine M. Aouad, Nhien-An Le-Khac and Tahar M. Kechadi, “Distributed Frequent Itemsets Mining in Heterogeneous Platforms”. Journal of engineering, computing and architecture, Volume 1, Issue 2, School of Computer Science and Informatics University College Dublin, 2007.

[27] M. Deypir and M. H. Sadreddini, “Distributed Association Rules Mining Using Nonderivable Frequent Patterns” Iranian Journal of Science & Technology, , Vol. 33, No. B6, pp 511-526,2009

[28] Walid Adly Atteya, Keshav Dahal and M. Alamgir Hossain, “Distributed BitTable multi-agent Association Rules Mining Algorithm” , Springer- Verlag, KES 2011, Part I, LNAI 6881.

[29] O. Ogunde, O. Folorunso, A. S. Sodiya and G. O. Ogunieye, “ A review of some issues and challenges in current agent based distributed association rule mining”, Asian Journal of Information Technology,2011.

[30] G. S. Bhamra, A. K. Verma and R. B. Patel, “Agent Enriched Distributed Association Rule Mining: A Review”. Springer Verlag Berlin Heidelberg, 2012.

L. Zeng, L. Li, L. Duan, K. Lu, Z. Shi, M. Wang, W. Wu, and P. Luo. “Distributed data mining: a survey”. Information Technology and Management, pages 403–409, 2012.