International Journal of Computer Science and Engineering (IJCSE)
ISSN(P): 2278-9960; ISSN(E): 2278-9979
Vol. 3, Issue 1, Jan 2014, 57-62

International Academy of Science, Engineering and Technology
Connecting Researchers; Nurturing Innovations

ISSUES AND CHALLENGES ASSOCIATED WITH ASSOCIATION RULES MINING ALGORITHMS

RUCHIKA YADAV[1], KANWAL GARG[2] & PRIYANKA KHURANA[3]

[1]Research Scholar, Department of Computer Science and Application, Kurukshetra University, Haryana, India

[2]Assistant Professor, Department of Computer Science and Application, Kurukshetra University, Haryana, India

[3]Assistant Professor, Department of Information Technology, HCTM, Kaithal, Haryana, India

ABSTRACT

The process of discovering frequent item sets using association rules is the most significant tasks in data mining. Various proficient algorithms are available in the literature for mining such frequent item sets. A critical review of these algorithms emphasized on designing of an efficient multiple level association rules algorithm and data structure that is able to reduce number of iterations, which further may achieve time efficiency. This paper provides a comprehensive survey for number of the state-of-the-art algorithms on mining frequent item sets using single or multiple association rules. The premise of this paper is to identify various interesting issues and challenges associated with existing research and to propose substantive areas for future research for upcoming researchers.

KEYWORDS: Association Rule Mining, Frequent Item Sets

INTRODUCTION

Association rule mining is an important method for asserting important pattern/trend. Data mining techniques are broadly classified into two categories i.e. traditional method and next generation method. Association rule mining come under the preview of next generation method. Association rule mining meets the problem related to Market Basket Analysis. This technique is used to predict frequent item sets, pattern and trend amongst those items. For a meaningful and accurate prediction; high volume of preprocessed data in the form of data warehouse is prerequisite. Support and Confidence are two key parameters in association rule mining. Association rule mining is divided into two phases. First is to find the item sets whose occurrences exceed a predefined threshold in the database; those item sets are called frequent or large item sets. And the second is to generate association rules from large item sets with the constraints of minimal confidence.

Mining association rules is particularly useful for discovering relationships among items from large databases [23]. The extraction of interesting correlations, frequent patterns, associations or casual structures among sets of items in the transaction databases or other data repositories is the main objective of ARM [11]. The implementation of ARM procedures helps in ascertaining the descriptive models of data mining.

LITERATURE SURVEY

To carry out, the present review work, the survey has been divided in to two sections i.e. section I and section II. Section I emphasized on outcomes based on single level association rules and section II anticipated the outcome based on multiple level association rules. The upcoming paragraphs gives a detailed overview of single level and multiple level association rules follow by performance of these algorithms along with their memory perception.

Section I

Initially Agrawal, R., Imielinski, T., and Swami, A. N. 1993 [1] proposed an algorithm for frequent item set mining is called Apriori. This algorithm consists of two phases; frequent item sets generation and the association rules generation. Agrawal Rakesh, and Srikant Ramakrishnan, 1994 [2] along with Apriori gives the enhancement i.e. AprioriTid and AprioriHybrid algorithms as well. Park J.S., Chen M.S, and Yu P.S., 1995 [13] explored another optimization Direct Hashing and Pruning (DHP). DHP is used to reduced the size of candidates (k+1) - item sets by hashing alike apriori. Savasere Ashok, Omieinski Edward and Navathe Shankant, 1995 [17] proposed completely different methodology from the above mentioned algorithms.

In this case the database partitioned in to blocks in which local frequent itemsets were found. In partition algorithm the support count is determined by the intersection of the transaction ids. Brin S., Motwani R., Ullman J. D., and Tsur S., 1997 [3] described the problem associated with Apriori i.e. large number of scanning of database. A new algorithm Dynamic Itemset Counting (DIC) was introduced to decrease number of scans as well as time. Hidber C., 1999 [7] presented a new algorithm Continuous Association Rule Mining Algorithm (CARM). CARM is another method which uses the DIC like approach in order to restrict the interval size M to 1. Zaki M.J., Parthasarathy S., Ogihara M. and Li W. 1997 [20] introduced a new algorithm based on depth-first search is Eclat used a vertical database representation and counted the item set supports using the intersection of tids. Another enhancement of this is dEclat developed by Zaki M. J. and Gouda K. in 2003 [21]. A new algorithm FP-growth presented by Han J., Pei J., and Yin Y., 2000 [8] is the well-known and extensively used. The FP-growth algorithm employs a divide-and-conquer approach to decompose the mining problem into a set of smaller problems. Recursive construction of the FP-tree, however, affected the algorithm's performance.

Section II

The literature studied which had been previously undertaken in the field of Multiple Level Association rules is started from 1995 to 2012. A number of algorithms are given below.

Firstly the concept of multiple level association rules was introduced by Jiawei Han and Fu Yongjian, 1995 [9]. New algorithms ML_T2L1, ML_T1LA, ML_TML1 and ML_T2LA had been developed, which were based on the top-down progressive deepening technique. Cheung W. David, Ng incent T. and Benjamin W. Tam, 1996 [5] presented an efficient algorithm named MLUp (multiple level association rules update) that updating of discovered multi-level association rules. This algorithm is applicable only to a database which allows frequent or occasional updates restricted to insertions of new transactions. A methodology that is entirely different from the previous approaches was proposed by Show-Jane and Chen Arbee L.P., 2001 [18].

The Multiple Level Association Pattern Generation (MLAPG) algorithm is generated by graph base approach. In this approach association graphs are constructed after only one scan of data base and then all large item sets are generated by traversing of that association graph. Rajkumar N., Karthik M. R. and Sivanandam S. N., 2003 [15] presented algorithm Apriori New Multi, for multilevel association rules in large database respectively.

The algorithms introduced a new concept called multi minimum support i.e. minimum support varies for different length of the item set. The frequent item sets of various lengths can be finding with the help of multi minimum support. Sharma L.K., Vyas O.P., Tiwary U.S. and Vyas R., 2005 [19] proposed a proficient approach of mining positive and negative association rules. In this approach multiple level spatial mining methods are applied to extract interesting patterns in spatial and/or non-spatial predicates. Kumar Kishore B. and Jotwani Naresh, 2006 [10] proposed a new algorithm,

efficient hierarchical online rule mining (HORM). The proposed new algorithm incorporates two specific enhancements: hierarchy-aware counting and transaction reduction. The new algorithm is modified in a natural way to model the generation of hierarchical association rules. The construction of an explicit adjacency lattice avoided by modified algorithm. Pater Mirela and Popescu Daniela E., 2008 [14] clarified that mining rules at single concept levels lead some difficulties of finding desired knowledge in databases.

This study explored new method ADA-AFOPT can be used to resolve the multilevel frequent pattern mining problem. This algorithm traverses the trees in top-down depth-first order and the items in the prefix trees are stored in ascending frequency order. Gautam Pratima and Pardasani K.R., 2010 [6] explored a new model (MLBM) for mining multilevel association rules which is based on Boolean matrix. It adopts Boolean vector relation calculus to generate frequent patterns at lower level. In this algorithm Boolean logic operations are used to generate the multilevel association rules. Mittar Vishav, Yadav Ruchika and sirohi Deepika, 2010 [12] presents an enhancement of existing algorithm MLT2_L1. This method is based on top-down progressive deepening technique.

This study is based on patterns counting inference approach, which is based on the notion of key patterns of equivalence class of pattern. Ramanaiah Preethi Kolluru, 2011 [16] developed a new hybrid algorithm for mining multilevel association rule called AC Tree (Apriori COFI), at multiple concept levels of concept hierarchy. Kousari Alireza Mirzaei Nejad, Mirabedini Seyed Javad, and Ehsan Ghasemkhani, 2012 [22] adopted a new multilevel fuzzy association rule mining model for extraction of implicit knowledge. In this method different support values at each level and different membership function each item is used. The proposed method incorporates fuzzy boundaries instead of sharp boundary intervals.

The several algorithms presented are categorized based on their performance and memory requirements and discussed concisely with comparison and other related works in the following sub-sections.

PERFORMANCE AND MEMORY EMPHASIZED WORKS

Section I

As above point out that Apriori perform better than AprioriTid in case of large size but in case of smaller size problem AprioriTid performed consistently well as Apriori. AprioriHybrid algorithm combines both Apriori and AprioriTid to perform better on problems of various sizes. Brin Sergey et al., 1997 [3] presented dynamic item set counting (DIC) algorithm which uses fewer passes over the database than traditional algorithms. They also presented a new way of finding "implication rules". These rules are normalized based on both the antecedent and the consequent. They produced more perceptive outcome than other methods.

Hidber C., 1999 [7] proposed a new algorithm named CARMA (Continuous Association Rule Mining Algorithm), is used to compute large item sets online. The memory efficiency of CARMA was greater than apriori due to restriction of the interval size to 1. The algorithm dEclat outperform Eclat due to use of difference of tids called. However, when the database is sparse, diffset will lose advantage over tidset. The FP-growth algorithm proposed by Jiawei Han et al., 2000 [8] needs two database scans as compare to Apriori [1] and its variants when mining all frequent item sets. The FP-growth method is faster than the Apriori algorithm and some other variations of Apriori. It is efficient and scalable for mining both long and short frequent patterns.

A novel method with vertical database representation is proposed by Zaki Mohammed J. et al., 1997 [20]. Instead of the tids intersection set difference of tids called diffset between a candidate k-item set and its prefix k-1- frequent items ets is stored. The size of memory required for intermediate results is significantly cut down due to diffsets. The running

time of vertical algorithm Eclat [21] was improved with the support of Diffsets. Tidset based methods are outperformed in several orders of magnitude by diffset algorithms. Bastide Yves et al., 2000 [4] considered a problem of several scans of large database during the process of frequent patterns mining. The PASCAL algorithm was the optimization of Apriori algorithm. This algorithm uses pattern counting inference, using the key patterns in equivalence classes to reduce the number of patterns counted and database passes.

Section II

Jiawei Han and Fu Yongjian, 1995 [9] proposed a new concept of multiple level association rules. New algorithm ML_T2L1 had been developed, which was based on the top-down progressive deepening technique. Some variations of this algorithm were ML_T1LA, ML_TML1 and ML_T2LA which perform better in different environments. Cheung David W., Ng incent T. and Tam Benjamin W., 1996 [5] presented an efficient algorithm named multiple level association rules update (MLUp) has superior performance in multi-level environment. HORM [10] optimizes the time requirements of the earlier reported algorithm MLT2, by hierarchy-aware counting and transaction reduction. In MLBM [6] method only one scan of database is required.

This method requires less memory space due to storage of transaction data in bits. But the complexity of multilevel association rules generation is large due to Boolean logic operations. In LWFT [12] algorithm the no. of passes over database at each concept level is reduced due to patterns counting inference approach. The size of database is reduced at each concept level due to filtration of encoded table. The experimental result show or proves that AC Tree method works faster in mining rule from large text documents. On the basis of running time and memory requirements, it can be concluded that the LWFT is better than the algorithms which are discussed in literature survey.

ISSUES AND CHALLENGES

From this comprehensive survey and study of various methods some problems and heated discussion are raised. In the light of the above, the issues and challenges in this field are:

- In single level or multiple level association rules; the first and most important issue is concerned with accurate data source in appropriate data format. Which encoding method should be used to convert the transaction tables is main issue because these encode tables are used to support the concept hierarchy of multiple levels.
- Another issue to develop/design algorithms for multiple level association rules to reduce the number of iteration and to achieve time efficiency. The time efficiency can be achieved by reduction of database scans at each level. The redundancy of association rules is a main issue in association rule discovery.
- If the interestingness parameters i.e. support and confidence thresholds are small, the number of frequent item sets increases, the number of rules presented to the user typically increases proportionately. Many of these rules may be redundant. So selection of appropriate values of interestingness parameters may be an important issue in association rule mining.
- There are so many measures of the interestingness of an association. Several interestingness metrics including support, confidence, gain, Laplace value, conviction, lift, entropy gain, gini, and chi-squared value. These measures are indicators of the degree to which items in an association are related to each other. The challenge to the users which concentrates on finding associations is to choose the user specified constraints.

- The algorithms which have lower CPU overhead and reduce the I/O overhead associated with previous algorithms are desirable.
- An another issue related to the association rules mining is to determination of usefulness of association patterns the task of decision making is found to be incorporated flawlessly within the association mining procedure.

CONCLUSIONS

Frequent item set mining and association rule mining are the two important tasks of data mining. Various proficient algorithms are available in the literature for mining frequent item sets and association rules. Discovering association rules used to ascend the business of an enterprise has long been recognized in data mining community.

In this paper, a comprehensive survey of single level and multiple level association rules mining algorithms has presented with their performance and memory perception. A brief discussion of a number of algorithms was presented along with some burning issues and challenges in the field of mining multiple level association rules. This paper focuses on frequent item set mining and has tried to cover both early and recent literature related to mining frequent item sets at each conceptual level. It is concluded that there are various issues and challenges are associated with single and multiple level association rules.

REFERENCES

1. Agrawal R., Imielinski T., and Swami A., 1993, "Mining Association Rules Between Sets of Items in Large Databases". In proceedings of the ACM SIGMOD Int'l Conf. on Management of data, pp. 207-216.
2. Agrawal Rakesh, and Srikant Ramakrishnan, 1994, "Fast Algorithms for Mining Association Rules". In Proceedings of the 20th Int. Conf. Very Large Data Bases, pp. 487-499.
3. Brin Sergey, Motwani Rajeev, Ullman Jeffrey D. and Tsur Shalom, 1997, "Dynamic Item set Counting and Implication Rules for Market Basket Data". In proceedings of the 1997 ACM SIGMOD international conference on management of data, vol. 26, issue 2, pp 255-264.
4. Bastide Yves, Taouil Rafik, Pasquier Nicolas, Stumme Gerd and Lakhal Lotfi, 2000, "Mining Frequent Patterns with Counting Inference". In proceeding of ACM SIGKDD, pp 68-75.
5. Cheung David W., Ng incent T. and Tam Benjamin W., 1996, "Maintenance or Discovered of Knowledge: A Case in Multi-Level Association Rules". In preceding the Hong Kong research grant council.
6. Gautam Pratima and Pardasani K.R., 2010, "A Fast Algorithm for Mining Multilevel Association Rule Based on Boolean Matrix". Proceeding in International Journal on Computer Science and Engineering, Vol. 02, No. 3, pp 746-752.
7. Hidber C., 1999, "Online Association Rule Mining". In proceedings of the 1999 ACM SIGMOD International Conference on Management of Data, volume 28(2) of SIGMOD Record, pp. 145-156.
8. Han J., Pei J., and Yin Y., 2000, "Mining Frequent Patterns without Candidate Generation". In ACM SIGMOD International Conference on Management of Data.
9. Jiawei Han and Yongjian Fu, 1995, "Discovery of Multiple-Level Association Rules from Large Databases". In proceedings of the 21st VLDB Conference Zurich, Switzerland, pp 420-431.

10. Kumar Kishore B. and Jotwani Naresh, 2006, "Efficient Algorithm for Hierarchical Online Mining of Association Rules". In Proc. 13th International Conference on Management of Data COMAD.

11. Kotsiantis S., Kanellopoulos D., 2006, "Association Rules Mining: A Recent Overview", GESTS International Transactions on Computer Science and Engineering, Vol.32, No. 1, pp.71-82.

12. Mittar Vishav, Yadav Ruchika and sirohi Deepika (2010), "Mining Frequent Patterns with Counting Inference at Multiple Levels". Proceeding in International Journal of Computer Applications Vol. 3, No.-10.

13. Park J.S., Chen M.S., and Yu P.S., 1995, "An Effective Hash Based Algorithm for Mining Association Rules". In Proceedings of the 1995 ACM SIGMOD International Conference on Management of Data, volume 24(2), pp. 175–186.

14. Pater Mirela and Popescu Daniela E., 2008, "Multi-Level Database Mining Using AFOPT Data Structure and Adaptive Support Constrains". Proceeding in Int. J. of Computers, Communications & control, ISSN 1841-9836, Vol. III, pp 437-441.

15. Rajkumar N., Karthik M.R. and Sivanandam S.N., 2003, "Fast Algorithm for Mining Multilevel Association Rules". Proceeding in IEEE Transactions on Knowledge and Data Engineering, vol. 02.

16. Ramanaiah Preethi Kolluru, 2011, "Hybrid Association Rule Mining Using AC Tree". Proceeding in Journal of Information Engineering and Applications, Vol. 1, No. 02.

17. Savasere Ashok, Omieinski Edward and Navathe Shankant, 1995. "An Efficient Algorithm for Mining Association Rules in Large Databases", Proceedings of the 21st International Conference on Very Large Data Bases, pp. 432 – 444.

18. Show-Jane and Chen Arbee L.P., 2001, "A Graph-Based Approach for Discovering Various Types of Association Rules". Proceeding in IEEE Transactions on Knowledge and Data Engineering. Vol. 13 No. 5, pp 839-845.

19. Sharma L.K., Vyas O.P., Tiwary U.S. and Vyas R., 2005, "A Novel Approach of Multilevel Positive and Negative Association Rule Mining for Spatial Databases". Proceeding in Springer-Verlag Berlin Heidelberg, pp 620-629.

20. Zaki M.J., Parthasarathy S., Ogihara M., Li W., 1997, "Parallel Algorithm for Discovery of Association Rules". Data mining knowledge discovery, pp 343–374.

21. Zaki M.J. and Gouda K., 2003, "Fast Vertical Mining Using Diffsets". In Proceedings of the 9th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Washington, D.C., New York, pp. 326-335.

22. Kousari Alireza Mirzaei Nejad, Mirabedini Seyed Javad, Ehsan Ghasemkhani, 2012, "Improvement of Mining Fuzzy Multiple-Level Association Rules from Quantitative Data". In Proceedings Journal of Software Engineering and Applications Vol. 5, No. 3, pp 190-199.

23. Yu-Chiang Li, Jieh-Shan Yeh, Chin -Chen Chang, 2005. "Efficient algorithms for Mining Share-Frequent Itemsets", In Proceedings of the 11th World Congress of Intl. Fuzzy Systems Association.