International Journal of Advance Research In Science And Engineering http://www.ijarse.com
IJARSE, Vol. No.3, Issue No.5, May 2014 ISSN-2319-8354(E)

A Review on Association Rule Mining

Meenakshi

Asst. Prof., Dept. of CSE, GGIAET, Gurgaon, Haryana

ABSTRACT:

ARM is the main curiosity area for plenty of researchers from many years. It is the backbone of data mining. Relationships are discovered among different items in the Database. The purpose of this paper is to provide a analysis on the ARM basic concepts technique in addition to the recent correlated work in this field. Additionally the paper also discusses the problems and challenges pertaining to the field of association rule mining.

***Keywords*- Association rule mining, Apriori, Weka.**

I. INTRODUCTION

Data mining is the step by step study and scrutiny of the KDD process (Knowledge Discovery and Data Mining). It is the process to extract exciting and useful (understood, formerly unidentified and constructive) information or patterns from mega information repositories such as: data of warehouses, relational database etc. The motto of the data mining process is to take out information from a data set and alter it into an comprehensible and clear structured manner for further use. Due to its broader applicability and acceptability, Data mining has attracted much interest in database communities. The issues of mining association rules from transactional database were introduced in [1]. The theory aims to find regular patterns, exciting correlations and links among sets of items in the data repositories or transaction databases.

Association rules are broadly used in controlling of inventory, diagnosis in medical field, market and risk management industry, drug testing industries etc.[4]

Association rule are the statements that find the association between data in any database. Association rule consist of two parts. First is “Antecedent” and second is “Consequent”. For instance: {Chassis} => {Engine}. Here Chassis is the antecedent and engine is the consequent. Antecedent is the item that is found in the database, and consequent is the item that is found in grouping with the first.

I.I Generalized Association Rule Mining Algorithm

Over a period algorithms for generating association rules are offered in abundance. Few of the finely recognized algorithms are AIS, Apriori, Partitioning algorithms Fp-growth, Apriori-TID, Apriori Hybrid, Tertius Apriori Algorithm and many more. Few of the parallel association rule mining algorithms which are based on Data and Task comprise of HPA (Hash based parallel Mining of Association Rules) ,CD(Count Distribution), PAR(Parallel Association Rules), PDM (Parallel Data Mining) plenty of others.

Universally, Item set is nothing but a set of items (such as antecedent (LHS) or the consequent (RHS) of a rule). The length of an item set is provided as the number of items enclosed in an item set. Item sets of some length J are called J-itemsets. Usually, an association rules mining algorithm has the following steps [11]:

a) The set of candidate J-item sets is generated by 1-extensions of the large (J-1)-item sets generated in the previous iteration.

b) Support for the candidate J-item sets are generated by a pass over of the database.

c) Item sets that do not have the minimum support are useless and the left over itemsets are called large J-itemsets.

This process is recurring until there are no larger item sets in the database. The most frequently used approach for finding association rules is based on the Apriori algorithm. The competence of the level wise generation of frequent itemsets is enhanced by using the Apriori property which states that all non-empty subsets of a frequent itemset must also be frequent [7].

II. LITERATURE REVIEW

A detailed study of journals and articles related to association rule mining algorithms has been carried out. Few papers compared association rule mining algorithms; others tailored the existing algorithms to improve the performance. Huaifeng Zhang et al [5] projected an algorithm to determine combined association rules. In Comparison with the existing association rule, this combined association rule technique permits various users to directly perform actions. In detailed study, rule generation and interestingness measures have been paid attention in combined association rule mining. The frequent itemsets among itemset groups are discovered to improve efficiency in combined association rule generation. An competent version of Apriori algorithm for mining multilevel association rules in large databases has been presented for finding maximum frequent itemset at lower level of abstraction by Praima Gautan and K.R. Pardasani[8]. A new, quick and an well-organized algorithm (SCBF Multilevel) with single scan of database for mining complete frequent itemsets has been proposed. The multiple-level association rules under different supports in simple and effective way can be derived by projected algorithm.

Xunwei Zhou and Hong Bao [12] planned for an algorithm for double connective association rule mining using a three table relational database. The rules are established among the primary keys of the two entity tables and the primary key of the binary relationship table. Under a Grid computing environment Raja Tlili and Yahya Slimani [9] proposed a dynamic load balancing strategy for distributed association rule mining algorithms. Experiments proved that the proposed strategy success in getting better use of the Grid architecture assuming load balancing. Basic concepts of negative association rule and an enhancement in Apriori algorithm for mining negative association rule from frequent absence and presence itemset proposed by Anis Suhailis Abdul Kadir et al[2]. Guimei Liu et al [3] proposed different methods to handle the false positive errors in association rule mining. Three multiple testing correction approaches- the direct adjustment approach, the holdout approach and the permutation-based approach have been used and widespread experiments have been conducted to examine their performances. Among the three permutation–based approach has the highest power of detecting real association rules, but it is costly computationally. Somboon Anekritmongkol and M. L. Kasamsan [10] proposed a time reducing technique (Boolean Algebra Compress Technique) in reading data from the database. It has been concluded that the time had reduced noticeably.Jesmin Nahar et al [5] predicted heart diseases data by comparing healthy heart and sick heart data utilizing various association rule algorithms. The three association algorithms used were Predictive apriori, Apriori and tertius algorithm. Based on the results it was concluded that Apriori algorithm is the best matched algorithm for this task. A comparable work was done by Jyoti Arora et al

[6] who proposed a comparison of various association rule mining algorithms on Supermarket data and achieved the results using Weka data mining tool.

III. ISSUES AND CHALLENGES

Although massive research work has been carried out in association rule mining field and various authors have projected different algorithms, yet there subsist many problems and challenges which must be resolved in order to get absolute benefit of this method. The list illustrating major shortcoming of the association rule mining algorithms

Is as under10]:

a) Enormous discovered rules

b) Attaining non interesting rules

c) Low algorithm performance

Users of association rule mining tools face many issues like the algorithms do not always return the results in practical time. Also the set of association rules can rapidly grow to be unwieldy, especially when we lower the frequency necessities. Fetching all association rules from a database requires counting all achievable and potential combinations of attributes. Support and confidence factors can be used for achieving interesting rules which have values for these factors higher than a threshold value. In most of the methods the confidence is determined once the relevant support for the rules is computed. The key constituent that makes association rule mining realistic is the minimum support specified by the user i.e. *minsup*. It is used to trim the uninteresting rules. But using only a single *minsup* means that all the items in the database are of the identical nature. Every time this may not be correct or perfect approach. For example, in business of retailing , customers frequently buy less priced items, while the higher priced items may not be bought very often. In this conditon, if the *minsup* is set high, the generated rules will have only those rules having only low price items and hence it all contribute to the firm's less profit . On the contrary, if the *minsup* is set too less, a lot of meaningless frequent patterns will be generated that will burden the decision makers unnecessarily.

Such situation is called as rare item problem [20].

IV. PERFORMANCE REVIEW

Over a time many algorithms for generating association rules have been offered. few of the well recognized algorithms are AIS, Apriori-TID, Apriori, Fp-growth, Partitioning algorithms, Apriori Hybrid, , FP-growth Algorithm, Tertius Algorithm and several others. The AIS algorithm was the first algorithm to generate all hefty itemsets in a transaction database. The algorithm is used in finding qualitative rules. This technique is restricted to only item in the subsequent. The AIS algorithm makes numerous passes over the database. The primary issue of the AIS algorithm is that it generates plenty of candidates that afterwards turn out to be small[1]. One more drawback of this algorithm is that the data structures obligatory for keeping huge candidate itemsets are not precise. The

Apriori algorithm developed [1] is the most well known association rule algorithm. Meaning of Apriori is "from what comes before". Its accomplishment is easier than other algorithms and utilizes less memory comapritively. But it has certain demerits too. It only details about the presence and absence of an item in transactional databases and needs a large number of database scan. However the minimum support threshold used is consistent and the number of candidate itemsets produced is massive.To solve few of the holdups of the Apriori algorithm, Fp-growth algorithm was brought into existence which is based on tree configuration or structure.

Association Rule Mining Algorithm Advantages and Disadvantages

AIS

Advantages:

1. A judgment is used in the algorithm to trim those candidate itemsets that have no scope to be large.
2. It is appropriate for low cardinality sparse transaction database.

Disadvantages:

1. It is limited to only one item in the subsequence.
2. It needs Multiple passes over the database.
3. Data structures required for maintaining large and applicant itemsets is not specified.

Apriori

Advantages:

1. This algorithm has smallest amount of memory usage.
2. Simple Execution.
3. For trim and snip ,it uses Apriori property hence, itemsets left for additional support scrutiny remain less.

Disadvantages:

1. It needs lot of scans of database.
2. It permits a single minimum support Threshold only.
3. It is constructive for small database only.
4. For an item in database, it details the presence or absence only .

FP- growth

Advantages:

1. It is quicker as compared to other association rule mining algorithm.
2. It utilizes compressed representation of original database.
3. Elimination of Repeated database scan.

Disadvantages:

1. The memory utilization is more.
2. Not useful for interactive mining and incremental mining.
3. Compressed representation of the database is used by FP –growth and hence the irrelevant information are trimmed. However if we use FP-tree method, it cannot be used for interactive and incremental mining system as changes in threshold value or new insertions in database may result in repetition of the entire process.

V. CONCLUSION

Association rules are extensively used in a variety of areas such as risk and market management, telecommunication networks, medical diagnosis, inventory control etc. This paper demonstrates a review on association rule mining. Initially a concise foreword about association rule mining is given which is the process of finding patterns, co-relations, associations or informal structures among sets of items in the transaction databases or other data repositories. A generalized association rule mining algorithm has been proposed. The paper surveys the research work done by many authors in this area. Some of the problems related to this field have also been highlighted which can be a support for upcoming researchers. The advantages and disadvantages of some of the mining algorithms have also been presented .

REFERENCES

1. Agrawal R., Imielinski, T., and Swami, " Mining association rules between sets of items in large databases", In Proceedings of the 1993 ACM SIGMOD International Conference on Management of Data,1993.

2. Anis Suhailis Abdul Kadir, Azuraliza Abu Bakar and Abdul Razak Hamdan, "Frequent Absence and Presence Itemset for Negative Association Rule Mining ", IEEE,2011.

3. Guimei Liu, Haojun Zhang and Limsoon Wong, "Controlling False Positives Iin Association Rule Mining" In Proceedings of the VLDB Endowment ACM,2011.

4. http://en.wikipedia.org/wiki/Data_mining 5. Huaifeng Zhang, Yanchang Zhao, Longbing Cao and Chengqi Zhang, "Combined Association Rule Mining", PAKDD 2008, LNAI 5012, pp. 1069-1074, 2008 © Springer-Verlag Berlin Heidelberg 2008

5. Jesmin Nahar,Kevin S.Tickle,Shawkat Ali and YI-Ping Phoebe Chen, "Diagnosis Heart Disease using an Association Rule Discovery Approach" In Proceedings of the IASTED International Conference Computational Intelligence August 2009.

6. Jyoti Arora, Sanjeev Rao and Shelza, "An Efficient ARM Technique for Information Retrieval In Data Mining " In International Journal of Engineering Research and Technology Vol 2,Issue 10, October 2013.

7. Nitin Gupta, Nitin Mangal, Kamal Tiwari and Pabitra Mitra, "Mining Quantitative Association Rules in Protein Sequences" Data Mining, LNAI 3755, pp. 273-281, 2006 © Springer- Verlag Berlin Heidelberg 2006.

8. Pratima Gautam and K.R. Pardasani, "Algorithm for Efficient Multilevel Association Rule Mining" In (IJCSE) International Journal on Computer Science and Engineering, Volume 02, No. 05, 1700-1704, 2010.

9. Raja Tlili and Yahya Slimani, "Executing Association Rule Mining Algorithm under a Gird Computing Environment" In PADTAD,July 2011.

10. Somboon Anekritmongkol and M. L. Kulthon Kasamsan , " The Comparative of Boolean Algebra Compress and Apriori Rule Techniques for New Theoretic Association Rule Mining Model" In IEEE,2009

11. Sotiris Kotsiantis and Dimitris Kanellopoulos, "Association Rule Mining: A Recent Overview" In GESTS International Transactions On Computer Science And Engineering, Vol. 32(1), pp. 71-82, 2006.

12. Xunwei Zhou and Hong Bao ," Mning Double-Connective Association Rules from Multiple Tables of Relational Databases " In IEEE,2008

13. Ya-Han Hu, Yen-Liang Chen, "Mining Association Rules with Multiple Minimum Supports: A New Mining Algorithm and a Support Timing Mechanism" © 2004 Elsevier B.V. Gurneet Kaur et al, / (IJCSIT) International Journal of Computer Science and Information Technologies, Vol. 5 (2) , 2014, 2320-2324