INTERNATIONAL JOURNAL OF ADVANCED INNOVATIVE TECHNOLOGY IN ENGINEERING

Published by Global Advanced Research Publication House Journal

Home page: www.ijaite.co.in

Study of Association Rule Mining

[1]Devyani Parate, [2]Bharti Shende, [3]Tinal Ganeshkar, [4]Pratiksha Meshram, [5]Megha Gaikwad, [6]Tejasvi Moon, [7]Prof. Vaishali Gedam

[1,2,3]Department of Computer science and Engineering, Nagpur Institute of Technology, Nagpur, Maharashtra, India

[1]dparate.2001@gmail.com, [2]shendebharti08@gmail.com, [3]tinalganeshkar52010@gmail.com, [4]meshpratiksha008@gmail.com, [5]meghagaikwad173@gmail.com, [6]moonteju11@gmail.com

Article History

Received on: 01 Jan 2022

Revised on: 28 Jan 2022

Accepted on: 05 Feb 2022

Keywords: Association rule mining, Data Mining, Apriori, Frequent Items, Itemset

e-ISSN: 2455-6491

Production and hosted by
www.garph.org

ABSTRACT

Association Rule Mining is the main curiosity area for plenty of researchers for 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 an analysis of the ARM basic concepts technique in addition to the recent correlated work in this field. Additionally, the paper also discusses the problems and challenges of the field of association rule mining.

1. 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 databases, etc. The motto of the data mining process is to take out information from a data set and alter it into a 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 the 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 inventory, diagnosis in the medical field, market and risk management industry, drug testing industries, etc. [4]

A. Itemsets

Item sets are a group of items together in a single transaction. It is the accumulation of item sets. If the database contains „n" items then there are 2n item sets.

B. Support

Support is defined as no. of the transaction containing that item. Support of rule can be defined as no. of the transaction containing both antecedent and consequent. Support of $A \rightarrow C$ is defined as No. of transactions that contain both A and C.

C. Confidence

Confidence of rule is defined as no. of the transaction containing both antecedent Ann sequent the divided by No of the transaction containing antecedent. Support of $A \rightarrow C$ is No. of transaction that contains both A and C divided by No. of the transaction containing A.

D. Frequent itemsets

If support of itemsets is greater than or equal to minimum support then it is called frequent itemsets otherwise item sets are infrequent.

Association rule mining is given in terms of support and confidence where strong rules basing on their interestingness. Association rule mining is given in terms of support and confidence where they are filtered i.e., biased if they don't satisfy the value considered.

$$Support\ (x, y) = \frac{Transactional\ support\ (xy)}{Total\ no.\ of\ Transactions\ in\ D}$$

$$Confidence\ (xy) = \frac{Support\ (xy)}{Support(x)}$$

The support reflects the frequency of the item and confidence reflects the no of transactions containing the if/then pattern. An association rule satisfying a predefined threshold (breakeven) value for support and confidence is only considered for rule exploration.
Association rules are the statements that find the association between data in any database. Association rules consist of two parts. The first is "Antecedent" and the second is "Consequent". For instance: {Chassis} => {Engine}. Here Chassis is the antecedent and engine is the consequent. The antecedent is the item found in the database, and the consequent is the item found in grouping with the first.

2. GENERALIZED ASSOCIATION RULE MINING ALGORITHM

Over a period, algorithms for generating association rules are offered in abundance. A 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 HPA (Hash-based Parallel Mining of Association Rules), CD (Count Distribution), PAR (Parallel Association Rules), PDM (Parallel Data Mining) plenty of others.
Universally, Itemset 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 leftover 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].

3. 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. Haifeng 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 a detailed study, rule generation and interestingness measures have

been paid attention to in combined association rule mining. The frequent itemsets among itemset groups are discovered to improve efficiency in combined association rule generation. A competent version of the Apriori algorithm for mining multilevel association rules in large databases has been presented for finding maximum frequent itemset at a lower level of abstraction by Prima Gautam and K.R. Pardasani [8]. A new, quick, and well-organized algorithm (SCBF Multilevel) with a single scan of the database for mining complete frequent itemsets have been proposed. The multiple-level association rules under different supports simply and effectively can be derived by the 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]. Guide 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 approaches 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 the Predictive apriori, Apriori, and Tertius algorithms. Based on the results it was concluded that the 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 the Weka data mining tool.

4. LITERATURE REVIEW

Although massive research work has been carried out in the association rule mining field and various authors have projected different algorithms, there subsist many problems and challenges which must be resolved to get the absolute benefit of this method. The list illustrates the major shortcoming of the association rule mining algorithms is as under [10]:

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., mins. It is used to trim the uninteresting rules. But using only a single min up means that all the items in the database are identical. Every time this may not be the correct or perfect approach. For example, in the business of retailing, customers frequently buy less-priced items, while the higher-priced items may not be bought very often. In this condition, if the mins up are set high, the generated rules will have only those rules having only low-price items and hence it all contributes to the firm's less profit. On the contrary, if the mins up are set too less, a lot of meaningless frequent patterns will be generated that will burden the decision-makers unnecessarily. Such a situation is called a rare item problem.

5. PERFORMANCE ANALYSIS

Over time many algorithms for generating association rules have been offered. A 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 the 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 afterward 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. The meaning of Apriori is "from what comes before". Its accomplishment is easier than other algorithms and utilizes less memory comparatively. But it has certain demerits too. It only details the presence and absence of an item in transactional databases and needs a large number of database scans. However, the minimum support threshold used is consistent and the number of candidates itemsets produced is massive. To solve a few of the holdups of the Apriori algorithm, the 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 are not specified.

Apriori Advantages

1. This algorithm has the 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 a lot of scans of the database.
2. It permits a single minimum support Threshold only.
3. It is constructive for small databases only.
4. An item in the database details the presence or absence only.

FP- growth Advantages

1. It is quicker as compared to other association rule mining algorithms.
2. It utilizes a compressed representation of the 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 the FP-tree method, it cannot be used for interactive and incremental mining systems as changes in threshold value or new insertions in the database may result in the repetition of the entire process.

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 of association rule mining. Initially, a concise foreword about association rule mining is given to find patterns, correlations, 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.

CONFLICT OF INTEREST

The authors declare that they have no conflict of interest.

FUNDING SUPPORT

The author declares that they have no funding support for this study.

REFERENCES

[1] M. Supriyamenon and Dr. P. Rajarajeswari, "A Review on Association Rule Mining Techniques concerning their Privacy-Preserving Capabilities", International Journal of Applied Engineering Research ISSN 0973-4562 Volume 12, Number 24 (2017) pp. 15484-15488

[2] Meenakshi, "A Review on Association Rule Mining", International Journal of Advance Research In Science And Engineering, IJARSE, Vol. No.3, Issue No.5, May 2014

[3] Yazgana, Pinar & Kusakci, Ali. (2016). A Literature Survey on Association Rule Mining Algorithms. Southeast Europe Journal of Soft Computing. 5. 10.21533/journal.v5i1.102.

[4] X. Chi and Z. W. Fang, "Review of association rule mining algorithm in data mining," 2011 IEEE 3rd International Conference on Communication Software and Networks, 2011, pp. 512-516, DOI: 10.1109/ICCSN.2011.6014622.

[5] Kaushik, M., Sharma, R., Peious, S.A. et al. A Systematic Assessment of Numerical Association Rule Mining Methods. SN COMPUT. SCI. 2, 348 (2021). https://doi.org/10.1007/s42979-021-00725-2

[6] S. Vijayarani and S. Sharmila, "Comparative analysis of association rule mining algorithms," 2016 International Conference on Inventive Computation Technologies (ICICT), 2016, pp. 1-6, DOI: 10.1109/INVENTIVE.2016.7830203.

[7] M. Poundekar, A. S. Manekar, M. Baghel, and H. Gupta, "Mining Strong Valid Association Rule from Frequent Pattern and Infrequent Pattern Based on Min-Max Sinc Constraints," 2014 Fourth International Conference on Communication Systems and Network Technologies, 2014, pp. 450-453, DOI: 10.1109/CSNT.2014.95.

[8] Prithiviraj, P & Dr. R, Porkodi. (2015). A Comparative Analysis of Association Rule Mining Algorithms in Data Mining: A Study.

[9] P. Prithiviraj and R. Porkodi, "A Comparative Analysis of Association Rule Mining Algorithms in Data Mining: A Study", American Journal of Computer Science and Engineering Survey, AJCSES [3][1], 2015, pp-098-119

[10] Győrödi, Cornelia & Gyorodi, Robert & Dr, Prof & Ing, Stefan & Stefan, Holban. (2004). A Comparative Study of Association Rules Mining Algorithms. 10.13140/2.1.1450.3365.

[11] K. Vani, "Comparative Analysis of Association Rule Mining Algorithms Based on Performance Survey", International Journal of Computer Science and Information Technologies, Vol. 6 (4), 2015, 3980-3985