International Journal of Advanced Technology & Engineering Research (IJATER)
www.ijater.com

REVIEW OF ASSOCIATION RULE MINING

Rupinder Kaur[1] Rajeev Kumar Bedi[2] Sunil Kumar Gupta[3]

Abstract

Association rule mining is an important data mining task that used to find out correlations, association between a set of transactions in the databases, data warehouses. This paper presents the basic concepts of association rule mining, the basic Apriori algorithm used for association rule mining and a brief overview of approaches used with basic Apriori algorithm to mine association rules.

Introduction

Association rule mining is an important technique used in data mining proposed by Agrawal et.al. in 1993. Association rule mining is used for discovering interesting patterns and associations between a set of transactions in the databases. Association rules are basically used in areas like market analysis and inventory control [8].

A. Basic concepts

Association rules are if/then statements that helped to find out relationships between data in databases or other information repositories. An association rule has two parts: an antecedent (if) part and consequent (then) part. An antecedent is an item found in the data. Consequent is an item that is found in combination with the antecedent.

The two important basic measures of association rules are confidence and support. Since database consists of large number of transactions and user is only interested in frequently purchased items, threshold of two parameters support and confidence known as minimum support and minimum confidence respectively are predefined by the user to drop out the transactions that are not useful or are not of user interest.

For example: - 90% of customers that purchase bread and butter also purchase milk
Antecedent: bread and butter
Consequent: milk
Confidence factor: 90%
$I = i_1, i_2, \ldots, i_m$: set of items
D : database of transactions
$T \in D$: a transaction. $T \subseteq I$
TID: unique identifier, associated with each T
X: a subset of I
T contains X if $X \subseteq T$.
Association rule X=>Y
Here $X\subset I$, $Y\subset I$ and $X\cap Y=\varnothing$.
Rule X=>Y has a confidence con in D
If con% of transaction in D that contain X also contain Y
Rule X=>Y has a support sup in D
If sup% of transactions in D contain $X\cup Y$

B. Basic Apriori Algorithm

Apriori algorithmic rule is basic algorithmic rule for association rule mining. It takings by distinctive the frequent individual things within the data and lengthening them to larger and bigger item sets as long as those item sets seem sufficiently usually within the data. The frequent item sets verified by Apriori are often used to determine association rules that highlight general trends within the data.

Apriori uses a “bottom-up” approach, wherever frequent subsets are extended one item at a time(a step called candidate generation), and tested against the data. Algorithmic rule terminates once no winning extension units are found. Apriori algorithmic rule generates frequent item sets. If association item satisfies a definite minimum support and minimum confidence then it’s thought about as a frequent item. This whole algorithmic rule relies on plan of looking out level by level.

Association rule mining is a 2 step process:-
i) Find all the frequent item sets from the data. If support of associate item set A is larger than the minimum support i.e., support(A)>=minsup, them itemset a is thought as frequent itemset otherwise not a frequent itemset.
ii) Generate association rules from the frequent itemsets.

Issues in finding association rules with Apriori Algorithm

1. To find out the frequent item sets, one need to scan the database many times. This multiple scan leads to the wastage of time and space.

2. Scalability is another problem that is encountered in the algorithm used for mining association rules i.e., with the

rise in number of transactions, performance get decreased.

3. Apriori algorithm used for mining association rules uses minimum threshold to select frequent item set. This value is provided by the user and must be set very precisely not too high so that new item sets in the database are omitted and not too low so that it leads to item set explosion.

Approaches used with basic Apriori algorithm

Association rule mining is one of the main concerns of the researchers working in the field of Data mining. Many researchers analyzed the existing algorithms to find out the methods to improve the performance of existing algorithms while many researchers proposed new algorithms for association rule mining. Apriori is the basic classical algorithm used for mining the association rules. This literature survey presents the overview of approaches used with the basic Apriori algorithm.

A. APRIORI IMPROVE

APRIORI-IMPROVE algorithm [14] addressed the time and space concern of the basic APRIORI algorithm. In this algorithm, traditional horizontal data approach which stored each transaction with tid was replaced with the mixed type structure which was the combination of the item id set and its complement set. APRIORI-IMPROVE algorithm generated L2 directly from the one scan of the database by using the hash function.

This algorithm used the hash table rather than hash tree to reduce the searching costs. This algorithm made use of two pruning strategies: - Dataset global pruning and Dataset local pruning. Dataset global pruning was based on the idea that t may contain a frequent k-item set I only if all the (k-1) subsets of I belongs to Lk-1.

B. CAPRIORI

CAPRIORI algorithm [2] was used for the fault analysis of CRH EMU. The basic idea was performing coding on each item and then used AND operation to different coded item to get frequent k-item sets. The coding length was judged by the amount of transaction in the database. If an item existed in the transaction then the corresponding location of in the item code was set to 1 otherwise 0. This coding system was used to calculate the support degree of each item.

For eg:- There are 4 records= t01,t02,t03,t04 in a database. Suppose item {X1} exists in the record 1 and 4 item {X2} exists in record 1, 2, 3 and then the code for {X1} is (1001) and code for {X2} is (1110). By this we get support degree of {X1}=2 and support degree of {X2}=3.

Then set L1 of frequent item sets was generated directly by selecting the item satisfying minimal support frequency. AND operation was then performed on L1 to get new codes. If in new codes the amounts of 1 are larger than or equal to minimal support degree then a new rule is generated and frequent item set- 2 was obtained and this process was iterated to get all the frequent item sets.

This algorithm addressed the multiple database scan issue of the basic APRIORI algorithm as it required only single scan to database and efficiency was improved by reducing system I/O and the candidate item sets were reduced.

C. APRIORI WITH PAMMS

One of the issue associated with the basic APRIORI algorithm as mentioned above is the use of single threshold to select frequent item sets. The probability Apriori Multiple Minimum Support (PAMMS) approach [15] made use of multiple minimum support to discover rare association rules.

Rare association rules are rules applicable to rare items but can provide useful knowledge. With the single minimum support it was not possible to mine rare association rules because rare association rules failed to satisfy minimum support if it was set too high. If minimum support was set too low it would lead to combination explosion. Multiple minimum support approach used the notion of “item-to-pattern difference” to discover rare association rules efficiently. This approach assigned minimum support values for frequent as well as rare items based on their item supports and overcome the problem of rule explosion and rule missing. Each item was specified with minimum item support (MIS).

A pattern was declared frequent if its support was found greater than or equal to minimal MIS value among all items. Rare items were specified with relatively lower MIS value. IPD (item pattern difference) was to differentiate the pattern to item. According to multiple minimum support basis the support of the pattern must be greater than minimum MIS value among all items in pattern.

D. APRIORI WITH WEIGHTED APPROACH

Weighted approach [7] with the basic APRIORI was introduced to address the problem of using single minimum support for selecting the frequent item sets. In the transactional databases items are not uniformly distributed. Use of single minimum support lead to either missing of rare association rules if set too high or lead to combination explosion if set too low. Weighted association rules deal with this issue. To reflect different importance to different items, weights were assigned to different items.

Consider D- transaction database

I= {i1, i2, i3……} = set of items. Each transaction is subset of I with transaction id-TID.

Then W= {w1, w2, w3….} is the weight set corresponding to I.

Classical algorithm was first used to obtain the frequent item sets without weights. After weight assigning approach, attributes with weighted support less than minimum weighted support were removed.

Conclusion

To enhance the performance of an algorithm time and space are two main parameters. Algorithms analyzed in this paper provided different strategies to deal with the different issues in the association rule mining. There is still need to enhance the performance of association rule mining algorithm. This paper aims toward the development of heterogeneous approach to be used with the basic APRIORI to enhance the performance of association rule mining.

References

[1] Chun-Hao Chen, Guo-Cheng Lan, Tzung-Pei Hong, and Yui-Kai Lin, "A High Coherent Association Rule Mining Algorithm" in the proceedings of IEEE international Conference on "Technologies and Applications of Artificial Intelligence", Nov. 2012 , pp.1 – 4.

[2] Chun Zhang, Dezan Xie, Ning Zhang, HonghuiLi, "The Improvement of Apriori Algorithm and Its Application in Fault Analysis of CRH EMU" in the proceedings of IEEE international conference on "Service Operations, Logistics, and Informatics (SOLI)" , July 20l1, pp. 543 – 547.

[3] Hangbin LI, Shuhua CHEN, Jianchen LI, Shuo WANG, Yihang FU "An Improved Multi-Support Apriori Algorithm Under the Fuzzy Item Association Condition" in the proceedings of IEEE international conference on "Electronics, Communications and Control (ICECC)" Sept 2011, pp: 3539-3542.

[4] Huiying Wang, Xiangwei Liu, "The Research of Improved Association Rules Mining Apriori Algorithm" in the proceedings of IEEE Eighth International Conference on "Fuzzy Systems and Knowledge Discovery", July 2011, pp. 961-964.

[5] Idheba Mohamad Ali O. Swesi, Azuraliza Abu Bakar, Anis Suhailis Abdul Kadir, "Mining Positive and Negative Association Rules from Interesting Frequent and Infrequent Itemsets" in the proceedings of IEEE 9th International Conference on "Fuzzy Systems and Knowledge Discovery (FSKD)", May 2012, pp: 650-655.

[6] Jia Baohui,Wang Yuxin, Yang Zheng-qing, "The Research of Data Mining in AHM Technology based on Association Rule" in the proceedings of IEEE conference on "Prognostics & System Health Management" May 2011, pp:1-8.

[7] Lei Chen, "The Research of Data Mining Algorithm Based on Association Rules" in the proceedings of ICCASM 2nd International Conference on "Computer Application and System modeling", 2012, pp: 0548-0551.

[8] Luo Fang, Qiu Qizhi, "The Study on the Application of Data Mining Based on Association Rules" in the proceedings of IEEE International Conference on "Communication Systems and Network Technologies" May 2012, pp: 477-480.

[9] Park J S, Chen M S, Yu P S. "Efficient parallel data mining of association rules" New York: ACM, 1996, pp: 134-145.

[10] Punit Mundra, Amit K Maurya, and Sanjay Singh, "Enhanced Mining Association Rule Algorithm with Reduced Time & Space Complexity" in the proceedings of IEEE annual "Indian conference (INDICON)" Dec 2012, pp: 1105-1110.

[11] Qiang Yang, Yanhong Hu, "Application of Improved Apriori Algorithm on Educational Information" in the proceedings of IEEE Fifth International Conference on "Genetic and Evolutionary Computing", Sept 2011, pp:330-332.

[12] R Agrawal, T Imielinski, A swami, "Mining association rules between sets of items in large databases" in the proceedings of the ACM SIGMOD conference on "management of data" 1993, pp: 207-216

[13] Rina Raval, Prof. Indr Jeet Rajput , Prof. Vinitkumar Gupta, " Survey on several improved Apriori algorithm" IOSR Journal of Computer Engineering (IOSR-JCE) e-

ISSN: 2278-0661, p- ISSN: 2278-8727 Volume 9, Issue 4 (Mar. - Apr. 2013), pp: 57-61.

[14] Rui Chang, Zhiyi Liu ,"An Improved Apriori Algorithm" in the proceedings of IEEE International Conference on "Electronics and Optoelectronics (ICEOE)", July 2011,pp: 476-478.

[15] Sandeep Singh Rawat, Lakshmi Rajamani, "Probability Apriori based Approach to Mine Rare Association Rules" in the proceedings of IEEE 3rd Conference on "Data Mining and Optimization (DMO)" June 2011, pp: 253-258.

[16] Savasere A, Omiecinski E, Navathe S, "An efficient algorithm for mining association rules in large databases" New York: ACM, 1995, pp: 432-443.

[17] S.Suriya , Dr.S.P.Shantharajah , R.Deepalakshmi , " A Complete Survey on Association Rule Mining with Relevance to Different Domain" International journal of advanced scientific and technical research issue2, volume1 , Feb 2012, pp:163-168.

[18] Suraj P. Patil, "A Novel Approach for Efficient Mining and Hiding of Sensitive Association Rule" in the proceedings of Nirma university international conference on "engineering", December 2012, pp:1-6.

[19] Suraj P . Patil, U. M. Patil and Sonali Borse, "The novel approach for improving apriori algorithm for mining association Rule" World Journal of Science and Technology 2012, pp:75-78 .

[20] Toivonen H, "Sampling large databases for association tules" 1996, pp: 134-145

[21] Xu Chil, ZHANG Wen Fang, "Review of Association Rule Mining Algorithm in Data Mining" in the proceedings of IEEE 3rd International Conference on "Communication Software and Networks" May 2011, pp. 512-516.

[22] Zhuang Chen, Shibang Cai, Qiulin Song and Chonglai Zhu ,"An Improved Apriori Algorithm Based on Pruning Optimization and Transaction Reduction" in the proceedings of IEEE 2nd international conference on "Artificial Intelligence, Management Science and Electronic Commerce (AIMSEC)" Aug 2011, pp: 1908-1911.

Biographies

RUPINDER KAUR is a student of M.Tech (CSE) at BCET, Gurdaspur. She did her B.Tech in Information technology from BCET, Gurdaspur. Her current research area of interest is data mining. She may be reached at rupinder_kahlon89@yahoo.co.in.

RAJEEV KUMAR BEDI is Assistant Professor in the Department of Computer Science & Engineering at BCET, Gurdaspur. His current research area of interest is cloud computing. Mr. Rajeev Kumar Bedi may be reached at rajeevbedi@rediffmail.com.

SUNIL KUMAR GUPTA is Associate Professor in the Department of Computer Science & Engineering at BCET, Gurdaspur. His current research area of interest is distributed computing. Mr. Sunil kumar Gupta may be reached at skgbcet@gmail.com.