E: ISSN No. 2349-9443 RNI No.UPENG/2012/42622 VOL.-8, ISSUE-1,January (Part-1) 2019
Asian Resonance

A Survey on Association Rules Mining

Abstract

In Data mining, Association Rules Mining (ARM) is one of the most important areas in research fields. It is used to identify the strong association rules or relationships between the objects. Many real time applications use this Association Rule Mining. In this paper, we present different algorithms that are used to find Association rules among the itemsets. Also, we provide an idea which gives more significant to the users.

Keywords: Association Rule Mining, Candidate Generation, Pattern Growth, Support, Confidence, Apriori.

S.J.Vivekanandan
Research Scholar,
Dept. of CSE,
Sathyabama Institute of Science and Technology,
Chennai, India &
Assistant Professor,
Dhanalakshmi College of Engineering,
Chennai, India

G.Gunasekaran
Research Supervisor,
Dept. of CSE,
Sathyabama Institute of Science and Technology,
Chennai, India &
Principal,
J.N.N Institute of Engineering,
Chennai, India

Introduction

From the huge amount of data, it is necessary to convert the huge amount of data into useful information, Data mining is used for this purpose. Data mining [13] [15] is used to find hidden information or unknown pattern or interesting rules from the large amount of data. There are different data mining techniques [12] like classification, clustering, temporal data and so on. Initially, data mining faced with different requirements and challenges [12] to satisfy its need and goals. Data mining is a key role in the process of knowledge discovery. KDD process is the backbone of data mining; it follows the sequence of steps to extract knowledge from the data [15]. Those steps are data cleaning, data integration, data selection, data transformation, data mining, pattern evaluation and knowledge representation. Data mining has various applications such as fraud detection, financial analysis, medical field, CRM, scientific applications and other applications.

Association Rule Mining

Association Rules Mining (ARM) is one of the most important research areas among the researchers. The objective of ARM is to identify the strong association rules or the relationship between the itemsets from the mass amount of data. It can be understand in simple phrase, “what goes with what” and “the purchase of one product when we purchase another product”.

Association rules are denoted as E→F, where E is antecedent and F is consequent. It means, if E occurs, then F also possible to occur. Association rule mining is mainly used in market basket analysis to identify the customer purchase habits.

Basic Terminologies used in ARM

Let I = $\{i_1, i_2, i_3, \ldots, i_n\}$ be a set of items and Tran_DB = $\{t_1, t_2, t_3, \ldots, t_n\}$ a set of transactions where every transaction is also a set of items. Association rules can be measured by two interesting measures such as support and confidence.

Support is the number of times an item available in the transaction. Support (i_1) = Number of times i_1 appears in the database divided by the total number of transactions. Support (i_1i_2) = Number of times i_1i_2 appears together in the database divided by the total number of transactions. Confidence is the probability of A occurs when B also occurs, where A and B are two different itemsets. Confidence ($i_1 \rightarrow i_2$) = ratio of Support (i_1i_2) to the Support (i_1). An item or itemsets are greater than or equal to the minimum support value, then the itemsets are called frequent item or frequent itemsets. An item or itemsets are less than the minimum support value, then the item or itemsets are called infrequent item or infrequent itemsets Association mining is a process of finding strong association rules [8] [9]. It can be completed in two steps

Step 1: Frequent Itemsets Generation

Compute all itemsets that are greater than or equal to support value.

Step 2: Association Rule Generation

Based on step 1, pull out all the rules that are greater than or equal to confidence value. Those rules are called strong association rules.

Different Algorithms Used In ARM

The second step of association rule, i.e. generating association rule is a very direct method. So, more research work only in the first step of association rule, i.e. finding frequent itemsets. In general, Association rule mining algorithms can be classified into two categories

1. With the Candidate generation approach
2. Without Candidate generation approach (pattern growth)

With Candidate Generation approach

A Naïve brute force algorithm [13] was used to find the association rule mining. In this algorithm, it considers all the itemsets in candidate generation, even that itemset count is zero. Therefore the performance of this algorithm is very low, i.e. Number of combinations of itemsets are very high. Therefore an improved naïve algorithm[13] was used to identify an association rule mining. In this algorithm, it considers all the itemsets in candidate generation; even that itemset is not frequent. Although, this algorithm is better than the Naïve brute force algorithm, it is not efficient when it handles a large number of transactions.

Apriori algorithm [8] is a famous algorithm to find the association rules. It works based on apriori property.

Property 1

if an itemset is an infrequent itemset, then all its supersets also infrequent itemset.

Property 2

if an itemset is a frequent itemset, then all its subsets also frequent itemset.

The transactional database can be Boolean transaction i.e. it contains only 0 or 1 entries. 0 represent an item or itemset has not purchased. 1 represent an item or itemset has purchased.

The classical Apriori algorithm [8] is taken as a backbone for many research works. So we have explained the algorithm in the next section.

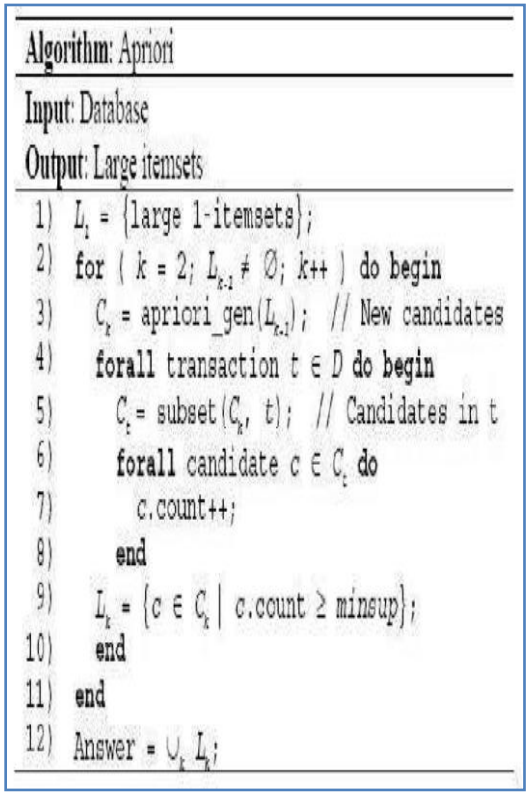
Pseudo code for Apriori Algorithm

Join Step

Join Step is generated by linking with itself.

Prune Step

Any (k-1) itemset that is infrequent cannot be a subset of a frequent k-item set



Although, it is better than previous algorithm, it scans the database many times due to estimate the support count of numerous different candidate itemsets. If 'n' is the length of the longest itemset, then 'n+1' scans are required to compute its support value.

Apriori-TID algorithm [8] was proposed to find the frequent itemsets. Instead of scanning the database for candidate itemset support, it scans the database T_k which is smaller than the original database. So that it is better than original algorithm but it can be efficient only when the database is small.

Direct Hashing and Pruning (DHP) algorithm [11][13] is another method to improve the original apriori algorithm. It is effective in trimming the database by discarding itemset from the transactions that do not need to be scanned. This algorithm uses a Hash table and bit vector to reduce the number of candidate itemsets generated in the first pass. It uses the hash table in the next pass to reduce the number of candidate itemsets. But this algorithm also works well only if the size of the database is small; if the database is large, it takes a large amount of memory for hash table and the time consumption is very high.

Dynamic Itemset Counting (DIC) [13] is an approach to improve the apriori algorithm. This algorithm divides the database into multiple partitions. It scans the first partition for 1-frequent itemset and combines with the next partition until it completes entire partitions. It works well when the database is in same format. If not in same format, it will not work well.

Association rule mining applied in direct marketing application[14][9] to yield profits in their business based on customer purchase history. The association rule used to build a model to predict the value of a customer group.

High-Dimension Oriented Apriori algorithm (HDO algorithm) [16] is used to find association rule in the high-dimensional data. This algorithm adopts a new method to prune candidate itemsets with infrequent itemsets instead of frequent itemsets, candidate itemsets can be pruned validly with the infrequent itemsets with lower dimension

A novel improved algorithm [6] was used to improve the mining efficiency of apriori algorithm. In this algorithm, they introduced some concepts called interest items and frequency threshold which are used to reduce the times of database search. Also dynamic mining used to improve the efficiency of the algorithm.

Another improved algorithm [7] was used to find association rules based on utility weighted score (UW – score) which are extracted from weightage constraint (W-gain) and utility gain (U-gain). In this algorithm, association rules are generated based on frequency as well as significant of the itemsets.

Another important approach to prune mined association rules were discussed in this novel algorithm [18]. In this algorithm, the post processing method is used to prune association. It uses ontology, taxonomy and rule selection schema. It uses matching operator (M) and user-constraint template (UC) to select the most interesting rule.

Another improved apriori algorithm for association rules [5] was used to improve the efficiency of the original apriori algorithm. It reduces the number of times the database has been scanned. In the original apriori algorithm, they have done a slight variation in the logic that gives more efficiency for this improved algorithm.

Completely different approach of association rule mining is secure mining and data privacy mining of association rules [4]. That approach was proposed with different protocols like multi-party algorithms. It uses s-items (secure items) and s-rules (secure rules) which helps to find secure mining rules.

A matrix based apriori algorithm [3] was used to improve the efficiency of the algorithm. In this method, transactional Boolean matrix was used to find the different candidate itemset generation. Initially, it uses the original apriori algorithm to find 1-frequent itemset and then it uses a Boolean matrix to find 2-itemsets, 3-itemsets with the transactional weight and Boolean matrix. So it is highly efficient than the original approach.

An improved apriori algorithm based on time series [2] was used to improve the efficiency of the algorithm. In this method, Boolean matrix as well as different new concepts like sequence association rule, frequent item sequence generation, etc. Those concepts were used to find the time series association rule.

An improved apriori algorithm based on support weight matrix [1][3] was used to find association rules. It uses 0-1 transaction matrix and it gives association rues as well as significance to the user.

In this candidate generation approach, still there were many algorithms which were efficient. But we have done our survey with few important algorithms.

Without Candidate Generation approach

Like the apriori algorithm, FP-Growth algorithm also very famous algorithm to find association rules. In this approach, only frequent itemsets are needed to find the association rules. It does not generate the candidate generation. FP-Growth algorithm [10] [13] is key role in many research works, so we have explained the algorithm in the next section.

Algorithm – FP Growth

The algorithm steps:

1. Scan the transaction once, to find all frequent items and their support count.
2. Arrange the frequent items in non-decreasing order.
3. Initially, start a tree with a root ‘null’.
4. Take each transaction from database, remove all infrequent items and arrange the remaining items according to the arrangement in sorted frequent items
5. Use transactions to construct branch of the tree with each node denotes its frequent items and representing its count.
6. Insert the transaction in the tree using any prefix that may appear. Increase the support count by 1 whenever it appears again.
7. Continue the steps 4 to 6, until all transactions are processed.

A FP-tree consists of many nodes [10][13]. A node has three fields: an item name, a support count and a node link. It also has header table with an entry for each itemset. This algorithm starts with the least frequent item, i.e. the last item in the header table. Then it finds the entire path from the root to the items and updates its support count. By applying condition pattern tree, it finds the frequent itemset.

An improved algorithm called an IFP algorithm [17] was used to find the frequent itemset. In this algorithm, they use new compact data structure IG and also they use different structure like an item table (IT) and item Link (IL). With the help of these structures, IG directly points to the same item of different transactions. It is more space efficient than a FP-Growth algorithm.

Conclusion

In this paper, we have explained two powerful algorithms which are used to find association rules, i.e. Apriori algorithm (candidate generation approach) and FP-Growth algorithm (pattern growth approach). These two algorithms are significant role in many research activities in the area of Association Rule Mining. Also, we gave a summary of various improved algorithms used to find the association in the Association Rule Mining. But we found that all the algorithms which yield association rules based on only frequent Itemsets. Any infrequent itemset may be more significant (profit) to the user which is neglected by these algorithms. So we require an Association rule Mining algorithm which is based on both frequent itemsets and significance to the user.

References

1. *Li-na Sun, "An improved apriori algorithm based on support weight matrix for data mining in transaction database", Journal of Ambient Intelligence and Humanized Computing(2020) Springer 11:495-501.*
2. *Chunxia Wang, Xiaoyue Zheng, "Application of improved time Series Apriori algorithm by frequent itemsets in association rule data mining based on temporal constraint" Evolutionary Intelligence(2020) Springer 13:39-49.*
3. *Qinliu Yang, Qunchao Fu, "A matrix based Apriori Algorithm Improvement" IEEE 3rd International Conf. on Data science in Cyberspace 2018.*
4. *Tamir Tassa, "Secure Mining of Association rules in Horizantally Distributed Databases", IEEE Transaction on Knowledge and Data Engg., vol. 26. No.4 April 2014..*
5. *Mohammed Al-Maolegi, Bassam Arkok, "An Improved Apriori Algorithm for Association Rules", IJNLC vol. 3, No.1, February 2014.*
6. *Libing Wu, Kui Gong, Yanxiang, "A Study of Improving Apriori Algorithm" IEEE 2010.*
7. *Pavinder S, Dalvinder S, S.N.Panda, Atul, "An Improvement in Apriori algorithm using profit and quantity", IEEE 2nd international conference on computer and network technology 2010.*
8. *Agrawal R, Srikant R, " Fast algorithms for mining association rules", Proceedings of 20th International Conf. on Very large Databases, Santiago, Chile, pp 487-499, 1994.*
9. *Agrawal R, Imielinski T., Swami A, "Mining association rules between set of items in large databases", Proceedings of the ACM SIGMOD Intl. Conf. on Management of Data, Washington, D.C.. may 1993, pp 207-216.*
10. *Han J, Pei J, Yin Y, "Mining Frequent Patterns without candidate generation", Proc. Of ACM-SIGMOD, 2000.*
11. *J.-S. Park, M.-S Chen, and P.S. Yu, " An Effective Hash based algorithm for mining association rules", Proc. of ACM-SIGMOD, pp. 175-186, May 1995.*
12. *M.-S Chen, Han J, " Data Mining: An Overview from a Database Perspective", IEEE transaction on Knowledge and Data Engineering, Vol.8, No. 6 December 1996.*
13. *G.K.Gupta, "Text Book: Introduction to Data Mining with Case Studies" 3rd edition, pp. 91-151 PHI Learning Pvt. Ltd. 2019*
14. *Wang K, Zhou Q, Yeung, "Mining Customer value: From Association Rules to Direct marketing", Data Mining and Knowledge Discovery, Vol. 11, pp. 57-79 2005.*
15. *J Han, M kamber, "Text Book: Data Mining: Concepts and Techniques", 2nd edition, Morgan Kaufmann publisher, An imprint of Elsevier 2006*
16. *Lei Ji, Baowen Zhang, jianhua Li, " A new improvement on Apriori Algorithm" IEEE 2006*
17. *D Liyan, Liu Z, Shi Mo, " A novel method of mining frequent item sets", Proceedings of IEEE Intl. Conf. on Information and Automation, june 2010, pp 173-178*
18. *D.Narmada, G.Naveen, S.Geetha, " An efficient approach to prune Mined association rules in large databases", IJCSI, Vol.8, Issues 1, jan 2011.*