



Review: Apriori Algorithms and Association Rule Generation and Mining

Minal Ingle¹, Nitin Suryavanshi²

¹ Research Student, ² Assistant Professor,

Department of Computer Engineering, SSBT's College of Engineering and Technology
Bambhori, Jalgaon. North Maharashtra University
INDIA

Abstract: The prototypical application of Association Rule Mining is market-basket-analysis. However, Apriori algorithm is efficient to generate and mine association rules from large database or data repository. Apriori algorithm generates interesting frequent or infrequent candidate itemsets with respect to support count. Apriori algorithm can require to produce vast number of candidate sets. To generate the candidate sets, it needs several scans over the database. Apriori acquires more memory space for candidate generation process. While it takes multiple scans, it must requires lot of I/O load. The approach to overcome the difficulties is to get better Apriori algorithm by making it reverse. Also will improve Pruning Strategy as it will reduce the scans required to generate candidate item sets and accordingly and a valence or weightage to strong association rule. So that, memory and time required to generate candidate item sets in Apriori will reduce. And the Apriori algorithm will get more effective and efficient.

Keywords: Association Rules, Apriori algorithm, frequent itemsets

I. Introduction

A. Association Rules and Association Rule Mining

Association Rule Mining (ARM) is one of the most important and well researched technique of data mining. ARM was first introduced by Agrawal et al. 1993 [1]. ARM aims to extract interesting correlations, frequent patterns, association or casual structures among set of items or database or other data repositories. Association Rules are if/then statements that help to discover relationships among unrelated data in a data repository. An Association rules consist of two parts, First is an antecedent which represents the if part and Second is a consequent which represents the then part. An antecedent is an item occurred in a data. A consequent is an item that is found in combination with the antecedent. Association rule uses two criteria support and confidence to identify the relationships and rules are generated by analyzing data for frequent if/then pattern. Association rules are generally needs to satisfy a user specified minimum support and a user specified minimum confidence at the same time.

B. Association Rule Mining

Association rule mining usually split into two separate steps:

1. First, minimum support is applied to find all frequent item sets in a database.
2. Second, these frequent item sets and the minimum confidence constraint are used to form rules. While the second step is straight forward, the first step requires more attention.

Support(S) of an association rule is described as the percentage of records that holds union of X and Y to the total number of records in the database.

Confidence(C) of an association rule is defined as the percentage of the number of transactions that contain union of X and Y to the total number of records that include X. Confidence is a measure of strength of the association rule.

$$\begin{array}{l} \text{Rule : } X \Rightarrow Y \\ \swarrow \quad \searrow \\ \text{Support} = \frac{\text{frq}(X, Y)}{N} \\ \text{Confidence} = \frac{\text{frq}(X, Y)}{\text{frq}(X)} \end{array}$$

Figure 1.1: Association Rule

Association Rule Mining is to find out association rules that satisfy the user defined minimum support and confidence from a given database. ARM is normally decomposed into two sub problems. First is to locate those item sets whose occurrences go beyond a user defined threshold in the database, these item sets are known as Frequent item sets. The Second sub problem is to produce association rules from those large item sets with a limitation of minimal confidence.

C. Association Rule Generation Steps

Normally, an association rules generation contains the following steps :

1. The set of candidate k-item sets is generated by 1-extensions of the large (k-1) item sets generated in the previous iteration.
2. Supports for the candidate k-items sets are generated by a scan over the database.
3. Item sets that do not have the minimum support are discarded and the remaining itemsets are called large k-item sets.

D. Model of Association Rule Generation

Association Rules Generation contains many process, they can easily understand by the following model. The conceptual model of Association Rule Generation is in Fig. 1.2.

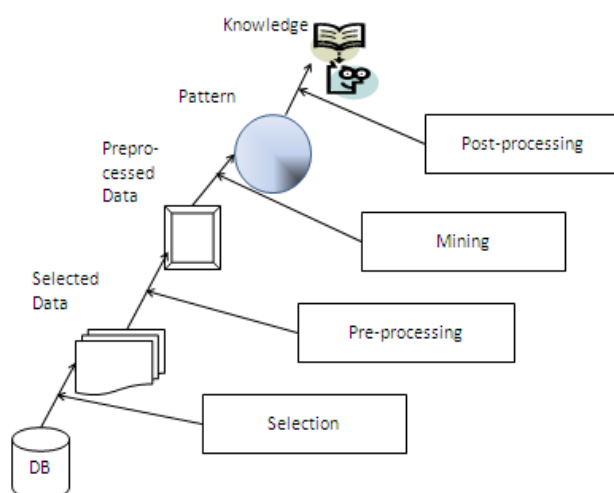


Figure 1.2: Conceptual model of Association Rule Generation

Association Rule Generation consist of processes as selection of database from the large repository, then preprocessing on selected data, after that mine the candidate frequent itemsets from the preprocessed data, then prune the frequent itemsets according to a given threshold. Such a way rules are generated then association rules are mined according to given support and confidence.

ARM is mainly used for mining the frequent and infrequent itemsets from the large databases. It is based on two principles - support and confidence. Association rules are the if/then sentences to show the relationship among the item sets. One of the most important ARM algorithm is Apriori algorithm. Section 2, it gives the survey of papers on Apriori algorithms with their merits and demerits. In Section 3, it gives short description about the improved Apriori algorithm.

II. Literature Survey

To understand the tradeoffs in today's ARM algorithms as Apriori algorithm, it is helpful to briefly examine their history. Apriori algorithm needs pruning techniques in such a way it can reduce the multiple scans over the database.

A. Related Work

To recognize the Apriori algorithm, it must needed to know about their variations.

1) Apriori Algorithm

Rakesh Agrawal et al. proposed Apriori algorithm in 1994. Apriori is one of the most improvement in the history of association rule mining. It solved the problem of association rule mining little bit, in generating many

candidate itemsets which requires more space and wastage of efforts. Apriori is more capable during the candidate generation process because it uses a different candidate generation process and an efficient pruning strategy. Apriori exploits the Downward Closure property. It uses Join and Prune steps to mine frequent item sets. Apriori takes the merit as any subset of a frequent item sets is also a frequent item set. Still, it has the drawback that it takes more time and more memory for candidate generation process but less as compared to original problem. To generate the candidate set it requires multiple scan over the database [2].

2) 3DApriori Algorithm

With the quick increase in log data, there was a need that how to handle such logging data. SHAO Xiao-dong et al. proposed 3DApriori algorithm in 2009. The authors tried to introduce association rules behind the logging data transactions. 3D-Apriori is to interpret the logging data. Attribute data discretion and Spatial predicate extraction are the main concepts of 3DApriori algorithm for association rule generation. It increases the efficiency of ARM for the logging data transformation. Time complexity of the 3D-Apriori algorithm is mainly caused by the space predicates [4].

3) 2.1.3 Apriori Mend Algorithm

The association rules are used to discover interesting rules from large collections of data which states an association between items or sets of items. D. Magdalene Delighta Angeline et al. proposed Apriori Mend algorithm in 2012. Many of the databases systems too long in size, so that, the efficiency of the algorithm could not get good. To improve the mining efficiency, to reduce the computation scans, the author proposed Apriori Mend algorithm. It generates item sets using a Hash function. Apriori Mend for association rule mining based on the concept of closed frequent itemsets. Apriori Mend algorithm is found to be more admirable than the traditional method Apriori algorithm in terms of efficiency. But it has disadvantage as its execution time is increased [6].

4) Parallel Apriori Algorithm

Exploring frequent patterns from transactional databases is considered as one of the association rule mining trouble. Apriori is one of the classic algorithm for this mining process. So, it is a challenging task to develop rapid and efficient algorithm that can handle large volume of transactional databases. Hence, Ning Li et al. proposed Parallel Apriori Algorithm Based on MapReduce in 2012. It is a structure or framework for processing huge databases from transactional databases. Map function and Reduce function are used to generate the association rules. It can scale well and efficiently processed large datasets on service hardware. Still, It has a drawback as, It requires more computation power and memory to find association rules [7].

5) Apriori Algorithm for Multidimensional Data

Feri Sulianta et al. used Apriori Algorithm for Multidimensional Data in 2013. Multidimensional Data Reduction process is a pre-processing. Validation levels are implemented to verify the reliability of the association rules - Data training after reduction, Data training without reduction, Data testing. It explores multidimensional data handling methods to build association rules more specific to product effectively. It must requires the data reduction [8].

6) Apriori-TFP Algorithm

Z. Yang et al. proposed Apriori-TFP algorithm in 1999. This algorithm works like taking or extracting Total from Partial. In an ARM process, raw data are firstly pre-processed and stored in a partial support tree (P-tree). Then, Association Rule Generation is done. ARM process time is reduced, due to efficient pre-processing. It takes number of scans for dense data [3].

7) GPAPriori Algorithm

Fan Zhang et al. proposed GPAPriori Algorithm in 2011. GPAPriori performs a parallelized version of the support counting step on the GPU(Graphical Processing Unit). Support counting procedure is optimized for GPU execution. It gives speed up on modern GPU compared with CPU Based Apriori implementations. But the Complexity is increased in terms of ARM due to GPU [5].

Apriori algorithms have various variations. Apriori algorithm is a classic algorithm of association rules, which enumerate all of the frequent item sets. When this algorithm come across dense data due to the huge number of lengthy patterns appear, this algorithm's performance declined dramatically. Apriori algorithm inherits the drawback of number of scans over the database and more memory space required to candidate generation process in each variance. It highly needs to reduce the scans over the database and reduce memory size, in such a way that Apriori will work efficiently and accurately.

III. Proposed System

Apriori algorithm was first introduced by Agrawal et al. in 1994 [5]. It uses breadth first search strategy to count the support of item sets and uses a candidate generation function. Apriori algorithm uses downward closure property. The Apriori algorithm based on Apriori principle states that if an itemset is frequent, then all of its subsets must be frequent. Apriori algorithm gets the advantage that any subset of a frequent item set is also a frequent item set. However, Apriori algorithm still have the drawback of requirement of more time, space and memory needed for candidate generation process. When this algorithm come upon dense data due to the large number of long patterns emerge, this algorithm's performance declined dramatically. With the intention of generating more valuable rules, the paper proposes the concept of an improved apriori algorithm of association rules.

A. Improved Apriori Algorithm

Apriori algorithm must to improve in such a way it provides efficiency in Association rule generation and mining. The paper proposes the concept of Improved Apriori algorithm. It will work exactly reverse than the traditional Apriori algorithm. It will generate large candidate k-itemsets first and then decreasing one by one, it generates candidate itemsets. Improved Apriori algorithm will use a pruning strategy to reduce the space required to store the candidate itemsets.

Pruning strategy based on reducing the number of candidates itemsets are generated which are infrequent. In such a way, only frequent item sets are taken out, and then it requires less memory. Improved Apriori algorithm provides the valence to a strong association rule. So that, highly strong association rule will generate and find out.

Improved Apriori algorithm will process in a reverse fashion. Also will improve Pruning Strategy as it will reduce the scans required to generate candidate item sets and accordingly find a valence or weightage to strong association rule. So that, memory and time required to generate candidate item sets in Apriori will reduce. And the Apriori algorithm will get more effective and efficient.

IV. Conclusion

In this paper, authors have studied the various variations of Apriori algorithm. Apriori algorithm catches the merit that any subset of a frequent item set is also a frequent item set. However, Apriori algorithm still have the drawback of requirement of more time, space and memory needed for candidate generation process. This review is focused on how to solve the efficient problems of Apriori algorithm and lift another improved apriori association rule generation algorithm which may give solution in order to improve the efficiency of Apriori algorithm. Apriori algorithm used in various applications such as Market Basket Analysis, Market and Risk management, Inventory control, Telecommunication networks.

References

- [1] R. Agrawal, T. Imielinski, and A. Swami, "Mining association rules between sets of items in large databases," *Proceedings of ACM SIGMOD Record*, vol. 22, no. 2. ACM, 1993, pp. 207-216.
- [2] R. Agrawal, R. Srikant et al., "Fast algorithms for mining association rules," *Proceedings of 20th Int. Conf. Very Large Data Bases, VLDB*, vol. 1215, 1994, pp. 487-499.
- [3] Z. Yang, W. Tang, A. Shintemirov, and Q. Wu, "Association rule mining-based dissolved gas analysis for fault diagnosis of power transformers," *Proceedings of IEEE Transactions on, Systems, Man, and Cybernetics, Part C: Applications and Reviews*, vol. 39, no. 6, 2009, pp. 597-610.
- [4] X. dong Shao, "The application of improved 3dapriori three dimensional association rules algorithm in reservoir data mining," *Proceedings of CIS (1). IEEE Computer Society*, 2009, pp. 64-68. [Online Available : <http://dblp.uni-trier.de/db/conf/cis/cis2009-1.html#Shao09>].
- [5] F. Zhang, Y. Zhang, and J. D. Bakos, "GPApriori: Gpu accelerated frequent itemset mining," *Proceedings of CLUSTER. IEEE*, 2011, pp. 590-594. [Online Available: <http://dblp.uni-trier.de/db/conf/cluster/cluster2011.html#ZhangZB11a>].
- [6] I. Samuel Peter James, D. Magdalene Delighta Angeline, "Association rule generation using Apriori mend algorithm for students placement," *Proceedings of International Journal of Emerging Sciences*, vol. 2, no. 1, 2012, pp. 78-86.
- [7] N. Li, L. Zeng, Q. He, and Z. Shi, "Parallel implementation of apriori algorithm based on MapReduce," *Proceedings of Software Engineering, Artificial Intelligence, Networking and Parallel Distributed Computing (SNPD)*, 2012 13th ACIS International Conference on, 2012, pp. 236-241.
- [8] F. Sulianta, T. H. Liong, and I. Atastina, "Mining food industry's multidimensional data to produce association rules using apriori algorithm as a basis of business strategy," *Proceedings of 2013 International Conference of*, 2013, pp. 176-181.

Acknowledgments

Authors would like to express gratitude to Prof. Dr. Girish K. Patnaik, Head of the Computer Science and Engineering Department, SSBT's College of Engineering and Technology, Jalgaon, Maharashtra, India for his guidance and support in this work. The authors are also thankful to the Principal, SSBT's College of Engineering and Technology, Jalgaon, Maharashtra, India for being a constant source of inspiration.