ISSN: 2350-0328

International Journal of Advanced Research in Science, Engineering and Technology

Vol. 8, Issue 3 , March 2021

Algorithms for Association Rule Mining – An Overview and Comparison

Rakhimova L.S, Khalmuratov O.U.

Master, Urgench branch of Tashkent University of Information Technologies named after Muhammad al-Kharezmi, Khorezm, Uzbekistan.
Phd, Urgench branch of Tashkent University of Information Technologies named after Muhammad al-Kharezmi, Khorezm, Uzbekistan.

ABSTRACT: In this article, a review of six different association rule mining algorithms AIS, SETM, Apriori, Apriori TID, Apriori Hybrid and FP-Growth algorithms and a comparison between different association mining algorithms. Association rule mining is the one of the most important technique of the data mining. Its aim is to extract interesting correlations, frequent patterns and association among set of items in the transaction database.

KEYWORDS: Data mining, Association rule, Apriori, Apriori TID, Apriori Hybrid and FP-Growth algorithms.

I. INTRODUCTION

The data mining techniques are always highly appreciated by the researchers for extracting information and knowledge from large relational and non-relational datasets. The dataset in general contains huge amount of information and need to be analysed and summarized correctly in order to gain useful information and knowledge. Many researchers are demonstrated the use of various tools and algorithms to find the pattern and dependencies among the parameters of the datasets. These techniques were used to determine the trend in the chemical, financial, pharmaceutical, insurance industries [1] [2].

Y. Hamuro have demonstrated the improvement in profit for the medical storages located in Japan [2]. The study demonstrated that the increment in sales of pain relievers by 50%. Strong associations such as "People with children tend to buy life insurance policies more often than others" and "Owners of sports utility vehicles are more likely to have wireless phones" can be extremely useful for target marketing [3].

Another popular trend of using data mining techniques is in supermarket or retail chains for improving the sales [4] [5]. R.Agrawal have demonstrated that, finding the relationships between items purchased by the customer can improve the sales [6]. Also similar type of study proves that the arrangements of frequently bought together items can be made next to each other's in order to make it simple for the customers.

II. ASSOCIATION RULE

Association rule learning is a rule-based machine learning method for discovering interesting relations between variables in large databases. It is intended to identify strong rules discovered in databases using some measures of interestingness.

Following are the steps of association rules calculation process [7].

- ✓ The system scans the database to get a 1-itemset candidate (set of items consisting of 1 item) and calculates the support value. Then the support value is compared with the specified minimum support, if the value is greater or equal to the minimum support then the itemset is included in a large itemset.
- ✓ Itemset which is not included in large itemsets is not included in the next iteration in pruning.
- ✓ In the second iteration the system will use large itemset results in the first iteration (L1) to form the second itemset candidate (L2). In the next iteration, the system will use the results of large itemset in the iteration and then will use large itemset results in the previous iteration (Lk-1) to form the following itemset candidate (Lk). The system will

ISSN: 2350-0328

International Journal of Advanced Research in Science, Engineering and Technology

Vol. 8, Issue 3 , March 2021

join Lk-1 with Lk-1 to get Lk, as in the previous iteration the system will delete / prune the itemset combination which is not included in the large itemset.

- ✓ After the join operation, the new itemset result from the join process is calculated for support.
- ✓ The candidate formation process which consists of a join and prune process will continue to be carried out until the candidate itemset set is null, or there are no more candidates to be formed.
- ✓ After that, the result of frequent itemset was formed an association rule that met the specified support and confidence values.
- ✓ In the formation of association rule, the same value is considered as one value.
- ✓ The association rule that is formed must meet the specified minimum value.

Minimal Support: a value determined by the researcher to cut the combination of set items into fewer. Minimal Confidence is a value that is also determined by the researcher to cut the combination of each k-item set (the result of minimal support trimming) to form association rules [8].

The basic methodology of association analysis is divided into two stages:
Support. Support from an association rule is the presentation of the combination of items in the database, where if have item A and item B then the support is the proportion of transactions in the database containing A and B. [9].

The support value of an item is obtained by the formula[10].

$$Support(A)=\frac{(The\ number\ of\ transactions\ containing\ A)}{Transaction\ total}$$

While the support value of 2 items is obtained from the following formula:

$$Support\ (A,B)=P(A\cap B)$$

$$Support\ (A,B)=\frac{\sum Transactions\ contain\ A\ and\ B}{\sum Transaction}$$

*Confidence.*Confidence of association rule is a measure of the accuracy of a rule, which is the presentation of a transaction in a database containing A and containing B [9].

$$Confidence\ (A|B)=\frac{\sum Transactions\ contain\ A\ and\ B}{\sum Transaction}$$

III. ASSOCIATION RULE MINING ALGORITHMS

AIS algorithm

The AIS algorithm makes multiple passes over the entire database. During each pass, it scans all transactions. In the first pass, it counts the support of individual items and determines which of them are large or frequent in the database. Large itemsets of each pass are extended to generate candidate itemsets. After scanning a transaction, the common itemsets between large itemsets of the previous pass and items of this transaction are determined. The AIS algorithm was the first published algorithm developed to generate all large itemsets in a transaction database. It focused on the enhancement of databases with necessary functionality to process decision support queries. This algorithm was targeted to discover qualitative rules. This technique is limited to only one item in the consequent.

SETM algorithm

Similar to the AIS algorithm, the SETM algorithm makes multiple passes over the database. In the first pass, it counts the support of individual items and determines which of them are large or frequent in the database. Then, it generates the candidate itemsets by extending large itemsets of the previous pass. In addition, the SETM remembers the TIDs of

ISSN: 2350-0328
International Journal of Advanced Research in Science, Engineering and Technology
Vol. 8, Issue 3 , March 2021

the generating transactions with the candidate itemsets. The relational merge-join operation can be used to generate candidate itemsets.

Apriori algorithm

Apriori Algorithm usually contains or deals with a large number of transactions. For example, customers buying a lot of goods from a grocery store, by applying this method of the algorithm the grocery stores can enhance their sales performance and could work effectively. It is also very effective in the field of healthcare for the detection of adverse drug reactions. To perform this algorithm, one should know about association rules because it is the most important and well-explored method for knowing the weak or the strong relationships among variables in a huge database and information. Apriori Algorithm has the property that helps to improve the efficiency level by reducing the search space [7].

The Apriori algorithm developed by is a great achievement in the history of mining association rules. This technique uses the property that any subset of a large itemset must be a large itemset. Apriori generates the candidate itemsets by joining the large itemsets of the previous pass and deleting those subsets which are small in the previous pass without considering the transactionsin the database. An association rule is validif its confidence and support are greater than or equal to corresponding threshold values.

The basic Apriori algorithm is a 3 step approach:

1. Join. Scan the whole database for how frequent 1-itemsets are.
2. Prune. Those itemsets that satisfy the support and confidence move onto the next round for 2-itemsets.
3. Repeat. This is repeated for each itemset level until we reach our previously defined size.

Apriori TID algorithm

AprioriTID algorithmic rule uses the generation operate so as to work out the candidate item sets. The sole distinction between the two algorithms is that, in AprioriTID algorithmic rule the info isn't referred for investigating support once the primary pass itself. Once a group action doesn't have a candidate k-item set in such a case the set of candidate item sets won't have any entry for that group action. This can decrease the quantity of group action within the set containing the candidate item sets Compared to the information. As worth of k will increase each entry can become smaller than the corresponding transactions because the variety of candidates within the transactions can persevere decreasing. Apriori solely performs higher than AprioriTID within the initial passes however a lot of passes area unit given AprioriTID definitely has higher performance than Apriori. Database isn't used for count the support of candidate itemsets once the primary pass. The method of candidate itemset generation is same just like the Apriori rule. Another set C' is generated of that every member has the TID of every dealing and therefore the massive itemsets gift during this dealing. The set generated i.e. C' is employed to count the support of every candidate itemset.

Apriori Hybrid algorithm

Apriori Hybrid algorithm was initially introduced by R. Agrawal in 1994. Apriori and AprioriTID use constant candidate generation procedure and computation constant item sets. Apriori examines each dealing within the information. On the opposite hand, instead of scanning the information, AprioriTID scans candidate item sets utilized in the previous pass for getting support counts. Apriori Hybrid uses Apriori within the initial passes and switches to AprioriTid once it expects that the candidate item sets at the tip of the pass are going to be in memory. As Apriori will higher than AprioriTid within the earlier passes and AprioriTid will higher than Apriori within the later passes.

FP-Growth algorithm

In the first pass, the algorithm counts the occurrences of items (attribute-value pairs) in the dataset of transactions, and stores these counts in a 'header table'. In the second pass, it builds the FP-tree structure by inserting transactions into a tree. Items in each transaction have to be sorted by descending order of their frequency in the dataset before being inserted so that the tree can be processed quickly. Items in each transaction that do not meet the minimum support

ISSN: 2350-0328
International Journal of Advanced Research in Science, Engineering and Technology
Vol. 8, Issue 3 , March 2021

requirement are discarded. If many transactions share most frequent items, the FP-tree provides high compression close to tree root.

Recursive processing of this compressed version of the main dataset grows frequent item sets directly, instead of generating candidate items and testing them against the entire database (as in the apriori algorithm).

IV. COMPARISON OF ASSOCIATION RULE MINING ALGORITHMS

In comparative study, all six algorithms has been compared with respect to three important criteria such as data support, speed and accuracy (Table 1). In data support aspect, AIS, SETM work well on small database Apriori work good for medium size data bases, and Apriori TID, Apriori Hybrid and FP-Growth well suited for large data bases. In speed in initial phase AIS, SETM and Apriori TID work slow speed in the first phase, Apriori, Apriori Hybrid, FP-Growth work well on fast speed in the starting phase. In speed in later phase AIS, SETM and Apriori work slow speed in final phase. AprioriTID, Apriori hybrid and FP-Growth well work on fast speed in final phase. In accuracy, AIS is very less accurate. SETM and Apriori are less accurate measurement. AprioriTID are medium accurate measurement, it's more accurate than Apriori. Apriori Hybrid are high accurate, and it's more accurate than Apriori TID. FP-Growth work well on very high accurate in all six algorithms.

Characteristics	AIS	SETM	Apriori	Apriori TID	Apriori Hybrid	FP-Growth
Data support	Less	Less	Limited	Often suppose large	Very large	Very large
Speed in initial phase	Slow	Slow	High	Slow	High	High
Speed in later phase	Slow	Slow	Slow	High	High	High
Accuracy	Very less	Less	Less	More accurate than Apriori	More accurate than Apriori TID	More accurate

(Table 1. Comparison of six algorithms.)

V. CONCLUSION

There are various association rule mining algorithms. In this paper we have discussed six association rule mining algorithms: AIS, SETM, Apriori, Apriori TID, Apriori Hybrid, FP-growth. Comparison is done based on the above performance criteria. Each algorithm has some advantages and disadvantages. Based on speed, the Apriori Hybrid algorithm equally good as FP-Growth. However, the FP-Growth algorithm outperforms well than the Apriori Hybrid with respect to Accuracy. The comparative result shows that the FP-Growth algorithm is more suitable for obtaining significant associations from very large datasets in a speedy and accurate manner.

REFERENCES

[1]. K. M. Decker and S. Focardi, "Technology Overview: A Report on Data Mining", Technical Report CSCS TR-95-02, Swiss Scientific Computing Center, 1995.
[2]. Y. Hamuro, N.Katoh, Y.Matsuda and K. Yada, "Mining Pharmacy Data Helps to Make Profits", Data Mining and Knowledge Discovery, Vol. 2, 1998, pp. 391–398.
[3]. Doddi, AchlaMarathe, SS Ravi, David C. Torney, Srinivas. "Discovery of association rules in medical data." Medical informatics and the Internet in medicine 26.1 (2001): 25-33.
[4]. R. Agrawal, T.Imielinski and A. Swami, "Mining Association Rules Between Sets of Items in Large Databases", Proc. 1993 ACMSIGMOD,Washington, DC, May 1993, pp. 207–216.
[5]. H. Toivonen, "Discovery of Frequent Patterns in Large Data Collections", Ph.D. Thesis, Report A-1996-5, Department of Computer Science, University of Helsinki, Finland, 1996.
[6]. R. Agrawal, T.Imielinski and A. Swami, "Mining Association Rules Between Sets of Items in Large Databases", Proc. 1993 ACMSIGMOD,Washington, DC, May 1993, pp. 207–216.
[7]. Bhandari, Akshita, et al. "Improvised Apriori Algorithm Using Frequent Pattern Tree for Real Time Applications in Data Mining." Procedia - Procedia Computer Science, vol. 46, no. Icict 2014, Elsevier Masson SAS, 2015, pp. 644–51, doi:10.1016/j.procs.2015.02.115.
[8]. Al-maolegi, Mohammed, and Bassam Arkok. AN IMPROVED APRIORI ALGORITHM FOR. Vol. 3, no. 1, 2014, pp. 21–29.

ISSN: 2350-0328

International Journal of Advanced Research in Science, Engineering and Technology

Vol. 8, Issue 3 , March 2021

[9]. Dutt, Shalini, et al. “An Improved Apriori Algorithm Based on Matrix Data Structure.” Global Journal of Computer Science and Technology: C Software & Data Engineering, vol. 14, no. 5, 2014, pp. 1–5.
[10].DommaLingga, "Application of Apriori Algorithms in Predicting Book Inventory at the Dwi Tunggal TanjungMorawa High School Library", Information and scientific technology, 2016, XI (1).