IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

ISO 3297:2007 Certified ⁜ Impact Factor 7.39 ⁜ Vol. 11, Issue 5, May 2022

DOI: 10.17148/IJARCCE.2022.115152

# Research on Association Rule Mining Algorithms

**Hirali Devendra Wadaskar[1], Vijay M. Rakhade[2], Lowlesh N. Yadav[3]**

Student, Computer Science & Engineering Department, Shri Sai College of Engineering & Technology, Bhadrawati, India[1]

Assistant Professor, Computer Science & Engineering Department, Shri Sai College of Engineering & Technology, Bhadrawati, India[2]

Assistant Professor, Computer Science & Engineering Department, Shri Sai College of Engineering & Technology, Bhadrawati, India[3]

**Abstract:** Association rule mining is one of the important part in the field of data mining. The scope of association rule mining is very broad. In association rules mining, frequent item sets mining is essential. Apriori algorithm, Eclat algorithm and FP-growth algorithm are famous algorithms to find frequent item sets. This algorithm can reduce the database and improve mining efficiency.

**Keywords:** Association Rule, Itemset, Data Mining, Algorithms.

## I. INTRODUCTION

One of most important data mining functionalities is association mining, and it is the most prevalent technique that researchers have studied. The essence of data mining is the collection of association rules. It is the search for association rules in a database of sales transactions between products, which is an important area of dataset research. These rules have the advantages of finding unknown relationships and giving data that may be used as a basis for making decisions and making predictions. It was the start of association rule mining research. Since then, association rule mining has been an attractive topic or research. Many researchers have contributed in this area. The finding of interesting associations and relations among huge sets of data objects is made possible by association rule mining. This rule indicates how well a given itemset appears in a transaction. It is works on the concept of if and else statements, such as if A then B. Here the if element is called antecedent, and then statement called as consequent. Relationships in which we can find an association or connection between two objects are known as single cardinality. It's all about setting rules, and cardinality increases as the number of items increases. So, to measure the associations between thousands of data items, there are some metrics, these metrics are support, confidence and lift.

To better understand the topic, we will use an example. You've probably seen that the pizza shop owner serves a pizza, soft drink, and breadstick combo. Customers who purchase these combos receive a discount from him. Do you ever think why he does what he does? Customers who buy pizza, he thinks, also buy soft drinks and breadsticks. He makes it easier for the clients, however, by creating combos. He also improves his sales performance at the same time. Similarly, you may purchase biscuits, chips, and chocolate packed together at Big Bazar.

It shows that the seller makes it easier for customers to purchase these items in the same location. The best examples of Association rules in data mining are these two.

## II. ALGORITHMS

Association rule mining can be divided into some algorithms and these algorithms are Apriori, Eclat, FP-Growth, AIS and SETM, but they only do half the job, since they are algorithms for mining frequent itemsets. Another steps must be done after to get rules from frequent itemsets found during a database.

### A. Apriori Algorithm

To generate association rules, this algorithm involves a large number of datasets. The algorithm used to generate the association rules between objects is known as an apriori algorithm. It describes the interaction of two or more items. In other word, we can say that the apriori algorithm is an association rule that analyses that people who bought product

 

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

ISO 3297:2007 Certified ⁘ Impact Factor 7.39 ⁘ Vol. 11, Issue 5, May 2022

DOI: 10.17148/IJARCCE.2022.115152

A also bought product B. In most cases, the apriori technique is used on a database with a large number of transactions. Consider a Big Bazar example with P = Rice, Pulse, Oil, Milk, and Apple as the product set. The database has six transactions, with 1 representing that the product is present and 0 indicating that it is not.

TABLE I: Transaction Database with TID

| Transaction ID | Rice | Pulse | Oil | Milk | Apple |
|---|---|---|---|---|---|
| t1 | 1 | 1 | 1 | 0 | 0 |
| t2 | 0 | 1 | 1 | 1 | 0 |
| t3 | 0 | 0 | 0 | 1 | 1 |
| t4 | 1 | 1 | 0 | 1 | 0 |
| t5 | 1 | 1 | 1 | 0 | 1 |
| t6 | 1 | 1 | 1 | 1 | 1 |

The Apriori Algorithm makes the given assumptions

- A frequent itemset's subsets must also be frequent.
- An infrequent item sets of subgroups must also be infrequent.
- Set a minimum support level. In our situation, we've set the percentage to 50%.

**Step 1:** Make a frequency table with all of the products that appear in each transaction. Shorten the frequency table to only include products that have a threshold support level of more than 50 percent. We find the given frequency table.

TABLE II: Frequent 1_itemset

| Product | Frequency (Number of transactions) |
|---|---|
| Rice (R) | 4 |
| Pulse (P) | 5 |
| Oil (O) | 4 |
| Milk (M) | 4 |

The products commonly purchased by customers are listed in the table above.
**Step 2:** Create pairs of products such as RP, RO, RM, PO, PM OM. The frequency table will be given to you.

TABLE III: Frequent 2_itemset

| Itemset | Frequency (Number of transactions) |
|---|---|
| RP | 4 |
| RO | 3 |
| RM | 2 |
| PO | 4 |
| PM | 3 |
| OM | 2 |

**Step 3:** Implementing the same 50 percent support criteria and considering products with more than 50 percent support. It is more than three in our case.
As a result, we have RP, RO, PO, and PM.

**Step 4:** Now go for a group of three products that customers purchase as a package. We get the given combination.

1. RP and RO give RPO
2. PO and PM give POM

**Step 5:** Calculate the frequency of the two itemsets to create the frequency table provided.

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

ISO 3297:2007 Certified ⁜ Impact Factor 7.39 ⁜ Vol. 11, Issue 5, May 2022

DOI: 10.17148/IJARCCE.2022.115152

TABLE IV: Frequent 3_itemset

| **Itemset** | **Frequency (Number of transactions)** |
|---|---|
| RPO | 4 |
| POM | 3 |

If you implement the threshold assumption, you can figure out that the customers set of three products is RPO i.e. Rice, Pulse and Oil. We have considered an easy example of the apriori algorithm in association rule mining.

### B. Eclat Algorithm

Eclat stands for Equivalence Class Transformation. The Eclat algorithm is a data mining algorithm for finding commonly occurring things. For the production of frequent itemsets, some association rule mining algorithms apply a horizontal data structure while others use a vertical data format. The vertical database is used by Eclat. If a horizontal database exists, it must be converted to a vertical database. The Eclat algorithm is faster than the Apriori because of its vertical approach. Only one scan of the database is required by Eclat. Let us now understand this algorithm with an example, consider the following transactions record:

TABLE V: Transaction Database With TID

| **Transaction ID** | **Bread** | **Butter** | **Milk** | **Coke** | **Jam** |
|---|---|---|---|---|---|
| T1 | 1 | 1 | 0 | 0 | 1 |
| T2 | 0 | 1 | 0 | 1 | 0 |
| T3 | 0 | 1 | 1 | 0 | 0 |
| T4 | 1 | 1 | 0 | 1 | 0 |
| T5 | 1 | 0 | 1 | 0 | 0 |
| T6 | 0 | 1 | 1 | 0 | 0 |
| T7 | 1 | 0 | 1 | 0 | 0 |
| T8 | 1 | 1 | 1 | 0 | 1 |
| T9 | 1 | 1 | 1 | 0 | 0 |

The above data is in a Boolean matrix, 1 means true while 0 means false. We now call the function for the first time and arrange each item with it's transaction id set (tidset) in a tabular form:
k = 1, minimum support = 2

TABLE VI: Items with their Transactions

| **Item** | **Tidset** |
|---|---|
| Bread | T1, T4, T5, T7, T8, T9 |
| Butter | T1, T2, T3, T4, T6, T8, T9 |
| Milk | T3, T5, T6, T7, T8, T9 |
| Coke | T2, T4 |
| Jam | T1, T8 |

We now call the procedure recursively until no more item-tidset pairs can be joined: k = 2
TABLE VII: Transaction with 2_itemset

| **Item** | **Tidset** |
|---|---|
| Bread, Butter | T1, T4, T8, T9 |
| Bread, Milk | T5, T7, T8, T9 |
| Bread, Coke | T4 |
| Bread, Jam | T1, T8 |
| Butter, Milk | T3, T6, T8, T9 |
| Butter, Coke | T2, T4 |
| Butter, Jam | T1, T8 |
| Milk, Jam | T8 |

International Journal of Advanced Research in Computer and Communication Engineering
ISO 3297:2007 Certified ⋇ Impact Factor 7.39 ⋇ Vol. 11, Issue 5, May 2022
DOI: 10.17148/IJARCCE.2022.115152

k = 3
TABLE VIII: Transaction with 3_itemset

| Item | Tidset |
|---|---|
| Bread, Butter, Milk | T8, T9 |
| Bread, Butter, Jam | T1, T8 |

k = 4
TABLE IX: Transaction with 4_itemset

| Item | Tidset |
|---|---|
| Bread, Butter, Milk, Jam | T8 |

Because there are no more item-tidset pairs to merge, we stop at k = 4.
We derive the following rules from the given dataset since the minimal support = 2:

TABLE X: Output

| Items Bought | Recommended Products |
|---|---|
| Bread | Butter |
| Bread | Milk |
| Bread | Jam |
| Butter | Milk |
| Butter | Coke |
| Butter | Jam |
| Bread and Butter | Milk |
| Bread and Butter | Jam |

The Eclat method requires less memory than the Apriori algorithm since it employs a Depth-First Search strategy. In most cases, the Eclat method is faster than the Apriori algorithm. The Eclat method does not require repetitive data scanning to calculate individual support values.

### C. FP-growth Algorithm

FP stands for Frequent Pattern. The Apriori approach has been improved by this algorithm. The FP growth algorithm only generates frequent itemsets based on the user-defined minimum support. The FP growth method, on the other hand, does not scan the entire database several times, and the scanning time increases linearly. As a result, the FP growth algorithm surpasses the Apriori algorithm. The FP growth method represents the database as a tree called as a frequent pattern tree or FP tree. This tree structure will retain the connection between the itemsets. The goal of the FP tree is to discover the most common pattern. A node in the FP tree represents each item in the itemset. Null is represented by the root node, and itemsets are represented by the lower nodes. The link of the nodes with the roots, i.e., the objects with other itemsets, is preserved while creating the tree. This algorithm's steps are as follows:

1. The first step is to search the database for instances of the itemsets. The support number or rate of 1-itemset refers to the amount of 1-itemsets in the database.
2. The FP tree is built in the second stage. To do so, start creating the tree's root. Null is used to represent the root.
3. Scanning the database and examining the transactions is the next step. Check the first transaction to see what itemset is contained within. The highest-counting itemset is placed first, then by those with lower counts, and so on. It denotes that the branch of the tree is made up of transaction frequent item sets in descending order of count.
4. The database's next transaction is examined. The itemsets are arranged in ascending order by count. If any item of this transaction already exists in another branch, this transaction branch will have a common prefix to the root.
5. The itemset's count is increased as transactions are completed. As transactions are created and linked, the count of both the common node and unique node increases by one.
6. The next step is to extract the FP tree that has been generated. The lowest node, as well as the relationships between the lowest nodes, are evaluated first. The frequency pattern length's lowest node is indicated by 1. Follow the FP tree's path from there. This path or paths are referred to as a conditional pattern base.

  

ISO 3297:2007 Certified ⋇ Impact Factor 7.39 ⋇ Vol. 11, Issue 5, May 2022

DOI: 10.17148/IJARCCE.2022.115152

7. Create a conditional FP tree based on the number of data items in the path. The itemsets that match the threshold support are considered by the conditional FP tree.
8. The conditional FP tree generates frequent patterns.

Two kinds of improved algorithms are N Painting-growth algorithm and Painting-growth algorithm. The N Painting-growth algorithm creates two item permutation sets to locate all of the frequent item sets' association sets, and then digs up all of the frequent item sets based on the association sets. The painting-growth method creates an association image based on the two item permutation sets in order to locate association sets for all frequent items, then digs up all the frequent item sets based on the association sets. Both of the improved algorithms scan the database only once, reducing the overhead of scanning the database twice in the traditional FP-growth algorithm, and completing the mining only according to two-item permutation sets, resulting in faster processing, smaller memory footprints, lower complexity, and easier maintenance. Improved algorithms clearly serve as a benchmark for future association rule mining research.

### D. AIS Algorithm

The AIS technique was the first algorithm to produce all huge itemsets in a transaction database, and it was published in 1997. As the AIS algorithm scans the data, itemsets are created and counted. The AIS method detects which large itemsets in transaction data contained a transaction, and new candidate itemsets are formed by extending the large itemsets with additional items in the transaction data. It focused on the enhancement of database with necessary functionality to process decision support queries. This algorithm was used to find if there was an association between departments in the customer's purchasing behavior.

### E. SETM Algorithm

Although the SETM algorithm generates candidate itemsets as it searches a database, it accounts for them at the end of the scan. New candidate itemsets are created in the same way as with the AIS method, but the generating transaction's transaction ID is recorded in a sequential data structure with the candidate itemset. The support count of candidate itemsets is calculated by aggregating the sequential structure at the end of the pass. In the first run, the SETM algorithm counts the support for individual objects and decides which of them are large or frequent in the database. The candidate itemsets were then created by extending the previous pass's huge itemset. With the candidate itemsets, the SETM remembers the TIDs of the generating transactions. While building candidate sets, the SETM approach saves a copy of the applicant itemsets in a sequentially, as well as the TID of the creating transaction.

## III. CONCLUSION

We covered many sorts of association rule mining algorithms in this research article. In the discipline of data mining, association rule mining algorithms is an interesting study topic. The association rule mining algorithm is still being researched and developed. These techniques should be modified so that they can be used in a wider range of situations.

## REFERENCES

[1]. Agrawal R and Srikant R, "Fast Algorithms for Mining Association Rules," In Proc. Of the 20th VLDB Conference, June 1994.
[2]. Agrawal R, Imielinsky T and Swami A, "Mining Association Rules in sets of items in large database," In Proc. of SIGMOD Conference on Management of Data, May 1993.
[3]. Bing Liu, Wynne Hsu and Yiming Ma, "Mining Association Rules with Multiple Minimum Supports," Decision Support Systems Volume 42, October 2006.
[4]. Jian Pei, Jiawei Han, "Mining frequent itemsets with convertible constraints." The 17th International Conference on Data Engineering was held in 2001, and the proceedings were published in 2001.
[5]. Han J, Pei J, Yin Y, Mao R, "Mining Frequent Patterns without Candidate Generation: A Frequent-Pattern Tree Approach," Data Mining and Knowledge Discovery, 1 Jan 2004.
[6]. Jian Pei, "Pattern-Growth Methods for Frequent Pattern Mining," Ph.D. Thesis, 2002.
[7]. Mueller A, "Fast Sequential and Parallel Algorithms for Association rule mining: A comparison," Technical report, Faculty of the graduate school of university of Maryland,1995.
[8]. Mr. Vijay M. Rakhade, "Reducing Routing Distraction in IP Networks using Cross-Layer Methodology" ICRTEST Volume 5 Issue 1 (21-22 January 2017).
[9]. Lowlesh Nandkishor Yadav, "Predictive Acknowledgement using TRE System to reduce cost and Bandwidth" IJRECE VOL. 7 ISSUE 1 (JANUARY- MARCH 2019).

 