Implementation of Apriori and FP Growth Algorithm using WEKA

[1]Anushree Raj, [2]Rio D'Souza

[1]Assistant Professor, [2] Professor
[1]M.Sc. Big Data Analytics, [2] Department of CSE
[1] St Agnes Centre for PG Studies and Research, [2] St Joseph Engineering College, Mangalore, India

Abstract*: *In Data Mining finding the frequent patterns from large database is being a challenging task. In Data Mining, Association Rule Mining is a standard and well researched technique for locating fascinating relations between variables in large databases. WEKA provides applications of learning algorithms that can efficiently execute any dataset. In WEKA tools, there are many algorithms used to mining data. In this paper, we find the best association rules using WEKA data mining tools and make a comparative study between the Apriori algorithm and FP growth algorithm. Apriori algorithm generates candidate itemset and discovers the itemset which is frequent. FP growth technique uses pattern fragment growth to mine the frequent patterns from large database. Frequent patterns, are patterns that occur frequently in data. There are many kinds of frequent patterns, including itemsets, subsequences, and substructures.

Keywords*: *Apriori algorithm, Confidence, Data Mining, FP-growth algorithm, Support.

1. INTRODUCTION

Data mining refers to extraction of information from large amount of data. Extracting important knowledge from a very large amount of data can be crucial to organizations for the process of decision-making [1].

Some data mining techniques are Association, Classification, Clustering, Sequential patterns and Decision tree.

Association rule mining finds interesting associations and relationships among large sets of data items. This rule shows how frequently an itemset occurs in a transaction. Given a set of transactions, we can find rules that will predict the occurrence of an item based on the occurrences of other items in the transaction. Association Technique helps to find out the pattern from huge data, based on a relationship between two or more items of the same transaction.

Association Mining aims to extract attention-grabbing correlations, frequent patterns, and association structures among set of things or objects in transaction data based relational databases or different data repositories. Two statistical measures that govern Association Rule Mining are Support and Confidence. Support should be measured as to how often it should occur in the database. Confidence may well be gauged to seek out the strength of the rule. The Association rules are interesting if they satisfy each a minimum Support threshold and a minimum Confidence threshold [2].

2. ASSOCIATION MINING RULE

Association rule mining, at a basic level, involves the use of machine learning models to analyze data for patterns, or co-occurrence, in a database. It identifies frequent if-then associations, which are called *association rules*.

An association rule has two parts: an antecedent (if) and a consequent (then). An antecedent is an item found within the data. A consequent is an item found in combination with the antecedent.

An Association rule is an expression of the form X → Y means that whenever X seems, Y also tends to appear. X and Y are itemsets. An itemsets is nothing but a collection of database items. X is usually stated as the rule's antecedent and Y as the consequent of the rule [3].

A. APRIORI ALGORITHM

Apriori is a bottom-up and breadth first approach. Apriori's principle: If an itemset is frequent, then all of its subset must also be frequent [4].

Drawbacks of Apriori Algorithm

- The Apriori algorithmic program takes longer time for candidate generation technique.
- The Apriori algorithmic program needs many scans of the database.
- Many trivial rules are derived and it will be hard to extract the most interesting rules.
- Rules can be inexplicable and fine grained.
- Redundant rules are generated.

Association rule generation is usually split up 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 needs more attention. Finding all frequent item sets in a database is difficult since it involves searching all possible item sets [5].

Support-

The support for a rule X => Y is obtained by dividing the number of transactions which satisfy the rule, N (X=>Y), by the total number of transactions, N

Support (X=>Y) =N (X=>Y) / N

The support is therefore the frequency of events for which both the LHS and RHS of the rule hold true. The higher the support the stronger the information that both type of events occur together.

Confidence-

The confidence of the rule X => Y is obtained by dividing the number of Transactions which satisfy the rule N (X=>Y) by the number of transactions which contain the Body of the rule, X.

Confidence (X=>Y) = N (X=>Y) / N (X)

The confidence is the conditional probability of the RHS holding true given that the LHS Holds true. A high confidence that the LHS event leads to the RHS event implies causation or Statistical dependence.

Lift-

The lift of the rule X => Y is the deviation of the support of the whole rule from the Support expected under independence given the supports of the LHS (X) and the RHS (Y).

Lift {X=>Y} = confidence (X=>Y) / support (Y) = support (X=>Y) / support (X). support (Y)

Lift is an indication of the effect that knowledge that LHS holds true has on the probability of The RHS holding true [6]. Hence Lift is a value that gives us information about the increase in Probability of the "then" (consequent RHS) given the "if" (antecedent LHS) part.

B. FG GROWTH ALGORITHM

FP growth algorithmic program is an efficient algorithm for producing the frequent itemsets without generation of candidate itemsets. It adopts a divide and conquer strategy and it needs two database scans to seek out the Support count. It can mine the items by using lift, leverage and conviction by specifying minimum threshold. [7]

FP-Growth algorithm follows two steps first it generates an FP- tree and next directly extracts the frequent items from the FP- tree. FP-tree represents the information of frequent datasets. Every path of FP –tree represents a frequent itemset and node in the path are arranged in the decreasing order of the frequency [8]. FP-tree first scans the data and then find support for each item. It removes the infrequent itemsets and sorts the frequent item set in decreasing order of their support. It reads each transaction and maps it to a path. The frequent items are extracted from the FP-tree.

3. IMPLEMENTATION OF APRIORI AND FP GROWTH USING WEKA TOOL

WEKA term is a set of modern machine learning ways and data pre-handling tools. It is identified as a set of machine learning approaches for data extraction tasks. [9] WEKA offers applications of learning algorithms that you can efficiently use to your dataset. It contains a diversity of tools for converting datasets, such as the algorithms for discretization and sampling [10] WEKA makes it easy to compare different solution strategies based on the same evaluation method and identify the one that is most appropriate for the problem at hand [11]. We use the weka tool to study the outcomes of Apriori and FP Growth algorithm.

```
@relation Products
@attribute Wine {Wine}
@attribute Chips {Chips}
@attribute Bread {Bread}
@attribute Butter {Butter}
@attribute Milk {Milk}
@attribute Apple {Apple}

@data
Wine,?,Bread,Butter,Milk,?
?,?,Bread,Butter,Milk,?
?,Chips,?,?,?,Apple
Wine,Chips,Bread,Butter,Milk,Apple
Wine,Chips,?,?,Milk,?
Wine,Chips,Bread,Butter,?,Apple
Wine,Chips,?,?,Milk,?
Wine,?,Bread,?,?,Apple
Wine,?,?,Butter,Milk,?
?,Chips,Bread,Butter,?,Apple
Wine,?,Bread,Butter,Milk,Apple
Wine,Chips,Bread,Butter,Milk,?
Wine,?,?,?,Milk,Apple
Wine,?,Bread,Butter,Milk,Apple
Wine,Chips,Bread,Butter,Milk,Apple
?,Chips,Bread,Butter,Milk,Apple
?,Chips,?,Butter,Milk,Apple
Wine,Chips,Bread,Butter,Milk,Apple
Wine,?,Bread,Butter,Milk,Apple
Wine,Chips,Bread,?,Milk,Apple
?,Chips,?,?,?,?
```

Fig 1. Arff data set used:

The data set used for the implementation is shown in Fig 1.

Fig 2 Load the data file

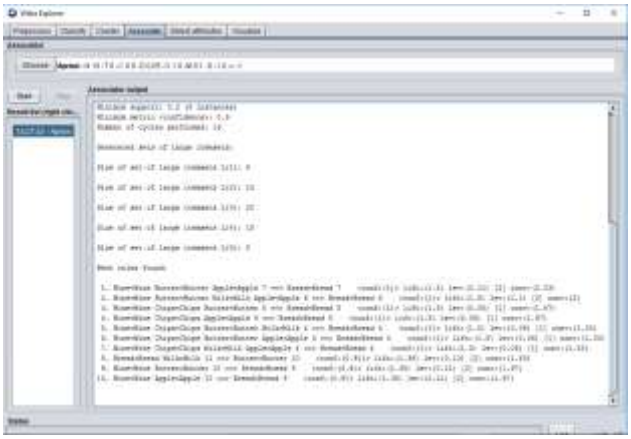

Fig 3. Invoke Apriori algorithm for given data set

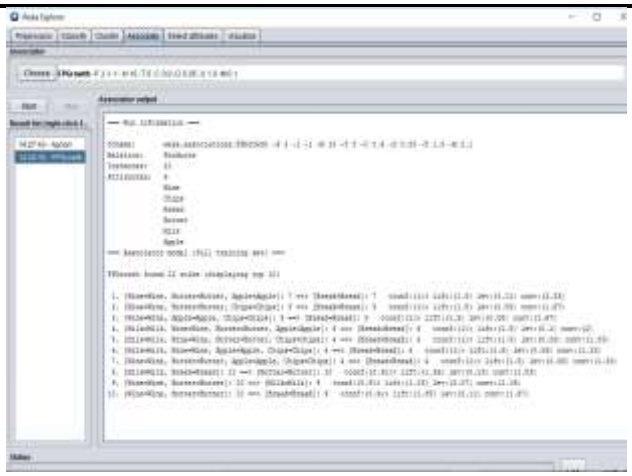

Fig 4. Invoke FP Growth for given data sets

4. COMPARATIVE STUDY

Both Apriori and FP Growth algorithm are used to mine the frequent patterns from database. Both the algorithm uses some technique to discover the frequent patterns. As a result of the experimental study, we understand that the performance of FP-growth algorithm is better than the Apriori algorithm. Apriori is easy to understand and popular but exhibits scalability issues and exhausts available memory much faster. FP Growth algorithm requires less memory due to its compact structure they discover the frequent itemsets without candidate itemset generation. Apriori uses breadth first search method and FP Growth uses divide and conquer method. Time taken by FP Growth is less compared to that of Apriori. Apriori algorithm performs multiple scans for generating candidate set. FP Growth algorithm scans the database only twice.

5. CONCLUSION

In data mining, association rules are useful for analyzing and predicting customer behavior. It is concluded that. Applying the algorithms to supermarkets, the scientists were able to discover links between different items purchased, called *association rules*, and ultimately use that information to predict the likelihood of different products being purchased together. Association Rule Mining is an interesting pattern mining problem. The algorithms used are conceptually clear and ensuing results are perceivable. The experiment was conducted on various data sets of varying sizes. From the experimental data conferred, it was examined that the FP-growth algorithm performs better than the Apriori algorithm. In future, it is possible to extend the research for implementation of different clustering techniques and also the Association Rule Mining for large number of databases.

REFERENCES

[1] C. Borgelt, “Frequent item set mining,” Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, vol. 2, no. 6, pp. 437–456, 2012.

[2] N.P. Gopalan and B. Sivaselvan, “Data Mining Techniques and Trends”, PHI Learning Pvt limited, New Delhi, 2009.

[3] Han, J., Kamber, M.,“Data Mining concepts and techniques”, Elsevier Inc., Second Edition, San Francisco, 2006.

[4] M.S. Chen, J. Han, P.S. Yu, “Data mining: an overview from a database perspective”, IEEE Transactions on Knowledge and Data Engineering, 1996, 8, pp. 866-883.

[5] Agrawal, R., Imielinski, T., and Swami, A. 1993. Mining association rules between sets of items in large databases.In Proc. 1993 ACM-SIGMOD Int. Conf. Management of Data (SIGMOD’93), Washington, DC, pp. 207–216.

[6] S. Moens, E. Aksehirli and B. Goethals, “Frequent Itemset Mining for Big Data,” in Proceedings IEEE International Conference on Big Data, pp. 111–118, (2013).

[7] Sumit Aggarwal andVinay Singal,"A Survey on Frequent pattern mining Algorithms", International Journal of Engineering Research & Technology (IJERT), ISSN: 22780181,Vol. 3 Issue 4,pp 2606-2608, April 2014

[8] Khurana Kand Sharma.S, ―A comparative analysis of association rule mining algorithms.‖, International Journal of Scientific and Research Publications, Volume 3, Issue 5, pp 38-45, May 2013.

[9] Seppelt, R., Voinov, A. A., & Lange, S. (2012). Tools for environmental data mining and intelligent decision support. *Iemss. Org*

[10] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten, “The weka data mining software: an update,” ACM SIGKDD explorations newsletter, vol. 11, no. 1, pp. 10–18, 2009

[11] Witten, I. H., Frank, E., Hall, M. A., & Pal, C. J. (2016). *Data mining: Practical machine learning tools and techniques*. Cambridge, MA: Morgan Kaufmann.