A Study on FP-Growth Algorithm for Association Rule Mining using Supermarket Data

N. Mounika

PG Scholar, Dept. of Computer Science Sri Venkateswara University, Tirupati

Abstract— *Association rule mining plays a crucial role in revealing meaningful patterns and relationships in large-scale transactional datasets, such as supermarket sales data. In this study, we apply the FP-Growth algorithm to mine association rules from a real-world supermarket dataset. The dataset comprises 1388 instances and represents customer transactions over a specific period. We set a minimum support threshold of 0.3 and a minimum confidence threshold of 0.85 to ensure the discovery of significant associations. After 14 cycles of analysis, we obtained sets of large itemsets with varying sizes, including L(1) with 25 itemsets, L(2) with 69 itemsets, and L(3) with 20 itemsets. The best association rules discovered are presented, along with their corresponding support, confidence, lift, leverage, and conviction values. These findings offer valuable insights for supermarket managers to optimize product placement, promotions, and enhance the overall customer experience*.

I. INTRODUCTION

Data mining is a process to obtain potentially useful, previously unknown, and ultimately understandable knowledge from the data. Information disclosure in data sets (KDD) is characterized as the non-paltry extraction of substantial, verifiable, possibly helpful and eventually reasonable data in enormous data sets [1]. For quite a long time, a large number of uses in different spaces have profited from KDD strategies and many works has been directed on this theme. The issue of mining incessant itemsets emerged first as a sub-issue of mining affiliation rules [3]. Association rules mining is one of the important portions of data mining and is used to find the interesting associations or correlation relationships between item sets in mass data [2]. Discovering frequent item sets is a key technology and step in the applications of association rules mining [4].

II. ASSOCIATION RULES MINING

Association rules mining is a function of data mining research domain and arise many researchers interest to design a high efficient algorithm to mine association rules from transaction database [6]. Generally all the frequent item sets discovery from the database in the process of association rule mining shares of larger, the price is also spending more.

2.1 Frequent Item Sets.

Set $I = \{i1, i2,...,in\}$ as an assortment of all kinds of things in the data set, every exchange T is a subset of I, or at least, $T \subseteq I$, and data set D is an assortment of exchanges. For a given exchange data set D, the complete number of exchanges it contains is N. Characterize the help count(X) of thing set $X(X \subseteq I)$ as the quantity of exchanges T in D making $X \subseteq T$ and the help support(X) of thing set X as count (X)/N [9]. The quantity of things in a thing set is called aspect or length of this thing set, on the off chance that the length of the thing set is k, called k-thing set [4][7].

Definition 1: For a given least help, minsup, if the thing set meets support(X) ≥ minsup, thing set X is known as a regular thing set and on the other hand thing set X is called a rare thing set. A set shows relationship between an incessant thing with different things, calling this set a successive thing affiliation set. The base help count, minCount, meets minCount = minsup$*N$. When count(X) ≥ minCount, one says support(X) ≥ minsup [4][5].

Definition 2: At the point when the length of the thing set X is k and support(X) ≥ minsup, one calls thing set Xk-thing regular set. In the event that k≥3, one can call thing set X multi-thing regular set[4][5].

III. FP-GROWTH ALGORITHM

The FP-Growth Algorithm is an elective method to discover continuous itemsets without utilizing applicant ages, accordingly further developing execution. The FP-Growth Algorithm, proposed by Han in [4], is a productive and adaptable technique for mining the total arrangement of successive examples by design piece development, utilizing an all-encompassing prefix-tree structure for putting away compacted and pivotal data about incessant examples named continuous example tree (FP-tree). FP-development calculation is an effective strategy for mining all continuous itemsets without competitor age. FP-development uses a blend of the vertical and even information base design to store the data set in primary memory.

The calculation mines the continuous itemsets by utilizing a gap and-vanquish methodology as follows: FP-development first packs the data set addressing successive itemset into a regular example tree, or FP-tree, which holds the itemset affiliation data also. The subsequent stage is to separate a compacted data set into set of restrictive All hubs relate to things have a counter.

The FP-development calculation comprises of the accompanying advances:

1. Scan DB once, discover incessant 1-itemset (single thing design)
2. Sort continuous things in recurrence diving request, f-list
3. Scan DB once more, develop FP-tree
4. Construct the restrictive FP tree in the succession of opposite request of F - List - produce incessant thing set

IV. EXPERIMENTAL RESULTS

The experiment was conducted using Weka. Weka stands for Waikato Environment for Knowledge Analysis. The software is written in the Java language and contains a GUI for interacting with data files.Weka implements algorithms for data pre-processing, classification, regression and clustering and association rules. It also includes visualization tools.

This section comprises the experimental analysis of Supermarket dataset was gathered from the UCI machine learning repository [8]. This dataset contains 4627 instances and 217 attributes. There are two classes of transactions i.e., Low containing 2948 records and High contains 1679 records. The summary and Statistical summary of Supermarket dataset are shown in the figure-1.

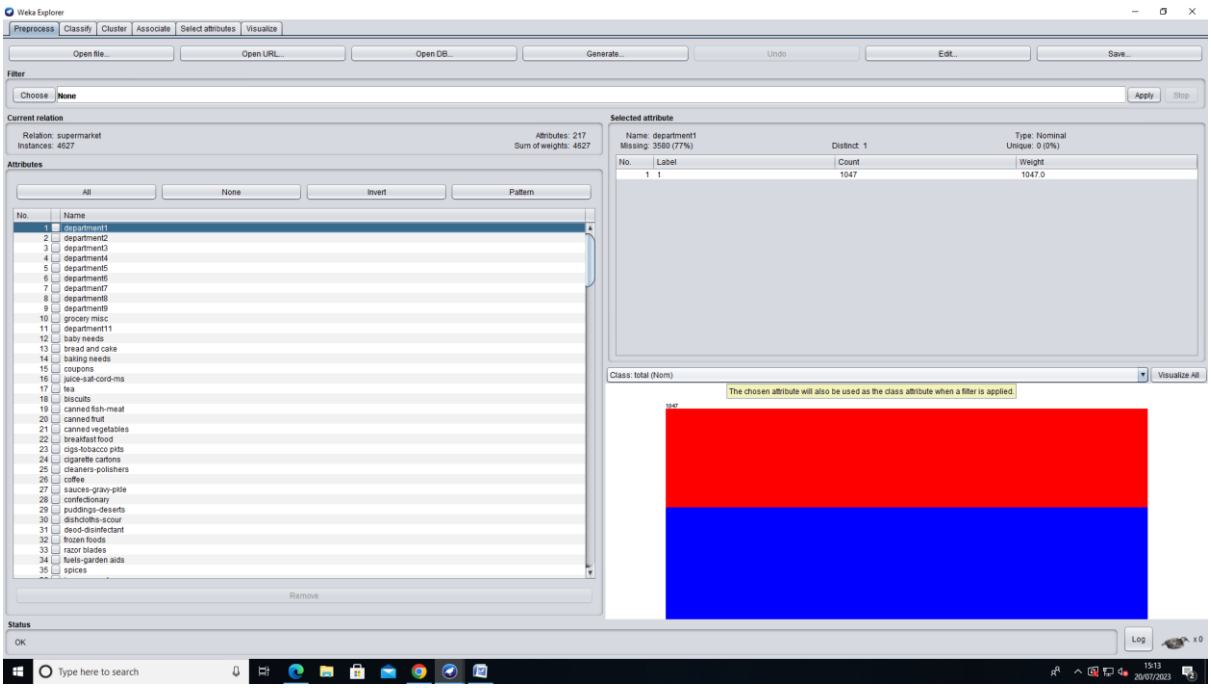

Figure-1: Statistical summary of Supermarket dataset

V. RESULTS AND DISCUSSION

The FP-Growth algorithm was applied to a supermarket dataset with a minimum support of 0.3 and a minimum confidence of 0.85. After 14 cycles of analysis, the algorithm generated sets of large itemsets with varying sizes, including L(1) with 25 itemsets, L(2) with 69 itemsets, and L(3) with 20 itemsets.

The top 5 association rules with the highest confidence values are presented below:

1) If a customer purchases biscuits and vegetables together, there is an 84% chance they will also buy bread and cake in the same transaction. This association has a lift value of 1.17, indicating a positive correlation between these items.

2) When the total purchase value is high (total=high), there is an 84% probability that the customer will purchase bread and cake as well. This association also has a lift value of 1.17.

3) Customers who buy both biscuits and milk-cream have an 84% likelihood of purchasing bread and cake in the same transaction, with a lift value of 1.17.

4) If a customer buys biscuits and fruit together, there is an 84% chance they will also purchase bread and cake. The lift value for this association is 1.17.

5) When customers purchase biscuits and frozen foods together, there is an 83% probability of them also buying bread and cake. The lift value for this association is 1.16.

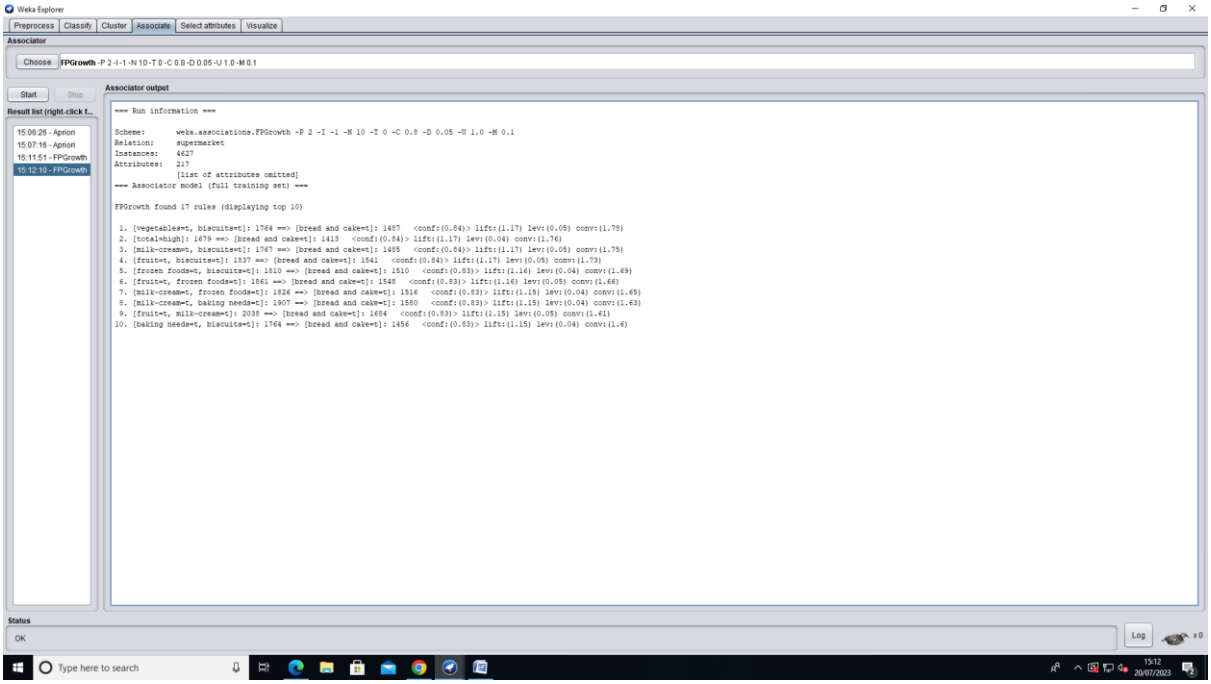

Figure-2: Screen Shot of Experimental Results

By analyzing the generated sets of large itemsets, we discovered several high-confidence association rules that highlight strong relationships between different product categories. For instance, customers who purchased biscuits and vegetables were highly likely to buy bread and cake in the same transaction. Similarly, when the total purchase value was high, customers tended to purchase bread and cake as well. These associations can be leveraged by supermarkets to optimize product placement, design targeted promotions, and enhance cross-selling opportunities.

VI. CONCLUSION

In this study, we conducted association rule mining using the FP-Growth algorithm on a real-world supermarket dataset with a minimum support of 0.3 and a minimum confidence of 0.85. The results revealed interesting and valuable patterns in customer purchasing behavior, which can have significant implications for supermarket management and marketing strategies.

In conclusion, the association rule mining results obtained in this study have practical implications for supermarket management, enabling them to make data-driven decisions to boost sales, improve customer satisfaction, and optimize store operations. As the retail industry continues to evolve, association rule mining techniques like FP-Growth will remain valuable tools for extracting actionable insights from vast transactional datasets, leading to more efficient and effective retail strategies.

REFERENCES

[1] D. Hand, H. Mannila, P. Smyth.: Principles of Data Mining. The MIT Press. (2001)

[2] F. H. AL-Zawaidah, Y. H. Jbara, and A. L. Marwan, “An Improved Algorithm for Mining Association Rules in Large Databases,” Vol. 1, No. 7, 311-316, 2011

[3] Ian H. Witten and Eibe Frank. Data Mining: Practical machine learning tools and techniques.2nd ed. San Francisco: Morgan Kaufmann, 2005.

[4] J.Han and M.Kamber,”Data Mining concepts and Techniques”, the Morgan Kaufmann series in Data Management Systems, 2nd ed.San Mateo, CA; Morgan Kaufmann, 2006.

[5] N.Michael, “Artificial Intelligence – A Guide to Intelligent Systems”, 2nd Edition, Addison Wesley 2005

[6] R. Agrawal, T. Imieliński, and A. Swami, "Mining association rules between sets of items in large databases," in ACM SIGMOD Record, vol. 22, pp. 207–216, 1993

[7] R. Srikant, "Fast algorithms for mining association rules and sequential patterns," UNIVERSITY OF WISCONSIN, 1996.

[8] UCI Machine Learning Repository. https://archive.ics.uci.edu/ml/ datasets.