CUSTOMER PREFERABLE PRODUCT SETS GENERATION USING DATA MINING

[1]Abdurashitova Muniskhon, [2]Khamidulla Khabibullaev, [3]Ugiloy Saidova

[1]Department of Control and Computer Engineering, Turin Polytechnic University in Tashkent

[2]Phd at Department of Control and Computer Engineering, Turin Polytechnic University in Tashkent

[3]Phd at Department of Taxes and Taxation, Tashkent State University of Economics

https://doi.org/10.5281/zenodo.10547029

Abstract. *The process of extracting and identifying patterns from massive data sets using techniques that lie at the confluence of database systems, statistics, and machine learning is known as data mining. The overall objective of data mining, an interdisciplinary branch of computer science and statistics, is to extract information from a data collection (using clever techniques) and organize it into a structure that can be understood and used further. In this work, we will analyze the online retail data using data mining techniques to generate customer preferable sets to improve the company's overall income.*

Keywords*: Data mining, Association rules, Apriori, FP-Growth, Market basket analysis.*

I OBJECTIVE

Generating customer-preferable product sets can improve the overall profit of trading organizations by suggesting compelling products to their clients. Data analysts use several data mining techniques to inspect transactional data collection to create frequently bought item sets. In this paper, we will apply Association rules algorithms to invent frequent patterns and compare the results of several techniques.

II INTRODUCTION

The rapid growth of e-commerce applications has led to increased data accumulation. Data mining, also known as Knowledge Discovery in Databases, is the process of identifying variations, connections, patterns, and trends in data to forecast outputs. In data mining, a frequently used approach is referred to as Apriori. It determines which items in a dataset appear most frequently and which relationships are significant. For instance, products that customers bring into a store might all be utilized as inputs in this system. An efficient market basket analysis is essential because it makes it easier for customers to buy their products, which promotes market sales. Moreover, it has been utilized in the medical field to assist in identifying adverse drug reactions. The Apriori approach is improved by the FP-Growth algorithm. It is not necessary to generate candidates to generate a frequent pattern. The FP growth algorithm represents the database as an FP tree, also known as a frequent pattern tree. This tree structure will preserve the relationships between the item sets. One frequently occurring item is used to fragment the database. "Pattern fragment" refers to this broken section. These fragmented patterns' item sets are examined. Therefore, this method reduces the relative search time for frequent item sets.

III MAIN PART

The Apriori algorithm was first introduced by Agrawal and Srikant in 1994. Its main function is to operate on databases that contain transactions. These transactions can be in different formats, such as market basket, textual, and structured data. Each transaction is viewed as a set of items, which is also known as an itemset. If an itemset has k elements, it is called a k-itemset. This technique employs several parameters such as "confidence" and "support" to define it completely. Given a transactional database T, item sets A and B are any non-empty set of items within T. The association rule:

$$A \Rightarrow B$$

defines the co-occurrence of item sets. The frequency of occurrence of items in T is commonly referred to as support 3.1, whereas confidence 3.2 is the frequency of B in transactions containing A. The Apriori algorithm uses a threshold to identify the item sets, which are subsets of at least T transactions in the database following the principle: *"if an itemset is frequent, then all of its subsets must also be frequent"*.

$$support = \frac{\#\{A,B\}}{|T|} \quad (3.1)$$

$$confidence = \frac{support(A,B)}{support(A)} \; (3.2)$$

The Apriori algorithm is a level-based approach used for extracting association rules. At each iteration, it extracts item sets of a specific length, k. At each level, the first step involves candidate key generation by either generating candidates of length k+1 by joining frequent item sets of length k or by applying the Apriori principle, which involves pruning length k+1 candidate item sets that contain at least one k-itemset that is not frequent. The second step involves frequent itemset generation by scanning the database to count the support for k+1 candidates and pruning candidates below the minimum support [1].

The Apriori algorithm is a data mining technique used for generating association rules. However, the process of joining and pruning item sets as well as testing a portion of each transaction against candidates can be computationally expensive, especially when dealing with a huge number of candidates, which requires a large amount of memory during algorithm execution. To address this issue, an alternative algorithm called FP-Growth was proposed, which is more efficient and scalable, and provides better performance in the analysis of generating association rules. The FP-Growth algorithm uses an efficient and scalable method for mining the whole set of frequent patterns by pattern fragment construction. This method employs an extended prefix-tree structure for storing compressed and critical information about frequent patterns, which is known as the frequent-pattern tree (FP-tree) [2]. Several studies have demonstrated that the FP-Growth algorithm outperforms other prominent methods for mining frequent patterns, such as the Apriori Algorithm and TreeProjection. Later investigations have further revealed that FP-Growth surpasses other approaches like Eclat and Relim in terms of efficiency and popularity. Because of the FP-Growth Algorithm's popularity and efficiency, several studies have proposed variants to improve its performance over time.

The FP-Growth Algorithm is a technique that uses a divide-and-conquer approach to find frequent patterns in a large dataset. This algorithm stores the item set association information in a data structure called the frequent-pattern tree (FP-tree). Firstly, the input database is compressed, and an FP-tree is generated to represent frequently occurring objects. Next, the compressed database is divided into smaller databases, each containing a single frequent pattern. Finally, the algorithm mines each of these smaller databases independently. It's worth noting that the FP-tree cannot be kept in main memory for larger databases. To address this issue, the database is divided into smaller databases (called projected databases), and an FP-tree is constructed for each of these smaller databases.

IV EXPERIMENTAL ANALYSIS

In this experiment, we are using two algorithms to analyze market basket data from The Online Retail Data Set [3], found in the UCI Machine Learning repository. This data set covers all transactions made by a UK-based online retailer between 01/12/2010 and 09/12/2011[4]. Our objective is to generate association rules between several products. To achieve this, we first preprocess the dataset to extract all possible item sets, which are collections of items contained in a single invoice. Then, we extract a list of frequent item sets using FP-Growth and Apriori implementations. Finally, we derive several different association rules from these frequent item sets. We will experiment with different minimum support threshold values, ranging from 0.01 to 0.75, in increments of 0.025. For each threshold value, both the Apriori and FP-Growth algorithms generate a corresponding number of frequent patterns.

	Minimum support threshold							
	0.75	**0.5**	**0.25**	**0.1**	**0.075**	**0.05**	**0.025**	**0.01**
FP-Growth	0	0	0	1	4	23	168	1472
Apriori	0	0	0	1	4	23	168	1472

Table 1. Number of frequent item sets obtained by Apriori and FP-Growth algorithms for different support thresholds

Consider the scenario where minimum support is set to 0.025. As per the data presented in Table 1, there are 168 item sets. A majority of these item sets will be single items that appear in at least 2.5% of transactions. Therefore, we will focus on item sets that have at least two items that are groupings of items that customers tend to purchase together. This is an important insight as it can be used to suggest relevant products to customers, such as the "frequently bought together" feature on Amazon. With the list of commonly occurring item sets, we can now focus on automatically extracting all relevant association rules based on confidence metric (3.2) obtain the overall results, we establish two intervals. The first interval involves adjusting the support values to ascertain frequent item sets, which are [0.01, 0.015, 0.02, 0.025, 0.03, 0.035]. The second interval involves determining association rules, considering the confidence metric within [0.5, 0.55, 0.6, 0.65, 0.7, 0.75]. In Figure 1, we can see the behavior of the association rules based on two metrics. When we set the minimum support threshold and confidence values to smaller values, we find a great number of association rules. However, as we increase the minimum support threshold, the number of frequent item sets decreases, which also leads to a decrease in the number of association rules in that interval.

V CONCLUSION

In conclusion, market basket analysis is a powerful technique in data mining that helps businesses gain insights into customer purchasing behavior. By analyzing the items that customers tend to buy together, businesses can identify patterns and make informed decisions about product placement, pricing, and marketing strategies. This information can help increase sales and customer satisfaction while also improving business efficiency. However, it is important to note that market basket analysis is just one of many techniques in data mining, and it should be used in conjunction with other methods to gain a complete understanding of customer behavior and preferences. By applying two famous data mining algorithms Apriori and FP-Growth, we are able to obtain significant results providing interesting association rules.

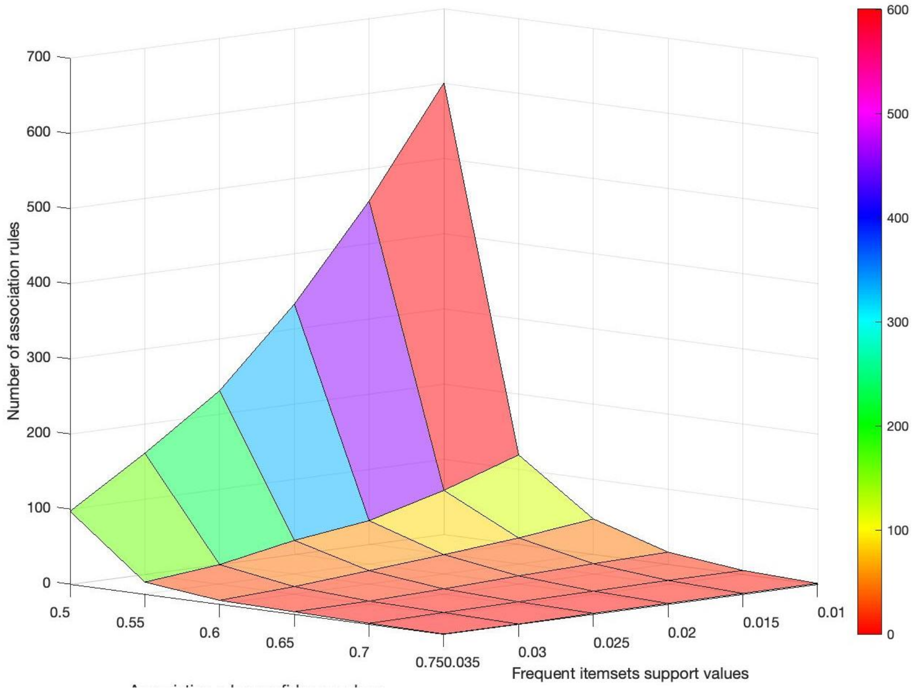

Figure 1: Variation of the number of association rules based on two metrics support and confidence.

REFERENCES

1. https://www.educative.io/answers/what-is-the-apriori-algorithm
2. https://www.javatpoint.com/fp-growth-algorithm-in-data-mining
3. D. Chen, S. L. Sain, and K. Guo, “Data mining for the online retail industry: A case study of rfm model-based customer segmentation using data mining,” Journal of Database Marketing & Customer Strategy Management, vol. 19, no. 3, pp. 197–208, 2012.
4. https://archive.ics.uci.edu/dataset/352/online+retail