**Into The Shopping Basket**

**Sarath Ravi**

*Data Mining*

*University of Detroit Mercy*

Detroit, USA

[ravisa@udmercy.edu](mailto:ravisa@udmercy.edu)

***Abstract***—Into the Shopping Basket utilizes different Data Mining algorithms to evaluate two current methods of the Market-Basket analysis. The Market-Basket analysis is a widely used strategy to increase knowledge of consumers purchasing behavior. Due to the importance of this analysis in the strategy of many companies it is equally important that the techniques utilized produce reliable results. The purpose of this paper is to compare different methods and conclude which method is most suitable using the following measures of Support and Confidence and by comparing their efficiency and accuracy of the algorithms. This paper involves selecting appropriate datasets and mining the hidden association rule in the dataset, then subsequently comparing the execution time and accuracy of the results of FP-Growth and Apriori algorithm.

**Keywords:** Market-Basket analysis, Support, Confidence measure, FP-Growth algorithm, Apriori algorithm, execution time.

**I. Introduction**

The algorithms chosen for evaluation are as follows:

1. Apriori Algorithm
2. FP-Growth

Apriori algorithm utilizes association rule learning, which is a machine learning tactic that checks for the dependency of one variable being true on another being true. This is helpful in Market-Basket analysis because it identifies the products that are purchased together and puts them into larger data sets.

FP-Growth produces similar results; however, the algorithm goes through two passes over the data and while doing so counts the single-most frequent variables and then seeks correlation between them and then constructs an FP-tree from the data obtained.

The measurements of assessment utilized are as follows:

1. Support count
2. Confidence measure
3. Execution time

Support Count “refers to how often a given rule appears in a dataset.”

Confidence measure refers to the overall accuracy of the rule during implementation of the algorithm.

Execution time refers to the length of time it takes for an algorithm to run.

# **II. Related Work**

## **A.** **Related Work #1**

The blog entitled “Market Basket Analysis: A Comprehensive Guide for Businesses,” explains how data mining techniques like Market Basket Analysis helps industries such as Retail, Hospitality, E-commerce, etc. Furthermore, it explains “what is Market Basket Analysis and how it works”. Association rule mining algorithms utilized are Apriori algorithm, SETM algorithm, FP Growth and AIS algorithm. Overview of all algorithms were given and finally explained implementation of Market Basket Analysis using the article's preferred method of analysis, the Apriori algorithm.

### Link: https://www.analyticsvidhya.com/blog/2021/10/a-comprehensive-guide-on-market-basket-analysis/

## 

## **B.** **Related work #2**

This paper implements a set-oriented algorithm for mining frequent itemsets in relational databases. In this paper the sort and merge scan algorithm SETM is implemented for a supermarket data set. This paper compares the frequent itemset and its execution time and the results with the traditional Apriori algorithm. Concluded that the SETM algorithm works well and stable and provides similar performance with the traditional Apriori algorithm.

Link: https://indjst.org/download-article.php?Article\_Unique\_Id=INDJST6661&Full\_Text\_Pdf\_Download=True

**C.** **Related work #3**

In this paper, initially discussed “architecture and scope of data mining” and various association rule algorithms and finally compared two algorithms: Apriori algorithm and Filter Associator. They have analyzed the frequent itemsets generation and number of cycles performed over the Apriori algorithm and Filter Associator in the context of association analysis. According to the comparison of the above two algorithms on weka tool, they conclude that Filter Associator is a more efficient algorithm than Apriori algorithm based on the two factors: number of cycles performed, large item sets. The Apriori algorithm generates greater numbers of cycles performed and generates extra-large itemsets which degrades the performance of the algorithm.

Link: https://www.ijcsmc.com/docs/papers/June2015/V4I6201552.pdf

## **D.** **Related work #4**

The article, “Fast Algorithms for Mining Association Rules,” refers to the following algorithms as being significant in Market-Basket strategy: AIS, SETM, Apriori, and AprioriTid. The first two were compared to the latter. The results of this comparison concluded that the Apriori methods outperformed the AIS and SETM algorithms when utilizing either “synthetic or real-world data.” Furthermore, the article took a step that produced an additional algorithm which was entitled the “AprioriHybrid” - a mix of both Apriori algorithms; this hybrid strategy was the article's most favored tactic.

Link: https://citeseerx.ist.psu.edu/document?repid=rep1&type=pdf&doi=608667cc9f2a5517a7bf2995beb5bbc4e1709eef

## **E.** **Related work #5**

The following article, “A Survey of Association Rules,” emphasizes the importance of association rules in Market research. Defining the use of association rules as, “are used to identify relationships among a set of items in a database.” Furthermore, the process of association rules was defined as: initially finding all of variables that occur a greater frequency than the support (variable chosen by the analyst); then the second step is finding the associations between frequent variables within the larger datasets provided by the first step. The following algorithms were compared: AIS, SETM, Apriori, Apriori-TID, Apriori-Hybrid, Off-line Candidate Determination (OCD), Partitioning, Sampling, and CARMA. After thorough analysis of each above the deciding analysis was that of the Apriori family, deeming that they are the “most commercially available implementations to generate association rules.” However, it can and should be noted that the article found that different algorithms were deemed more suitable than others depending on the datasets/variables used.

Link:

<https://s2.smu.edu/~mhd/pubs/00/ar.doc>

# **III. Methodology**

# **Datasets used:**

The following dataset was utilized for the comparative analysis of the Apriori and FP-Growth algorithms:

1. Purchase data from December 2018 - December 2021 from a medium sized jewelry store.

Link:

https://www.kaggle.com/datasets/mkechinov/ecommerce-purchase-history-from-jewelry-store

**Data preparation:**

Sampling was deemed the most reasonable tactic for data selection. This process involves selecting a subset of data from the population (that is representative of the population) and then using the subset for analysis. To ensure that the data was being represented accurately a larger sample than necessary was utilized for analysis; and 100 selections were made from each dataset.

The following preprocessing steps were deemed necessary for these datasets:

1. Data Cleaning
2. Data Transformation

Data cleaning entails making sure that the data we are using is complete and correct. Each dataset sample was scanned for incomplete/inconsistent variables and corrections were made on an as needed basis.

Data Transformation entailed making sure that the data from each sample was formatted correctly from the original database to the software utilized for analysis.

**Data Mining techniques:**

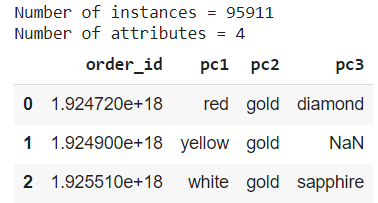
In purchase data from medium sized jewelry store, we have data under the column names such as “datetime, product\_id, quantity, order\_id, category\_id, category\_alias, Brand\_id, price, user\_id, gender and colors of the jewels purchased in each order (pc1, pc2,pc3)”.

We are finding out what color of jewels are related together.

As a part of data preprocessing, we are dropping some unnecessary columns.

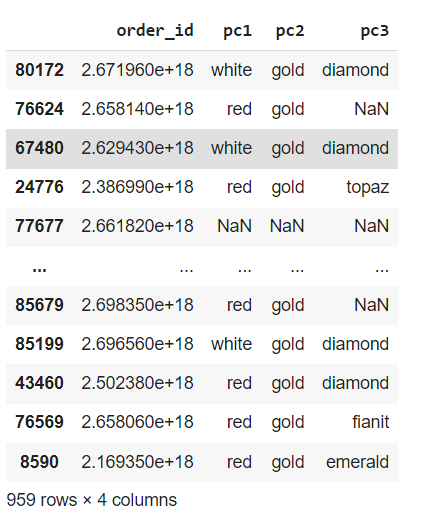
We first downloaded the data using Pandas read\_csv() function and displayed its first 3 data points. Dropped some of the columns and printed the number of rows and attributes in the table.

Output:



Sampling:

We are taking one percent of the total number of instances as a sample, without replacing the data.



As shown above, there are 959 rows and 4 columns in the sample data.

**Apriori:**

Once we have read the dataset completely, we are required to get the list of items in every transaction. So, we are going to run two loops. One will be for the total number of transactions, and the other will be for the total number of columns in every transaction. Executing apriori algorithm and printing the generated association rules.

Output:

frozenset({'sapphire', 'white'}) frozenset({'gold', 'sapphire', 'white'})

The above colors are related. Let us see the rules and their support, confidence level.

Output:

Rule: sapphire -> white

Support: 0.007306889352818371 Confidence: 0.5384615384615384

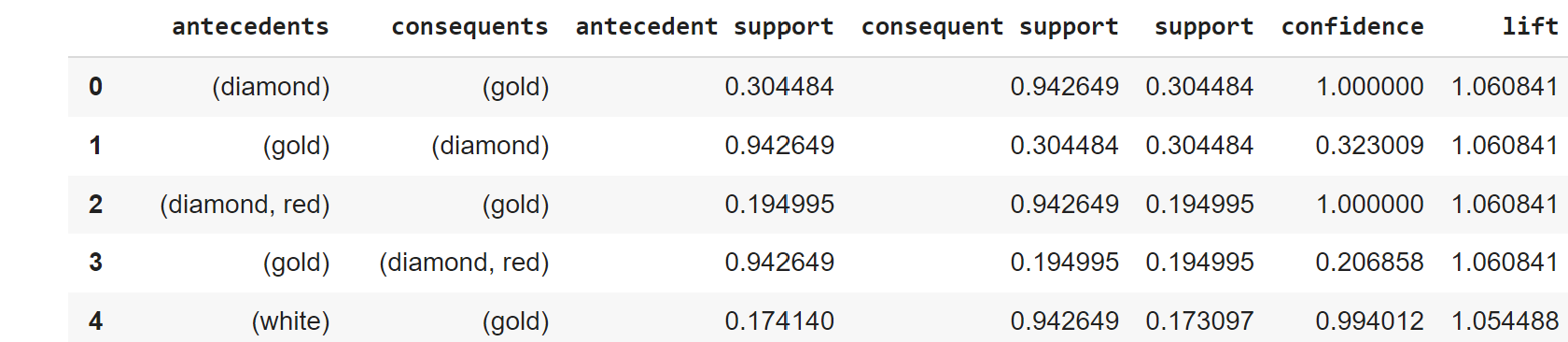
-------------

Rule: gold -> sapphire

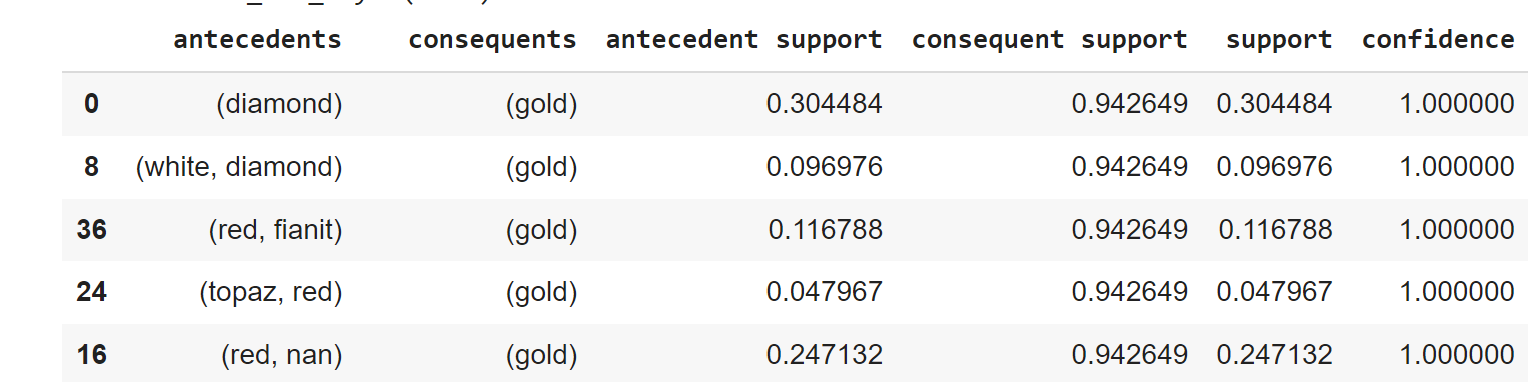
Support: 0.007306889352818371 Confidence: 0.5384615384615384

**FP-Growth:**

The process for FP-Growth initially began much like that of Apriori. The algorithm's initial pass over the data identifies the most frequently occurring variables. Furthermore, the algorithm filters out the non-frequently occurring variables with minimum support. See the below examples of support produced:



Below is the second pass in which the variables with the greatest occurrences and connections were pinpointed:



# **IIII. Conclusion**

Based on the results of the Apriori and FP-Growth algorithms we determined that the FP-Growth algorithm is the most suitable for Market-Basket analysis. As seen from the tables displayed (in the previous sections) the FP-Growth algorithm displayed important units of measure. For instance, FP-Growth had short execution times and higher support and confidence measures. This indicates that not only were the algorithms suitable for the dataset/purpose of the outcomes, but that the data supported a high degree of accuracy regarding the FP-Growth algorithm.

##### **References**

[1]<https://indjst.org/download-article.php?Article_Unique_Id=INDJST6661&Full_Text_Pdf_Download=True>

[2] https://www.ijcsmc.com/docs/papers/June2015/V4I6201552.pdf

[3] https://citeseerx.ist.psu.edu/document?repid=rep1&type=pdf&doi=608667cc9f2a5517a7bf2995beb5bbc4e1709eef

[4] Kadlaskar, A. (2023, April 26). *Market basket analysis: A comprehensive guide for businesses*. Analytics Vidhya. https://www.analyticsvidhya.com/blog/2021/10/a-comprehensive-guide-on-market-basket-analysis/

[5] Mining Association rule - SMU. (n.d.). https://s2.smu.edu/~mhd/pubs/00/ar.doc