A Visual Exploration of Groceries Dataset

Nirmal Dahal[1] and Lokesh Sapkota[1]
[1]Department of Computer Science and Engineering, KU

***Abstract*— This paper explores the application of two prominent algorithms, Apriori and FP-Growth, in Market Basket Analysis (MBA) using a comprehensive dataset. While Apriori utilizes a traditional support-based approach for identifying frequent itemset, FP growth employs a tree structure. Two metrics of support and confidence are discussed highlighting their ability to reveal patterns and the reliability of association rules. A comparative analysis of Apriori and FP-Growth is presented considering their strengths and weaknesses along with insights into the impact of support and confidence thresholds. Leveraging a grocery dataset with 9835 transactions and 169 unique items, the research provides practical insights for the retail sector. In addition, this paper highlights the integration of Streamlit as a powerful tool for presenting and visualizing data related to Market Basket Analysis. The report showcases key findings from Exploratory Data Analysis (EDA), uncovering transaction patterns and notable association rules.**

***Index Terms*— Apriori, Confidence, FP-Growth, Frequent Itemsets, Market Basket Analysis, Support**

I. INTRODUCTION

MARKET basket analysis is a data mining technique that helps identify products that customers frequently buy together. It uses association rules to find patterns in purchase history and optimize product placement, pricing, and marketing. For example, if a customer buys chicken, they are likely to buy chicken seasonings as well. This is often represented as an association rule: “Chicken” -> "Seasonings".

It can help to improve customer understanding, inventory management, pricing strategies, and sales growth. This paper explores market basket analysis through a thorough analysis of Apriori and FP-Growth algorithms.

The Apriori algorithm is a popular data mining technique used for frequent itemset mining and association rule learning over relational databases [1]. It was proposed by R. Agrawal and R. Srikant in 1994. The Apriori algorithm can highlight general trends in the database.

In Apriori, at each step, the algorithm must scan the database to build the candidate sets. These steps can be redundant, and a new association-rule mining algorithm was developed named Frequent Pattern Growth Algorithm. FP algorithm stores all the transactions in a trie data structure.

Three ways to measure association are used here. They are:

1. Support
2. Confidence
3. Lift

Support is a measure used to identify itemsets that are interesting or frequent in the transaction dataset. It is calculated as the number of transactions containing a particular item set divided by the total number of transactions. For example, in a dataset of 1000 transactions, the itemset {Chicken, Seasonings} appearing 100 times would mean that the itemset {Chicken, Seasonings} has a support of 10%.

The formula to calculate support is:

$$\text{Support(itemset)} = \frac{\text{No. of transactions with the itemset}}{\text{Total number of transactions}}$$

Confidence is a measure of the reliability or support for a given association rule. It is defined as the proportion of cases in which the association rule holds true [2]. For example, in a dataset of 1000 transactions, if the itemset {Chicken, Seasonings} appears 100 times and the itemset {Chicken} appears in 200 of those transactions, the confidence of the rule “If a customer buys Chicken, they will also buy Seasonings” would be 50%.

The formula to calculate confidence is:

$$\text{Confidence(rule)} = \frac{\text{Support(itemset} \cup \text{consequent)}}{\text{Support(itemset)}}$$

Lift is a measure used to evaluate the performance of an association rule. It compares the probability of occurrence of the itemsets in the rule together to the probability of occurrence of the itemsets independently.

The formula to calculate lift is:

$$\text{Lift(rule)} = \frac{\text{Confidence(rule)}}{\text{Support(consequent)}}$$

If the lift is greater than 1, it means that the items in the rule are more likely to be bought together than at random. Conversely, a lift of less than 1 indicates that the items are unlikely to be bought together. A lift of exactly 1 implies that the probability of occurrence of the antecedent and that of the consequent are independent of each other.

II. WORKING

The Apriori algorithm finds (n + 1) itemsets from n items by using an iterative level-wise search technique. E.g., let's take a sample example of transaction details of 5 items as shown in Table 1.

The Apriori algorithm treats each item as an itemset and determines support based on their frequency in the dataset. Itemset with support equal to or more than the minimum threshold are retained. This process of scanning the database continues until no more itemsets are left with more than the minimum threshold support.

TABLE I
SAMPLE TRANSACTION DETAILS

Transaction ID	List of items
T100	I1, I2, I5
T200	I2, I4
T300	I2, I3
T400	I1, I2, I4
T500	I1, I3
T600	I2, I3
T700	I1, I3
T800	I1, I2, I3, I5
T900	I1, I2, I3

If minimum threshold support is 2:

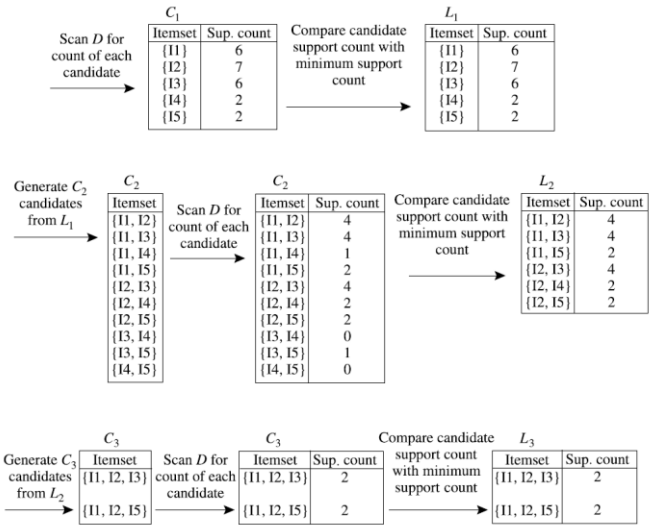

Fig. II-1. Generation of the candidate itemsets and frequent itemsets, where the minimum support

Initially, items with a support of 2 are chosen, and in subsequent steps, item sets with a minimum support count of 2 are consistently processed further.

Frequent Pattern growth algorithm represents data in a tree structure that maintains association information on frequent items and the tree is referred as FP-tree. Once the FP-Tree is constructed, it is divided into a collection of conditional FP-Trees, each associated with a frequent item. These conditional FP-Trees can be individually mined and analyzed separately.

E.g., let's take a sample example of transaction details of 5 items as shown in Table 2.

TABLE II
SAMPLE TRANSACTION DETAILS

Transaction ID	List of items
T1	I1, I2, I3
T2	I2, I3, I4
T3	I4, I5
T4	I1, I2, I4
T5	I1, I2, I3, I5
T6	I1, I2, I3, I4

If Support threshold = 50%, Confidence = 60%, then minimum support = 0.5*6 =3

TABLE III
COUNT OF EACH ITEM

Item	Count
I1	4
I2	5
I3	4
I4	4
I5	2

TABLE Iv
SORT THE ITEMSET IN DESCENDING ORDER

Item	Count
I2	5
I1	4
I3	4
I4	4

Now, building Frequent Pattern tree:

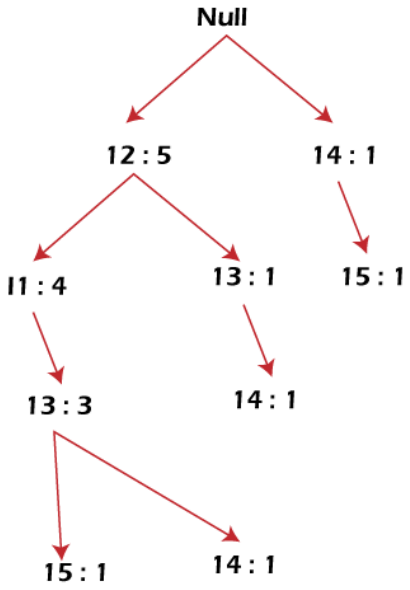

Fig. II-2 Generated FP tree.

Frequent Pattern is now generated in Table VI.

TABLE VI
FREQUENT PATTERN GENERATED

Item	Conditional Pattern Base	Conditional FP-tree	Frequent Patterns Generated
I4	{I2, I1, I3:1}, {I2, I3:1}	{I2:2, I3:2}	{I2, I4:2}, {I3, I4:2}, {I2, I3, I4:2}
I3	{I2, I1:3}, {I2:1}	{I2:4, I1:3}	{I2, I3:4}, {I1:I3:3}, {I2, I1, I3:3}
I1	{I2:4}	{I2:4}	{I2, I1:4}

III. EXPLORATORY DATA ANALYSIS (EDA)

Grocery dataset from Kaggle is used. The portion of the database is given below:

```
citrus fruit,semi-finished bread,margarine,ready soups
tropical fruit,yogurt,coffee
whole milk
pip fruit,yogurt,cream cheese,meat spreads
other vegetables,whole milk,condensed milk,long life bakery product
whole milk,butter,yogurt,rice,abrasive cleaner
rolls/buns
other vegetables,UHT-milk,rolls/buns,bottled beer,liquor (appetizer)
potted plants
whole milk,cereals
tropical fruit,other vegetables,white bread,bottled water,chocolate
citrus fruit,tropical fruit,whole milk,butter,curd,yogurt,flour,bottled water,dishes
beef
frankfurter,rolls/buns,soda
chicken,tropical fruit
butter,sugar,fruit/vegetable juice,newspapers
fruit/vegetable juice
packaged fruit/vegetables
chocolate
```

Fig. III-1Portion of the database

This database was transformed into a transaction matrix where 1 denotes that the item was present in the specific transaction and 0 denotes that the item was absent in the specific transaction. Below is a snapshot of the sample transaction matrix derived from the dataset:

	citrus fruit	semi-finished bread	margarine	ready soups	tropical fruit	yogurt	coffee
0	1	1	1	1	0	0	0
1	0	0	0	0	1	1	1
2	0	0	0	0	0	0	0
3	0	0	0	0	0	1	0
4	0	0	0	0	0	0	0
5	0	0	0	0	0	1	0
6	0	0	0	0	0	0	0
7	0	0	0	0	0	0	0
8	0	0	0	0	0	0	0
9	0	0	0	0	0	0	0

Figure III-2 Sample transaction matrix

The dataset consists of 9835 transactions as rows and 169 numbers of items as columns.

From Fig III-3, it can be observed that items with support of more than 500 are most frequent in the transaction. Thus, the support value is established at 500 / 9835 = 0.05.

Moving on to Fig III- 4, the illustration highlights itemsets having a length equal to 2 with each itemset support being greater than or equal to minimum support.

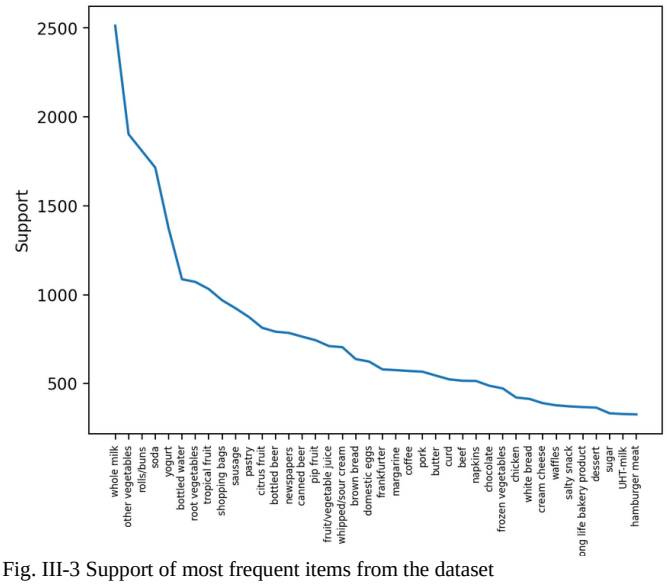

Fig. III-3 Support of most frequent items from the dataset

	support	itemsets	length
28	0.0560	frozenset({'whole milk', 'yogurt'})	2
29	0.0748	frozenset({'other vegetables', 'whole milk'})	2
30	0.0566	frozenset({'whole milk', 'rolls/buns'})	2

Fig. III-4 Result when n-length itemsets with each itemset support are≥ minimum support

	antecedents	consequents	antecedent support	consequent support	support	confidence	lift	leverage	conviction	zhangs_metric	antecedent_length	consequent_length	length
0	frozenset({'yogurt'})	frozenset({'whole milk'})	0.1395	0.2555	0.056	0.4016	1.5717	0.0204	1.2441	0.4227	1	1	2
1	frozenset({'other vegetables'})	frozenset({'whole milk'})	0.1935	0.2555	0.0748	0.3868	1.5136	0.0254	1.214	0.4208	1	1	2
2	frozenset({'rolls/buns'})	frozenset({'whole milk'})	0.1839	0.2555	0.0566	0.3079	1.205	0.0096	1.0757	0.2085	1	1	2

Fig. III-5 Result when P(Consequent | Antecedent) ≥ threshold confidence of 0.30

In Figure III-5, the interrelation between itemsets becomes apparent, particularly when the confidence threshold is set at ≥ 0.30. The strongest rule is (Yogurt -> Whole Milk). From the figure, the likelihood of purchasing whole milk alongside yogurt stands at 40%. Thus, it is advisable to strategically position both items together in the store to potentially enhance overall sales.

In Fig. III-6, Subplot 1,1 describes a linear relationship among items with the highest support, indicating associations with other items. To reveal this relationship, the dataset was filtered with a support threshold of 0.02 and a confidence threshold of 0.4.

Some rules in Fig III-7 have "whole milk" as the resulting item (consequent) with high confidence. This means that if other items in the rule are bought, it's very likely that whole milk will also be bought. However, because whole milk is already frequently purchased (high support) and often appears in transactions with other items (positive correlation shown in the support-lift curve), these rules might not be as insightful as they seem.

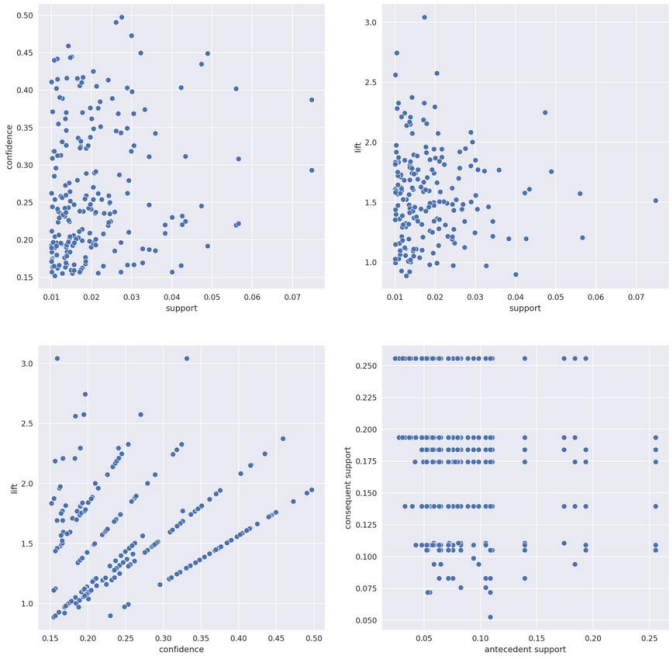

Fig. III-6 Relationship between metrics for determining associativity.

	antecedents	consequents	antecedent support	consequent support	support	confidence	lift	leverage	conviction	zhangs_metric
0	frozenset({'margarine'})	frozenset({'whole milk'})	0.0586	0.2555	0.0242	0.4132	1.6171	0.0092	1.2687	0.4053
1	frozenset({'tropical fruit'})	frozenset({'whole milk'})	0.1049	0.2555	0.0423	0.4031	1.5776	0.0155	1.2473	0.409
2	frozenset({'yogurt'})	frozenset({'whole milk'})	0.1395	0.2555	0.056	0.4016	1.5717	0.0204	1.2441	0.4227
3	frozenset({'butter'})	frozenset({'whole milk'})	0.0554	0.2555	0.0276	0.4972	1.9461	0.0134	1.4808	0.5147
4	frozenset({'curd'})	frozenset({'whole milk'})	0.0533	0.2555	0.0261	0.4905	1.9195	0.0125	1.4611	0.506
5	frozenset({'beef'})	frozenset({'whole milk'})	0.0525	0.2555	0.0213	0.405	1.5852	0.0078	1.2513	0.3896
6	frozenset({'root vegetables'})	frozenset({'whole milk'})	0.109	0.2555	0.0489	0.4487	1.756	0.0211	1.3504	0.4832
7	frozenset({'whipped/sour cream'})	frozenset({'whole milk'})	0.0717	0.2555	0.0322	0.4496	1.7598	0.0139	1.3527	0.4651
8	frozenset({'domestic eggs'})	frozenset({'whole milk'})	0.0634	0.2555	0.03	0.4728	1.8502	0.0138	1.412	0.4906
9	frozenset({'frozen vegetables'})	frozenset({'whole milk'})	0.0481	0.2555	0.0204	0.4249	1.6631	0.0081	1.2946	0.4189
10	frozenset({'root vegetables'})	frozenset({'other vegetables'})	0.109	0.1935	0.0474	0.4347	2.2466	0.0263	1.4267	0.6228
11	frozenset({'whipped/sour cream'})	frozenset({'other vegetables'})	0.0717	0.1935	0.0289	0.4028	2.0819	0.015	1.3506	0.5598

Fig. III-7 Relationship between 2 items

Two interesting rules involving "other vegetables" can also be observed in Fig III-7:

- **Root vegetables -> other vegetables:** If customers buy root vegetables, there's more than a 40% chance they will also buy other vegetables.
- **Whipped/sour cream -> other vegetables:** Similarly, if customers buy whipped cream or sour cream, there's a 40% chance they will also buy other vegetables.

These rules suggest an association between these specific items and other vegetables.

	antecedents	consequents	antecedent support	consequent support	support	confidence	lift	leverage	conviction	zhangs_metric
12	frozenset({'other vegetables', 'yogurt'})	frozenset({'whole milk'})	0.0434	0.2555	0.0223	0.5129	2.0072	0.0112	1.5283	0.5246
13	frozenset({'other vegetables', 'root vegetables'})	frozenset({'whole milk'})	0.0474	0.2555	0.0232	0.4893	1.9148	0.0111	1.4577	0.5015
14	frozenset({'root vegetables', 'whole milk'})	frozenset({'other vegetables'})	0.0489	0.1935	0.0232	0.474	2.4498	0.0137	1.5333	0.6222

Figure III-8 Relationship between 3 items

Fig III-8 shows the relationship of itemsets having length 3. When a customer buys just root vegetables, there's a 40% chance they'll also purchase other vegetables. However, this probability jumps to 47.4% if they buy both root vegetables and whole milk. This suggests that whole milk, in combination with root vegetables, further increases the likelihood of buying other vegetables.

IV. Benchmark

TABLE VI
Benchmark Table

Algorithm	(50, 5000)	(50, 9835)	(169, 5000)	(169, 9835)
Apriori	0.35	0.55	0.95	2.5049
FP growth	0.08	0.107	0.13	0.26

[Here (50, 5000) denotes: Number of Unique items = 50 and Number of transactions = 5000]

Table VI compares the time it takes for the Apriori and FP-Growth algorithms to identify all frequent itemsets in a dataset with a given number of unique items and a given number of transactions. As shown, FP-Growth has a significant performance advantage for this task.

References

[1] R. Agrawal, V. Profile, R. Srikant, and O. M. A. Metrics, "Fast algorithms for Mining Association rules in large databases: Proceedings of the 20th International Conference on Very Large Data Bases," DL Hosted proceedings, https://dl.acm.org/doi/10.5555/645920.672836 (accessed Dec. 14, 2023).

[2] "What is support and confidence in data mining?" GeeksforGeeks, https://www.geeksforgeeks.org/what-is-support-and-confidence-in-data-mining/ (accessed Dec. 14, 2023).