Market Basket Analysis of Instacart

Group members: Suhasini Kalaiah Linagiah, Bharath kumar karre, Sumanth Pobala, and Ari Sagherian

Class: CIS 5700 Intro to Big Data

Abstract

The utility of Big Data is transforming the shopping experience in remarkable ways, particularly by suggesting items to users based on their prior purchases and similarity to others. For this project, we used the Instacart dataset which has over 3 million online grocery purchases. From this data, we generated frequent itemsets and association rules with the Apriori algorithm, then clustered users using the K-means algorithm, and finally made recommendations of items using the bigram frequencies of items. The combination of these three steps allows us to predict the item bundles that a user will buy next time, offering immense benefit to businesses.

Introduction

The shopping experience has been transforming with advances in computational abilities and techniques. With the advent of Big Data analytics, similar items can now be grouped near each other to facilitate the shopping experience of customers. For example, without Big Data analytics, grocery stores would not put diapers and beer near each other because these seem to be wholly separate items on the surface until associations are found between them in the data.

The impacts of Big Data analytics are not restrained to brick and mortar stores though. Many companies are now delivering food directly to the front door of customers' homes. One of these companies, Instacart, provides a publicly available, large dataset of 3 million customer purchases along with the day of the week, time of purchase, and relative time in between purchases for each user [4]. The size and nature of this dataset make it an ideal candidate to test Big Data approaches in order to discern patterns and potentially find creative ideas to improve profits for food delivery companies.

This project focuses on the ability to predict and recommend items for users. These will be based on the association rules based on the similarity of items as well as the similarity in features of users, to give as accurate and personalized of a prediction for purchases with the ultimate goal of increasing profits for a prospective company.

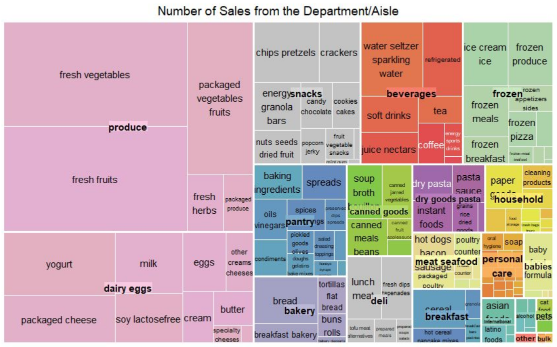
Related work

The inspiration to do this project came from a mixture of class lectures and research papers that explore methods to form association rules and make item predictions. The conceptual framework that inspired the use of Apriori for market basket analysis came from the foundational paper by Brin *et al.* [3]. They explored the implications of using Apriori and other dynamic algorithms for forming association rules. They found that Apriori works better than others on high support thresholds where support is the number of occurences of an itemset. Most importantly, this paper showed that the use of market basket analysis is much more suited for grocery store transactions than for census data and similar datasets with high levels of correlations and redundancies. This narrowed our focus to the Instacart dataset since they proved Market Basket analysis is well-suited for this type of dataset.

Having decided on the dataset, we searched for papers that performed market data analysis on retail purchases. The approach used by Annie *et al.* provided a practical framework upon which our approach was built. They utilized a K-Apriori approach to form association rules and then clustered the users in groups to predict future items [2]. Similar to the notes from the CIS 5570 lecture on Frequent Itemsets [1], Annie *et al.* took advantage of the monotonic property of support measures in which the support of an itemset never exceeds the support of its subsets. This allowed Annie *et al.* to form frequent itemsets using the Map/Reduce architecture in an efficient manner even on sparse datasets. The customers were then grouped into clusters based on the similarity of their purchased items using the K-means algorithm and then recommending items to certain clusters [2].

Methodology

This project analyzed the Instacart dataset in four phases. The first phase explored the features of the Instacart dataset to find general patterns. Exploratory data analysis (EDA) was conducted to visualize patterns and further used in prediction models. Examples can be seen in Figures 1 and 2 in which the number of sales per department and number of sales per day can be visualized.



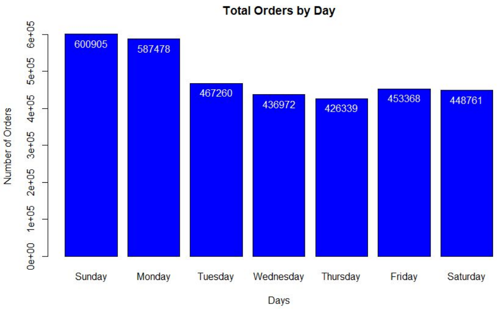


Figure 1: Sales per department depicted by size of boxes

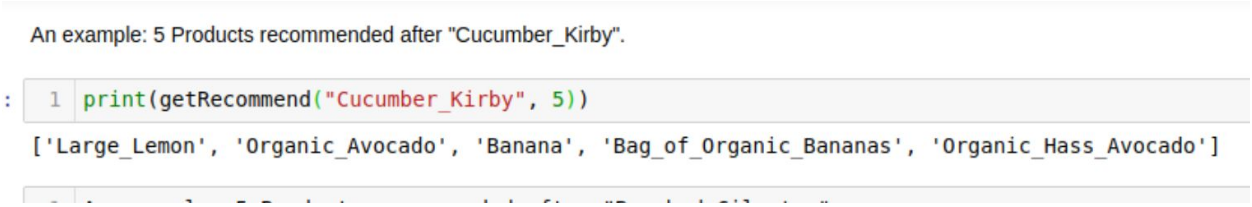
Figure 2: Number of orders by day

The second phase of our project performed association rule mining on the entire dataset using the Apriori algorithm. As described in slides 30-35 of the Frequent Itemsets presentation from CIS 5570 [1], the Apriori algorithm is implemented on Hadoop using Map/reduce to find frequent itemsets. It accomplishes this by iteratively generating candidate frequent itemsets and pruning non-frequent itemsets. This repeats *k* times for whatever size *k* itemsets are desired. The efficiency of this algorithm is that it utilizes the monotonicity rule and the contrapositive of it. The monotonicity rule states that if an itemset is frequent (meaning its support is greater than the threshold), then all subsets of it must also be frequent. The contrapositive states that if an itemset is not frequent, then all supersets of it cannot be frequent [2]. The combination of these rules allows for efficient pruning of itemsets so that the bottleneck of main memory can be managed [1]. From the frequent itemsets, association rules were generated in the form of A -> B, which translates to the probability that item B will be bought given item

A is bought. To calculate this, the confidence is computed by the formula: confidence(A -> B) = support(A, B) / support(A) and support of any itemset is the number of occurences in the dataset.

The third part of this project was to cluster the users based on similar habits that were found from part 1, the exploratory data analysis phase. The K-means algorithm was employed to measure the similarity of users and cluster them into groups. The K-means algorithm we used was adopted from Spark.mlib, which is a parallelized version of the standard K-means algorithm adapted for use in Big Data analytics. We decided to cluster the users based on the weekdays of purchases, days since last order, number of total products and orders, and hours of purchases. By implementing the elbow method, 40 clusters were created from which we identified purchasing patterns and grouped them. Since we used our own machines, we were restricted to using 10% and 20% of the dataset for clustering.

The fourth part of this project aimed to recommend product bundles based on bigram frequencies from 33% of the dataset. Bigrams are co-occurences of 2 items (Item A, Item B) and their probabilities are derived from the association rules using Apriori. After clustering users into groups with K-means, we then recommended bundles of 5, 10, and 15 items to groups based on their frequently purchased items. Finally, to evaluate the accuracy of the recommended items the number of recommended items found in the order are divided by the total number of items in the order. For example, if an order contains 10 products, we start by recommending a bundle of 10 items similar to item 1 based on the top 10 associations for item 1. Then, for each occurrence of the recommended item in the original order, the score is incremented by 1. This process is repeated for each item in the order and the final score is divided by the size of the order. This process is then repeated for all orders and an average score is reached.



An example: 5 Products recommended after "Cucumber_Kirby".

```
print(getRecommend("Cucumber_Kirby", 5))
['Large_Lemon', 'Organic_Avocado', 'Banana', 'Bag_of_Organic_Bananas', 'Organic_Hass_Avocado']
```

Figure 3: Example of a bundle of 5 items recommended based on bigram frequencies of “Cucumber_Kirby”

Experimental Discussion

The goal of our experiment was to find patterns in the data and determine the accuracy of recommending bundles of items to clusters of users based on their past purchases. The EDA revealed that Sundays and Mondays had the highest number of orders, most purchases are from 8 A.M. to 6 P.M., most healthy products are ordered during the daytime, and most unhealthy items are ordered at night. For the recommendation system, we began by computing the association rules for the entire Instacart dataset. Then, we clustered the users into groups of 10 based on their similarities in number of transactions. Following this, we iterated through orders placed in different clusters, generated a recommendation bundle of items similar to the size of each order based on the top associations (bigrams), and then computed the accuracy of recommended items being in the order. We were restricted by the memory of our machines to only 33% of the dataset and produced a final accuracy of 17.94% after iterating through the different clusters and orders.

Contribution of each member

1. Sumanth Pobala - Performed Exploratory Data analysis, Implemented Apriori Algorithm

2. Suhasini Kalaiah Linagiah - Implemented Apriori Algorithm, Performed Clustering of users based on behavioral patterns for recommendation.

3. Bharath Kumar Karre - Post grouping the users by the behavioral patterns and association rules, Implemented the product bundles and generated recommendation list based on the purchase frequency.

4. Ari Sagherian – Helped find the project/dataset, wrote the report and presentation, and minorly assisted with code implementation.

Conclusion

The approaches used produced an accuracy of 17.94% for recommended items being purchased. This shows promise with this approach and potential applicability to companies, particularly online retailers that simply need to recommend items instead of stock different items near each other. Going forward, more computers can be utilized to nullify the bottleneck that our computer's memory played and the Apriori algorithm could be expanded for trigrams and larger itemsets than just bigrams.

References

1. Abouelenien, Mohamed; Retrieved from CIS 5570 Introduction to Big Data Lecture, Chapter 6 Frequent Itemsets. (Dec 2019), Dearborn, Michigan, USA
2. Annie, Loraine Charlet M C; Kumar, Ashok D.; Market Basket Analysis for a Supermarket based on Frequent Itemset Mining. International Journal of Computer Science Issues (IJCSI); Mahebourg Vol. 9, Iss. 5, (Sep 2012): 257-264.
3. Brin, Sergey; Motwani, Rajeev; Ullman, Jeffrey D.; Tsur, Shalom; Dynamic itemset counting and implication rules for market basket data, Proceedings of the 1997 ACM SIGMOD international conference on Management of data, p.255-264, (May 11-15, 1997), Tucson, Arizona, USA [doi>10.1145/253260.253325]
4. "The Instacart Online Grocery Shopping Dataset 2017", Accessed from https://www.instacart.com/datasets/grocery-shopping-2017 on 12/7/2019