Redesigning a Retail Store Based on Association Rule Mining

Dalal Al Attal, Mariam Naser, Nourah AlBaghli, Nourah Al Muhaimeed and Sarah Al Awadh
American University of the Middle East
250 St, Eqaila, Kuwait
dalattal_32@yahoo.com , Nalmuhaimeed@hotmail.com , Noraalbaghli@hotmail.com , sarah.alawadh@gmail.com

Abstract

In this study, we perform Market Basket Analysis in a supermarket, and find associations between different product categories by using Apriori and Éclat algorithms, and represent the findings by using Tableau program was used to represent the data graphically data visualization tool. We first analyzed the point of sales data (POS) data to see which SKUs are purchased together, and we categorize the SKUs into main product categories to find the associations between products. After finding the associations among product categories, we redesign the store layout based on these findings, and calculated the distance traveled by a typical customer to choose the best store layout. It turns out that the first layout alternative tends to improve the walking distance by an average of 13.3%, for a typical customer, the second one will improve the distance traveled by 8% on average.

Keywords Data Mining, Association Rule Mining, Market Basket Analysis, Facility Layout

1. INTRODUCTION

Market Basket Analysis is a data mining technique that is used widely to find the associations among products. Market Basket Analysis provides the retailer useful information on products are brought together by its customers. This information may be used to influence customers' purchases, to re-organize the current layout of the retail shop, allocating the products on the shelves, etc. Market Basket Analysis results allows companies make easier decisions on how to bundle products and how to reorganize stores in a better way to increase sales and increase customer satisfaction. Whenever a customer enters a retail store she walks around with the intention of buying the items in mind. During grocery shopping, some unlisted items might grab the customer's attention, and she will most probably get the item, which is most likely not directly related to any of the other items in the basket. There's a high probability that amongst various customers we will find similar associations between items. When checking out at the counter the items bought look unrelated to each other, but that may not be the case. The items bought may be highly dependent on each other, where if a customer buys item "A" there is a high chance she would buy item "B". Market Basket Analysis is done for decision making in choosing the appropriate allocation of products in the retail's layout. This is done by finding how the items are dependent on each other.

There are many hidden and interesting relationships between the products in the supermarkets that cannot be spotted easily. These relationships are represented in the form of association rules by having a left and right hand side. The association rule consists of two parts, the antecedent, A and the consequent, B. Association rules are considered interesting if they satisfy a minimum support threshold. Given A is the antecedent and B is the consequent, the support is the probability that a transaction contains both A and B. The confidence on the other hand is a conditional probability, where the probability of having both A and B is divided by the probability of having A.
There are also various algorithms that are available for performing Market Basket Analysis. Some of the many known algorithms include the Apriori, Éclat, FP Growth, Decision tree algorithm and the Apriori prefix tree

algorithm. Sometimes algorithms are combined into a hybrid algorithm, one of them is the AprioriHybrid algorithm. Since the data being dealt with in this project is large, Éclat can be used since it deals with large data in less time. The documentation of such association rules can help and support sellers in decision making regarding marketing strategies. As a result, customer satisfaction is achieved which will cause an increase in revenue.

2. LITERATURE REVIEW

Several studies have been conducted on data mining in market basket analysis talking about the algorithms, post analysis, how to work with large data, how to get the interesting relationships and how association rules are used in different fields.

Association mining can be applied in many fields such as Market Basket Analysis, bio-medical literature, census data, logistic regression, medical diagnosis, and fraud detection in web and in other several fields (Rajak, 2008). Market Basket Analysis is a widely used technique to improve product allocation. Tanusondjaja, Nenycz-Theil, and Kennedy, (2015) studied the structure and size of the shoppers' baskets. They used the co-occurrence method in the Market Basket Analysis to record the coincidence of pairs of items. The predictable patterns gave what are expected for the use of the store layout management, found out which categories compete in the store and product category interaction.

In another related situation based on basket purchases, Hilage, and Kulkarni (2011) applied association rule mining technique, rule induction technique and Apriori algorithm to the database of shopping malls. From the calculations, they observed that Sunflower Oil, Sugar and Rava are highly correlated to each other.

There are many different algorithms that can be used to obtain the association rules. Rahman et al. (2016) used the K-means algorithm, to analyze large volumes of data in different ways, relationships were found in order to put them together into useful information to be used later. Annie and Kumar (2012) used the K-Apriori algorithm to divide the customers into different clusters. After implementing the algorithm on Anatha stores the related items are placed near each other and promotions were made. This achieved customer satisfaction, and in return increased the overall revenue. Another paper discussed the LIPI algorithm, which is a more improved algorithm compared to the traditional algorithms (Zhi, Yan, & Yan, 2016).

Since there are different types of algorithms, the case studies above show that some prefer one over the over. Algorithms are chosen based on memory utilization, number of scans and the time required to obtain the association rules. The Apriori algorithm, for example requires more memory in comparison to the FP-Growth (Pawar, 2013).

Most of the time the amount of data that requires analysis is large. Venkatachari, and Chandrasekaran (2016) wanted to see how different items interrelate to take advantage of them. She has chosen to apply the Apriori algorithm and FP growth, which were very slow due to the large data set. When she used the R programming Apriori algorithm multiple rules were generated in milliseconds. Agrawal and Srikant (1994) introduced two new algorithms that are quite different from the known algorithms which are combined into a hybrid algorithm called AprioriHybrid.

Sometimes new methods are used instead of new algorithms. In addition to the data mining techniques, Papavasileiou and Tsadiras (2013) suggested a new method to classify the most valuable association rules among the large variety set of the known rules using Apriori algorithm. They used the assessment of the lift, support, and confidence, as well as their time variation by introducing Overall Variability of Association Rules and Overall Variability of Products measures. Based on these variations the Overall Variability Association Rule Method was developed which identifies inflexible rules.

One of the problems in data mining is the post-analysis of the found rules. Adomavicius and Tuzhilin (1999) aimed to show a method of validating the rules with the help of a human expert. He approached the issue by taking transactional histories and building personalized profiles of the users using various mining techniques. The system he worked with was able to validate 98.5% of the rules.

Other than looking at profiles sometimes customer behavior is categorized as what Prokeinova (2014)'s main focus was, which was sustainable consumption calculated through Apriori algorithm. The results were conducted as categorizing the consumers into student/non-student categories in terms of either consumption or behavior when purchasing items. Looking into their transactions it shows that younger people have more tendency of buying products from retail shops in a sustainable way more than older people. Parents, usually value products based on money not on personal preferences. Roodpishi and Nasshtaei (2015) used K-Means clustering method for a big amount of data obtained in an insurance company in Anzali, Iran to find patterns. They picked the ideal number of clusters in order to obtain the essential data for grouping the customers. Singh (2013) talked about sport stores used FP-Growth algorithm on the collected data that lead to the sale optimization of the sport equipment.

Deshpande, Shirpada, Patole and Bulsara (2010) showed how different methods of data mining can be applied to different fields. To conclude, it is very challenging to design and improve a data mining system, which can work correctly for different domains. Similarly Geetharamani, Revathy, and Jacobs (2015) aimed to predict patterns of how many times a user visits a web page from the categories of msn.com. According to their navigating behavior by using Apriori prefix tree algorithm to create the association rules. This research had a confidence of 100% and with the high value of lift that the frontpage, news, local, opinion, on-air, weather, living, business, sports and bbs were the pages that had many visits compared to the others. In another case study a categorizing technique was used as a solution, Yen et al. (2016), used mining consumption pattern algorithm and consumption association rule in mining the sequential data of purchasing behaviors patterns in Taiwan, to help the seller endorse a product that the customer might need in the future based on the patterns they have gathered.

Most association rules may seem too obvious, Raeder and Chawla (2010) found that it is quite difficult to isolate interesting relationships, to differentiate between the obvious and interesting rules. The only methods the authors have found were interestingness and redundancy and they were of limited use. Thus, they addressed this problem by modeling transactional data as a network, which helped to find strong relationships not found in traditional association rules. In another study Surjandari, Rachman, and Lusia (2012) used multi-level association rules as a way to find interesting rules. They studied the factors affecting the profitability of several mini markets located in close regions in Jakarta, Indonesia. They used multi-level association rules and the zero-integer programming to optimize the model. It resulted in allocating the most frequent item set sold on the shelves level.

Other than applying the mining techniques in markets, they are also applied in several fields for example, Grady (2017) talked about how data mining empowers the legal services organizations. He concludes that legal service firms have to build data management strategies and begin applying them through their establishments. In another field, Burton, Morris, Giraud-carrier, West, and Thackeray (2014) used the association rule mining methods to a big set of questionnaires and surveys in the USA. They showed that these rules were able to test the hypothesis, group the data to find highly related questions, and then terminate the item sets that contain questions from the same group. Likewise Dimitrijević and Krunić (2013) used association rule mining of an educational institution website out of the log files. They calculated the confidence and lift in two different years, along with that they compared the different two outcomes, which helped in enhancing the websites.

In our literature review we reviewed several papers on association mining algorithms. Some of the common algorithms include Apriori, K-Apriori and FP growth. As a second group of papers we concluded that large datasets require either using R Programming coding or hybrid algorithms. After attaining the associated rules, some of the papers stated that post-analysis is required to validate the constructed results. The final set of papers talked about how Market Basket Analysis is used in many fields other than in retail stores, like websites, legal organizations and questionnaires.

3. DESIGN

a. **Data Collection and Analysis**

b. The data collected for our project was a set of transactions from Al-Shuhada supermarket for a period of 6 months. The reason behind obtaining that amount of data was to avoid any trends, errors and bias assumptions. Since the data was very large, random transactions were chosen from the six month period. These transactions will provide the number of frequent items that occur in this specified time, later association assumptions can be made

Figure 1: Procedure of Market Basket Analysis

based on these transactions. After that, we analyzed the obtained data by first of all, translating all of the data from Arabic to English and inputted them manually, as this process cannot be done by a computerized system. While inputting the transactions, we cleaned them by extracting the unnecessary information such as the general stock items and promotions. Meanwhile, we categorized the items using the merchandise hierarchy that we have prepared. After that, we implemented the association mining algorithms on the data. In order to have an effective plan of study, we have designed Figure 1.

c. Theory behind the different modules

As shown in Figure 3 1, the first thing we did was that we collected P.O.S data for a selected period of time from a supermarket in Kuwait, specifically Al-Shuhada Supermarket. After that, we cleaned the P.O.S data, which consists of detecting and removing the inaccurate records from the transactions, removing repeated items from the same transaction and translating the Arabic transactions to English. Since the type of items in the transactions were very specific, the items are then categorized into general categories. For example milk, chips, nuts and etc. Several association algorithms are then implemented such as Apriori and Éclat on the gathered P.O.S. After that, we will extracted association rules based on the support measure. Finally recommendations are made based on the association rules, which include cross-selling, and an improved idea of the layout of the items.
The importance of the association rules is quantified by calculating either of the support, confidence and lift measures. The support is the probability that both items would be found in a random transaction, as shown in equation 1.

$$s(A \Rightarrow B) = P(A,B) = P(A \cap B) \quad (1)$$

The confidence is the probability that a transaction contains the antecedent also contains the consequent. It is the probability of B given A, as shown in equation 2.

$$c(A \Rightarrow B) = P(B|A) \quad (2)$$

The lift is calculated to find the degree to which the antecedent and the consequent are dependent on one another. The lift takes into consideration both the whole data and the confidence. The lift is calculated by dividing the confidence by the probability of the consequent, as shown in equation 3. The lift is a value that is either 1 or greater than 1. So if the value equals one then the antecedent and consequent are independent of each other. When the lift is greater than 1, the value is equivalent to the degree to which both items are dependent on each other.

$$L(A \Rightarrow B) = \frac{c(A \Rightarrow B)}{P(B)} = \frac{P(B|A)}{P(B)} = \frac{P(A \cup B)}{P(A)P(B)} \quad (3)$$

A merchandise hierarchy is a way of providing a wide-range of categories of the available products in a retail store. The function of the hierarchy is to categorize products. The merchandise hierarchy consists of various levels. The highest level of the hierarchy is the store, which is the customer’s destination. The second level is the general category of the products. As the hierarchy goes down, the product category narrows to classes, followed by sub-classes and ending with a stock keeping unit level (SKU) as seen in Figure 3 2. SKU is the level where products are uniquely described based on their characteristics such as; brand, color, and size /volume.

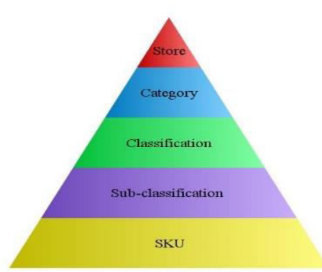

Figure 2: Merchandise hierarchy

d. Algorithm

Choosing a Market Basket Analysis algorithm required a main step, the minimum support threshold was adopted, which gave us a reasonable number of rules. The minimum support is then applied to find all the frequent itemsets in the database, to remove the items that are under the minimum support threshold. There are many different

algorithms that can be implemented to find association rules, such as, Apriori, FP Growth, ELCAT, FP tree, and the Decision tree algorithms.
The Apriori algorithm follows a downward closure property, where any subset of a frequent itemset must be frequent. For example if {nuts, shampoo, juice} exceeds the minimum support threshold in the current frequent itemset (k), then so has {nuts, shampoo} in the previous frequent itemset (k-1). That is so, because every transaction that contains {nuts, shampoo, juice} also contains {nuts, shampoo}. The frequent (k-1) itemsets is used to generate k itemsets candidates, by scanning the databases. Apriori is an efficient way to make a list of frequent itemsets from the collected data.
There are two well-known algorithms Apriori and ÉCLAT which are going to be implemented to find associations between different items. Apriori is defined as deduction. The Apriori algorithm was proposed in 1994 by Agrawal and Srikant to find the frequent item sets in a database, specifically transactions. On the other hand, ÉCLAT was proposed by Zaki (2000), which is efficient for discovering frequent itemset in different transactions. Unlike most algorithms, both algorithms are useful since the dataset being dealt with is transactional data.

4. METHODOLOGY

Alteryx Designer is a program used to merge and investigate data in simple steps to help in reducing the time needed to get useful association rules, which is important for decision making. Since analyzing a large set of data can be a time-consuming procedure for the team working on it, using this computer program can be beneficial. Alteryx Designer uses different tools to deliver insight of customer's behavior through finding the highly correlated items. The first step in using Alteryx Designer was to have the obtained transactions into excel, the trim function was used to eliminate spaces before and after the invoice number and categories. The categories that were used are shown in Table 1.

Table 1: The List of Categories Used

Categories				
Butter	Creamer	Garbage Bags	Pastry	Beans and Grains
Soup	Powdered Milk	Baby Care	Rusk and Breadstick	Refrigerated Coffee
Body Care	Fresh Milk	Juices	Frozen Beef	Honey
Hair Care	Carton Milk	Frozen Meals	Frozen Chicken	Chips
Shower Essentials	First Aid	Syrups	Air Refreshers	Biscuits
Pudding	Pet Supplies	Chocolate	Sauces	Canned Food
Healthy Snacks	Spread Cheese	Cereal	Cotton Buds	Feminine Products
Instant Food	Coffee	Solid Cheese	Tissue	Shaving Appliances
Oral Care	Home Cleaning Supplies	Frozen Vegetables	Eggs	Laundry Essentials
Frozen Seafood	Labnah	Yogurt	Popcorn	Pest Control
Refrigerated Meat	Spices	Snack Bars	Pita Bread	Spreads
Cream and Whipped Toppings	Croissant	Gum	Gummy and Chewy Candy	Toast
Salad Dressings	Water	Energy Drinks	Iced Tea	Hard Candy
Buns	Pasta	Flour and Others	Baking	Ciders
Nuts	Tea	incense	Hand Soap	Oils
Ice Cream	Arabic Sweets	Vegetables and Fruits	Fizzy Drinks	Rice
Leban	Food Jars	Cakes	Shredded Cheese	Refrigerated Meals
Oats	Sugar	Cotton	Flavored water	Facial Care

To input the data, the Text Input tool is used instead of Input Data tool, since the Input Data tool is limited to a total of a thousand cells. Whereas we are dealing with 6426 cells. The data in the Text Input tool is placed in two fields,

where one field is the transaction key field, which is the invoice number. The other field contains the item identifier, which is the categories. After each tool is inputted, the tool has to be run by clicking on the run button. From the predicative grouping tools, the Market Basket Rules tool is placed after the Text Input tool, where it is the first step of Market Basket Analysis. This tool provides information about the transactions, such as the most frequent items that has occurred. There are two different algorithm options that can be used, Apriori and Éclat. Both algorithms were selected to create analysis with the same chosen minimum support and confidence threshold.
After that, the minimum support threshold was chosen as 3% and the minimum confidence threshold was chosen as 5%. The higher the minimum support and confidence thresholds, the less rules generated. Most Predictive tools have 2 output streams, the report "R" and the output "O".

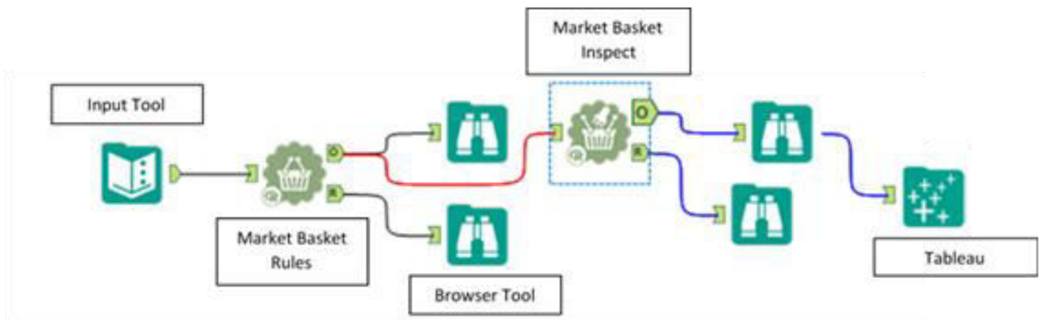

Figure 3: Alteryx work flow

The report stream summarizes both the transaction data and the itemset rules. On the other hand, the output stream produces a model that can be further investigated. It can be used to continue the analysis downstream as it contains all the necessary data. The output stream is usually connected to one or two tools, where in our case is connected to the Browser tool and the Market Basket Inspect tool. The Browser tool is used to review and verify the data (Alteryx, 2017). The Market Basket Inspect tool further investigates the output of the MB Rules tool, to show which items are related to each other by identifying and analyzing rules in the transactions. Therefore, those rules can be further filtered to minimize the number of rules of itemset.
When using Alteryx Designer, the Apriori Algorithm was the algorithm chosen first, which is selected in the Market Basket Rule tool to find association rules. The report stream provided by the Market Basket Rule tool in Table 2 consists of the most frequent items found in the transactions. The vegetables and fruits category has occurred the most, specifically in 576 of the total transactions, followed by chocolate, fizzy drinks, chips and then pita bread categories. The count indicates how frequent each category has happened out of the overall transactions.

Table 2: The Count of the Items in the Transactions

Item	Count
Vegetables and fruits	576
Chocolate	306
Fizzy drinks	297
Chips	235
Pita bread	225

The report provided by the Market Basket Inspect tool, shows the association rules generated, which is shown in Table 3.

Table 3: Association Rules Generated from Alteryx

LHS	RHS	Support	Confidence	Lift
{chips}	{chocolate}	0.0722	0.5234	2.91
{chocolate}	{chips}	0.0722	0.4020	2.91

LHS	RHS	Support	Confidence	Lift
{chocolate}	{veg and fruits}	0.0664	0.3693	1.09
{veg and fruits}	{chocolate}	0.0664	0.1962	1.09
{solid cheese}	{veg and fruits}	0.0646	0.6044	1.79
{veg and fruits}	{solid cheese}	0.0646	0.1910	1.79
{toast}	{veg and fruits}	0.0599	0.5340	1.58
{veg and fruits}	{toast}	0.0599	0.1771	1.58
{juices}	{veg and fruits}	0.0558	0.4798	1.42
{veg and fruits}	{juices}	0.0558	0.1649	1.42
{chips}	{veg and fruits}	0.0534	0.3872	1.14
{veg and fruits}	{chips}	0.0534	0.1580	1.14
{juices}	{chocolate}	0.0505	0.4343	2.42
{chocolate}	{juices}	0.0505	0.2810	2.42
{juices}	{chips}	0.0423	0.3636	2.64
{chips}	{juices}	0.0423	0.3064	2.64
{fizzy drinks}	{chips}	0.0417	0.2391	1.73
{chips}	{fizzy drinks}	0.0417	0.3021	1.73
{fizzy drinks}	{chocolate}	0.0411	0.2357	1.31
{chocolate}	{fizzy drinks}	0.0411	0.2288	1.31
{biscuits}	{chocolate}	0.0382	0.5556	3.09
{chocolate}	{biscuits}	0.0382	0.2124	3.09
{sauces}	{veg and fruits}	0.0370	0.7326	2.17
{veg and fruits}	{sauces}	0.0370	0.1094	2.17
{toast}	{solid cheese}	0.0370	0.3298	3.09
{solid cheese}	{toast}	0.0370	0.3462	3.09
{eggs}	{veg and fruits}	0.0335	0.7215	2.13
{veg and fruits}	{eggs}	0.0335	0.0990	2.13
{fizzy drinks}	{juices}	0.0335	0.1919	1.65
{juices}	{fizzy drinks}	0.0335	0.2879	1.65
{solid cheese}	{chocolate}	0.0335	0.3132	1.74
{chocolate}	{solid cheese}	0.0335	0.1863	1.74
{juices}	{solid cheese}	0.0329	0.2828	2.65
{solid cheese}	{juices}	0.0329	0.3077	2.65
{frozen chicken}	{veg and fruits}	0.0323	0.6471	1.91
{veg and fruits}	{frozen chicken}	0.0323	0.0955	1.91
{laundry essentials}	{veg and fruits}	0.0317	0.6835	2.02
{veg and fruits}	{laundry essentials}	0.0317	0.0938	2.02
{biscuits}	{veg and fruits}	0.0311	0.4530	1.34
{veg and fruits}	{biscuits}	0.0311	0.0920	1.34
{solid cheese}	{chips}	0.0311	0.2912	2.11
{chips}	{solid cheese}	0.0311	0.2255	2.11
{spices}	{veg and fruits}	0.0305	0.7761	2.29
{veg and fruits}	{spices}	0.0305	0.0903	2.29
{toast}	{chocolate}	0.0305	0.2723	1.52
{chocolate}	{toast}	0.0305	0.1699	1.52

In this table, chips and chocolate have the strongest correlation amongst the other categories. This rule can be read as, people who buy chips tend to buy chocolate at the same time with a support, confidence and lift of 0.0722,

0.5234, and 2.91, respectively. A support of 7.22% means that out of all the transactions, 7.22% of them have chips and chocolate purchased together. A confidence of 52%, indicates that given a customer bought chips, has a 52% chance of purchasing chocolate. Finally, since the lift is 2.91 which is greater than 1, it designates that chips and chocolate are highly dependent on each other.
Alteryx Designer was used again to apply the Éclat Algorithm to find the frequent itemsets. Similar to the Apriori Algorithm, two reports were provided, and in contrast frequent itemsets were found.
The Market Basket inspect tool provided a report that consists of the frequent itemsets, which is shown in Table 4. The most items that tend to re-occur together are chips and chocolate with a support of 0.722.

Table 4: Itemsets Generated from Altreyx

Items	Support
{chips,chocolate}	0.0722
{chocolate,veg and fruits}	0.0664
{solid cheese,veg and fruits}	0.0646
{toast,veg and fruits}	0.0599
{juices,veg and fruits}	0.0558
{chips,veg and fruits}	0.0534
{fizzy drinks,veg and fruits}	0.0523
{chocolate,juices}	0.0505
{pita bread,veg and fruits}	0.0440
{chips,juices}	0.0423
{chips,fizzy drinks}	0.0417
{chocolate,fizzy drinks}	0.0411
{cream and whipped toppings}	0.0405
{shower essentials}	0.0393
{biscuits,chocolate}	0.0382
{sauces,veg and fruits}	0.0370
{solid cheese,toast}	0.0370
{gummy and chewy candy}	0.0346
{eggs,veg and fruits}	0.0335
{fizzy drinks,juices}	0.0335
{chocolate,solid cheese}	0.0335
{juices,solid cheese}	0.0329
{frozen chicken,veg and fruits}	0.0323
{laundry essentials,veg and fruits}	0.0317
{biscuits,veg and fruits}	0.0311
{chips,solid cheese}	0.0311
{spices,veg and fruits}	0.0305
{chocolate,toast}	0.0305

Figure 4, shows the comparison between the Apriori and Éclat based on support value. It is shown that the rules or itemsets that were generated from both the Apriori and Éclat algorithms have the same support value. There are only a few rules that are in common and that is because the Apriori algorithm works in a way that the original dataset is scanned, while the Éclat scans the currently generated dataset.

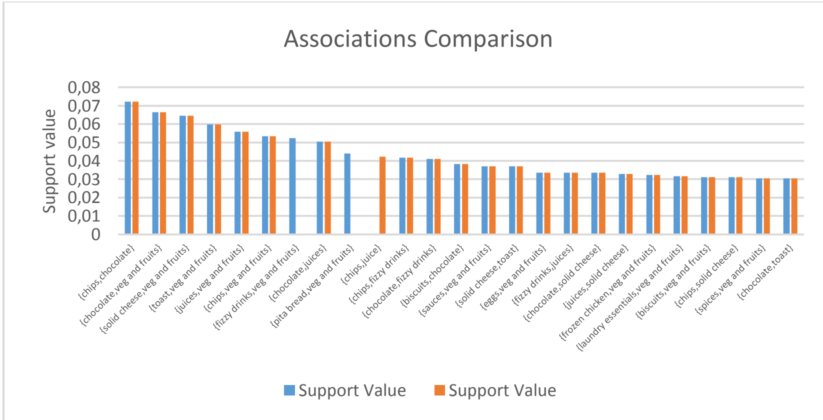

Figure 4: Associations Comparisons Based on Support Value

Tableau is a tool commonly placed at the end of the Alteryx workflow, to transfer the results to a server where we can represent it. After getting the associations of the products through the two algorithms, Apriori and Éclat, the results were represented through a graphical representation using the Tableau tool, such as packed bubble chart and bar charts. These graphs would show us the obtained association rules and itemsets but in a graphical manner, including and excluding what is necessary, as well as arranging them with what is found most suitable and understandable. Simply, this was done by connecting Tableau to Alteryx, and the data will be transferred to Tableau to represent it with any chosen type of graph. This was done twice, once using the results of the Apriori algorithm and once using the Éclat results.

Figure 5 shows the packed bubble chart representing the itemsets gained from the Apriori algorithm through Alteryx. A packed bubble chart is used to show the relationships of products without the use of axes. The larger the circle, the higher the value it has with the others, as an example veg and fruits is the largest, is greatly associated with chocolate, as well as chocolate and chips.

Next, Figure 6 shows a bar chart for the Apriori algorithm results, in descending order, starting with the items that are highly associated with each other and their value, to the least associated items and their values. Using a bar chart

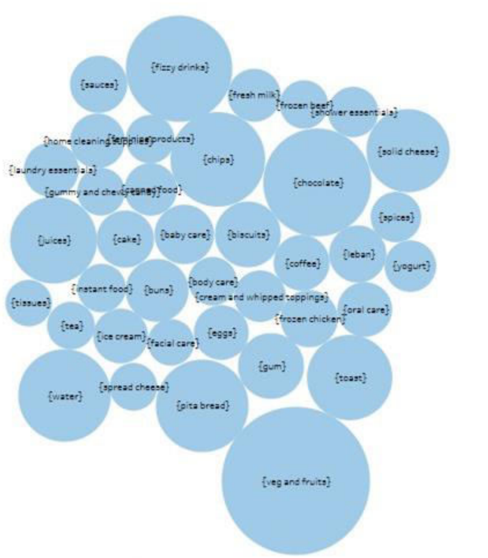

Figure 4: Apriori Software Bubble Chart

is useful as it gives a clearer representation of the associations along with their support values represented across the axis.

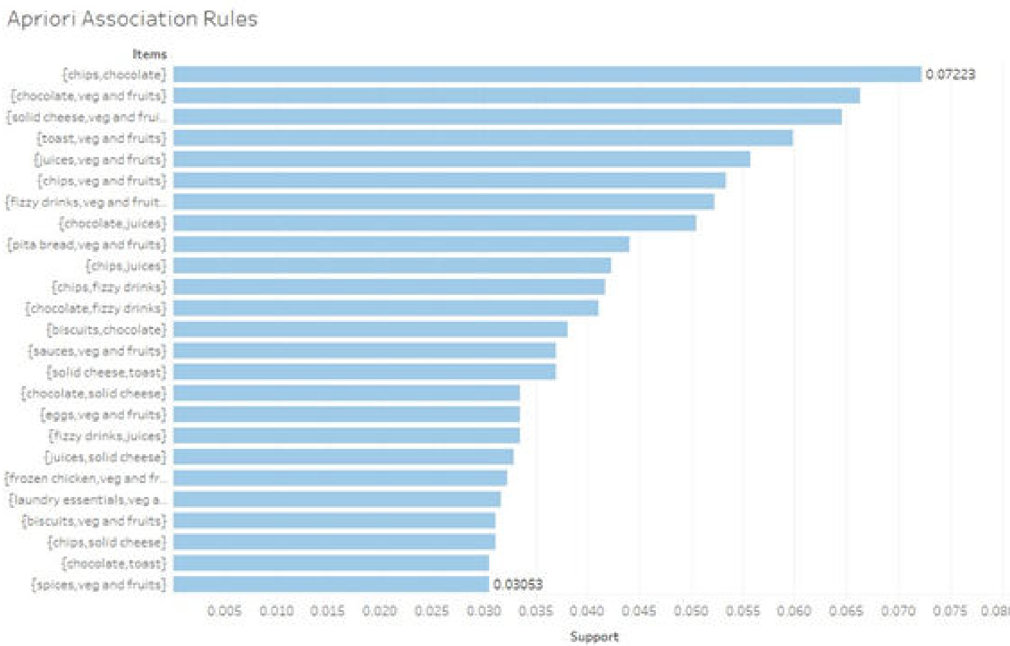

Figure 6: Apriori Association Rules in Descending Order

Figure 7 shows the packed bubble chart but for the itemsets obtained from the Éclat algorithm in Alteryx. These also show that the largest circle of veg and fruits is highly related with chocolate.

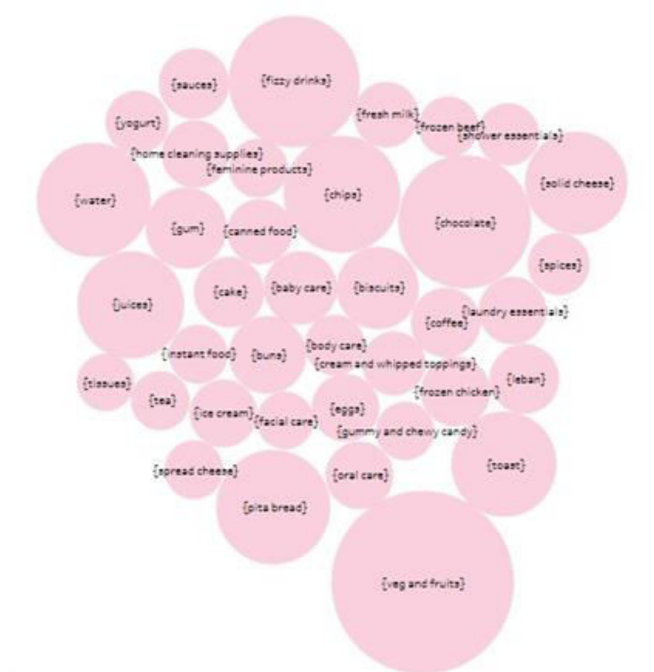

Figure 7: Éclat Packed Bubble Chart

Lastly, Figure 8 shows the bar chart representing the Éclat associated itemsets, arranged in a descending order as well, with the itemset that has the highest support value on top, being chips and chocolate.

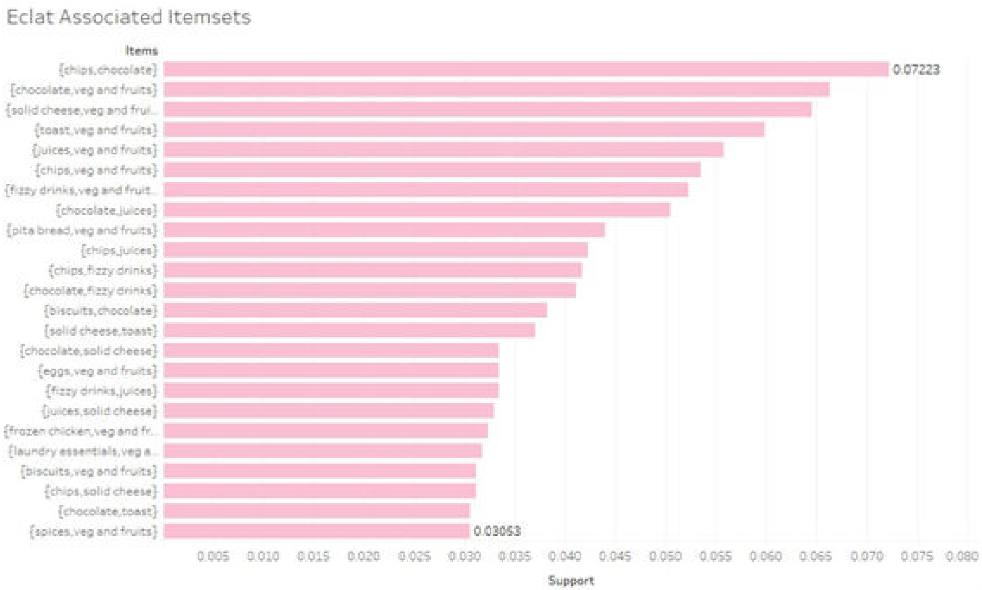

Figure 8: Éclat Association Rules in Descending Order

5. Results

About 60% of shoppers go into the supermarket without a list, thinking that they will remember what they need when they enter the supermarket. The supermarket can take advantage of this by placing associated products next to each other, which leaves them at risk of impulsive buying. Generally, customers do not only follow their purchase plans, they also tend to buy items that are not included in the plan; impulsive purchases. In return, this can increase the revenue of the supermarket as well as the customer's satisfaction, which will create a win-win situation (Schindly, 2017).

A store layout is a critical key element that every retailer uses to display their products in a way that serve both the financial state of the store and the satisfaction of the customers. A proper layout design allows the customer to walk around the store without losing the interest of what's presented on the shelves or feeling tired.

The most common store design used is the traditional layout design, where they are organized as departments, where products with similar functionalities are next to each other. For example, the bakery department consists of bread, biscuit, cakes, and so on. This approach is company oriented, which means they neglect the needs of the time-pressured customers. This type of product placement gives the customers more exposure to the whole store departments, searching for their specific products while walking around the aisles. This approach will consume the customer's time, making it impossible to have a quick trip to the supermarket. A new layout design is preferred to reduce the time in the supermarket, which will reduce the effort needed (Lingonblad, 2015). Having a layout with the minimum distance between products will avoid the time spent for the customer to search for specific items. Therefore, this should be taken into consideration since the customers have time constraints. (Cil, 2012)

A good store layout should have a positive influence on the customers purchasing behavior. Therefore, layouts should be customer oriented. There are several studies made suggesting several approaches, which tend to increase the customer satisfaction by reducing the time spent while increasing the profit of the supermarket (Aloysius & Binu, 2013). Therefore, two different alternative layouts were conducted based on different approaches.

The aim of the re-layout is to decrease the walking distance as it is an important factor that affects the performance of a supermarket. There are two main types of markets, supermarkets and hypermarkets. Al-Shuhada is a supermarket, since it is of a regular size and is located in Al-Shuhada living area. Loyal customers usually select the shortest shopping tour according to their purchase plan, since they are familiar with the layout. To maximize the impulsive purchases, the categories were located by reconstructing the layout of the supermarket, based on the association rules generated. This is done by reducing the time in the system, which is measured in this project via distance traveled in the supermarket. The current layout as shown in Figure 9 was conducted by visiting the supermarket, and writing down the location of each category and their count of occurrence. Figure 10 shows the current categories color coded with the previous figure.

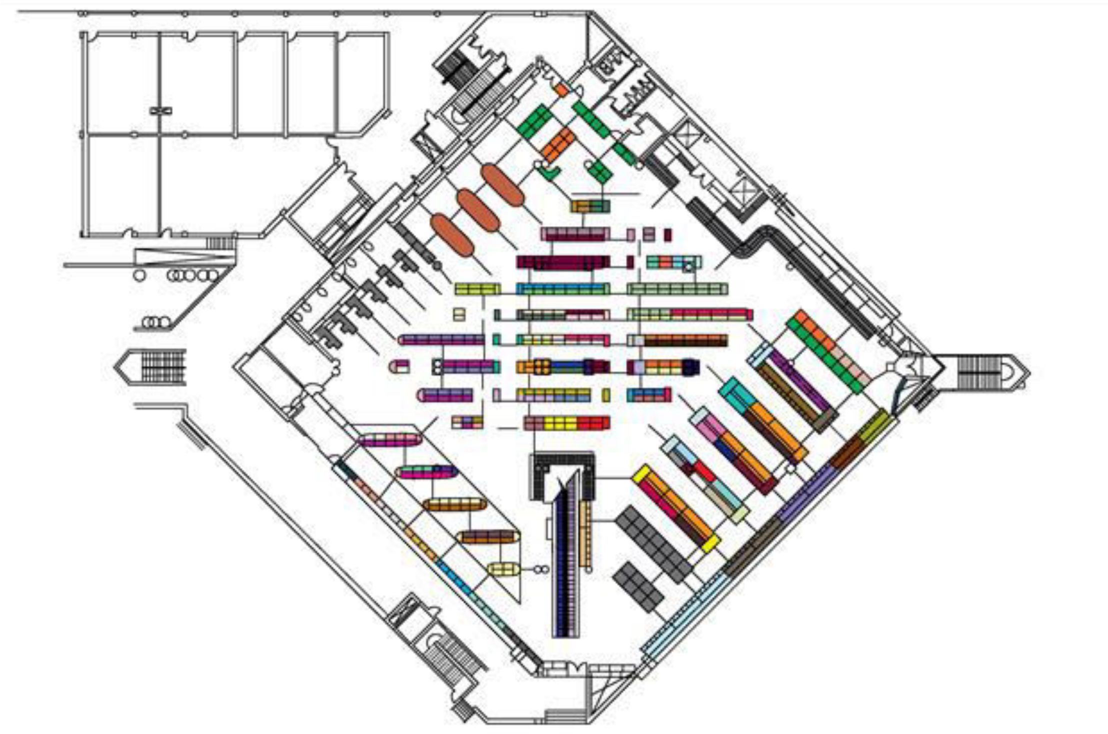

Figure 9: Current layout

Figure 10: Current categories

The following assumptions were made while conducting alternative layouts. The categories in the current layout are not organized, as they are scattered all over the supermarket. When reconstructing the layout each category has a specific place. For example, the chocolates were scattered, as shown in Figure 11 and are put in one place in the alternative layouts. Another assumption includes that the moving of categories are restricted to its set of potential positions. For example, the refrigerators and the freezers remain in their position, and the refrigerated categories are relocated within the refrigerators only.
The consumption universes layout approach was used to find the first alternative solution. This approach is customer oriented, where it clusters various categories based on customers purchasing behavior. In a consumption universes layout, instead of finding cornflakes in the cereal section, jam in the spreads section, milk in the dairy section, these products could be found in the breakfast universe. Other universes suggest the same structure to cluster different categories such as; self-care universe, baby universe, and beverage universe (Cil, 2012). Placing items that are repeatedly bought with each other in proximity can further encourage the sales of such products together.

Figure 11 illustrates the first proposed consumption universes layout solution, where categories were clustered into 20 different universes. The placement of products in each cluster was based on the associations of items within the cluster. Whereas associations between clusters were also taken into consideration. In Figure 11 each color represents a universe, and each universe has its own list of categories.

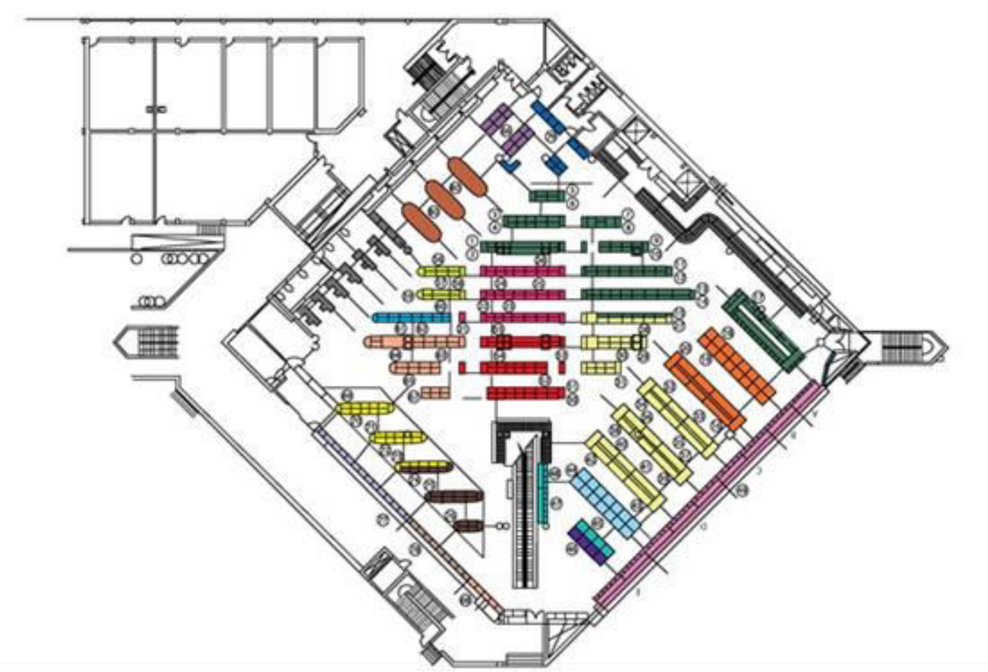

Figure 11: Universes approach layout

Figure 12: The Universes and their categories

Based on the support value, vegetables and fruits category is the most occurring category found in relation with the other categories. Therefore, the vegetables and fruit universe was remained close to the entrance, and close to the bakery universe since it is correlated with the pita bread and toast. Therefore, within the bakery universe, pita bread and toast were placed on the closest shelves to the vegetables and fruits universe. The oils and spices universe and the laundry universe are also placed close to the vegetables and fruits. Within the snacks universe the chocolates are placed between the chips and biscuits, since chocolate is highly associated to both. The snacks universe is

surrounded by the vegetables and fruits universe, since they have a high support value with chips and chocolate. Furthermore, the snack universe is also placed next to the bakery universe, specifically next the toast. As well as next to the beverages universe, precisely next to the juices and fizzy drinks. Where also, within the beverages universe the fizzy drinks and juices are located next to each other. Finally, in the breakfast universe, solid cheese was placed in the refrigerator closest to the vegetables and fruits, bakery, and beverages universes, since it is correlated with the toast in the bakery universe and the juices in the beverages universe.
The second alternative solution is based on the integration of the "cubes per order index for dedicated shelves" with the "counterclockwise customer orientation approach". This integrated approach was chosen to be named the level approach for short. The locations needed for each category are based on the number of shelves needed. The cube per order index of each category is calculated by dividing the number of needed location by its single support value. After that the hit rate was calculated by taking the reciprocal of the cube per order index. The category with the highest hit rate has the priority of being placed next to the door, so the higher the hit rate the closer it is to the door. The logic behind this is that the higher the support value of the category, and the less space it takes up, the closer to the door it should be placed. Categories that either have a low support value, or require a large number of shelves, are placed farthest from the door (Tompkins et al., 2010). The categories with no single support values obtained, due to the minimum support threshold, were placed after all the single support values of the categories were obtained, and were arranged in a logical order. Figure 13 shows the hit rate values of the categories that have a single support value. Nonetheless, the direction of the placement of the categories was based on the counterclockwise customer orientation approach. This approach states that since most of the population is right-handed, whenever a customer enters a retail store they unconsciously start their shopping trip from the right side of the supermarket.

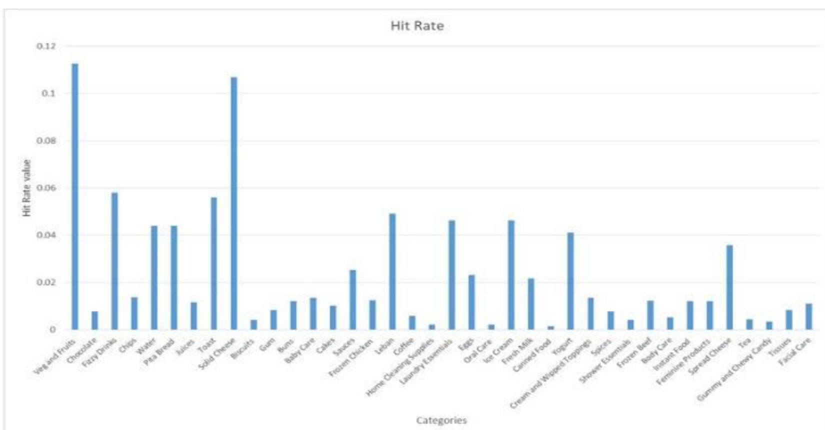

Figure 13: Hit rate

Figure 14 illustrates the second proposed alternative solution. The layout consists of four levels depending on the support value of each category. For example, once a customer enters the store the first level will be located on her right-hand side, which consists of the categories with the highest hit rate. The arrows in the figure show the order of the category placement based on their hit rate. The hit rate of the categories decrease as the customer descends from level to another. For instance, since toast has a high hit rate, it was placed on the closest shelves to the entrance. Even though the solid cheese has the second highest hit rate, it was placed in the closest refrigerator to the entrance. The hit rate value of the categories decreases as the customers walks from level to another.

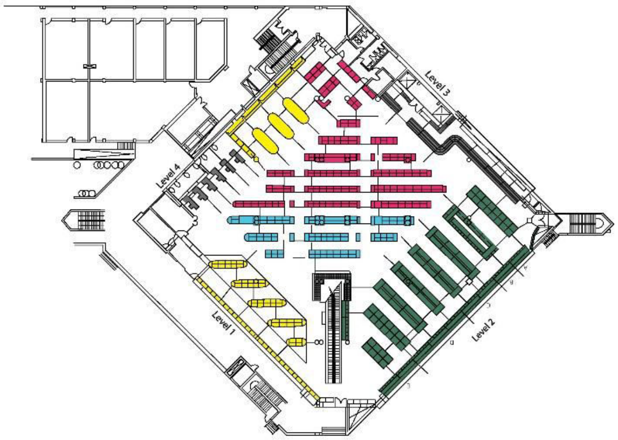

Figure 14: integration of the "cubes per order index for dedicated shelves" with the "counterclockwise customer orientation approach"

The walking distance was measured from the entrance to the midpoint of each category, and finally to the cashiers. Three random customers are chosen as a reference to calculate the distance travelled, to compare between the current layout and the two new alternative layouts. Customer A will purchase the highest six correlated items, which include toast, chips, chocolate, solid cheese, juices, and vegetables and fruits. In the current layout customer A tends to travel a distance of 153 m as shown in Figure 15, where the arrows show the customer A's shopping trip, while in the first alternative layout the distance is 120 m, as shown in Figure 16, with the arrows representing the walking distance of the customer. However, in the second alternative layout, she will travel a distance of 103.3 m, as shown in Figure 17, which show the arrows that indicates the purchasing trip of customer A.
On the other hand, Customer B will purchase the highest three correlated items, which include chips, chocolate and vegetables and fruits. The customer will travel a distance of 87.4 m in the current layout, 73 m in the first alternative layout, and 98.15 m in the second alternative layout.
Finally, as for customer C, she will be purchasing three random items, which include solid cheese, vegetables and fruits and toast. In the current layout the customer will travel a distance of 95.04 m, however in the first alternative layout she will be traveling a distance of 93.6 m, and 90.89 m in the second alternative layout.

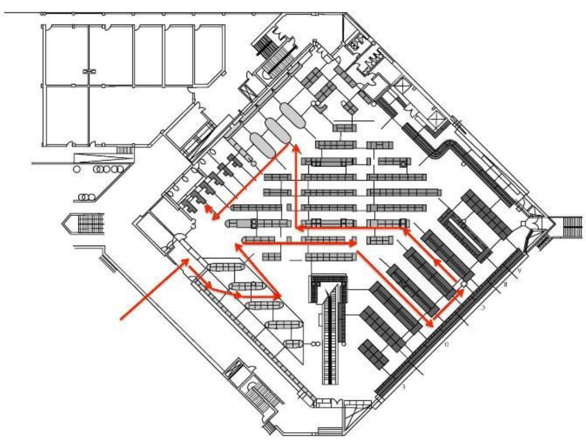

Figure 15: Customer A's path in the Universes layout

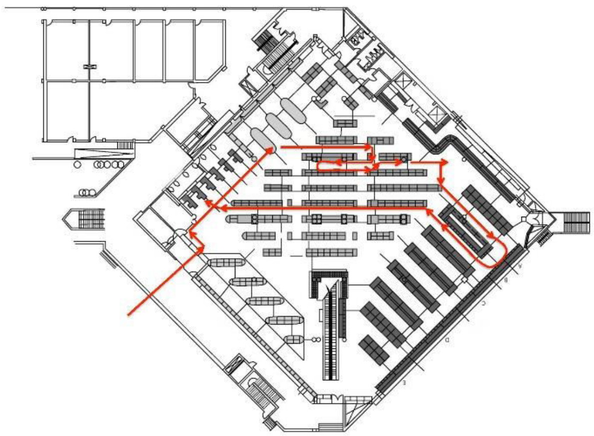

Figure 16: Customer A's path in the Universes layout

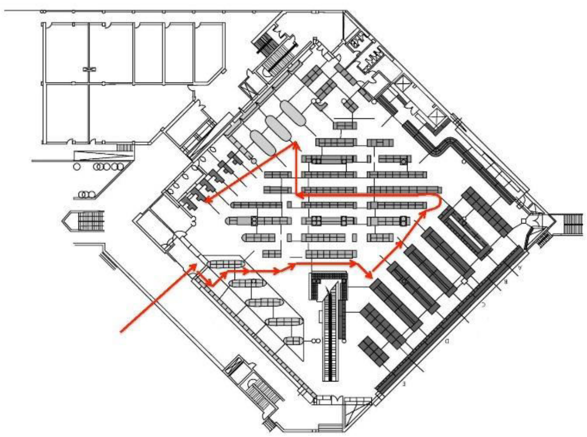

Figure 17: Customer A's path in the level's layout

The percent improvement of the walking distance was calculated to compare the current layout with both, the first and second alternative layouts for each customer, as shown in Figure 15. The improvement percentage was calculated for customer A to be 22% in the first alternative layout, and 32% in the second alternative layout. As for customer B, the walking distance improved by 16% in the first alternative layout, where it was calculated to be -12% in the second alternative layout, which indicates that there was increase in walking distance. Lastly, customer C's walking distance was found to be 2% better in the first alternative layout, and 4% in the second alternative layout.
To conclude, both alternative layouts, the consumption universes and the integration of the "cubes per order index for dedicated shelves" with the "counterclockwise customer orientation" approach quantify improvements based on the reduction in walking distance traveled by the proposed customers. The first alternative tends to improve the walking distance by an average of 13.3%, based on proposed customers. On the other hand, the second will improve the distance traveled by 8% on average.

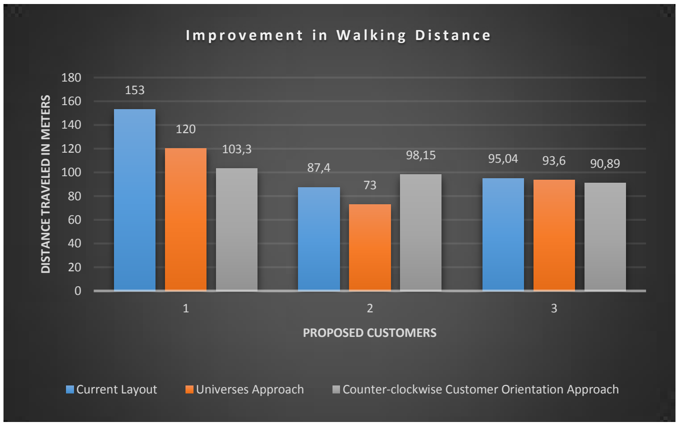

Figure 18: Improvement in walking distance

6. CONCLUSION

In this project, we addressed the technique of Market Basket Analysis, which is used in retailing to find relationships between products. The strength of the associations between products can be measured by the three measures: support, confidence and lift. Our literature review evaluates the several approaches that can be applied to perform Market Basket Analysis. The implementation of theses algorithms can help retailers in decision making. The steps in our methodology are to obtain point of sales data, clean data, categorize data, implement different algorithms, extract useful associations, and then make new layout recommendations to the supermarket.

7. RECOMMENDATIONS

We recommend to Al-Shuhada supermarket to change their current layout based on either of the alternative layouts, since the study was done on the customers in Al-Shuhada. In future references the project can be further analyzed were the support threshold can be decreased and more rules can be obtained. Additionally, more transactions can be inputted in order to increase the range of customers.

References

[1] Adomavicius, G., & Tuzhilin, A. (1999). User Profiling in Personalization Applications through Rule Discovery and Validation . 377-381.

[2] Agrawal, R., & Srikant, R. (1994). Fast Algorithms for Mining Association Rules. San Jose: IBM Almaden Reseach Center.

[3] Annie, L. C., & Kumar, A. (2012). Market Basket Analysis for a Supermarket based on Frequent Itemset Mining. International Journal of Computer Science Issues , 9 (5), 257-264.

[4] Aloysius, G., & Binu, D. (2013). An approach to products placement in supermarkets using PrefixSpan algorithm. Journal of King Saud University - Computer and Information Sciences, 25(1), 77-87. doi:10.1016/j.jksuci.2012.07.001

[5] Alteryx . (2017). Input and Output Tools. Retrieved October 22, 2017, from https://www.alteryx.com/input-and-output-tools

[6] Burton, S. H., Morris, R. G., Giraud-Carrier, C. G., West, J. H., & Thackeray, R. (2014). Mining useful association rules from questionnaire data. Intelligent Data Analysis , 18, 479-494.

[7] Chaturvedi, M. (2014, July 7). Merchandize Hierarchy and its relevance in planning. Retrieved April 15, 2017, from citixsys web site: https://www.citixsys.com/merchandize-hierarchy-and-its-relevance-in-planning/

[8] Cil, I. (2012). Consumption universes based supermarket layout through association rule mining and multidimensional scaling . Expert Systems with Applications , 39, 8611–8625.

[9] Deshpande, G., Shripada, R., Patole, S., & Bulsara, M. (2010). Update Meta-analysis of Probiotics for Preventing Necrotizing Enterocolitis in Preterm Neonates. Pediatrics , 125 (5).

[10] Dimitrijević, M., & Krunić, T. (2013). Association rules for improving website effectiveness: case analysis. Applied Knowledge Management , 1 (2), 56-63.

[11] Ertugrul, I., Oztas, T., Zeynep Oztas, G., & Ozcil, A. (2016). Shelf Layout With Integrating Data Mining And Multi-Dimensional Scaling. European Scientific Journal, 190-202.

[12] Geetharamni, R., Revathy, P., & Jacob, S. G. (2015). Prediction of users webpage access behaviour using association rule mining. Sadhana , 40 (8), 2353–2365.
[13] Grady, K. (2017). Mining Legal Data: Collecting and Analyzing 21st Century Gold. Internet Law , 20 (7), 9-21.
[14] Heaton, J. (2017, January). Comparing Dataset Characteristics that Favor the Apriori, Eclat or FP-Growth Frequent Itemset Mining Algorithms. Retrieved November 3, 2017, from https://arxiv.org/pdf/1701.09042.pdf
[15] Hilage, T. A., & KulKarni, R. V. (2011). Application of data mining techniques to a selected business organization with special reference to buying behavior. International Journal of Database Management Systems , 3 (4), 169-181.
[16] Journal of Information Science and Engineering , 32, 271-285.
[17] Zhi, D., Yan, H., & Yan, L. (2016). Analysis on a New Data Mining Algorithm of the Statistics Work. TELKOMNIKA , 14, 42-46.
[18] Kumar, A., & Annie, L. C. (2012). Web Log Mining using K-Apriori Algorithm. International Journal of Computer Applications , 41 (11), 16-20.
[19] Levy, M., Weitz, B., & Grewal, D. (2014). Retail Market Strategy. In Retailing Management (9th Edition ed., pp. 134-140). USA: McGraw-Hill.
[20] Lingonblad , M. (2015). Lost in the supermarket: Studying the effects of familiarity on consumer decision making. Retrieved December 14, 2017, from http://handelsradet.se/wp-content/uploads/2016/07/Lost-in-the-supermarket.pdf
[21] McNicholas, P., Murphy, T., & O'Regan, M. Standardising the Lift of an Association Rule. Department of Statistics. Dublin, Ireland: Trinity College.
[22] Morgan, R. (n.d). Six Levels of Merchandise Hierarchy. Retrieved April 15, 2017, from Chron web site: http://smallbusiness.chron.com/six-levels-merchandise-hierarchy-13949.html
[23] Papavasileiou, V., & Tsadiras, A. (2013). Evaluating time variations to identify valuable association rules in market basket analysis. Intelligent Decision Technologies. 7, pp. 81-90. Thessaloniki, Greece: IOS Press.
[24] Pawar, D. (2013, March 7). A Seminar on the Comparative Study of Apriori and FP-Growth Algorithm for Association Rule Mining.
[25] Prokeinová, R. (2014). Identification of the Patterns Behavior Consumptions by Using Chosen Tools of Data Mining - Association Rules. AGRIS online papers in Economics and Informatics , 6 (3), 3-10.
[26] Raeder, T., & Chawla, N. V. (2010, August 28). Market basket analysis with networks. Notre Dame, USA: Springer.
[27] Rahman, A., Noman, A., Mishu, H. M., & Sumi, F. (2016). Data Mining & It's Algorithms. American Journal of Engineering Research , 5 (11), 273-277.
[28] Rajak, A. (2008). Association rule mining- Applications in various areas. Krishna Institute of Engineering and Technology, Department of Computer Applications, Krishna, India.
[29] Roodpishi, M. V., & Nashtaei, R. A. (2015). Market basket analysis in insurance industry. Management Science Letters , 5, 393-400.
[30] Schindly, B. S. (2017). Advantages & Disadvantages of a Supermarket Layout. Retrieved December 20, 2017, from https://bizfluent.com/info-8651141-advantages-disadvantages-supermarket-layout.html
[31] Sears, J. (2004). Five Top Tips for Creating a Merchandising or Product Hierarchy . jlsears merchandise hierarchy . Seattle, USA: JL Sears Consulting, Inc .
[32] Singh, M. (2013). The Changing Collateral Space. International Monetary Fund.
[33] Surjandari, I., Rachman, A., & Lusia, M. (2012). Optimization of Products Shelf Space Allocation Based on Product Price Using Multilevel Association Rules. International Conference on Industrial Technology and Management. 49, pp. 62-65. Jakarta, Indonesia: IACSIT Press.
[34] Tanusondjaja, A., Nenycz-Thiel, M., & Kennedy, R. (2015). Understanding shopper transaction data. International Journal of Market Research , 58 (3), 401-419.
[35] Tompkins et al. (2010). Facilities Planning, 4th edition, John Wiley & Sons Inc.
[36] Venkatachari, P., & Chandrasekaran, I. D. (2016). Market Basket Analysis Using FG Growth and Apriori Algorithm: A Case Study of Mumbai Retail Store. BVIMSR's Journal of Management Research , 8 (1), 56-63.
[37] Yen, S.-J., Wang, C.-K., & Lee, Y.-S. (2016). Mining Consumption Association Rules.
[38] Zaki, M. J. (2000). Scalable algorithms for association mining. Transactions on Knowledge and Data Engineering, 12(3), 372–390.