Identifying customer buying behaviour in medium-scale supermarkets using data mining techniques

Sanduni, P.R.
Department of Information and Communication Technology
University of Ruhuna
sprarthana960@gmail.com

Jayaneththi, J.K.D.B.G.
Department of Information and Communication Technology
University of Ruhuna
buddika@ictec.ruh.ac.lk

ABSTRACT

Data mining with the assist of the most acceptable pattern returning and the frequent algorithm can majorly affect marketing and sales. Frequent pattern (FP) recognition is a considerably researched sector in data mining because of its significance in many day-to-day applications. Market basket analysis is one of its usages; applied by retailers/groceries to discover customers' buying behaviour from stores. The output of the analysis may increase the profitability of the retail owners, service quality, avoid empty shelves, and customer satisfaction. The research aims to focus on descriptive analysis of the customer buying patterns, items buying together, and item units that are highly bought from the supermarket and find the most efficient algorithm for this task. It facilitates reordering set the supermarket shelves layout. It was performed by identifying the current information to discover and analyse frequent itemsets to illustrate an association rule. Mainly, FP growth and apriori algorithms were used for market basket analysis. Both two algorithms are implemented using python by feeding pre-processed raw data. One algorithm was selected considering the efficiency. Since the FP growth algorithm consumes more execution time against the steps of the algorithm, the apriori algorithm is used to implement the recommendation model. The model which is implemented using the apriori algorithm, was developed to recommend the products using month-wise association rules.

Keywords: Apriori, Association Rules, Data mining, FP growth, Market basket analysis, Recommendation model

INTRODUCTION

Identification of customer buying behaviour is the most crucial factor in supermarkets and all kinds of organizations and institutions. Many kinds of research were conducted globally and in Sri Lanka to identify the customer buying behaviour using different factors, such as sample size and location. As mentioned above, most of the research is done through different factors, in different locations, and using different methodologies. The main objective of the research is to identify whether any frequent buying patterns occurred with the changing of time. The specific objectives related to the main objective can be mentioned: predict buying patterns according to different periods, Use different algorithms, com-

pare them after comparison, and select the most suitable algorithm. Using the algorithm implements the recommendation model. Predicting buying patterns helps to understand what products are sold together and what products are mainly sold. It helps to avoid out-of-stock. When avoiding the out-of-stock, it causes to reduce the other mentioned problems such as keeping customer loyalty towards the supermarket. Not only the out-of-stock problem but the overstock problem was also overcome with this pattern identification. When identifying the pattern helps to discover the buying pattern. According to that buying pattern, it can measure the quantities also. When reviewing the related research, customer buying behaviour is the investigation of what consumers buy or do not buy a product or service. The shopping details such as purchasing rate, what products are purchased, and socioeconomic information, such as gender, age, and salary, are the critical inputs used in data mining of customers buying behaviour.

Another research was conducted to check on determinants of organic food buying behaviour to check the effect of health domain concern, and reference group effect and realization on buying objective of organic food. Data were collected using a survey questionnaire that includes 400 persons from the Western Province of Sri Lanka. The multiple linear regression model was selected to evaluate the effect of each factor on purchase intention (Kapuge, 2016). When analyzing customer behaviour to investigate the necessary factors, which are displaying the success of a brand, this research identified the bridge between the dependent and independent factors. The baseline data has been collected using questionnaires, and for analysis reasons, SPSS statistical software has been used. The sample size of 150 customers has been gathered and tested for the accuracy of the model. As the result of research, found that packaging is one of the most critical factors. Here, only the products with the package have been considered (Raheem et al., 2014). Due to the spread of technology and information systems, although supermarkets are increasingly gathering tremendous amounts of data, they can still not release the power behind the data to its full potential. Since there are more data and more competitors, it is essential to identify marketing strategies with the patterns related to time. It also needs to know what factors affect customer buying behaviour, how the market should gain customers to the market or store, and how to keep the existing customers. Those are the most important things to consider when working with a huge number of data and many customers to get the most competitive advantage in this market industry. Not only

those details but also it needs to know how customers think, their intentions, feelings which affect the marketing, and their needs like those things. When understanding the factors, it is easy to tackle market strategies.

The main objective of the research is to identify the best algorithm to extract any frequent buying patterns that occurred with the changing of time and implement a recommendation model using those association rules as the specific objectives related to the main objective are, (i) to pre-process datasets as it can apply to the algorithms, (ii) to predict buying patterns according to different periods, (iii) to use different algorithms and compare them with the datasets, (iv) to select the most suitable algorithm, and (v) to create a model to recommend products.

This research aims to use the apriori algorithm and frequent pattern (FP) growth algorithm to identify any pattern of customer buying behaviour in Sri Lankan medium-scale supermarkets. According to the variety of time(monthly). After that, compare both algorithms and get the most efficient algorithm to implement a model. It is the next task of research. As well data gathering method is not the surveying method. Used raw data to complete the research.

METHODOLOGY

As a first task, we gathered data then followed the actions outlined the Figure 1.
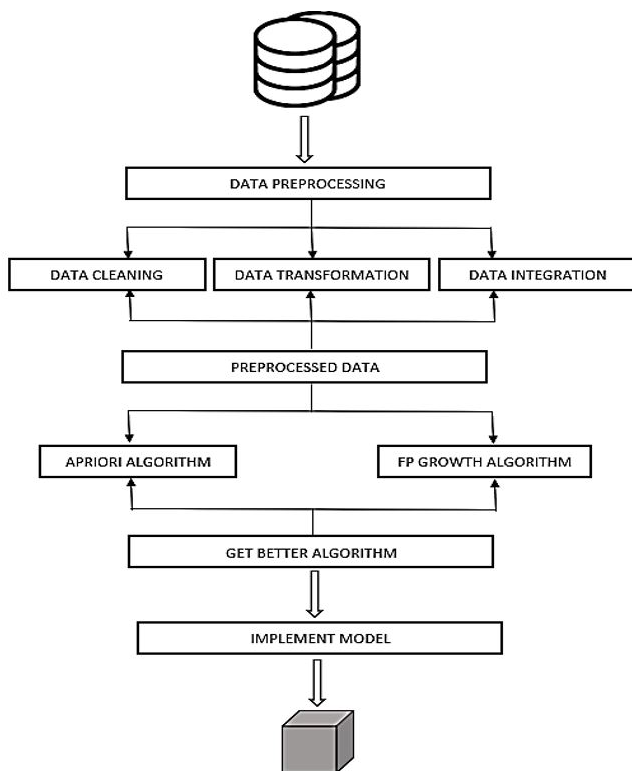
Figure 1: Flowchart of the proposed method

For this research, we needed data as transactions so that applicable to selected algorithms. Data collected from the MySQL database has been used, which contained all invoices data. It contains five years (2015-2021) data for twelve months (January-December).

Figure 2 shows the used tables from the database which is required. According to Figure 1, the next task is data pre-processing. Data is store in MySQL database, and it needs to create data as transactions month-wise. Before moving to the data like transactions needs to be pre-process, which is very important to get the best result. The requirement of data pre-processing occurs because every day/real-time data and many of the data of the database or files are often not completed and without consistency, which may produce improper and inaccurate results. Thus enhancing the quality of data on which the analysis is to be done. According to the collected data and the selected algorithms, it was challenging to apply raw data directly. Pre-process was done as Data Cleaning, Data Transformation and Data integration. Data cleaning was done by preparing the item name as identical, and It means some item names with inverted commas, some were not. So basically, they make non-inverted words when considering the item names. Collected data should be created as transactions so that they apply to the algorithms. To create data as transactions here used MySQL joins. Data integration was done by joining tables which are item table and invoice detail table. By using these joins, I created separate items as transactions. It means joining item names that contain the same invoice number. After that, it created an item list per invoice.it was a transaction. Like those, it creates transactions for twelve months of five years. Then those transactions can apply for the apriori and FP Growth algorithm After pre-processing and created data as transactions, it moves to algorithm implementation. Algorithms were implemented using the python programming language. Apriori and FP growth algorithms were implemented, and they were automated by feeding the pre-processed data. By executing both algorithms, it extracted the association rules according to month wise. Using these association rules, it can identify the customer buying behaviour change month-wise. After implementing both algorithms, it calculates the execution time for each step of the algorithms. According to the time comparison, it gets the better algorithm from the time measurement of each algorithm.

Implementing the recommendation model can be mentioned as the final output of the research. The model implementation here used association rules of the most efficient algorithm. It means an algorithm that has less low execution time was selected for the model creation. When getting the model creations, there were twelve models which used association rules of twelve months. Each month has a model to get recommendations. The model is also created using the python programming language. It can use as a backend for software that can use for the get recommendations. When giving a month with the specific item and to check what items will buy
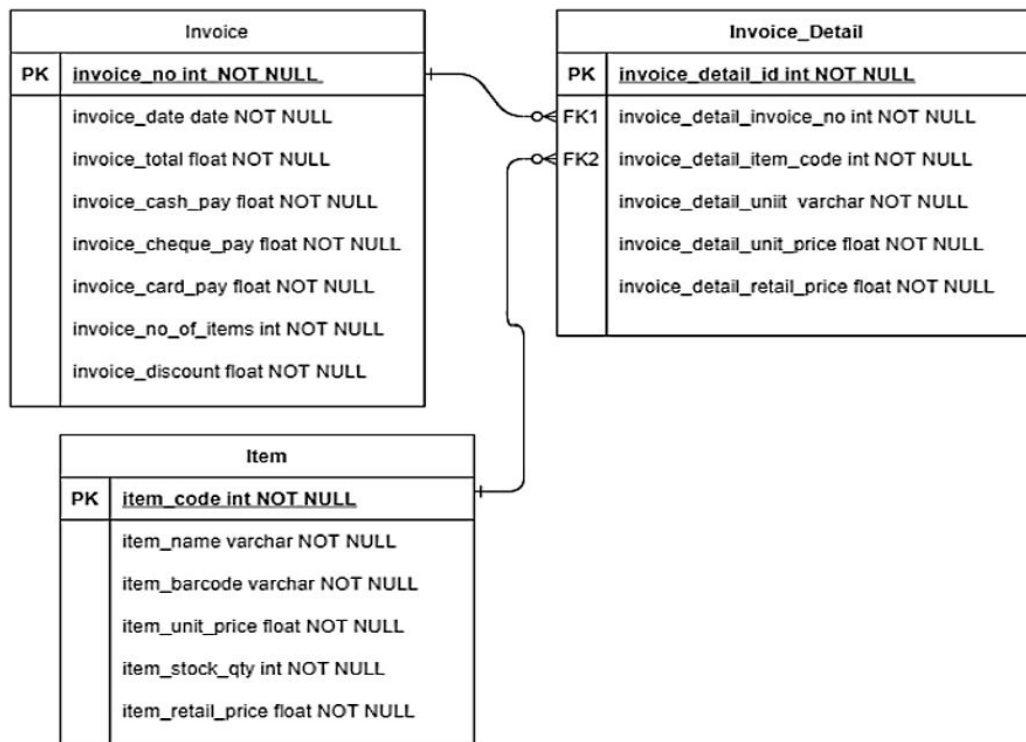
Figure 2: Database Structure

together in the relevant month can identify through this model. It was easy for supermarket owners to make store layout and the stock balancing.

RESULTS AND DISCUSSIONS

Socio-economic key variables

When moving to the research findings, in the pre-processing steps, it could be identified, and the raw data was not suitable for the selected apriori and FP growth algorithms. Because there are several noisy and incomplete data, it needs to format to apply to the algorithms. Furthermore, implementation of the apriori algorithm most suitable algorithm is apriori to model generation. Though there was candidate generation, the most suitable of these two algorithms are the apriori algorithm because of the execution time.

Table 1 shows the execution times for each step from milliseconds. When considering the time for each step, there were slight differences. After calculating the total time for each algorithm, apriori take minimum time to execute with the given data set. Figure 3 displays the more graphical results.

As shown in Figure 3, the apriori algorithm is more efficient than the FP growth algorithm. As the result of the research, It identifies that apriori takes minimum time rather than FP growth algorithm. Therefore this research has been used the apriori algorithm to implement

the recommendation model. Using the recommendation model can use to recommend items according to month-wise.
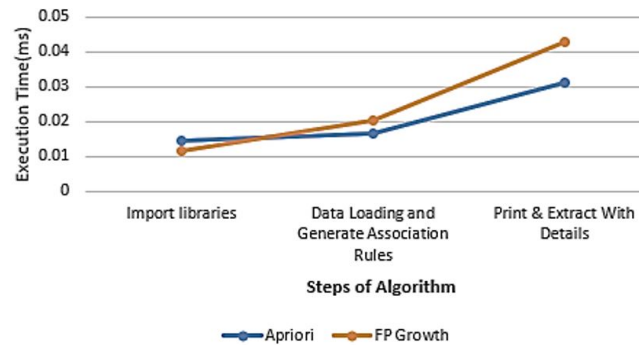

Figure 3: Execution time against steps

When moving to the association rule generation, the rules are different from the association rules of each month. For example, the association rule generated in march is different from the association rules generated in January. By that, it can identify that customer buying behaviour is changing according to the month according to the time. So it needs two to generate a twelve association rule list to twelve months. Model generation is also the same; there were twelve models generated for twelve months. When changing the month, recommended products are different; when a user needs to see recommendations, he/ she needs to open a transaction list with the relevant association rule file. To generate association rules for every month, there was

Table 1: Execution time against steps (apriori and FP growth)

Steps of algorithm	Apriori	FP growth
Import libraries	0.0145349	0.0114896
Data loading and generation of association rules	0.0163799	0.0202885
Print and extract details	0.0310876	0.0429028

an automated algorithm to pre-processed and generate association rules when giving the transactions. From this model, supermarket owners can identify what items will buy together in the relevant month. Not only one item, but the model can also recommend several items that users requested through the model. If the supermarket owner needs to know which two items will buy with another item, it can be done through this model. By that, this research can identify customer buying behaviour changes according to time, and their needs are changing from time to time; the apriori algorithm is the best algorithm from these two algorithms to model generation.

CONCLUSION

This research proves that the better algorithm of these two algorithms is the apriori algorithm to implement a recommendation model for the medium-scale supermarket. The association rules extracted from the apriori algorithm can help to reveal fascinating insights about the customers, which contribute to maximize the profit. Using data mining helps to identify customer behaviour using association rule mining; it is easy to identify customer buying behaviour. According to the supermarket data set most suitable algorithm is apriori to implement the model. It was proved by the comparison related to execution time. Created model paves a way to identify customer behaviour on various products quickly. Also it is easy to identify which products sold together in relevant months. The created model can recommend any count of products sold together with selected items; easy to create store layout and stock balancing.

REFERENCES

Bala, P.K. (2010) Mining changes in purchase behavior in retail sales with products as conditional part.*Proceedings of IEEE 2nd International Advance Computing Conference (IACC) 2010.*

Kapuge, K.D.L.R. (2016) Determinants of Organic Food Buying Behavior: Special Reference to Organic Food Purchase Intention of Sri Lankan Customers. *Procedia Food Science*, 6:303–308.

Raheem, A.R., Vishnu, P. and Ahmed, A.M. (2014) Impact of Product Packaging On Consumer's Buying Behavior. *European Journal Of Scientific Research*, 122(2):125-134