Association Mining Approach for Customer Behavior Analytics

D.M.R.M Dissanayake
Facluty of information Technology, University of Moratuwa, Sri Lanka
ranjan.md81@gmail.com

S. C. Premaratne
Facluty of information Technology, University of Moratuwa, Sri Lanka
samindap@uom.lk

***Abstract*—This research suggests a proper decision-making system using Data mining technique such as Frequent Pattern Growth (FP-Growth) Algorithm and K-means clustering, in order to identify consumer behavior on trending food items and conduct profitable marketing campaigns and promotions by comparing association rules of a particular date or day with the previous year. Once the system is developed, timely promotion creation can be done in a more consistent and straight forward way rather than promoting items in a senseless way by comparing the previous year's same season association behaviors.**

In order to create such promotions, the proposed system contains data which is undergone through data mining Algorithms. In developing the system, past transaction data is collected from the Point of sales system, and data preprocessing is done by data mining preprocessing techniques. One of the main activities in a food outlet is to determine associations, the inherent regularities in data such as products purchased together and what is the likelihood of buying a specific product after purchasing a certain product. *Niwa Sushi Pte Ltd,* Singapore as a Japanese food outlet has not yet been used such categorization and consideration in their sales system for processing associations, frequent itemset and subsequent items. The process is still not done even manually, but occasionally done by random observation & heuristic on sales data. The process conducted manually is also not the most accurate information but kind of near guesses.

***Keywords*-** Data mining, FP-Growth Algorithm, Frequent Item set Mining, K-means cluster analysis, Consumer behavior, Association Rules

I. INTRODUCTION

With the fast-moving lifestyle and competitive market in fast-food business, food outlets are more concerned about their promotion strategies. Due to significant rise of food industry in Singapore, fast food outlet business was selected as the core of this research. Various food items are expanded into locations in the country and offer various prices which will leads to huge competitiveness among the other sellers which results by winning the consumer attraction.

The success of an outlet is analysis on these consumer decisions resulting in enhanced services and conducting better marketing strategies. Also, the marketers can predict how consumers will respond to their strategies. There are many sources of customer attraction. For example, attributes such as Staff representation, tastiness of food, culinary art are few among them. Many of them are immeasurable. Therefore, we predict that the sources to identify such attraction are the transactional sales data occurring in the Point-Of-Sales system. The goal of this Decision Support System is to make use of these transactional data to identify decision patterns using Data Mining techniques to grow sales using timely promotions on particular trending products.

Many analyzing has been conducted on consumer purchase transaction data using various data mining techniques, methods and algorithms. Techniques such as FP-Growth Algorithm and K-Mean Clustering are common due to their accuracy. Using such approaches, exploration and examination of utilized marketing strategies and promotions is straight forward rather than heuristic promotion planning.

This algorithm uses bottom-up approach by generating candidate and group of candidates and are tested against the data. The advantages of using the FP-Growth algorithm are it can be applied in a large amount of item set properties, easily parallelized and easier to implement with the assumptions on the database is resided in memory.

A. Background and Motivation

The food industry in Singapore is a relatively stable industry with steady development potential. Be that as it may, unfortunately, the examination has not yet been so advanced in this industry, for example, retail, e-commerce, banks, etc. Other industries, there are customers who purchase the items and sellers who sell them, and each whenever an exchange happens, it is recorded to have computerized evidence. Previously, the utility of this data goes similarly to next month's sales forecast utilizing chronicled averages.

Big data can revolutionize the customer experience when they enter a food outlet. So as to achieve excellent marketing efforts, data from a food outlet in Singapore have been collected. This food court has 30 branches crosswise over the nation, with different cousins. A customer can have multiple exchanges per visit, so the average number of exchanges recorded per day is a critical number. The food court keeps track of its customer's purchases by means of a card or money.

This food court collects a large measure of data about the client and his purchases, however, Unfortunately, they don't use this data to increase sales or pull in new customers.

The data is wealthy in data because we can follow every move of a customer in the food court and use it to predict future events. This data can be shared with restaurants to help better sell to customers, or in the future, even begin profiling the customer in segments or gatherings for marketing purposes. This thesis uses this data to make recommendations to the food court management and marketing department and to lead advancements based on the results of the investigation as a dashboard that is easily understood by them.

B. Proposed Solution

As better decision making of consumer trends on proper analysis of Fast food chains, within this research, we proposed a system using results of data mining techniques. Researchers have identified data mining as the best solution for digging useful hidden patterns within large repositories of data using the support of different software tools with its ability to deal with a large number of dynamic variables simultaneously.

As the initial step, find out the factors that contribute to determine the marketing strategies.

After that define the sub research questions which provide the higher impact of Food marketing, then the data set obtained by Umisushi Pte Ltd, Singapore is preprocessed and prepared for further analysis. According to selected sub-research questions find out main attributes and items that are more likely to purchase by consumers and hence initiate marketing strategies like timely base promotions.

As data mining is the process of discovering correlations, patterns, trends or relationships by searching through a large amount of data stored in repositories, corporate databases, and data warehouses, it is selected as the basic methodology for this research. Data mining process had to follow basic life cycle subprocesses which mentioned in Figure 1 to build appropriate models and to generate predictions.

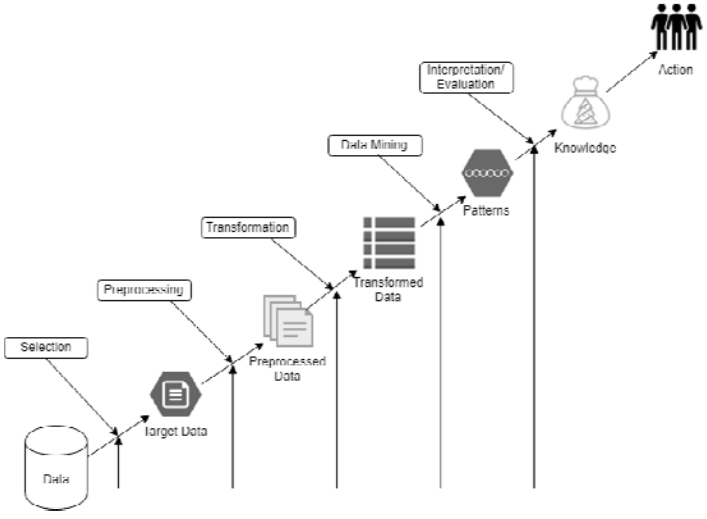


Fig. 1. Steps in Data mining process

The main requirement of this analysis is to identify the mining methods which cope with growing neediness in fast food industry marketing. This analysis can be used to inform Marketing people about the best methodologies to find general and specific consumer trends, patterns and series in an ongoing, timely manner in order to take the advantage of the information existing in Point of sales systems in outlets to maximize the sales of food items, to have an objective means to investigate lowest selling food items branch or country wise, detecting and preventing of food waste on specific days in a week and understanding the consumer behavior.

C. Thesis Structure

This thesis is sorted out into eight Chapters. Chapter I presents the problem and the inspiration for this thesis. Chapter II is a literature review where a bit of the paper that exchange about examination and data mining are reviewed and discussed. Chapter III is tied in with dealing with the data collection and how it is used for investigation so as to info and source to the Decision support system. Chapter IV is on the investigation and results, here every method is clarified in detail, and the outcomes are presented too. Chapter V presents results, where every one of the data is abridged, and recommendation for the marketing department is introduced. The last Chapter VI manages future work.

II. LITERATURE REVIEW

A. Consumer Behavior in purchased products

Mining frequent patterns in transaction databases, time series databases, and numerous different sorts of databases have been contemplated in data mining research. Most of the past studies were focused on Apriori-like candidate key generated approach. Nevertheless, candidate key generation is considered as costly when there are complex and long patterns.

Jiawei Han, Jian Pei and Yiwen Yin proposed[1] a novel pattern called Frequent Pattern Growth which consists of Frequent Pattern Tree(FP-Tree) structure, for mining the complete set of frequent patterns. This is an alternative to Apriori-like algorithms in order to avoid memory and time consumptions for computation by not generating candidate keys and limiting to 2 process steps. That is to generate frequent items and then the FP-Tree. This research shows the performance of FP-Growth is efficient and scalable for big or small frequent patterns than Apriori Algorithm and much faster.

The efficiency is retrieved using 3 techniques. First, the massive database is compressed into compact and smaller data structure resulting in avoiding repeated scans. Secondly, FP_Tree based mining avoids generating large numbers of candidate keys. Finally, used a divide and conquer method based on partitioning break down the extraction task into a smaller set of tasks for extraction of confined models into conditional databases, which greatly reduces the search space.

The performance of study shows that the FP growth method is efficient and expandable for the exploitation of long and short frequent patterns and is about an order of magnitude faster than the Apriori algorithm[1] and also faster than some recently reported new and frequent mining methods.

There are numerous of interesting research challenges related to FP-Tree based mining, including implementation of SQL-based FP structure, constraint-based FP-Tree structure, constraint-based mining frequent pattern using FP-Trees and extension for mining sequential patterns, max-patterns, partial periodicity and other interesting frequent patterns.

[1] Frequent patterns are usually defined as subsequences that can be seen in a data set regularly. Searching such patterns is crucial in machine learning. Further, it is useful for data classification, clustering and other data mining approaches as well. Frequent itemset mining is the core of Market Basket Analysis which helps to predict customer behavior in certain ways. Market Basket Analysis processes the consumer buying habits by identifying associations among the different items brought by the customer.

Data Mining: Concepts and Techniques[2] give ideas and methods to manage accumulated data or data that will be used in different applications. In particular, it clarifies the extraction of data and the devices used to find information from the collected data. This book is referred to as the knowledge discovery from data (KDD). It focuses on the practicality, the ease of use, the suitability and the adaptability of the expansive data set procedures. As a result of representing data mining, this version clarifies the strategies to know, preprocess, manage and store data. At that time, it presents data on data distribution centers, online analytical processing (OLAP) and innovation of data blocks. At that time, the techniques related to the mining of examples, affiliations and incessant connections for extensive data sets are described. The book details the strategies for the order of the data and presents the ideas and techniques for grouping the data. The rest of the sections explain the identification of exceptions and the patterns, applications and investigations of the periphery in data mining.

Key features of this book are as follows. Presents many calculations and execution models, all in pseudo-code and reasonable for use in real-world, substantial scale data mining projects. Addresses propelled points, for example, mining multimedia databases, spatial databases, object-relational databases, time-series databases, content databases, the World Wide Web, and applications in a few fields. Gives a far-reaching, viable take a gander at the ideas and procedures you have to capitalize on the data. Rakesh Agrawal and Srikant introduced[3] the Apriori Algorithm in 1993 which is widely used in data mining and designed on a database containing transactions captured from the point-of-sale system in order to find associations among items. Purchased items indirectly represent consumer behavior and considering this as the base, identifying regularities among purchased items is a source to conduct marketing strategies. This algorithm is an efficient method of generating

all significant association rules between items in the database and supports buffer management and novel estimation and pruning methodologies.

As of not long ago, however, only global data about the sales transactions during a certain period was saved on the computer. Advancement in barcode innovation has made it conceivable to store the so-called basket data [3] that stores items purchased on a per-exchange premise. Basket data type transactions do not necessarily consist of items bought together at the same point of time. It may consist of items bought by a customer over a period of time. Precedents incorporate monthly purchases by individuals from a book club or a music club. A few associations have gathered a vast amount of such data. These data sets are generally put away on an external database and are gradually relocating to the database system. One of the fundamental purposes behind the constrained accomplishment of database systems around there is that present database systems don't give vital functionality for a client keen on exploiting of this data.

This paper presents the issue of "mining" a substantial gathering of basket data type transactions for affiliation rules between sets of things with some base indicated confidence and presents a proficient algorithm for this reason. A case of such an association decide is the explanation that 90% of sales data that buy flour and butter likewise buy egg. The predecessor of this rule comprises of flour and butter and the subsequent comprises of egg alone. 90% considered as the confidence measurement of the rule.

Apriori uses bottom up approach and hash tree structure to count candidate item sets efficiently. Without such algorithm, searching associations of items might be a time exhaustive task. Generally, not understanding the patterns discovered with data mining are unlikely to act on them.[4] Suresh[5] has been conducted by the help of association rules and market basket analysis and support and confidence metrics, to identify interrelationship between products or products of a particular category. This study is based on daily transactions in a Point of sales system of a retail shop. Analysis illustrates the most demanding product sets which emphasize the customer buying behavior which leads towards identifying profitable products and product categories and generating sales strategies to increase the sales hence more profit. Further, cluster analysis is used to make the relationship among products and to distinguish customer buying behavior factors.

This is an examination on use of retail analytics with respect to a particular retail outlet. The extent of this exploration is to distinguish the items or results of the specific category which has an association that is items which are probably going to purchase with different items as a collection of item sets. This is characterized by association rules with the support and confidence measurements, in this manner finding the frequent market basket package item sets. This examination was finished utilizing the point of sales data that is gathered in a retail shop of daily transactions.

Utilizing the association rules and market basket analysis, beneficial related item sets are discovered. This distinguishes consumer pattern of purchasing behavior and it is then used to devise strategically selling procedures to improve product sales. Cluster analysis is used to characterize the relationship among items and distinguish the factors that impact the purchasing conduct of the consumers. Also, the outcomes determined is utilized to draw a visual promoting method that will fulfill the consumers specifically in service and increase the willingness of product purchase, expanding the profit of the retail outlet and holding its consumers.

B. Unsupervised Exploration

Clustering is commonly used for grouping similar items around centroids. This helps to get an understanding of data[4] and how they are spread. k-means clustering is a widely used classification in machine learning[6] which was initially used by James MacQueen in 1967 originally from signal processing. The aim of the method is to partition n observations into k clusters in which each item belongs to a particular cluster with the nearest mean, working as a template of the cluster.

K means clustering is one of the least complex unsupervised learning algorithms that take care of the clustering problem. The system pursues a basic and simple approach to characterize a given data set through a specific number of clustering (expect k clusters) settled a priori. The fundamental idea is to define k centroids, 1 for each cluster. These centroids should be set in a specific way due to various locations causes the distinctive outcome. Segmentation of people into clusters upon their purchased items allows accurate recommendations[7] of new items for purchase. Similar interested items recommendation is useful in many domains but does not always work since data are always sparse. Therefore, accurate assumptions can be made by segmentation into clusters which tend to be interested in the same set of people.

Online and offline shopping centers purposely track who purchased what. These data can be used for future prediction of what shoppers may want to buy. These predictions are open to the large domain.[7]This the research presents a statistical model of collaborative filtering and a comparison of various algorithms to estimate the parameters in K-means and Gibbs sampling. On- and off-line shopping merchandisers consistently keep transaction data of buyer buy and their items. These data can be utilized to anticipate what future customers should need to purchase. Such predictions are not constrained to purchases: One can utilize records of what

movies, CDs or online documents individuals have enjoyed in the past to foresee which ones they will enjoy later on.

As a summary, this research shows that collaborative filtering is very much portrayed by a probabilistic model in which people and the products they view or purchase are each partitioned into (unknown) clusters and there are relationship probabilities between these clusters. Also shows that Expectation–Maximization (EM) is a conspicuous technique for assessing these models, yet does not work since it can't be effectively built to perceive the requirement that a film preferred by two unique individuals must be in a similar movie class each time. K-means clustering is quick yet ad hoc. Repeating clustering utilizing K means clustering or a "soft clustering" form of K-means might be valuable; however, for the most part does not enhance precision. Clustering films or people on other pertinent attributes can help - and helps for the instance of CD buy data.

Further illustrates, Gibbs sampling functions admirably and has the righteousness of being effortlessly reached out to substantially more intricate models, yet is computationally costly. Researches state that the current development of proficient Gibbs sampling methods for collaborative filtering problems, expanding repeated clustering and Gibbs sampling code to fuse numerous attributes, and applying them to all the more genuine data sets.

C. Enhancing Association algorithm

Su,Xu, Cheng,Li & Yang[8] tries to identify the possibility of developing a differentially private Frequent Item Dataset algorithm which achieves not only high data utility but also high level of privacy and time efficiency. The proposed algorithm is based on the FP-growth algorithm and consists of preprocessing phase and a mining phase. A novel smart splitting method is proposed to convert the database in preprocessing phase in order to improve the utility and privacy where the method is used once for a given database. To estimate the loss information due to transaction splitting, a run-time estimation method is invented the real dedication of the item set in original database.

By using the downward closure property, this research introduces a method to reduce dynamically amount of variations added to ensure privacy while the mining process continues. Through formal privacy analysis, this research shows that proposed PFP-growth algorithm is E-differentially private. Experiments on actual data show that proposed PFP-growth algorithm significantly performs the state-of-the-art techniques.

Kabir[9] proposes a data mining framework in order to make sales and marketing decisions. The proposed framework is based on association rules generated from transactional sales data where the data is captured from a raw database in which the scanned data is stored. The scanned data is captured from point of sales system which is located in different terminal locations. This research explains the usage of association rules in order to make improved decisions in marketing and sales. However, in this research, only a single relation of sales data is taken to consideration. The decision making could have been more intelligent, if the research is based on multiple relations of sales which will leads to customer satisfaction as an added advantage.

Yagi[10]Presented efficient algorithms for computing optimized sequential pattern for a highlighted optimization problem of sequential pattern mining which can calculate optimized sequential patterns for sales deviation events in point of sales transaction data. This research presents another algorithm for mining successive examples. Our algorithm is particularly proficient when the successive examples in the database are long. We present a novel profundity first pursuit technique that coordinates a profundity first traversal of the hunt space with successful pruning components. Our usage of the hunt system joins a vertical bitmap portrayal of the database with proficient bolster tallying. A striking component of our algorithm is that it steadily yields new frequent item sets in an on the web design. In an exhaustive test assessment of our algorithm on standard benchmark data from the writing, our algorithm beats past work up to a request of greatness. Finding consecutive examples is an imperative issue in data mining with a large group of use areas including drug, broadcast communications, and the World Wide Web. Traditional mining frameworks give clients just an exceptionally confined component (in view of least help) for determining examples of intrigue. In this paper, we propose the utilization of Regular Expressions (REs) as an adaptable requirement particular device that empowers client-controlled center to be fused into the example mining process. In this research, build up a group of novel algorithms (named SPIRIT– Sequential Pattern mining with Regular Expression Constraints) for mining frequent successive examples that too fulfill client indicated RE requirements. The fundamental distinctive factor among the proposed plans is how much the RE requirements are upheld to prune the seek space of examples amid calculation. This research answers give significant bits of knowledge into the tradeoffs that emerge at the point when imperatives that don't buy in to decent properties (like anti-monotonicity) are incorporated into the mining procedure. A quantitative investigation of these tradeoffs is led through a broad trial think about on engineered and genuine data sets.

With these recently created sequential mining algorithms, for example, Prefix Span, it is conceivable to mine sequential client get to patterns from Web-logs. While this data is exceptionally valuable while updating sites for simpler scrutiny and less organize traffic bottlenecks, it would be so a lot more extravagant on the off chance that we could join different measurements of data. For instance, knowing the referral site that clients frequently originate from, may have the capacity to figure out what data all alone site is important to them - and upgrade or separate this data as required.

Essentially, knowing what weekday and time certain entrance designs frequently happen at, could guarantee refreshed data is prepared and accessible for these clients. This research proposes and investigates two distinct systems, HYBRID and PSFP, to consolidate extra elements of data into the way toward mining sequential patterns. It explores the qualities and impediments of each methodology. The HYBRID technique first finds frequent measurement combination, and after that mined sequential patterns from the arrangement of groupings that fulfill every one of these mixes. PSFP approaches the issue from the other way. It mines the sequential examples for the entire dataset just once (utilizing Prefix Span), and mines the comparing frequent dimensions patterns close by each sequential pattern (utilizing existing association algorithm FP-Growth). Analyses demonstrate that HYBRID is best at low help in datasets that are inadequate concerning measurement value groups yet thick with regard to the sequential patterns. PSFP is the better option in each other case; including datasets that are thick concerning both measurements value groups and sequential products at low help.

[11]The main aim of this paper is the combination of the data analysis and to parameterize the simulation of sales receipt data which will be closer to reality.

Mirajkar[12] propose a methodology with the base of developing a reliable algorithm that output appropriate frequent pattern algorithms on many available dataset which will help the marketing and sales persons in order to invent mind catching cross-sells and related goods. There is concealed data secured up in the tremendous of information of the organizations' databases. The unidentified data is conceivably imperative for the organizations' prosperity.

The relations among the results of the retailers can be productively removed from extensive retail databases which impact the deals and gainfulness. The activities administrators can change both with by and large plans of the store and the assignment of the space to different items with this covered up data assembled by the information mining investigation. In this paper affiliation rules are connected to the informational index and guideline sets are accumulated to accomplish a superior format and retire course of action for the retailer. Design has key significance to a firm. Position of the power things which have a high introduction rate, augment gainfulness per square foot of the floor space.

They have developed the algorithms based on association mining and the combination method is tuned up for more interesting outcomes. Further shows that how basket analysis is important to identify the item arrangement in racks, designing and executing sales promotions in the appropriate time to improve customer satisfaction and eventually to increase the profit.

Ismail[13] tried to answer and avoid limitations in typical high-utility methods when using to identify profitable items in super markets, by giving new method to identify the Productive high-utility periodic patterns using consumer specific data which will informally shows high profit interrelated item groups. To avoid limitations, they have set a new pattern-growth algorithm including a tree structure.

D. Forecast Model

Ishigaki[14] illustrates a method of computational customer behavior modeling on top of real transactional datasets and present some of the results generated from the model. The model is developed using Bayesian network applied on a massive point of sales dataset which consists of customer details and questionnaire responses. Further, they have implemented an automatic categorization with the help of probabilistic latent semantic indexing (PLSI) due to the need of categorization for the development of a realistic model.

E. Novel Methods

Aravindan[15] propose an algorithm call Transaction id frequent sequence pattern and its techniques are used for mining item sets in the diverse multiple databases like E-Commerce website data sources, to overcome problems related to limitations of multiple database Sequential pattern mining like ApproxMap algorithm where this algorithm is not capable of mining frequent sequences to distinct and past queries from multiple databases of variant structure.

[16]This paper propose a novel procedure on top of item co-occurrences analysis in order to reduce the number of join operations in HUI-Miner in which having a vertical representation and conduct a dept-first search to identify patterns and investigate their utility without conducting high cost database searching. The final output of the research experiment is the new algorithm called Fast High-Utility Miner which reduces significant number of joins up to 95% and gives up to six times faster than the typical HUI-Miner.

[17]This paper presents a personalized product tree called purchase tree, based on leaf nodes (products to sell) and internal nodes (multiple product categories) to illustrate consumer transaction data by compressing these data in to set of purchase trees. Also suggest a partition clustering algorithm called PurTreeClust in order to fasten clustering of tree data. The paper also suggests a gap statistic-based method to form an idea of number of clusters. By conducting series of investigations on live datasets, proposed method shows significant performance.

F. Classification Comparison

[18]In this paper, different kind of data mining classification techniques are tested to verify any similarities or dissimilarities and to choose the best classifier which is the most suitable for investigating consumer online buying attitude and behavior on large scale e-commerce shopping data sets. According to the study, among the classifiers, decision table classifier and filtered classifier illustrated the highest accuracy and the clustering and simple cart was showing the lowest accuracy. Further this paper provides a system based on decision table classifier to help the customer on searching the products in some online shopping sites. This recommender system captures information from customers and products to provide suitable user wise suggestions to customers on buying appropriate products.

[21]This paper is upon clustering technology of web mining to output a user wise solution to develop an ecommerce recommendation system. Also, the paper presents the UserID-URL associated matrix related to log data. Users are clustered by calculating UserID-URL associated matrix and Distance matrix. The system can suggest goods to the users in a cluster browsed by another user in the same cluster and gain the target of customized product recommendation.

[22]This paper focuses on the interrelationship between purchased products in one transaction done by consumers in a retail shop and supermarket where the final objective is the coherence between inventory model and purchase dependencies and eventually to measure the impression of purchase dependencies on items availability. To illustrate the purchase dependencies, they have developed an inventory model where input will be purchase dependencies. Further, experiments done to identify the differences are applied to illustrate the interrelationship of purchase dependencies of inventory model. Results shows that by combining purchase dependency elements into inventory domain can reduce total cost in inventory towards reduce lost profit.

[23]This paper investigates the Long Tail theory where how it is utilized to become more extensive in enterprises product ranges. Also, this paper discusses how to invent new items or services to facilitate mass potential customers. Finally, a new business model is introduced where the enterprises are based on two distinct interests. First is, using big data to investigate new items, customers, accurate marketing, varied management and examination of customer loyalty. Secondly, using big data by itself as a new product and the transaction data as origination of enterprise profit.

G. Data mining in other domains

Raju,Bai & Chaitanya[24] provides a review and criticizes the fundamentals of Data mining and CRM in sectors like Banking and Retail Industries. Also, it illustrates and discusses the standard tasks used in data mining. Further, analyze various data mining applications in separate sectors.

The goal of this Decision Support System is to make use of these transactional data to identify decision patterns using Data Mining techniques to grow sales using timely promotions on particular trending products.

H. Decision Support System

Rok Rupnik and Matjaz Kukar have interpreted in Data Mining Based Decision assist gadget to support Association Rules which they have expanded DMDSS on Oracle platform. To aid the business location, many selection assist systems are based on OLAP. This allows customers to without problems and selectively choose and consider data from distinctive factors and permits studying database details from distinctive database systems straight away. OLAP, in particular, describes the databases which are especially designed to facilitate decision making analysis.

To perform a quick evaluation, OLAP makes use of cubs. The association of cubes allows overcoming the dilemma of relational databases. Cubes encompass numeric information called measures that are categorized by means of dimensions.

OLAP mining models are suitable whilst overall performance and drill down capability is taken to consideration. Cubes are pre-aggregated and it allows offering rapid responses to queries which handle tens of millions of data. Many reporting services will allow drill up and down whilst the report source is an OLAP dice.

In DMDSS the use of statistics mining strategies on different problems the usage of all of three records mining strategies referred to. Decision-assist systems (DSS) are described as interactive computer-primarily based systems supposed to help decision makers make use of facts and models in order to pick out issues, remedy problems and make choices. They include each statistics and models and they are designed to help decision-makers in semi-based and unstructured selection making tactics. They provide help for choice making; they do no longer update it. The mission of selection aid structures is to improve effectiveness, in place of the efficiency of selections.

I. An Overview of Classification Rule and Association Rule Mining

B. Srinivas, Gadde Ramesh and Shoban Babu Sriramoju have represented how information mining may be carried out in Market Basket evaluation to discover new tendencies and buying styles of customers. In this point in time big measure of data is generated each day and this fact is maintained in database in distinctive fields, as an instance, healthcare, schooling, market basket analysis, and so on. With this increasing statistics size, there may be a want to recognize large and complex information and attain important determinations. The process of extracting important facts from massive pre-current databases is known as Data Mining. It could be very troublesome for the neighborhood outlets to tug in clients, so it is their need to apprehend the shopping trends of the clients. Numerous purchasers choose on line buying. With the improvement of the e-trade websites, retailers have a tendency to overlook to drag in more and more clients. This problem can be eliminated by using making use of records mining techniques to investigate new patterns and developments. The information mining techniques are implemented to the accumulated facts associated to client conduct sample, with the goal that outlets will be able to know the brand-new styles and traits.

III. A NOVEL APPROACH FOR MARKET BASKET ANALYSIS

A. Introduction

Chapter III discussed the innovation for analyzing market basket to distinguish the relationship of bought items. This section exhibits our way to deal with examine association rules in detail utilizing data mining under a few headings, in particular, speculation, input, yield, process, clients and highlights. This section features the key highlights that recognize our novel methodology from the current methodologies for market crate investigation.

B. Input

As the underlying contribution for this procedure, data obtained from Food outlet Point of Sales system of Niwa Sushi Pvt Ltd, Singapore.

C. Output

As the output of this procedure, distinctive data designs identified with the bought items can be uncovered by the sub research question identified. The forecast will be given as output for the research question attached to predictive tasks. The outline will be given for the research question connects with descriptive tasks.

D. Process

In this procedure of breaking down purchased items with data mining all the standard steps in the knowledge discovery process which incorporate data selection to evaluate are carried out. All through the process the data set is preprocessed and prepared for mining and interpretation.

E. Data Selection

Point of sales data is the data gathered where cash transaction and charging happen in a food outlet which is a secondary data. In this research point of sales data was utilized which has gathered from food outlet. The data structure is comprising of Receipt ID, Item Name, no of units purchased, cost and payment type which gives information about the item.

Business utilization of market basket analysis has fundamentally expanded since the presentation of electronic point of sale. Supermarkets utilize association analysis for strategically pitching when it recommends items to buyers' dependent on their purchase history and the purchase history of other individuals who purchased a similar thing.

Market basket analysis may tell a shop that clients frequently purchase item A and Item B together, so putting the two things on advancement in the meantime would not make a huge increment in income, while an advancement including only one of the items would almost certainly drive sales of the other.

Market basket analysis may give the retailer data to comprehend the purchase behavior of a purchaser. This data will empower the retailer to comprehend the purchaser's needs and revamp the store's appearance appropriately, create cross-promotional events, or even catch new purchasers.

Market basket analysis can be utilized to partition clients into groups. An organization could take a gander at what different things individuals as an example purchase alongside eggs, and arrange them as baking a cake (in the event that they are purchasing eggs alongside flour and sugar) or making omelets (on the off chance that they are purchasing eggs alongside bacon and cheddar). This recognizable proof could then be utilized to drive different programs. Correspondingly, it very well may be utilized to partition items into regular groups. An organization could take a gander at what items are most as often as possible sold together and adjust their classification the board around these factions.

Transaction data mostly captured precisely in order to close sale end of the day. Sold items are saved and database backups are taken regularly to continue the account processes consistently. Any discrepancy on figures might cause accountability errors. Taking these concerns in to consideration, point of sales systems are developed carefully to maintain the accuracy from the beginning to the end.

F. Data Preprocessing

For this research data is taken from the food outlet point of sales system. Data is captured when the customer buys food items. There are many payment types and voucher redeem. For this research, we have omitted such areas and contain hidden knowledge and can be used for future research works. Further, since there are multiple branches and millions of records, we had to limit to one branch with maximum sales numbers. This filtering helped to increase computing performance.

Incomplete data may originate from not relevant data value when gathered, human or equipment or programming issues, diverse considerations between the occasions when the data was gathered and when it is analyzed. Noise in data is another issue which decreases the nature of data. Data preprocessing is an imperative advance in the data mining process.

Data-gathering techniques are regularly loosely controlled, bringing about out-of-range values, contradictive data combinations, missing qualities and so on. Investigating data that has not been cautiously screened for such issues can create misdirecting results. Hence, the representation and nature of data are most important before running an analysis. Frequently, data preprocessing is the most imperative phase of machine learning research.

G. Data Transformation

This is the step where data is transformed or consolidated into forms appropriate for mining by performing operations such as summary and aggregation. Due to availability of huge amount of data and immense need for tuning those data to useful information and knowledge to support management decisions, generalization, normalization, smoothing and aggregation like methods used within data transformation process. Smoothing is used to remove noise from data; aggregation is used to summarization and data cube construction; generalization is used to concept hierarchy climbing and normalization is used to scale within a small specific range. According to the selected data for the research question appropriate smoothing and aggregations are done.

Further, the date is changed or united with the goal that the subsequent mining procedure might be increasingly efficient, and the patterns found might be less complicated.

H. Evaluation/Interpretation

Evaluation is the most critical stage where as it illustrates unknown knowledge from the mined data. The overall idea of an evaluation is to give validation for the research being conducted. Every evaluation consists of opinions in order to do the research correctly. Moreover, good evaluation is based on facts which are unbiased and reasonable.

There are two main concerns. One is how to identify business value from pattern resulted in data mining. Secondly, which technique or visualization tool should be used to show the result.

Therefore, evaluation of created patterns should be done according to a goal or objective to increase efficiency. In order to interpret patterns, visualization tool is important. Many tools are available including chart, histograms, plots, trees and networks.

IV. ANALYSIS AND DESIGN OF THE PROPOSED SOLUTIONS

A. Introduction

Chapter IV presented the approach to analyze associations among items. This chapter illustrates the approach and focuses on high-level design and sub-sectors within the design. Latter part of this chapter focuses on the interactions among the sub sections.

B. Research Design

Science as a collective, aims to create a more precise natural explanation of how the world functions, what its parts are, and how the world got the chance to be how it is presently. Traditionally, science's principle objective has been building information and seeing, paying little mind to its potential applications.

A scientific theory is a clarification of a part of the common world that can be over and again tried and checked as per the logical strategy, utilizing acknowledged conventions of perception, estimation, and assessment of results. Where conceivable, speculations are tried under controlled conditions in an examination. In conditions not manageable to experimental testing, Theories are assessed through standards of thinking. Built up logical hypotheses have withstood thorough examination and encapsulate logical learning.

In this research of analyzing associations among purchased items using data mining, we wish to find how common specific types of associations happens and their examples by taking large dataset which comprise of transaction data of point of sales system. Hence, we need to do a quantitative data analysis using appropriate methodologies.

In this research of analyzing associations among purchased items using data mining, we have used scientific approach which includes common steps. Also, we provide a dashboard containing valuable data for marketing people in order to utilize the research outcome.

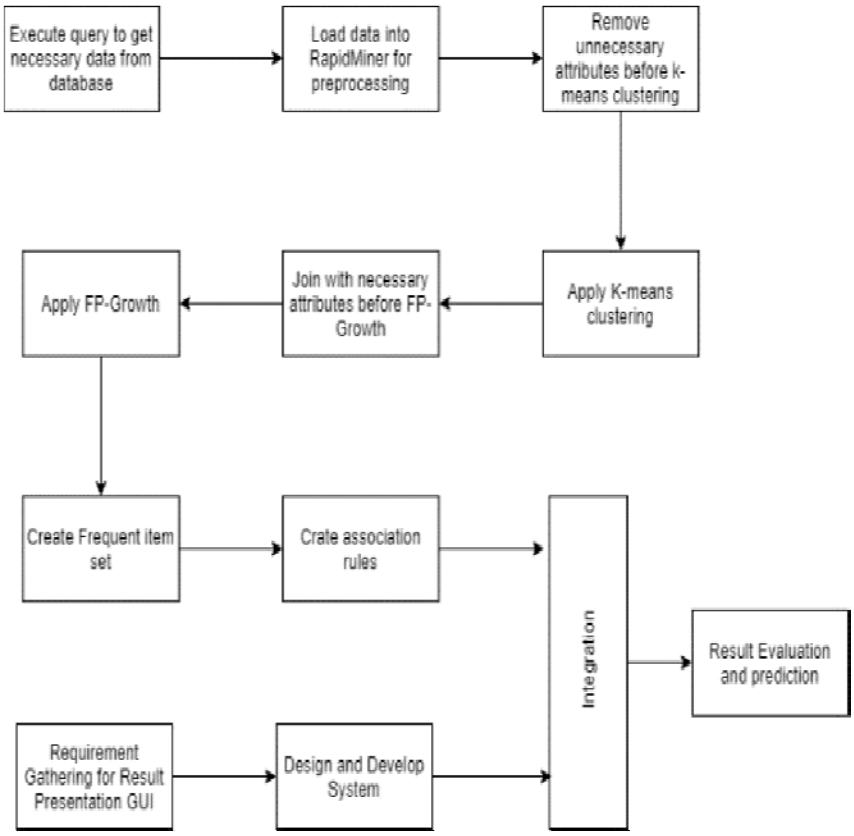


Fig. 4. Activity diagram

The data were obtained from *Niwasushi Pte Ltd*, Singapore. Due to the massiveness of item records (over 5 million) which belongs to 30 branches across the country, only data that belongs to branch WP, which has the highest sales were considered in this research.

Data preprocessing consists of data cleaning, data reduction and data transformation.

Here the author's aim is to use the K-means clustering and FP-Growth algorithm in order to identify frequent dataset patterns related to items and item groups upon consumer purchase transactional data. Assuming bill value represents customers' buying ability, the dataset will be classified using K-means by bill value. Once the classification is completed, transactions are filtered which belongs to a specific time period. Finally, the FP-Growth algorithm is applied on the filtered dataset in order to identify association rules among purchased items.

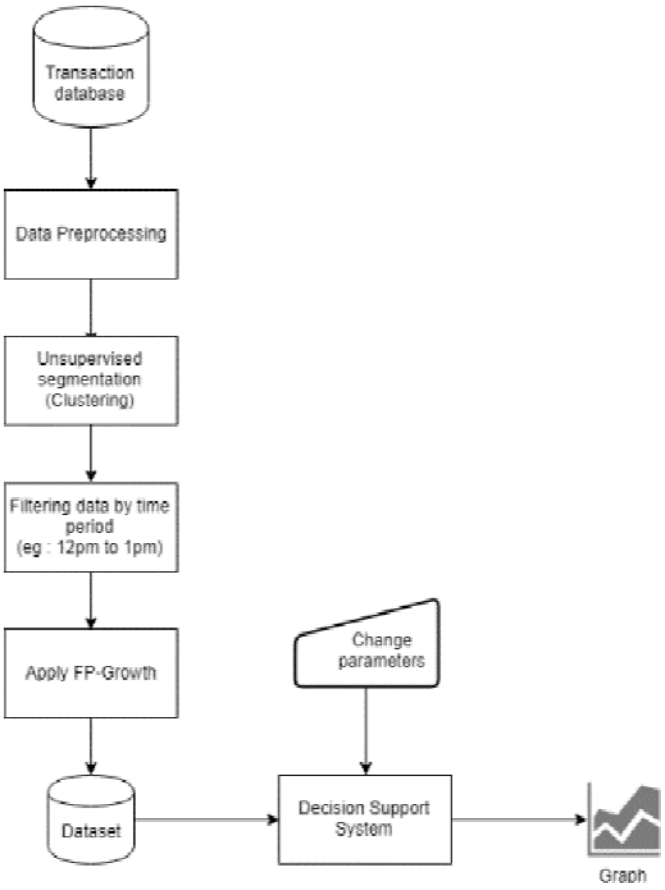


Fig. 5. Research design and method

In this research we focus on 2 areas of food item analysis which will finally be integrated into the Decision support system.

1) Clustering Analysis: This is used to group inter cluster homogeneous items together persisting intra cluster heterogeneity. Depending on research aims and objectives, various clustering algorithms can be performed, for example, K-Mean, Hierarchical are famous algorithms. In this research K-Mean clustering is used.
2) Association Rules Mining: This is an approach in order to find relationships among items in a collection of data. In this research, it is a transactional database. Minimum support and confidence are 2 criteria that are defined in order to filter out related items. There are many algorithms used in data mining: for example, FP-Growth and Apriori Algorithm are few among them. In this research we use the FP-Growth algorithm for the analysis of associations among food items.

C. Clustering Analysis

With the purchase of huge food items per day, there are similarities between the purchasing item groups by different people. Every purchase has a bill value and our objective is to cluster by bill value in order to formulate similar groups together. Items in the cluster are heterogeneous and clusters together are homogeneous. Depending on the research objective, different clustering methods can be used like hierarchical which is developed on distance and construct a hierarchy of clusters. Hierarchical clustering is performed under two types namely Agglomerative and Divisive. Agglomerative refers to a bottom-up approach and divisive follows a top-down mechanism.

Among clustering development that are based on minimizing a formal objective, the widely used and studied is K-Means clustering [25]. K-Means clustering intended to separate objects to clusters where each object included to a cluster with nearest centroids.

We mainly focused on K-Mean clustering is its usability with high resources and the ability to reduce hardware usage while performing the clustering by avoiding candidate key generation and reduce the number of iteration cycles.

In K-Mean clustering, the number k represents the number of clusters and it can be arbitrarily set according to the requirement of the experiment. Increasing the value of k will be giving many clusters resulting in reduced hardware performance in order to execute clustering. We set the K to 5 as an initial value to run the analysis. Thus, the data is set to 5 clusters and centers are calculated accordingly. Then the items are assigned to the closest cluster according to the Euclidean distance, Manhattan or Chebychev functions [26].

For these analyses, full 15 months data was used from 2017-01-01 to 2018-05-24 and a receipt count of 211574 which belong to branch "WP".

Figure 5 shows the analysis which is conducted in this research. Initially, data is extracted from database tables. Once extracted, data is uploaded to RapidMiner for preprocessing. Once preprocessed, in order to apply the k-means algorithm, unnecessary attributes are removed. Secondly the K-Means algorithm is applied. Then select a random cluster and join with another relevant attribute in order to perform the FP-Growth algorithm. Once FP-Growth is applied, save the data and integrate data with the developed system to visualize data.

In this research clustering algorithm is applied to the full dataset in order to find similar groups on food items. Data is extracted from the original database filtered by the branch code "*WP*" and extracted data columns as follows.

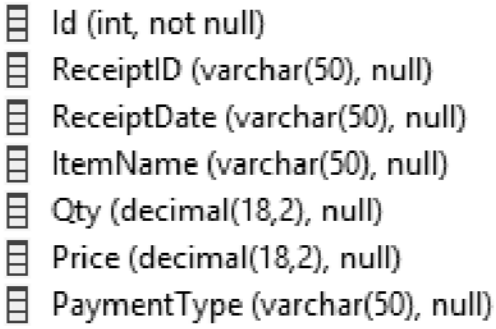


Fig. 6. Table structure

Once the dataset is preprocessed, we aggregate by Item Price and group by Receipt ID.

Then K-Means clustering is performed with Mixed Measures as type and Mixed Euclidean Distance as a measure. Based on the two-component descriptor, a label of distance for each point, it is illustrated that Euclidean distance maps can be developed by the effective sequential algorithm.[27]

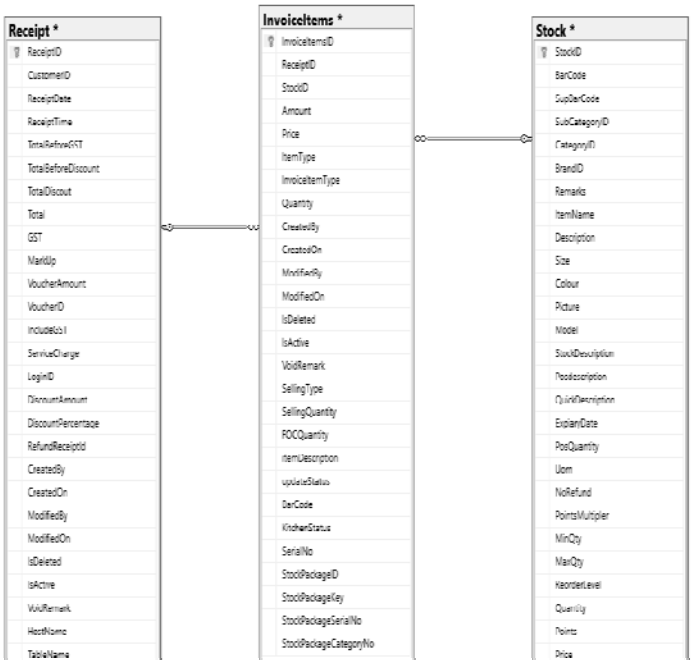


Fig. 7. Table diagram

V. IMPLEMENTATION OF SOLUTION

A. Introduction

In Chapter V, the approach of the research was summarized in a high level. In this chapter, illustrates detail steps which were carried out in order to get the desired results. Moreover, this chapter presents a comprehensive detail on algorithms and methods used in this research and how the output is linked with a developed Decision support system.

B. Data Collection and Preprocessing

Data were obtained from *Niwa Sushi Pte Ltd*, Singapore. Once the tables in the database are identified, we created an appropriate SQL query in order to extract data.

After filtering out for the branch "WP" using Figure 8 query, a sample data set is shown in Figure 9. Since the database was well structured, retrieving relevant data was made faster resulting in better performance.

```
select
r.ReceiptID,
CONVERT(VARCHAR(16),R.ReceiptDate, 120) as ReceiptDate
,st.ItemName
,item.SellingQuantity Qty
,item.Price
,payment.PaymentType
 from InvoiceItems item
inner join Receipt R on item.ReceiptID=R.ReceiptID
inner join Stock st on item.StockID=st.StockID
inner join PaymentType payment on item.ReceiptID=payment.ReceiptID
where
upper(r.BranchName)=upper('WP')
```

Fig. 8. SQL query

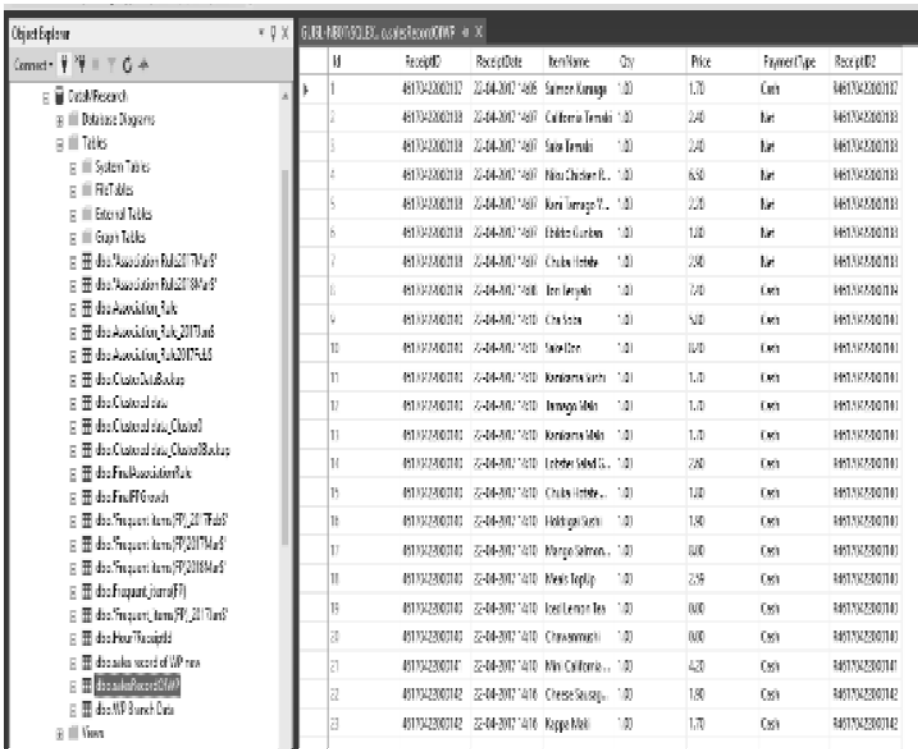


Fig. 9. Sales Records

A comprehensive data preprocessing was conducted while extracting data from tables, in order to remove any anomalies. This step included finding any incomplete, noisy and inconsistent data since real-world data commonly are with anomalies.

Data preparation for data mining consist of cleaning, integration, reduction and transformation. Due to the dirtiness of real-world data, incomplete, noisy and inconsistent data needs to be handled properly or else cannot gain quality result. Data cleaning consists of filling missing values, smoothing noisy data, handling outliers and inconsistencies. Missing values can be handled by ignoring them, treating as a separate value or filling with mean or median. Due to the accuracy of the point of sale system, most of the tuples are complete, consistent and accurate which will reduce the discrepancies in sales figures, end of the day.

This process mainly included the following steps.

1) Data cleaning
2) Data integration
3) Data transformation
4) Data reduction
5) Data discretization

Data cleaning includes filling missing values, treatment for noisy data, outlier removing and solve any inconsistencies. Data integration consists of joining multiple databases, tables or files. Data transformation includes normalization and aggregation. Data integration combines separate data sources. Integration is done carefully in order to avoid redundancies and inconsistencies in the final dataset which will affect the accuracy and processing speed of subsequent methods. Data reductions strategies need to conduct due to avoid too many instances and the curse of dimensionality. The much smaller dataset is easy to handle but yet produce the same analytical result. The reduction can be done by removing unimportant fields, grouping and clustering.

Dimensionality reduction is conducted by direct or indirect methods. The direct method selects the minimum set of attributes that is sufficient for applying data mining tasks.

Indirect methods consist of various methodologies like Principle component analysis and Singular value decomposition.

Data reduction includes minimizing the number of records without affecting the analytical results. Data discretization includes replacing the numeric fields with nominal values. Discretization divides the continuous attributes to a range of intervals since some algorithms accept only categorical attributes. There are techniques used in discretization. Binning methods like equal-width, equal-depth and Entropy-based are commonly used.

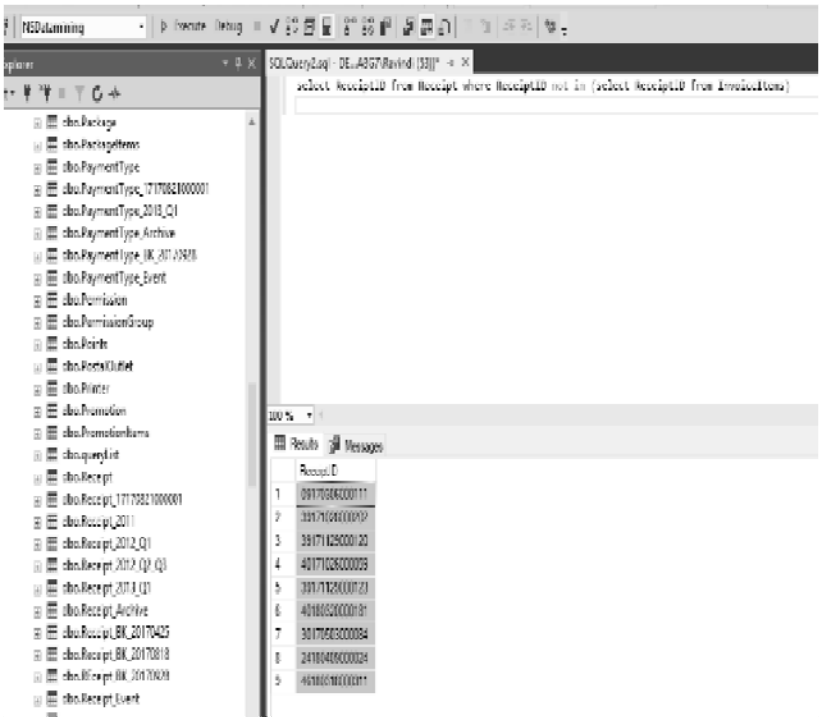


Fig. 10. Missing records

There were several Receipt numbers without sales. Even though is not affecting the final outcome, we tend to remove them. Figure 10 shows sample receipt numbers without items.

C. Clustering Analysis

Since our initial approach was to analyze hourly associations, we calculated the number of receipts per hour.

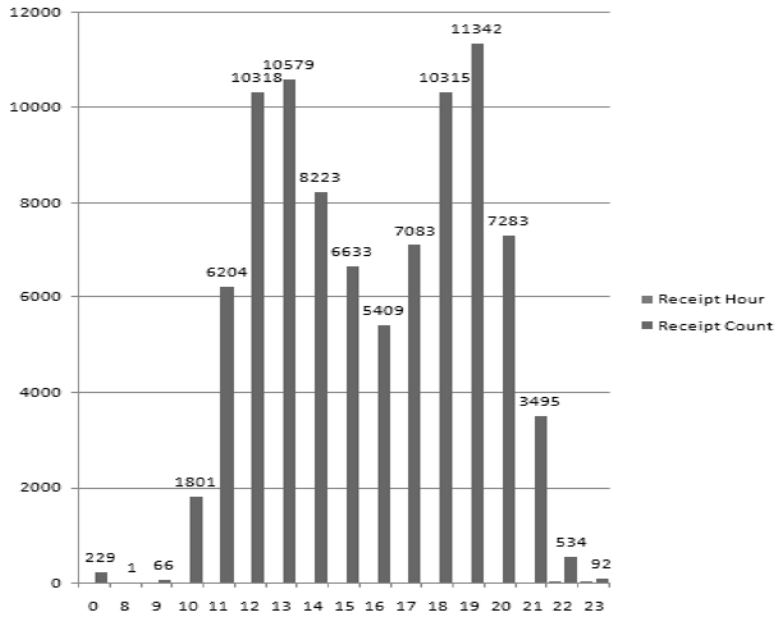


Fig. 11. Hourly item count

K- Means clustering was applied to the "WP" branch data set with the following parameters. The clustering parameters we set before execute as follows.

Number of clustering: 5

Maximum iterations: 10

Measure Type: Mixed Measures

Mixed measures: Mixed Euclidean Distance

Maximum Optimization steps: 100

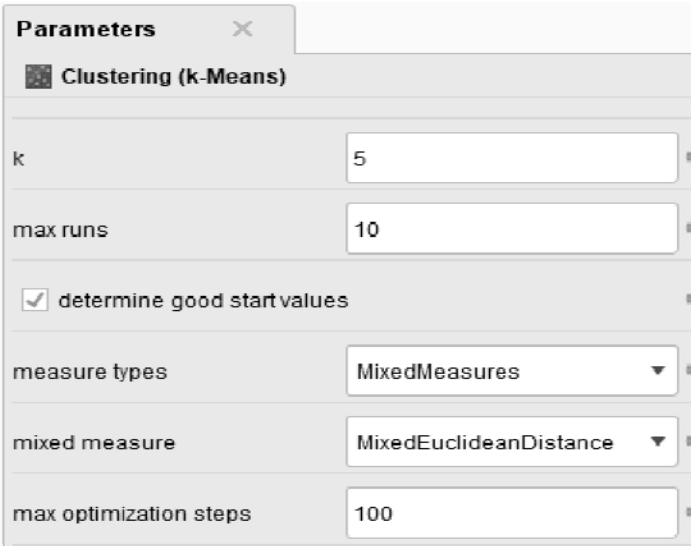


Fig. 12. Clustering parameters

Figure 11 shows the schema diagram of the Clustering process in Rapid Miner. We have followed the default layout of K-Means clustering. First, we imported the process data source which is saved in a spreadsheet. Secondly, we aggregated to sum the individual items prices to get the total price adjacent to Receipt number. Finally, K-Means clustering was applied to the aggregated items.

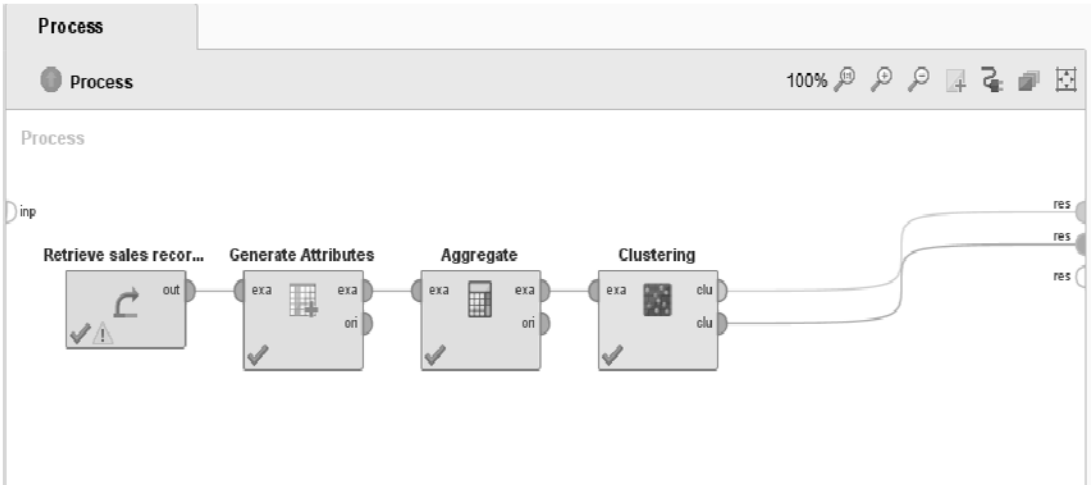


Fig. 13. Clustering diagram

Clustered output was as follows. Since the research needs to conduct in a comprehensive manner, assuming that more data output more results, we selected the cluster having the maximum number of items, by choosing "Cluster 0".

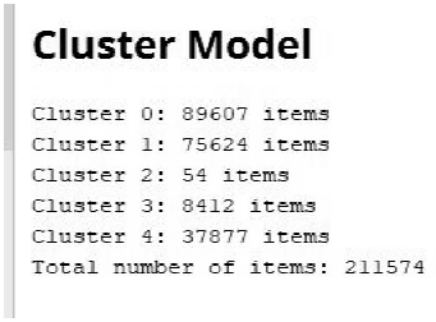


Fig. 14. Clustering summary

Figure 13 shows a sample collection of how the clustering was distributed and the Receipt number with the included cluster. This is the base of the next step of generating association rules using the FP-Growth algorithm.

ExampleSet (211574 examples, 3 special attributes, 1 regular attribute) Filter (211,574 / 211,574 examples):

Row No.	id	ReceiptID	cluster	sum(Price)
1	1	46170101000001	cluster_0	30.700
2	2	46170101000002	cluster_3	4.800
3	3	46170101000003	cluster_2	1.800
4	4	46170101000004	cluster_3	4.800
5	5	46170101000005	cluster_2	2.600
6	6	46170101000006	cluster_2	1.800
7	7	46170101000007	cluster_3	6.500

Fig. 15 Sample clustering data

Since the clustered records were not with receipt dates, we updated them by joining with initial dataset which was a time-consuming process.

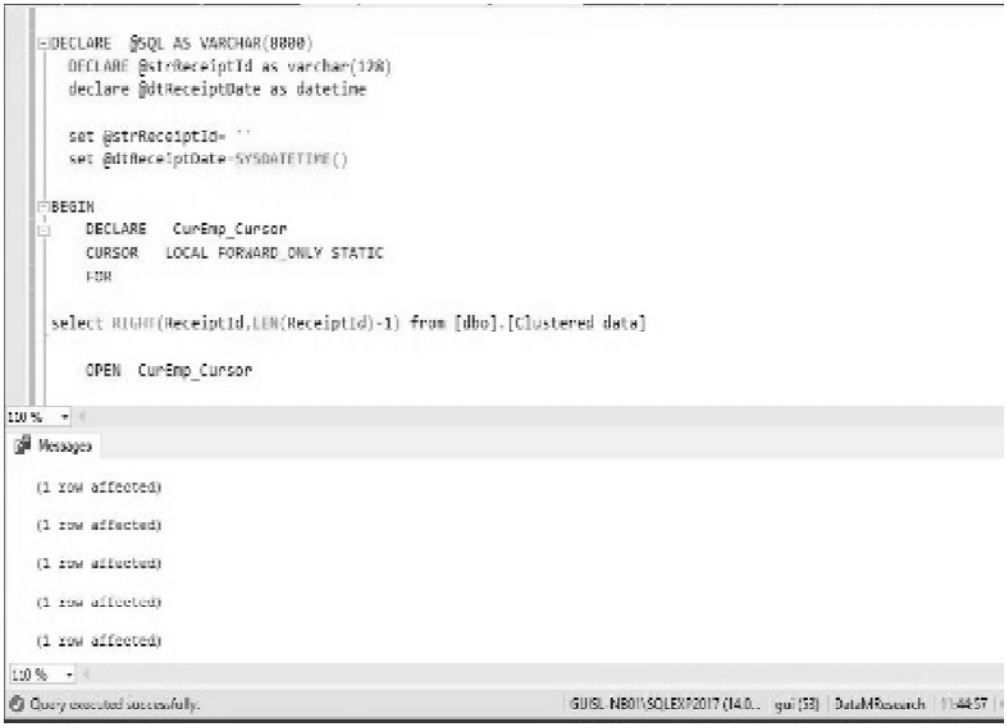


Fig. 16 Merged attributes

Further, we selected the time frame 12PM to 1PM in Cluster 0, where we can find the associations throughout the dataset belongs to the time frame 12PM to 1PM.

D. Association rule generation

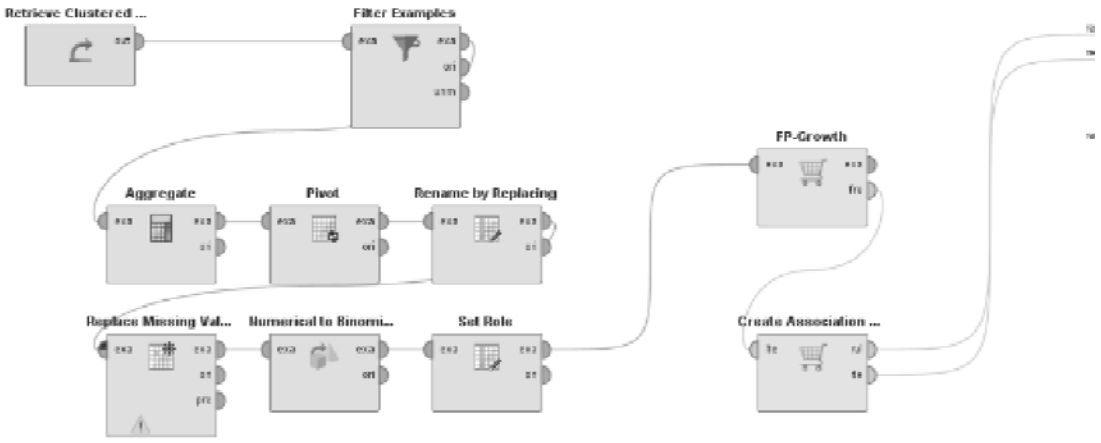


Fig. 17 Generating association rules

We applied FP-Growth to the Cluster 0 data in order to find the association rules which will be the input for a Decision support system. Figure 6.10 shows the schematic diagram of generating association rules before the approach change.

No.	Promises	Conclusion	Support	Confidence	LaPlace	Gain
1	Grilled Wings	Iced Lemon Tea	0.012	0.100	0.905	-0.225
2	Grilled Wings	Gyu Don	0.012	0.100	0.905	-0.225
3	Grilled Wings	Meals TopUp, Iced Lemon Tea	0.012	0.100	0.905	-0.225
4	Grilled Wings	Meals TopUp, Tori Teriyaki	0.012	0.100	0.905	-0.225
5	Grilled Wings	Meals TopUp, Gyu Don	0.012	0.100	0.905	-0.225
6	Grilled Wings	Iced Peach Tea, Tori Teriyaki	0.012	0.100	0.905	-0.225
7	Grilled Wings	Iced Peach Tea, Gyu Don	0.012	0.100	0.905	-0.225
8	Grilled Wings	Meals TopUp, Iced Peach Tea, Tori Teriyaki	0.012	0.100	0.905	-0.225
9	Grilled Wings	Meals TopUp, Iced Peach Tea, Gyu Don	0.012	0.100	0.905	-0.225
10	Meals TopUp	Chawanmushi, Tori Toji	0.036	0.105	0.774	-0.639
11	Meals TopUp	Iced Lemon Tea, Tori Toji	0.036	0.105	0.774	0.639
12	Meals TopUp, Grilled Wings	Iced Lemon Tea	0.012	0.105	0.910	-0.213
13	Meals TopUp, Grilled Wings	Tori Teriyaki	0.012	0.105	0.910	-0.213
14	Meals TopUp, Grilled Wings	Gyu Don	0.012	0.105	0.910	-0.213
15	Meals TopUp	Chawanmushi, Iced Lemon Tea, Tori Toji	0.036	0.105	0.774	-0.639
16	Meals TopUp, Grilled Wings	Iced Peach Tea, Tori Teriyaki	0.012	0.105	0.910	-0.213

Fig. 18 Sample data for association rules

Figure 16 shows sample data after applying the FP-Growth algorithm. Later, as instructed by the supervisor, the approach was changed from hourly calculation to a daily basis. This is due to make the research more comprehensive by comparing a particular date or day in a year with the corresponding day or date in the adjacent year. This helps the marketing team to compare and contrast significant relationships and use the information to generate marketing strategies.

1) Association rule generation by date

Association rules were calculated based on the date for 3 months namely January, February and march each in 2017 and 2018. This would help to compare the correspondent months by changing different parameters. We can further analyze how the final data is responding according to criteria. Main criteria are the support and confidence levels.

Support is the probability of the item set appearing in the data set.

Fig. 19 Generating association rules by date

Final generated data for each day were copied to a spreadsheet and finally imported to SQL Server for merging and to prepare the database for the Decision support system. Figure 18 shows the final stage of importing data containing association rules and frequent item sets belongs to a particular month.

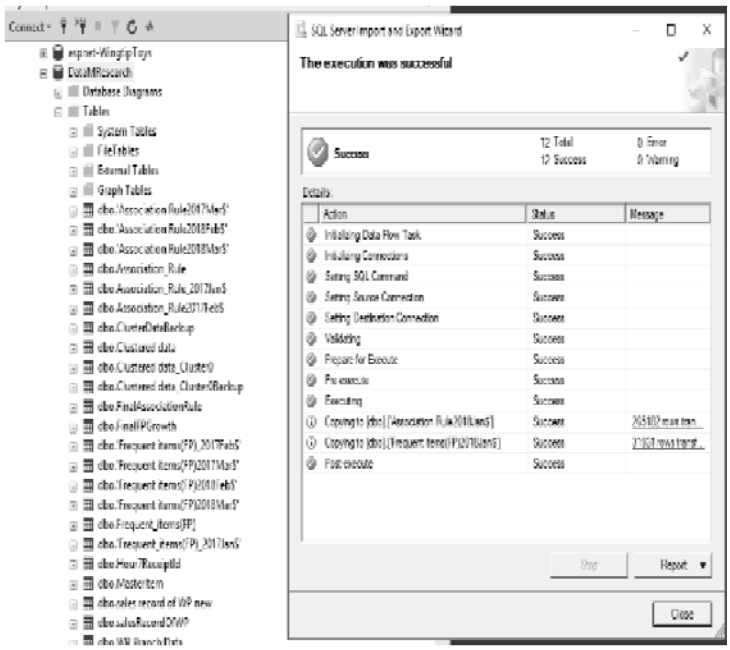


Fig. 20 Import data

E. Graphical user interface for analysis of associations

We have developed a graphical user interface in order to analyze the association rules with the ability to change important parameters such as support and confidence. Support and confidence play a major role in analyzing association rules and helps them to discover the hidden knowledge of data which is our ultimate goal.

A compact decision support system is introduced with a number of parameters to support advance decision making. In the system, we have provided the following parameters that can be changed and projects the summary of associations in a tabular format.

1) Compare by 2 dates of the year 2017 & 2018
2) Compare by weekdays of a month in 2017 & 2018
3) Search by multiple items
4) Search by a maximum value of support
5) Search by a maximum value of confidence

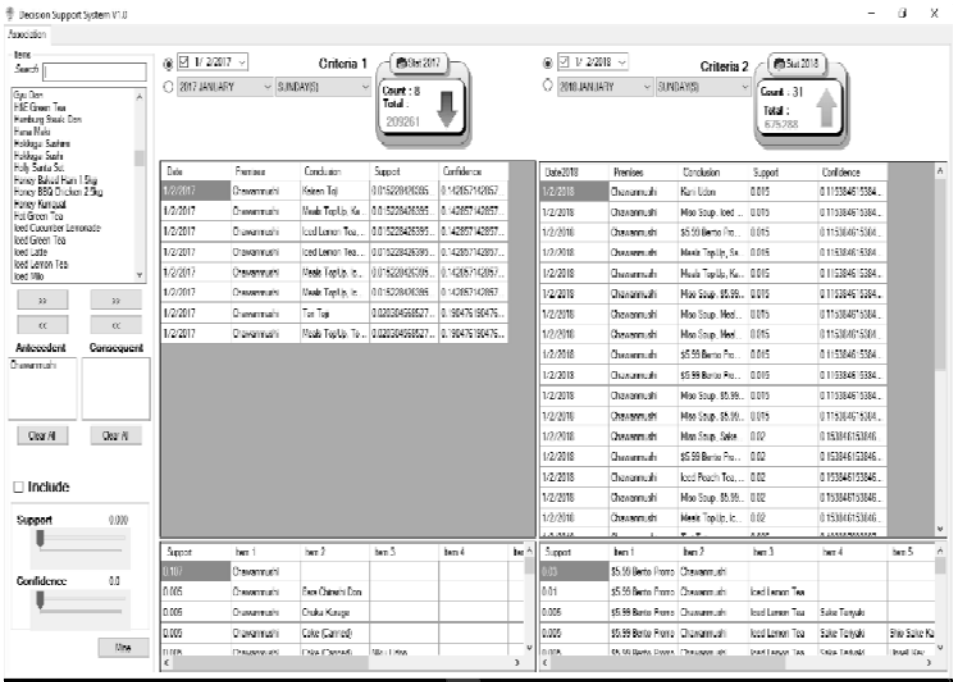


Fig. 21 Decision support system

Figure 19 shows the main user interface of which marketing personals mainly interact.

The user interface consists of 6 sections and they are as follows.

1) Date and week day filtering
2) Item, support & confidence filtering
3) 2017 Association rules
4) 2017 Frequent item set
5) 2018 Association rules
6) 2018 Frequent item set

F. Date and week day filtering

Fig. 22 Week day filtering

This filtering option provides the 2 options to compare either from a specific date with corresponding year or compare with a week day of a specific month with corresponding year. In week day comparison, all the days are first calculated and then the associations are calculated. The comparison can be done with parallel year week day or another week day.

G. Item Comparison

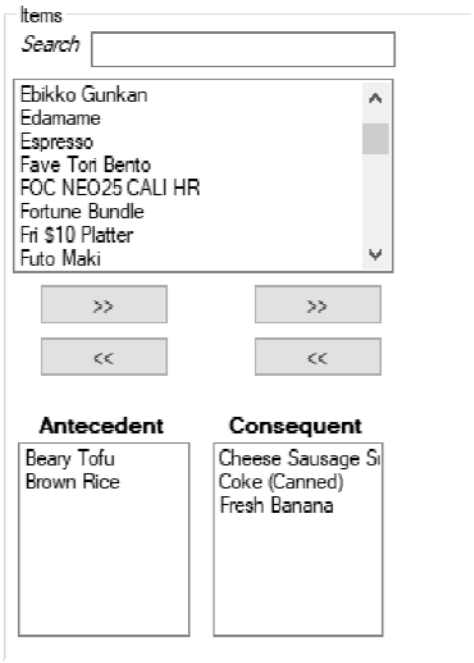
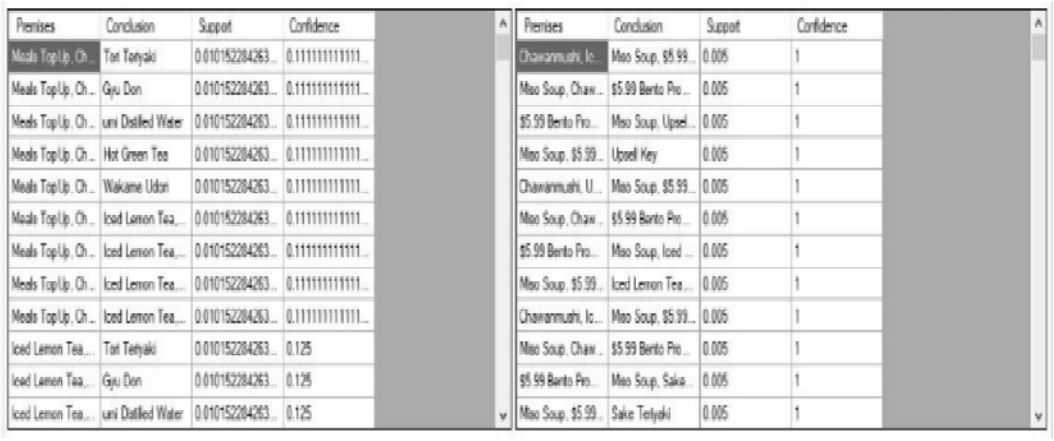


Fig. 23 Item comparison

Figure 21 shows the section to select a specific item or a collection of items used to analyze the associations. Further the support and confidence can be set by dragging the track bars and values are set as less than or equal of the track bar value.

H. Association rules (2017/2018)

Once the parameters are set, the corresponding association rules are visible in terms of support and confidence in tabular format. Mainly this section is compressed to premises, conclusion, support and confidence. Figure 23 shows 2017 & 2018 association rules to compare in a convenient visual.

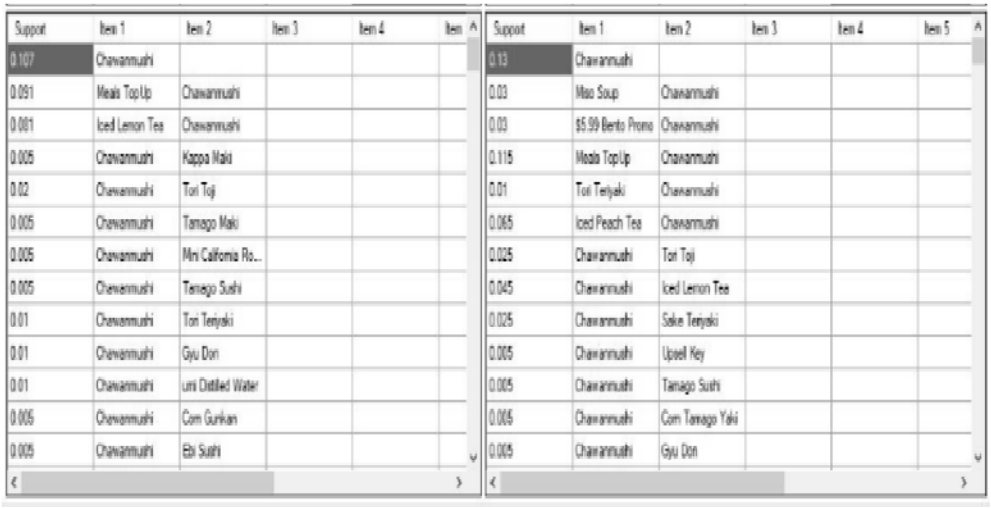


Premises	Conclusion	Support	Confidence
Meals Top Up, Ch...	Tori Teriyaki	0.010152284263...	0.111111111111...
Meals Top Up, Ch...	Gyu Don	0.010152284263...	0.111111111111...
Meals Top Up, Ch...	uni Distilled Water	0.010152284263...	0.111111111111...
Meals Top Up, Ch...	Hot Green Tea	0.010152284263...	0.111111111111...
Meals Top Up, Ch...	Wakame Udon	0.010152284263...	0.111111111111...
Meals Top Up, Ch...	Iced Lemon Tea,...	0.010152284263...	0.111111111111...
Meals Top Up, Ch...	Iced Lemon Tea,...	0.010152284263...	0.111111111111...
Meals Top Up, Ch...	Iced Lemon Tea,...	0.010152284263...	0.111111111111...
Meals Top Up, Ch...	Iced Lemon Tea,...	0.010152284263...	0.111111111111...
Iced Lemon Tea,...	Tori Teriyaki	0.010152284263...	0.125
Iced Lemon Tea,...	Gyu Don	0.010152284263...	0.125
Iced Lemon Tea,...	uni Distilled Water	0.010152284263...	0.125

Premises	Conclusion	Support	Confidence
Chawanmushi, Ic...	Miso Soup, $5.99...	0.005	1
Miso Soup, Chaw...	$5.99 Bento Pro...	0.005	1
$5.99 Bento Pro...	Miso Soup, Upsel...	0.005	1
Miso Soup, $5.99...	Upsell Key	0.005	1
Chawanmushi, U...	Miso Soup, $5.99...	0.005	1
Miso Soup, Chaw...	$5.99 Bento Pro...	0.005	1
$5.99 Bento Pro...	Miso Soup, Iced ...	0.005	1
Miso Soup, $5.99...	Iced Lemon Tea,...	0.005	1
Chawanmushi, Ic...	Miso Soup, $5.99...	0.005	1
Miso Soup, Chaw...	$5.99 Bento Pro...	0.005	1
$5.99 Bento Pro...	Miso Soup, Sake...	0.005	1
Miso Soup, $5.99...	Sake Teriyaki	0.005	1

Fig. 24 Association rules

I. Frequent item set (2017/2018)



Support	Item 1	Item 2	Item 3	Item 4	Item
0.107	Chawanmushi				
0.091	Meals Top Up	Chawanmushi			
0.081	Iced Lemon Tea	Chawanmushi			
0.005	Chawanmushi	Kappa Maki			
0.02	Chawanmushi	Tori Toji			
0.005	Chawanmushi	Tamago Maki			
0.005	Chawanmushi	Mini California Ro...			
0.005	Chawanmushi	Tamago Sushi			
0.01	Chawanmushi	Tori Teriyaki			
0.01	Chawanmushi	Gyu Don			
0.01	Chawanmushi	uni Distilled Water			
0.005	Chawanmushi	Corn Gunkan			
0.005	Chawanmushi	Ebi Sushi			

Support	Item 1	Item 2	Item 3	Item 4	Item 5
0.13	Chawanmushi				
0.03	Miso Soup	Chawanmushi			
0.03	$5.99 Bento Promo	Chawanmushi			
0.115	Meals Top Up	Chawanmushi			
0.01	Tori Teriyaki	Chawanmushi			
0.065	Iced Peach Tea	Chawanmushi			
0.025	Chawanmushi	Tori Toji			
0.045	Chawanmushi	Iced Lemon Tea			
0.025	Chawanmushi	Sake Teriyaki			
0.005	Chawanmushi	Upsell Key			
0.005	Chawanmushi	Tamago Sushi			
0.005	Chawanmushi	Corn Tamago Yaki			
0.005	Chawanmushi	Gyu Don			

Fig. 25 Association rules

Figure 23 shows the resulting frequent item set according to the filters. These frequent item sets are used for generating the association rules.

Support gives the number of occurrences (frequency) of an item in a dataset. This can be denoted as follows.

The criteria $X \Rightarrow Y$ keeps support C if percentage of transactions in D having $X \cup Y$. Rules that contain C larger than a specific support is mentioned to have minimum support.

TABLE I: SUPPORT

Transaction ID	Item	Support =Occurrence/Total support
1	ABC	Support {ABC}= 1/5 =20%
2	BCD	Total support = 5
3	AC	Support {AB} = 2/5 = 40%
4	BC	Support {BC} = 3/5 = 60%
5	ABD	

The second important factor is confidence. Confidence is the percentage of item A in transaction T, that also having item B. Further explaining, the probability of occurring item A with fact of item B is already in the basket. Using notations, the confidence is denoted as following way.

$$\text{Confidence}(A \rightarrow B) = P(B/A)$$

We can represent confidence with the following example

TABLE II: CONFIDENCE

Transaction ID	Item	Given X ⇒ Y Confidence = Occurrence {Y} / Occurrence {X}
1	ABC	
2	BCD	Confidence {A⇒B} = 2/3 = 66%
3	AC	Confidence {B⇒C} = 3/4 = 75%
4	BC	Confidence {AB⇒C} = 1/2 = 50%
5	ABD	

In this research, we use K-Means clustering in order to segment heterogeneous groups of transaction receipt by bill value. The bill values are clustered to 5 groups and each centroid are as follows. These centroids were selected randomly by the Rapid miner tool. According to cluster item count, Figure 7.1 shows that they fall into a broad range from a minimum 54 to maximum 89607. Results may be different and unique for one cycle of the clustering process. Therefore, results may totally different if the clustering process executes another time. The knowledge discovery was done by using the only 1/5th of the dataset and the clustering may different once the data size is increased.

TABLE I: CLUSTER SUMMARY

Cluster	Centroid	Item count
cluster_0	8.715	89607
cluster_1	2.573	75624
cluster_2	163.804	54
cluster_3	34.296	8412
cluster_4	18.152	37877
	Total	**211574**

J. Association Rules statistics

Since cluster_0 has more data, we selected cluster_0 to apply FP-Growth in order to find the association rules. Figure 27 illustrates sample associations generated.

Further, the proposed decision support system also emphasis the association rules among the generated result with user-friendly parameters variations. Figure 7.3 shows the summary of the association rules data which was created throughout the research. As it shows, over 1 million associations were generated for 6 months in 2 years.

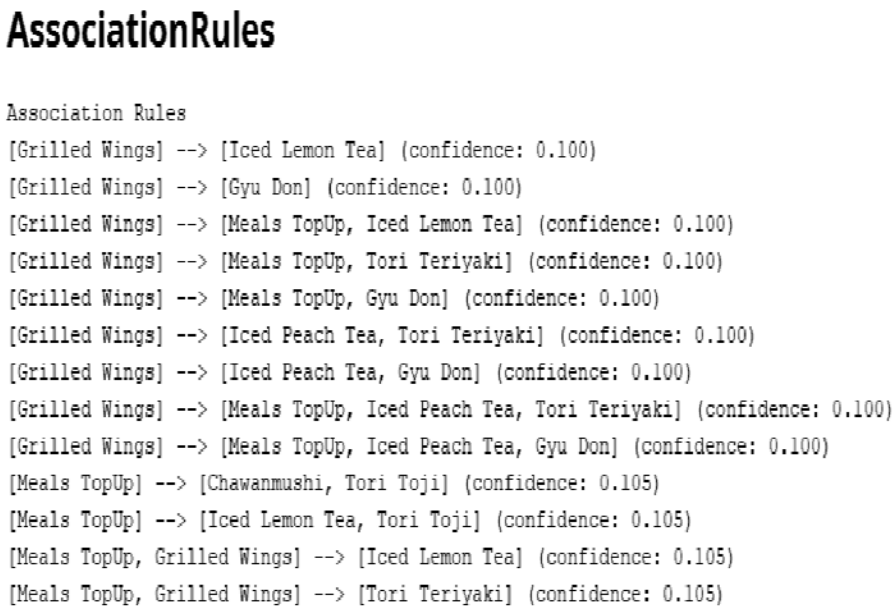


Fig. 26 Sample data for association rules

TABLE IV: ASSOCIATION RULES STATISTICS

Year	Month	Count	Min Support	Max Support	Confidence Min/Max	
2017	JAN	179081	0.005	0.27	0.1	1
2017	FEB	169757	0.0050	0.2222	0.1	1
2017	MAR	142500	0.0051	0.2368	0.1	1
2018	JAN	265182	0.005	0.3450	0.1	1
2018	FEB	225439	0.0051	0.2559	0.1	1
2018	MAR	184667	0.0050	0.1793	0.1	1
	Total	**1166626**				

TABLE V: FREQUENT ITEM SET STATISTICS

Year	Month	Count
2017	January	24891
2017	February	24263
2017	March	20403
2018	January	31931
2018	February	27817
2018	March	28618
	Total	**157923**

VI. EVALUATION

A. Introduction

The previous chapter described the implementation in detail of all the sections of this research. This chapter evaluates and justifies the results and the decision support system developed by the researcher.

B. Association Comparison

With the decision support system, we are now able to compare association rules among items or items sets with the date or week days of the corresponding year. This comparison helps to identify consumer behavior on particular item or items set which is our final goal. This knowledge further can be converted to marketing information in order to conduct a promotional campaign or adjust business decisions and approaches.

Further, our approach is unique and novel; this research makes traditional market basket analysis in a comprehensive manner.

Among over 1 million of association rules (Figure 28) which were generated from the research process makes the complexity of the analysis is probably high. Therefore, we select random items and item sets to illustrate knowledge discovery by comparing them.

As an example, in 2017 January, a promotion was conducted named *$5 Bento Promo* and the maximum support is 1% and the confidence is limited to 25%. In 2018 January, there was a similar promotion conducted and that shows the highest support throughout the association rules. The promotion was *$5.99 Bento Promo* and with *Miso Soup* gives the highest support value of 0.34. This emphasizes that buyers who bought *Miso Soup* are having 34% probability of buying *$5.99 Bento Promo*. Further, it shows that in some days the confidence level reaches a maximum of 100%.

This is an advantage for the marketing team in order to predict and conduct such promotions in upcoming years.

Another example was buying *Chawanmushi* on Sundays in January 2017, shows support of 3% along with *Tori Toji* and *Iced Lemon Tea*. Further among all transactions, when buying *Chawanmushi* there is a confidence of 26.9% of buying *Tori Toji* and *Iced Lemon Tea*.

If we compare this with Sundays of January 2018, this association is spread through 2 Sundays with the support of 2%, confidence of 10% and support of 1%, confidence 13.6% respectively.

Fig. 27 Comparison of knowledge

C. *Evaluation of Decision support system*

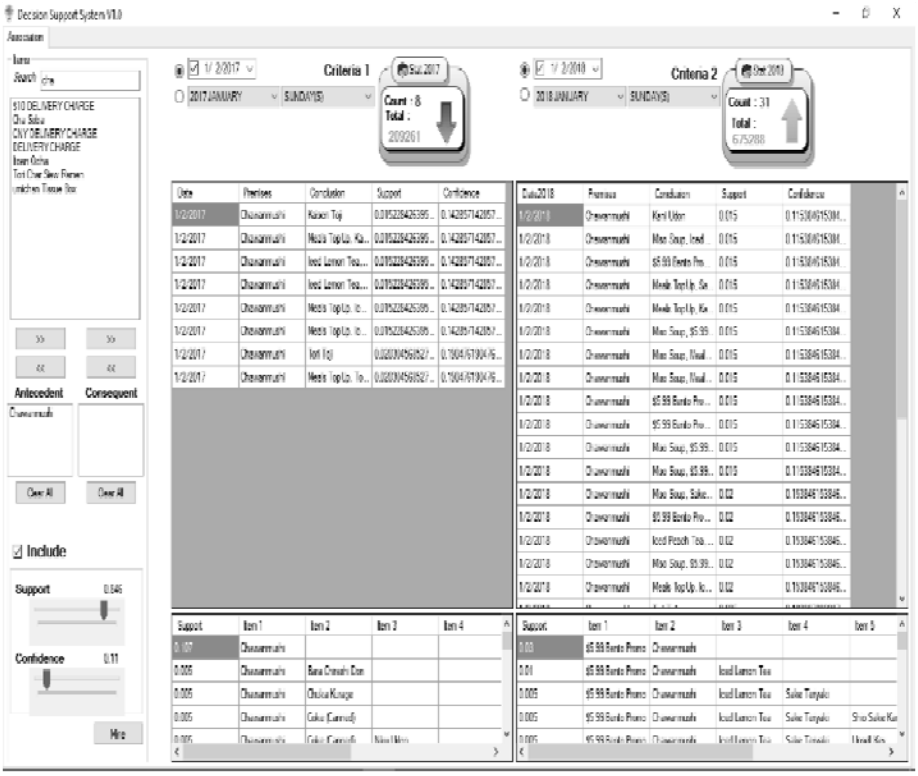


Fig. 28 System user interface

Proposed Decision support system can generate summarized results for association rules and the marketing people can check for least associations and frequent item sets and conduct marketing campaigns. Therefore, we have provided a very compact summary of the output of this research in terms of support and confidence.

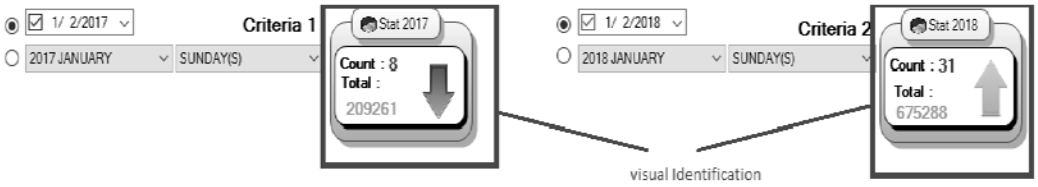


Fig. 29 Visual statistics

As an enhanced feature, the user can easily identify the year of having the most counts. This feature allows the user to visually identify the year which having most records without manually counting the number of records. This feature is especially helping when the record count exceeds the human counting limitations and not clearly visualized in the grid view.

It is very difficult to evaluate a system in terms of errors; bugs and we have managed to extract only the relevant and useful information and knowledge from the executed process. This knowledge is the key to the aim and objective of this research. Marketing people are now able to analyze these associations and then conduct marketing campaigns and strategies for less selling item sets with low support and confidence.

VII. CONCLUSION AND FUTURE WORKS

A. Introduction

Data Mining has played a very important role in Market basket Analysis and various other fields. The most important point to succeed in a marketing strategy is to create an accurate purchase analysis. The motivation for applying data mining approach on purchase item Analysis is to learn about buying patterns and retailers can use this information so more numbers of consumers are attracted towards them.

B. Limitations

There are few known limitations of the tool such as finding association rules of months other than provided months. Also, the correctness of the information is highly dependent on the accuracy of the clustering and performing association algorithm. Also, customer management is not involved with data; an accurate promotion targeting a specific customer does not take place. Further, the data is limited to a narrow time frame and hence conclusions cannot be suggested for other time frames.

C. Future work

As future works, we can propose to conduct a similar process with another association algorithm and conduct a comparison for the most efficient process. Also, by changing the clustering algorithm, comparison can be made comprehensively with the conducted research.

Another approach is to conduct a traditional market basket analysis in order to compare with the result of this research. This will lead to compare the and contrast the results of this research.

Further, there are other branches excluded from this research, which also may contain hidden knowledge totally different from this research which is rich knowledge discovery sources. Alternatively, the 4 other clusters which were avoided, can be used to follow the steps of this research and compare with the final results, where the 4 other clusters may contain full of unknown knowledge.

An interesting approach would be to design the research with test and train data in order to validate and verify the final result with a threshold error rate in order to accept or reject certain associations rules.

D. Summary

This chapter concluded the system functionality, limitations of the system and the future work that can be potential.

ACKNOWLEDGMENT

I dedicate the output of this research work to *Umi sushi Restaurant* chain who are pioneers in Japanese cuisine in Singapore.

REFERENCES

[1] J. Han, J. Pei, and Y. Yin, "Mining frequent patterns without candidate generation," in ACM sigmod record, 2000, pp. 1-12.

[2] J. Han, J. Pei, and M. Kamber, Data mining: concepts and techniques: Elsevier, 2011.

[3] R. Agrawal, T. Imieliński, and A. Swami, "Mining association rules between sets of items in large databases," in Acm sigmod record, 1993, pp. 207-216.

[4] K. Thearling, "An introduction to data mining," ed, 2017.

[5] S. Suresh, "Application of Retail Analytics Using Association Rule Mining In Data Mining Techniques With Respect To Retail Supermarket* A. Pappu Rajan," 2015.

[6] J. MacQueen, "Some methods for classification and analysis of multivariate observations," in Proceedings of the fifth Berkeley symposium on mathematical statistics and probability, 1967, pp. 281-297.

[7] L. H. Ungar and D. P. Foster, "Clustering methods for collaborative filtering," in AAAI workshop on recommendation systems, 1998, pp. 114-129.

[8] S. Su, S. Xu, X. Cheng, Z. Li, and F. Yang, "Differentially private frequent itemset mining via transaction splitting," IEEE Transactions on Knowledge and Data Engineering, vol. 27, pp. 1875-1891, 2015.

[9] M. H. Kabir, "Data Mining Framework for Generating Sales Decision Making Information Using Association Rules," INTERNATIONAL JOURNAL OF ADVANCED COMPUTER SCIENCE AND APPLICATIONS, vol. 7, pp. 378-385, 2016.

[10] K. Yagi, M. Soramoto, M. Mokuda, and Y. Morimoto, "Optimized Sequential Pattern Mining from Point Of Sales Data," in Data Engineering Workshops, 2005. 21st International Conference on, 2005, pp. 1223-1223.

[11] C. Schwenke, V. Vasyutynskyy, and K. Kabitzsch, "Analysis and simulation of sales receipt data in supermarkets," in Emerging Technologies & Factory Automation (ETFA), 2011 IEEE 16th Conference on, 2011, pp. 1-8.
[12] A. M. Mirajkar, A. P. Sankpal, P. S. Koli, R. A. Patil, and A. R. Pradnyavant, "Data Mining Based Store Layout Architecture for Supermarket," 2016.
[13] W. Ismail, M. M. Hassan, and G. Fortino, "Productive-associated Periodic High-utility itemsets mining," in Networking, Sensing and Control (ICNSC), 2017 IEEE 14th International Conference on, 2017, pp. 637-642.
[14] T. Ishigaki, T. Takenaka, and Y. Motomura, "Computational Customer Behavior Modeling for Knowledge Management with an Automatic Categorization Using Retail Service's Datasets," in e-Business Engineering (ICEBE), 2010 IEEE 7th International Conference on, 2010, pp. 528-533.
[15] V. Aravindan, "Mining Frequent Sequential Patterns From Multiple Databases Using Transaction Ids," University of Windsor (Canada), 2016.
[16] P. Fournier-Viger, C.-W. Wu, S. Zida, and V. S. Tseng, "FHM: faster high-utility itemset mining using estimated utility co-occurrence pruning," in International symposium on methodologies for intelligent systems, 2014, pp. 83-92.
[17] X. Chen, Y. Fang, M. Yang, F. Nie, Z. Zhao, and J. Z. Huang, "PurTreeClust: A Clustering Algorithm for Customer Segmentation from Massive Customer Transaction Data," IEEE Transactions on Knowledge and Data Engineering, 2017.
[18] R. A. E.-D. Ahmeda, M. E. Shehaba, S. Morsya, and N. Mekawiea, "Performance study of classification algorithms for consumer online shopping attitudes and behavior using data mining," in Communication Systems and Network Technologies (CSNT), 2015 Fifth International Conference on, 2015, pp. 1344-1349.
[19] X. Chen, S. Peng, J. Z. Huang, F. Nie, and Y. Ming, "Local PurTree Spectral Clustering for Massive Customer Transaction Data," IEEE Intelligent Systems, vol. 32, pp. 37-44, 2017.
[20] K. Lu and T. Furukawa, "A Framework for Segmenting Customers Based on Probability Density of Transaction Data," in Advanced Applied Informatics (IIAIAAI), 2012 IIAI International Conference on, 2012, pp. 273-278.
[21] X. Min-jie and Z. Jin-ge, "Research on personalized recommendation system for e-commerce based on web log mining and user browsing behaviors," in Computer Application and System Modeling (ICCASM), 2010 International Conference on, 2010, pp. V12-408-V12-411.
[22] S. S. R. Shariff and Z. Bakri, "Managing product stock keeping based on purchase association effect," in Technology Management and Emerging Technologies (ISTMET), 2015 International Symposium on, 2015, pp. 187-192.
[23] G. Chunfang and G. Zhongliang, "Innovation of enterprise profit patterns based on Big Data," in Logistics, Informatics and Service Sciences (LISS), 2015 International Conference on, 2015, pp. 1-5.
[24] P. S. Raju, D. V. R. Bai, and G. K. Chaitanya, "Data mining: Techniques for enhancing customer relationship management in banking and retail industries," International Journal of Innovative Research in Computer and Communication Engineering, vol. 2, pp. 2650-2657, 2014.
[25] T. Kanungo, D. M. Mount, N. S. Netanyahu, C. D. Piatko, R. Silverman, and A. Y. Wu, "An efficient k-means clustering algorithm: Analysis and implementation," IEEE Transactions on Pattern Analysis & Machine Intelligence, pp. 881-892, 2002.
[26] X. Cui and T. E. Potok, "Document clustering analysis based on hybrid PSO+ K-means algorithm," Journal of Computer Sciences (special issue), vol. 27, p. 33, 2005.
[27] P.-E. Danielsson, "Euclidean distance mapping," Computer Graphics and image processing, vol. 14, pp. 227-248, 1980.