Association Rule Mining on Customer's Data using Frequent Pattern Algorithm

Khaled H. Alyoubi
kalyoubi@kau.edu.sa
Information Systems Department, Faculty of Computing and Information Technology, King Abdulaziz University, Jeddah, Saudi Arabia.

Summary
Nowadays, organizations are generating immense amount of data dealing with multiple stakeholders. The collaborative environment and the use of latest technologies in the current market scenario, creating an additional pressure on the organization. The use of latest computing tools can help the enterprises to be competent, resourceful and to deal with huge data smartly. Therefore, this research shown the use of one of the promising computing strategies known as data mining. Data mining is commonly known for generating hidden patterns and for knowledge discovery. This research proposed a model for analyzing the customer's data to generate hidden patterns from it. The purpose is to extract hidden knowledge from the data generated form multiple situations while dealing with the customers. The model implementation performed using association mining algorithm called FP-Growth. The algorithm is famous for generating association between multiple products purchased by different customers. The results generated bunch of rules, based on those rules organization can take future decisions. The proposed model can work for the organizations to support their business development plans using hidden knowledge.

Key words:
Association mining; FP-Growth; rules generation; customer-oriented organization.

1. Introduction

Customers are an important factor that plays a vital role in the success of an organization. The use of information and communication technologies has improved the way to communicate with the customers. To deal with customers efficiently, organization are employing different strategies. Data mining is one of the common technologies proved to be very helpful in creating smart business environment. Therefore, this paper is dealing with one of the common data mining techniques known as "Association Mining" [1]. The main purpose of this technique is to create frequent patterns and to create association rules using the dataset of generated from multiple applications. The past researches highlighted that association mining has variety of implementation on different application such as large transactional databases [2], product sales pattern [3], customer's database [4] and others [5], [6].
This research aims to apply association mining technique on sales and customer's data. The implementation is useful for the retailers and the super stores to implement efficient business strategies for attaining more sales and customers. This techniques is useful for gaining competitive advantage for the business organization. There are many benefits suggested by the researchers of using this technique [7]. Furthermore, it can predict the organization in improving the number of sales by placing more purchased and dependent items together. In addition, from the customer's point of view, it will give them ease to pick the most sold items at one place, rather to do visit too many aisles [8].
The main problem statement considered in this research is how companies can improve the dealing with their customer using data mining techniques. Association mining is one of the common data mining techniques, which required enough amount of market data to predict and generate some rules. In the current scenario, the use of technology, social media and mobile communication are already playing supporting role for spreading the information. These technologies are generating large amount of dataset, the smart use of customer's related data can provide a new way for improving their business. Therefore, this research is focused on using those kind of sales data generated by the organization and apply association mining. The results can be beneficial for the companies to use the results and develop new plan for business growth and customer's satisfaction.
Moreover, the idea and output of this research can be interpreted in two ways; one from organization's and second from customer's point of view [4]. The research will be beneficial to both of the stakeholders directly. To put more stress, the algorithm's implementation will assist the organization to organize their product appropriately, whereas it will help the customers to get their product in a fastest way. Ultimately, the customer-oriented organizations are always planning to improve the business functionalities to get more customers as well as to improve the customer's satisfaction and loyalty [9]. The idea has been discussed and applied earlier as discussed in the next section.
The rest of the paper is organized as follows. The next section, describes the details and use of association mining algorithm. Section 3, discussed the proposed framework and its implementation steps. Section 4, gives the details of conducted experiment using data mining tool. After that, section 5 offers the detailed analysis on generated results.

Manuscript received May 5, 2020
Manuscript revised May 20, 2020

Finally, the paper ends by presenting conclusion and future work in the last section.

2. Frequent Patterns – General Overview

The idea of this research is based on one of the famous terminology known as "knowledge discovery" [10]. The term is useful and normally associated with all kind of data mining techniques. The most common approaches of data mining are known as classification, clustering, and association [11], [12], where each of them has different purposes and implications. Data mining can be implemented in different application such as hospitals, students, management systems, finance, and others [13]. The technique that focused in this research is known as association mining.

Association mining's main concept is same as data mining, which is used to extract the hidden information from the large dataset and generate association between the different attributes. The most common implementation of association mining is known as market-basket analysis. In which, the collection of data items (let suppose in the basket) selected to generate association between them [8]. The analysis perform on this basket by selecting one or more items and support by measuring the dependency of each item on others. Furthermore, the evaluation conducted using two measure criteria of association mining called "support" and "confidence".

The progression of association mining have discovered different algorithms. The most common algorithm used in association mining are (i) Frequent Pattern (FP) Growth, (ii) Apriori Algorithm, and (iii) AIS. The implementation of each algorithm is different from each other, but the purpose is same. According to the [14] Apriori algorithm has complex procedure of implementation than FP-Growth. Apriori uses multiple scan of the dataset to generate knowledge and extract new information, whereas FP-Growth can generate candidate dataset within two scans. In addition, the advantages and disadvantage of different algorithms associated with association mining [14] presented in figure 1. Based on the above discussion, the FP-Growth algorithm selected in this study.

FP-Growth has applied in various applications as found in previous work. A research presented to understand the consumer behavior using FP-Growth algorithm by generating association rules. The study highlighted the use of different data mining technique as well to understand the behavior of consumers [11]. Another research used this algorithm for intrusion detection system. After removing the noisy data and proper pre-processing techniques the algorithm generated 17 association rules using 6 attributes [15]. Several other researches also used the same approach for intrusion detection system [16], [17]. By generating the weighted support in two-dimensional table, the research proposed the new version of FP-Growth algorithm for fast searching process [18].

Association Rule Mining Algorithm	Advantages	Disadvantages
1. AIS	a) An estimation is used in the algorithm to prune those candidate itemsets that have no hope to be large. b) Memory management is used in AIS when memory is not enough.	a) Requires multiple scanning of the database . b) Only rules of the kind X∩Y=>Z can be generated. Rules of the kind X=>Y∩Z cannot be generated. c) Too many candidate itemsets that finally turn out to be small are generated.
2. Apriori	a) Easy implementation. b) It uses Apriori property for pruning therefore, itemsets left for further support checking remain less.	a) It explains only the presence or absence of an item in the database. b) It scans the complete database multiple times. c) It has a complex candidate generation process that uses most of the time, space and memory. d) It works well only for small databases with large support factor.
3. FP-Growth	a) It scales better than Apriori algorithm. b) It requires only two scans of the database without any candidate generation. c) It is not influenced by support factor.	a) The resulting FP-Tree is not unique for the same logical database. b) It cannot be used in interactive mining system. c) It cannot be used for incremental mining.

Fig. 1 An Overview of Association Mining Algorithms [14]

The implementation of FP-Growth is start by selecting frequent patterns from the large dataset. Here, frequent patterns means the more items repeated several times. Accordingly, the rules can be generated by using frequent dataset only, where the remaining data will be discarded. A research presented the modified version of this algorithm by using the strategy of linked list. The proposed framework in that research generates the hash table to link the frequent item set to process the data in a faster and shorter way [7]. Furthermore, the extraction of system failure mode and analysis of its impact presented by [19] using FP-Growth algorithm. They used the algorithm to extract the regular patterns to understand the reason behind each failure. Therefore, it can be suggested that association mining is commonly using for generating rules. The rule can be applied on a particular system for saving time and reducing errors in future transactions [20]–[22]. The following sub-section defined the step-wise approach to apply FP-Growth algorithm.

2.1 FP-Growth Algorithm

Firstly, the FP-Growth algorithm used to generate all frequent item set from the given data. The frequent item illustrate the behavior of the data in the form associating the listed attributed and their connection. The generation of frequent data is based on the criteria called "support" value. The next step after extracting the frequent items from the data set is to generate multiple rules. The algorithm then discard all irrelevant items from the dataset [23]. Furthermore, the results generates from the first phase are useable for rules generation. This phase requires a parameter named "confidence" value, which help to extract rules from the frequent items. Finally, this algorithm generate tree-like structure in finding out the association between the attributes. Therefore, all frequent items grow by using the same tree [24]. Researcher have used multiple tools to implement this algorithm such as oracle data miner [25], Weka [26], XLMiner [4], and Rapid Miner [27]. The step wise working of algorithm is presented below [28]:

Input: constructed FP-tree
Output: complete set of frequent patterns
Method: Call FP-growth (FP-tree, null).
procedure FP-growth (Tree, α)
{ 1. if Tree contains a single path P then
2. for each combination do generate pattern β α with support = minimum support of nodes in β.
3. Else For each header ai in the header of Tree
do {
4. Generate pattern β = ai α with support = ai.support;
5. Construct β.s conditional pattern base and then β.s conditional FP-tree Tree β
6. If Tree β = null
7. Then call FP-growth (Tree β, β)
}
}

3. Proposed Framework – An Overview

The study is focused on generating association rules from the customer related dataset to help the organizations in taking decisions and can use to forecast future sales. For this, as discussed earlier, the data mining technique, association mining considered for implementing the proposed model. Nowadays, organizations are dealing with large amount of data related to different stakeholders such as customers, employees, partners, and competitors [29]. The idea here is to use those data for the betterment of the organization. Therefore, this study conducted to do the experiment using data mining technique to assist the enterprises and to use the data in an efficient way. The association mining technique implementation on sales data applied by [3] generated different types of rules using support and confidence values. The purpose of the implementation discussed in that research is to generate frequent product purchased by the customers. As evident in that paper, the size of the dataset was not large enough, whereas the major requirement of the data mining implementation is large size of data.

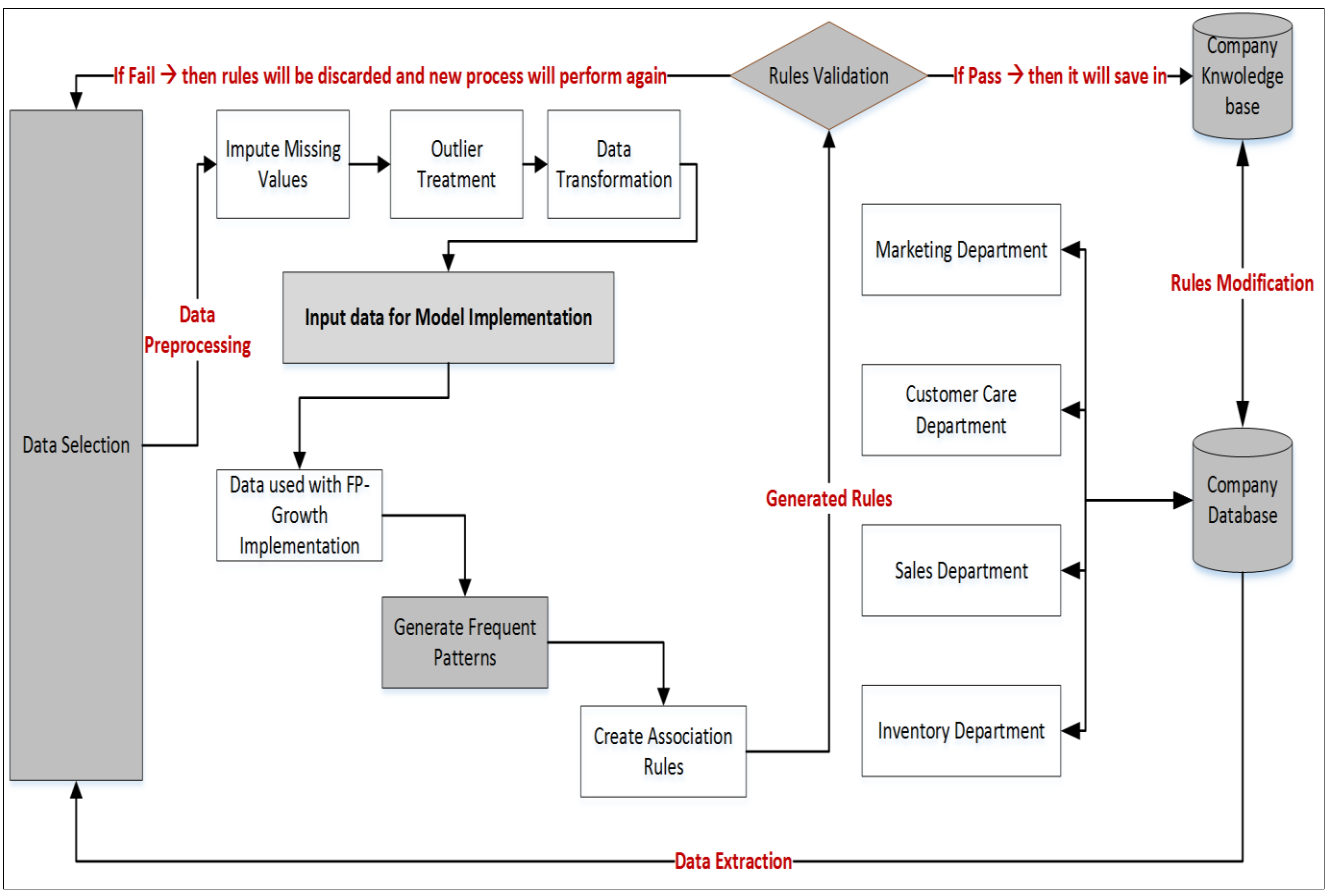


Fig. 2: The Proposed Framework for Association Mining

The experiment conducted in this research is based on the proposed framework shown in figure 2. As illustrated in the figure, the organization may have different department such as marketing, sales, inventory control, and customer care. The first step in this methodology is the selection of data, which extracted from the company's main database. The main database is connected with multiple departments and managing the transactional dataset. This study is related to deal with sales and customer related data, therefore, the extraction of the data will be limited to customers only.

Afterwards, the next step is to pre-process the data using different techniques commonly applied earlier in previous work [30], [31]. The first preprocessing method in the framework is "impute missing values". This method review the data and find out if there is any missing data, which was further filled using KNN method. Furthermore, "outlier treatment" executed for identification of data that is out of range. Finally, the last and important preprocessing method "data transformation" applied to convert the data set into proper format according to the experiment requirements. Further details of each method is discussed in later section 4.1.

After taking the proper steps for pre-processing the input file, now the data set is ready for implementation as shown in the framework. The dataset connected with FP-Growth algorithm to generate frequent patterns from the data file, based on the provided "support" value. In addition, the frequent patterns generation means to mine the data and extract the most recurrent products purchased by different customers in single or multiple transactions. At this stage, the data engine will only select those dataset that are frequently occurred. The remaining data will be discarded from the experiment. This is the major benefits of applying FP-Growth algorithm, that step by step the unnecessary data will be removed from the experiment, which will ultimately improve the performance of overall process.

Finally, as illustrated from the framework, the next step is to execute the frequent pattern and generate rules. This steps is critical as this will generate the final output after all processing applied on the selected dataset. At this point, the point to be remembered is that if any of the previous step encountered any mistake it will impact upon the generated rules. The rules generation required a parameter called "confidence". This value is required to know the strength of selected rules. The rules need to be verified before injecting to the company's knowledge base. Therefore, the generated rules will process from the validation phase (will be performed manually) to check its authenticity and applicability. If the decision will be in favor of rules, then it will be inserted in the company's knowledge base. If not, then rules will be discarded and can generate again with new process. As can be seen from the framework, the knowledge base is connected with company's main database. Company's managers and decision makers can access this knowledge base while building new business strategy, which is main purpose of proposing this framework. Using of this framework, benefits will undergone in the favor of customers and organization as well.

To conclude this section that explained the conceptual implementation of the framework. Whereas, the subsequent section, articulates the practical implementation of the proposed model. Before starting of the main experiment, the data description and preprocessing steps implementation presented to provide comprehensive overview and familiarity of the selected data.

4. Implementation of Proposed Framework

Organizations are generating large amount of data every day, while dealing with customers and other stakeholders. The framework proposed in this study is to process the organizational data using data mining tasks to extract hidden knowledge from the data. Therefore, the conceptual model presented in the previous section explained all methodological step. Whereas, this section explains the practical implementation of the framework using real world data. Before moving to the experiment, the subsequent section describes the description and background of the dataset. Furthermore, this section describes the preprocessing steps used before the main experiment.

4.1. Data Description and Preprocessing

This study used a customer based data having set of relational attributes. The dataset is publicly available by Kaggle repository, one of the famous dataset providing different kinds of datasets for academician, competitions, and for research purposes [32]. The dataset belongs to one of the famous grocery super store chain operating in USA and Canada. The name of the dataset is "Instacart Market Bastet Analysis" as mentioned by the publisher Kaggle [33]. Overall, the dataset providing the order history of anonymous customers purchased different products from the store. The dataset contains a data sample of around 3 million requests by almost 200,000 Instacart's consumers. There were total 7 different files provided in the dataset. The list of the files and their description is shown in the Table 1.

As shown in the table, there were several files in which the data was organized. It also highlights the overall description of the dataset. Regarding the data preprocessing, the first step was to review the data file, and extract the attributes which are related with this experiment. Therefore, the attributes are selected in this experiment are (i) order ID, (ii) product list (iii) purchasing history. The data was selected and organized according to the requirements of FP-Growth algorithm. In the beginning the data was not fully ready for implementation. Order ID, list of products, and purchasing history was written in the different files, which integrated in single input data file as can be seen in figure 3.

Table 1: List of Files in the Dataset

S.No.	File Name	File Description
1	Aisles.csv	The main sections list in the superstore
2	Departments.csv	Departmental wise distribution
3	Order_products_prior.csv	List of orders
4	order_products_train.csv	List of orders, training data
5	Orders.csv	Combined order list
6	Products.csv	Lis of products
7	sample_submission.csv	Sample distribution

The data preprocessing and implementation performed using an open source tool called Rapid Miner [24]. The tool has incorporated with all common operators required for association mining. Firstly, the selected data input file connected with the operator "select attribute" which is used for choosing number of required attributes in this process [34]. All the product transactions selected, while order ID excluded as was not required for this algorithm. Secondly, the selected attributes forwarded to the node of "impute missing values" as shown in figure 3. The purpose of this step is to reveal any missing values in the dataset. We used impute missing values using KNN approach to fill the missing values using nearest neighbor method [35]. Furthermore, the next step is to continue the process by forwarding dataset and connect it to the "outlier detection" operator. Outlier detection is the method used to extract data values that are out of the scope or range from the overall data [36]. It will help to validate the ranges of data values within the main boundary of the data. It will not change any data value, either it will just provide the reservations, which further depend on the researcher to remove or keep the particular value.

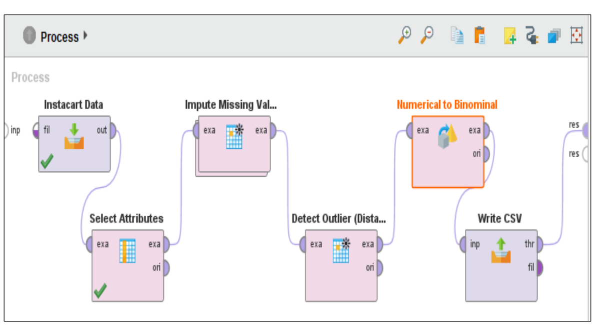


Fig. 3 An Overview of Association Mining Algorithms

To take a step forward in this phase, after processing and checking the outliers in the dataset, it connects with the next operator called "numerical to binominal" as evident from the above figure. This operator used for converting the data type from numbers to binominal values as defined in Rapid Miner [37]. The purpose of using this operator as it is the main requirement of FP-Growth in Rapid Miner that all data values ought to be binominal. Lastly, the finishing step of this phase is to export the final output in a "comma separated values" (CSV) format, so it can be used in the main experiment. For this, Rapid Miner provides an operator known as "Write CSV", to export the data in CSV format [38]. To conclude this section, all necessary and essential steps for preparing the data file applied, to make it useful for FP-Growth algorithm has been applied. The subsequent section explained about the implementation of the associating mining model and rules generation.

4.2. Conducting Experiment using Rapid Miner

The main experiment starts by importing the data set prepared in last phase. The step-wise approach of data preprocessing is discussed in previous section to understand the efforts applied for making dataset applicable according to the requirements of Rapid Miner. The preprocessed data imported using operator called "Read CSV" as the first step of the experiment as displayed in figure 4. As discussed earlier, the experiment used FP-Growth algorithm, which is commonly used for association mining. The purpose behind selecting this algorithm is a better, fasters, and intelligent implementation than other two association mining techniques as mentioned in figure 1. Moreover, the dataset selected in this study is famous for market basket analysis, which is the highly applicable data for the validation of proposed framework presented in figure 2.

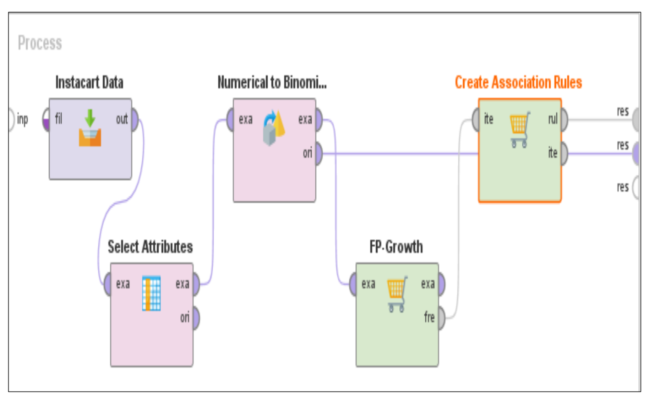


Fig. 4 FP-Growth Implementation

Association mining is mainly known for unsupervised learning approach [39], which implies that there is no class/category defined in the dataset. The dataset has all attributes as regular. After importing the data file it is asked for selecting the particular list of attributes specially required for association mining. After that, the data has been sent to the node associated with FP-Growth algorithm. The essential parameter required for running this algorithm is the value of "support". The minimum "support" value is described to search the probability of occurring any products with others. The main purpose of minimum support value is to exclude the products that are not frequently found in the basket. The two benefits of minimum support are; (i) the exclusion of infrequent products from the dataset, (ii) which will make the results statistically significance [40]. If the minimum support is high, the chances of frequent items selection will be low. On the other side, if the minimum support is low, the list of frequent items will be high, as the criteria for selecting the products is very minimal. In this experiment, the researcher tried to use different minimum support values, to validate the impact of minimum support on the rules generation. Therefore, the experiment performed 10 times, the details and impact of minimum support is discussed in the next section (result and discussion).

At this stage, the data has compressed as FP-Growth excluded all infrequent products from dataset. The list of frequent item will be executed to the next node called "create association rules" as shown in figure 4. This node is the last step in this experiment, which takes frequent items as an input and extract rules as an output. This operator further required next criteria for the rules generation known as "confident" value, which assigned as fixed as "0.5%" in this experiment. The value of minimum confidence identifies the selection of the rules based on the given value. The rule selection work based on if/then statement. For example, if the product A purchased, how many chances of product B will be purchased. As the support values measure the frequency of the product in the dataset, the confidence values evaluate the number of time if/then statement is found valid.

5. Results and Discussion

The experiment conduced successfully, and the extracted rules are shown in figure 5. The result is showing the generation of rules after first experiment, where the minimum support (0.2%) and minimum confidence (0.5%) entered. Based on the given values total 54 rules were generated, whereas sample of rules is showing in figure 5. The rules can be interpreted using if/then statement, as the example given below.

First Rule: if the "Mint Chocolate Flavored Syrup" purchased by a customer, then there are 61% chances that the same customer will also purchase "Organic Turkey Burgers".

Second Rule: If the "All-Seasons Salt" purchased by a customer, then there are 62% chances that the same customer will also purchase "Chocolate Sandwich Cookies".

[Mint Chocolate Flavored Syrup] --> [Organic Turkey Burgers] (confidence: 0.616)
[All-Seasons Salt] --> [Chocolate Sandwich Cookies] (confidence: 0.621)
[Dry Nose Oil, All-Seasons Salt] --> [Chocolate Sandwich Cookies] (confidence: 0.639)
[Fresh Breath Oral Rinse Mild Mint] --> [Organic Turkey Burgers] (confidence: 0.651)
[Overnight Diapers Size 6] --> [Organic Turkey Burgers] (confidence: 0.658)
[Robust Golden Unsweetened Oolong Tea] --> [All-Seasons Salt] (confidence: 0.658)
[Robust Golden Unsweetened Oolong Tea] --> [Chocolate Sandwich Cookies] (confidence: 0.678)
[Dry Nose Oil, Chocolate Sandwich Cookies] --> [All-Seasons Salt] (confidence: 0.688)
[Saline Nasal Mist] --> [Organic Turkey Burgers] (confidence: 0.694)
[Sparkling Orange Juice & Prickly Pear Beverage] --> [Organic Turkey Burgers] (confidence: 0.727)
[Small & Medium Dental Dog Treats] --> [Organic Turkey Burgers] (confidence: 0.747)
[Dry Nose Oil, Robust Golden Unsweetened Oolong Tea] --> [All-Seasons Salt] (confidence: 0.752)
[Light Strawberry Blueberry Yogurt] --> [Organic Turkey Burgers] (confidence: 0.764)
[Robust Golden Unsweetened Oolong Tea] --> [Dry Nose Oil] (confidence: 0.779)
[All-Seasons Salt] --> [Dry Nose Oil] (confidence: 0.796)
[Chocolate Sandwich Cookies, All-Seasons Salt] --> [Dry Nose Oil] (confidence: 0.820)
[All-Seasons Salt, Robust Golden Unsweetened Oolong Tea] --> [Dry Nose Oil] (confidence: 0.890)

Fig. 5 A Sample of Extracted Rules

The results suggested the rules based on the given data and product list. As the list of the products were selected randomly, there may be no proper association between the products in the generated rules. But the rules are only based on purchased history found in the data and the given minimum support and confidence values. The completed list of rules shown in figure 6. The figure is known as "circle" figure to show the association between all rules generated in this experiment. As can be seen that due to low minimum support the generated list are more.

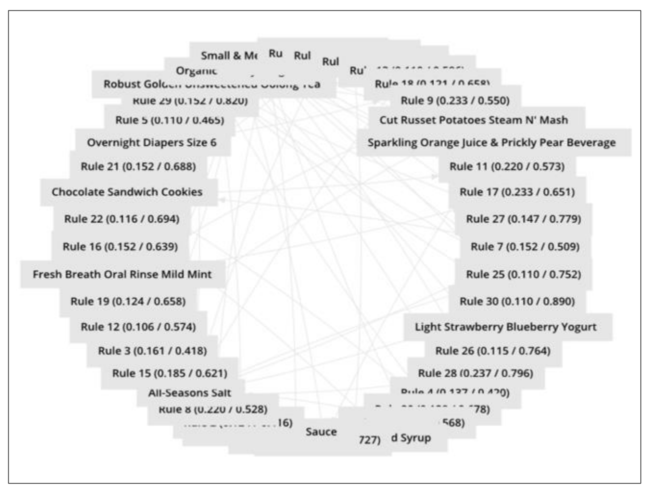


Fig. 6 Rules Circle Graph

To generate the variety of results and know about the impact of different minimum support values. The process executed 10 consecutive time by changing the minimum support values between 0.10% to 0.95%, whereas the minimum confidence value were kept fixed as 0.5%. The result of this process is shown in figure 7, highlighting the variety of results and association between minimum support value and

number of rules. It is proved as showing in the figure that number of rules are increasing as reducing minimum support and vice versa. It has been mentioned before that decreasing minimum support means infrequent items will be selected, which will automatically amplify number of rules. The graph also disclose that when minimum support given as 0.95%, there is “0” number of rules generated.
This research proposed a model for the customer based organization to do analysis on the research theme “market-basket analysis”. The model has implemented using association mining technique to analyze the real world data. For the experiment, a publicly available customer’s data selected that has the long purchasing history of various customers. The purpose of experiment to suggest some rules and patterns to the organization to create new business strategy and development plan. The generated rules should be considered as a recommended set of instructions, which can be undertaken during decision making process. There are several points proposed for different scenarios based on the major findings of this study as follows:

- The least frequent product: the item can be controlled by inventory management system.
- The most frequent product: this product should be provided and available in large quantity.
- The most associated product: the combination of products should be placed together.
- The low confident rule: the organization can discard those rules due to less efficiency.
- The high confident rule: company should put more emphasizes on these products in planning of the future business strategy.

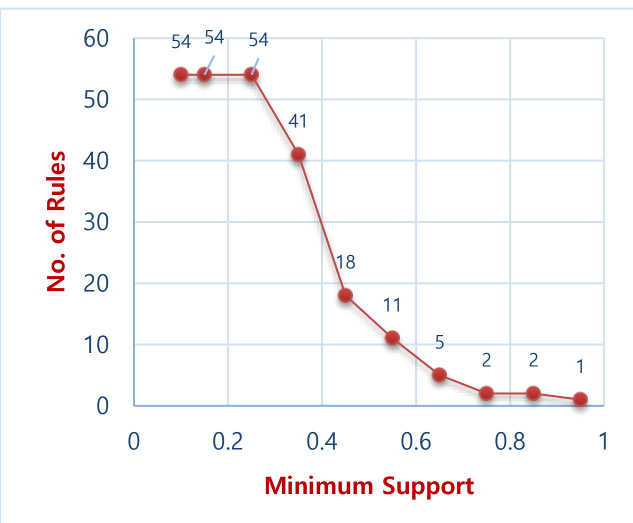


Fig. 7 Association between Minimum Support and Rules Generation

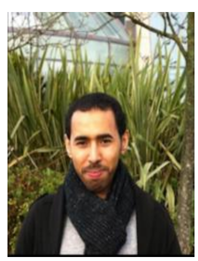
6. Conclusion and Future Work

The research focused on providing support to the customer-oriented organization that are dealing with customer on daily bases. The study proposed a framework to analyze the market-basket analysis. The purpose was to assess the data using customer’s purchasing history, buying multiple product in one or multiple orders. For model implementation, association mining technique applied for generating rules based on the real world data. The results generated 54 rules by giving minimum support and minimum confidence as 0.2% and 0.5% respectively. The extracted rules can be helpful for the organizations to build new business strategy. As this hidden patterns were not readable from the real data, whereas proposed rules can be viable for the organization in taking future decisions.

References

[1] S. Lessanibahri, L. Gastaldi, and C. G. Fernández, “A novel pruning algorithm for mining long and maximum length frequent itemsets,” Expert Syst. Appl., vol. 142, p. 113004, 2020.
[2] D. Sathyavani and D. Sharmila, “An improved memory adaptive up-growth to mine high utility itemsets from large transaction databases,” J. Ambient Intell. Humaniz. Comput., pp. 1–10, 2020.
[3] P. Prasad, “Using association rule mining for extracting product sales patterns in retail store transactions,” 2011.
[4] A. M. Khattak, A. M. Khan, S. Lee, and Y.-K. Lee, “Analyzing association rule mining and clustering on sales day data with xlminer and weka,” Int. J. Database Theory Appl., vol. 3, no. 1, pp. 13–22, 2010.
[5] Y.-D. Shen, Z. Zhang, and Q. Yang, “Objective-oriented utility-based association mining,” in 2002 IEEE International Conference on Data Mining, 2002. Proceedings., 2002, pp. 426–433.
[6] J. Ragaventhiran and M. K. Kavithadevi, “Map-optimize-reduce: CAN tree assisted FP-growth algorithm for clusters based FP mining on Hadoop,” Futur. Gener. Comput. Syst., vol. 103, pp. 111–122, 2020.
[7] S. S. Maw, “An Improvement of FP-Growth Mining Algorithm Using Linked list.” 2020 IEEE Conference on Computer Applications(ICCA), Yangon, Myanmar, 2020, pp. 1-4
[8] R. Agrawal, T. Imieliński, and A. Swami, “Mining association rules between sets of items in large databases,” in Proceedings of the 1993 ACM SIGMOD international conference on Management of data, 1993, pp. 207–216.
[9] M. Gammelgård, M. Ekstedt, and P. Gustafsson, “A Categorization of Benefits From IS / IT Investments,” Proc. 13th Eur. Conf. Inf. Technol. Eval., no. October, pp. 1–11, 2001.
[10] C. Romero and S. Ventura, “Data mining in education,” Data Min. Knowl. Discov., vol. 3, no. 1, 2013.
[11] A. Raorane and R. V Kulkarni, “Data mining techniques: A source for consumer behavior analysis,” arXiv Prepr. arXiv1109.1202, 2011.
[12] F. Saleem and A. Malibari, “DATA MINING COURSE IN INFORMATION SYSTEM DEPARTMENT–CASE

STUDY OF KING ABDULAZIZ UNIVERSITY," in 3rd International Congress on Engineering Education, 2011.
[13] J. Han, J. Pei, and M. Kamber, Data mining: concepts and techniques, 3rd ed. Elsevier, 2012.
[14] N. Sharma and C. K. Verma, "Association rule mining: An overview," Int. J. Comput. Sci. Commun., vol. 5, pp. 10–15, 2014.
[15] Y. Sun and X. Chen, "An Improved Frequent Pattern Growth Based Approach to Intrusion Detection System Alert Aggregation," in Journal of Physics: Conference Series, 2020, vol. 1437, no. 1, p. 12070.
[16] M. Fan and C. Li, "Mining frequent patterns in an fp-tree without conditional fp-tree generation," J. Comput. Res. Dev., vol. 40, no. 8, pp. 1216–1222, 2003.
[17] F. S. Tsai, "Network intrusion detection using association rules," Int. J. Recent Trends Eng., vol. 2, no. 2, p. 202, 2009.
[18] Y. Li and S. Yin, "Mining algorithm for weighted FP-tree frequent item sets based on two-dimensional table," in Journal of Physics: Conference Series, 2020, vol. 1453, p. 12002.
[19] Z. Xu, Y. Dang, P. Munro, and Y. Wang, "A data-driven approach for constructing the component-failure mode matrix for FMEA," J. Intell. Manuf., vol. 31, no. 1, pp. 249–265, 2020.
[20] A. A. Nanavati, K. P. Chitrapura, S. Joshi, and R. Krishnapuram, "Mining generalised disjunctive association rules," in Proceedings of the tenth international conference on Information and knowledge management, 2001, pp. 482–489.
[21] H. Wang and X. Liu, "The research of improved association rules mining apriori algorithm," in 2011 eighth international conference on fuzzy systems and knowledge discovery (FSKD), 2011, vol. 2, pp. 961–964.
[22] X.-W. Liu and P.-L. He, "The research of improved association rules mining Apriori algorithm," in Proceedings of 2004 International Conference on Machine Learning and Cybernetics (IEEE Cat. No. 04EX826), 2004, vol. 3, pp. 1577–1579.
[23] D. Hunyadi, "Performance comparison of Apriori and FP-Growth algorithms in generating association rules," in Proceedings of the European computing conference, 2011, pp. 376–381.
[24] R. M. Team, "Rapid Miner Documentation." [Online]. Available: https://docs.rapidminer.com/latest/studio/operators/modeling/associations/create_association_rules.html. [Accessed: 10-Mar-2019].
[25] Oracle, "Oracle Data Miner." [Online]. Available: https://www.oracle.com/technetwork/database/options/odm/dataminerworkflow-168677.html. [Accessed: 04-Oct-2018].
[26] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten, "The WEKA data mining software: an update," ACM SIGKDD Explor. Newsl., vol. 11, no. 1, pp. 10–18, 2009.
[27] R. M. Team, "Rapid Miner." [Online]. Available: https://rapidminer.com/. [Accessed: 10-Dec-2019].
[28] J. Han, H. Pei, and Y. Yin., "Mining Frequent Patterns without Candidate Generation," in Proc. Conf. on the Management of Data (SIGMOD'00, Dallas, TX). ACM Press, New York, NY, USA 2000., 2000.
[29] E. K. Hong, "Information technology strategic planning," IT Prof., vol. 11, no. 6, pp. 8–15, 2009.
[30] J. Liang, J. Yang, Y. Wu, C. Li, and L. Zheng, "Big data application in education: Dropout prediction in edx MOOCs," in Proceedings - 2016 IEEE 2nd International Conference on Multimedia Big Data, BigMM 2016, 2016, pp. 440–443.
[31] A. A.-M. AL-Ghamdi and F. Saleem, "Enterprise application integration as a middleware: Modification in data & process layer," in Proceedings of 2014 Science and Information Conference, SAI 2014, 2014, pp. 698–701.
[32] Kaggle, "Kaggle Repository." [Online]. Available: https://www.kaggle.com/.
[33] Instacart, "Instacart Market Basket Data," Kaggle, 2017. [Online]. Available: https://www.kaggle.com/c/instacart-market-basket-analysis/data.
[34] R. Miner, "Select Attributes." [Online]. Available: https://docs.rapidminer.com/9.5/studio/operators/blending/attributes/selection/select_attributes.html. [Accessed: 15-Mar-2020].
[35] R. Miner, "Impute Missing Values." [Online]. Available: https://docs.rapidminer.com/latest/studio/operators/cleansing/missing/impute_missing_values.html. [Accessed: 05-May-2020].
[36] R. Miner, "Outlier Detection Using Rapid Miner." [Online]. Available: https://docs.rapidminer.com/latest/studio/operators/cleansing/outliers/detect_outlier_distances.html. [Accessed: 20-Apr-2019].
[37] R. S. Core, "Numerical to Binominal," Rapid Miner Documentation. [Online]. Available: https://docs.rapidminer.com/latest/studio/operators/blending/attributes/types/numerical_to_binominal.html. [Accessed: 01-Apr-2020].
[38] R. Miner, "Write CSV." [Online]. Available: https://docs.rapidminer.com/latest/studio/operators/data_access/files/write/write_csv.html. [Accessed: 25-Mar-2020].
[39] M. J. Berry and G. Linoff, Data mining techniques: for marketing, sales, and customer support. John Wiley & Sons, Inc., 1997.
[40] R. Miner, "FP-Growth." [Online]. Available: https://docs.rapidminer.com/latest/studio/operators/modeling/associations/fp_growth.html. [Accessed: 08-May-2020].



Khaled H. Alyoubi received his Ph.D. degree in Database Query Optimization (Computer Science) from Birkbeck University of London, UK. His academic rank is an Assistant Professor at King Abdulaziz University, Jeddah, Saudi Arabia. His research interests include Data Management, Data Sciences, Data Mining, and Information Retrieval & Extraction.