JCST Journal of Data Acquisition and Processing

ISSN: 1004-9037
https://sjcjycl.cn/
DOI: 10.5281/zenodo.755230

K-MEANS CLUSTERING APPROACH BASED INTELLIGENT CUSTOMER SEGMENTATION TO INCREASE SALES USING CUSTOMER PURCHASE BEHAVIOR DATA

[1]Dr. S.Srinivas, [2]Mohammad Rezaul

[1]Associate Professor, Department of Computer Science and Engineering, Holy Mary Institue Of Technology and Science, Bogaram (V), Keesara (M), Hyderabad, Telangana, India

[2]M.Tech Scholar, Department of Computer Science and Engineering, Holy Mary Institue Of Technology and Science, Bogaram (V), Keesara (M), Hyderabad, Telangana, India

ABSTRACT: E-commerce system has become more popular and implemented in almost all business areas. E-commerce system is a platform for marketing and promoting the products to customer through online. Customer segmentation is known as a process of dividing the customers into groups which shares similar characteristics. The purpose of customer segmentation is to determine how to deal with customers in each category in order to increase the profit of each customer to the business. Using the large amount of data available on customers and potential customers, a customer segmentation analysis allows marketers to identify discrete groups of customers with a high degree of accuracy based on demographic, behavioral and other indicators. This paper presents, K-Means Clustering approach based Intelligent Customer Segmentation to increase sales using Customer Purchase Behavior Data. To process the collected data and segment the customers, learning algorithm is used which is known as K-Means clustering. We evaluate the prediction model through a set of evaluation metrics, Mean Squared Error (MSE) and coefficient of determination (R2). K-Means clustering gives better performance for a large number of observations.

KEYWORDS: Purchase behavior, K-Means clustering, customer segmentation, Cluster analysis.

I. INTRODUCTION

Over the years, increased competition among businesses and the availability of large-scale historical data has resulted in widespread use of data mining techniques to find critical and strategic information that is hidden in organizations' information [1]. Data mining is the process of extracting logical information from a dataset and presenting it in a human-accessible manner for decision support. Modern intelligent information technologies allow solving many data analysis tasks to improve business efficiency and support decision making [2]. Using intelligent technologies, you can analyze large amounts of data from various sources, identify patterns and predict economic performance. These tasks include analysis and forecasting of foreign economic activity, inventory management, customer basket analysis, and others [3].

When customers receive too much information or unwanted details which is not related to their regular purchase or their interest on the products, it can cause confusion on deciding their needs. This might lead their customers to give up on purchasing the items they required and effect the business to lose their potential customers. The clustering analysis will help to categorize the E-commerce customer according to their spending habit, purchase habit or specific product or brand the customers interested in [4]. One of the most useful techniques in business analytics for the analysis of consumer behavior and categorization is customer segmentation. By using clustering techniques, customers with similar means, end and behavior are grouped together into homogeneous clusters [5].

Customer Segmentation helps organizations in identifying or revealing distinct groups of customers who think and function differently and follow varied approaches in their spending and purchasing habits [6]. Clustering techniques reveal internally homogeneous and externally heterogeneous groups [7]. Customers vary in terms of behavior, needs, wants and characteristics and the main goal of clustering techniques is to identify different customer types and segment the customer base into clusters of similar profiles so that the process of target marketing can be executed more efficiently.

Highlight Clustering is a statistical technique much similar to classification. It sorts raw data into meaningful clusters and groups of relatively homogeneous observations. The objects of a particular cluster have similar characteristics and properties but differ with those of other clusters [8]. The grouping is accomplished by finding similarities among data according to characteristics found in raw data. The main objective was to find optimum number of clusters. There are two basic types of clustering methods, hierarchical and non-hierarchical. Clustering process is not one time task but is continuous and an iterative process of knowledge discovery from huge quantities of raw and unorganized data. For a particular classification problem, an appropriate clustering algorithm and parameters must be selected for obtaining optimum results. Clustering is a type of explorative data mining used in many application oriented areas such as machine learning, classification and pattern recognition [9]. In recent times, data mining is gaining much faster momentum for knowledge based services such as distributed and grid computing.

A customer segmentation strategy allows firms to target particular groups of consumers, resulting in more efficient marketing resource allocation and greater potential for cross and up-selling [10]. It's easier for firms to create unique offers to entice customers to spend more when they deliver customized communications to a group of customers as part of a marketing mix tailored to their requirements [11]. Consumer segmentation may help with customer loyalty and retention by improving customer service. Because of their individualized character, marketing materials that employ customer segmentation are more valued and appreciated by the consumer who gets them than impersonal brand communications that ignore purchase history or any type of customer relationship [12].

Customer segmentation has been demonstrated to benefit from clustering. Clustering is a sort of unsupervised learning that allows us to locate clusters in unlabeled datasets. Remaining paper is organized as follows: Section II elaborates the Literature survey, Section III explains the described methodology, Section IV describes results and discussions and finally paper is concludes with Section V.

II. LITERATURE SURVEY

X. Chen, W. Sun, B. Wang, Z. Li, X. Wang and Y. Ye, et. al. [13] proposes a PurTree subspace metric to measure the dissimilarity between two customers represented by two purchase trees, in which a set of level weights are introduced to distinguish the importance of different tree levels and a set of sparse node weights are introduced to distinguish the importance of different tree nodes in a purchase tree. Two-level subspace weighting spectral clustering (TSW) was compared with six clustering algorithms on ten benchmark data sets and the experimental results show the superiority of the new method. X. Chen, Y. Fang, M. Yang, F. Nie, Z. Zhao and J. Z. Huang, et. al. [14] propose the “personalized product tree”, named purchase tree, to represent a customer's transaction records. So the customers' transaction data set can be compressed into a set of purchase trees. We propose a partitional clustering algorithm, named PurTreeClust, for fast clustering of purchase trees. Finally, the clustering results are obtained by assigning each customer to the nearest representative. A series of experiments were conducted on ten real-life transaction data sets, and experimental results show the superior performance of the proposed method.

Maryani, D. Riana, R. D. Astuti, A. Ishaq, Sutrisno and E. A. Pratama, et. al. [15] aims to perform customer segmentation on Nine Reload Credit by utilizing data mining process based on RFM model and by using techniques Clustering. The algorithm used for cluster formation is K-Means algorithm. Furthermore, we analyzed cluster by using K-Means algorithm with the result of 63 Customers in Cluster 1 and 39 Customers in Cluster 2. The result of this research can be used by company to know customer category, and then the company will know how to maintain the customer owned.

R. Taniguchi, Y. Ohtaka and S. Morishita, et. al. [16] modeled Purchase behavior of customers in a store by Cellular Automata (CA) and a simulation. The transaction data of each customer, which often refers as "Point of Sales (POS)" data, has been recorded, and these data allow retailors to understand favorite or attractive items for each costumer. In CA algorithm, local neighbor rules are defined as interaction of elements which compose the phenomena. We modeled the movement of customers considering two ways of purchasing, the planned purchase and the unplanned purchase. As a result, the customer walked around in a store depending on the layout of items. J. Zhu, H. Wang, M. Zhu, B. K. Tsou and M. Ma, et. al. [17] studies aspect-based opinion polling from unlabeled free-form textual customer reviews without requiring customers to answer any questions. First, a multi-aspect bootstrapping method is proposed to learn aspect-related terms of each aspect that are used for aspect identification. Second, an aspect-based segmentation model is proposed to segment a multi-aspect sentence into multiple single-aspect units as basic units for opinion polling. Finally, an aspect-based opinion polling algorithm is presented in detail. Experiments on real Chinese restaurant reviews demonstrated that our approach can achieve 75.5 percent accuracy in aspect-based opinion polling tasks.

X. Zhang, G. Feng and H. Hui, et. al. [18] explores the unique features of the customer relationship management (CRM) system in Telecom industry and presents a customer-churn model based on customer segmentation. First, the improved Fuzzy C-means clustering algorithm is used to segment customer and conclude high value customer group characteristics.

Second, using the history data and SAS Enterprise Miner builds a prediction model of customer-churn based on SAS data mining technology. Last but not least, the result of customer segmentation is applied to customer-churn model and gotten accuracy list of lost customer. Experiment proves that this method can obtain a satisfactory result of customer-churn. T. Jiang and A. Tuzhilin, et. al. [19] present a direct grouping-based approach to computing customer segments that groups customers not based on computed statistics, but in terms of optimally combining transactional data of several customers to build a data mining model of customer behavior for each group. Then, building customer segments becomes a combinatorial optimization problem of finding the best partitioning of the customer base into disjoint groups. It is shown that the best direct grouping method significantly dominates the statistics-based and one-to-one approaches across most of the experimental conditions, while still being computationally tractable.

T. Iwata, K. Saito and T. Yamada, et. al. [20] present a novel recommendation method that maximizes the probability of the lifetime value (LTV) being improved, which can apply to both measured and subscription services. We infer a user's interests from the purchase history based on maximum entropy models and use the interests to improve recommendation. Since a higher LTV is the result of greater user satisfaction, our method benefits users as well as online stores. We evaluate our method using two sets of real log data for measured and subscription services.

III. INTELLIGENT CUSTOMER SEGMENTATION

The work flow of K-Means Clustering approach based Intelligent Customer Segmentation to increase sales using Customer Purchase Behavior Data is represented in below Fig. 1.

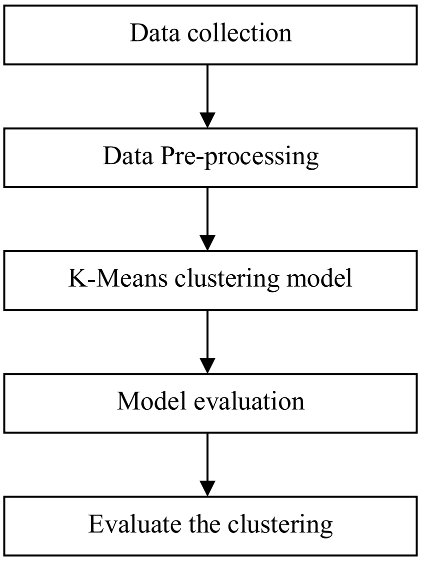

Fig. 1: WORK FLOW OF INTELLIGENT CUSTOMER SEGMENTATION MODEL

The first step is obtaining data regarding E-commerce purchase and check if the data obtained has clustering trend. The dataset used for this research is Malaysia’s E-commerce dataset from MDEC repository for machine learning based on customers purchasing behavioral. There in this project the dataset contains the E-Commerce behavior data from multi category store which comprises 285,000,000 customer purchase history.

The response of the questionnaire will be examined, and the analysis will be created based on the collected quantitative data. The questions generated for the questionnaire are known as the foundation of the research as it will result a statistical analysis using the collected data. For the data set chosen in this research pre-processing techniques will be done to check missing values, noisy data, and other inconsistencies before executing it to the algorithm. RStudio and Microsoft Office Excel are the tools that will be used to perform the data pre-processing and using K-Means algorithm.

The next step is to apply the K-Means algorithm to the dataset and divide the dataset into several clusters such as c1, c2, c3...cn. K-Means clustering is known as an unsupervised learning which is used to solve problems related to clustering. K-Means clustering is a process of classifying the dataset into certain number of clusters where each cluster will be defined with k centers. The k-centres should be strategically placed since various locations produce various results. The result will be better if each cluster is as far as possible. The ideal number of clusters k that leads to the maximum distance which can be calculated from the dataset. One of the ways to choose the optimum number of clusters is elbow method. A practical technique would be to compare the results of numerous runs with multiple k and select the best one based on a predetermined criterion. In general, a high k reduces error but raises the likelihood of overfitting.

Then, summary each class into one or more rules according to the characteristics of each class based on the data object characteristics in the class and analyze the clustering outcomes.

We evaluate the described models through a set of evaluation metrics, Mean Squared Error (MSE) and coefficient of determination (R^2). Finally, comparisons are made to show the robustness of the proposed method in relation to the previous methods, Means-Multiple Linear Regression (MLR) and Support Vector Regression (SVR). If the clustering result is extremely dependable, it is confirmed for the actual application. If not, the clustering analysis is repeated using different clustering techniques. As conclusion, K-Means algorithm helps to increase the quality of customer data clustering and increase the effectiveness of E-commerce activities enterprises.

IV. RESULT ANALYSIS

The clusters generated by k-means were used for customers clustering based on the inputs ratings. The dataset used for this research is Malaysia's E-commerce dataset from MDEC repository for machine learning based on customers purchasing behavioral. The data in each cluster were divided into training and test sets. The training sets (60% of the data) were used to construct the prediction models, and test sets (40% of the data) were used to evaluate the prediction models for their performance. We evaluate the described model through a set of evaluation metrics, Mean Squared Error (MSE) and Coefficient of determination (R^2). Data visualization and Dashboards are used for visualizing the results of customer segmentation to the end users.

Fig. 2 shows the overview of the dashboard. The diagram below illustrates the overview of all the e category code, brand, price, product details and event type. This page shows the overall distribution and analysis on the data for all the attributes or variables analyzed.

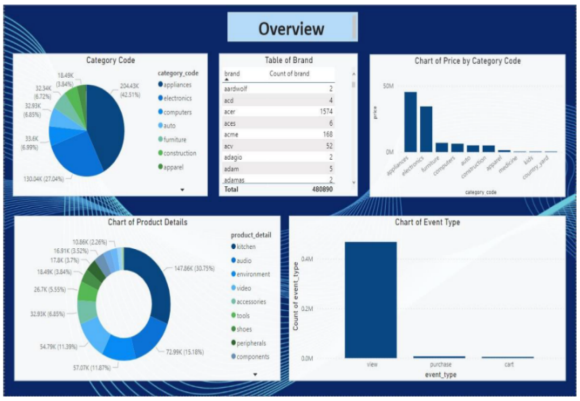

Table of Brand

brand	Count of brand
aardwolf	2
acd	4
acer	1574
aces	6
acme	168
acv	52
adagio	2
adam	5
adamas	2
Total	480890

Fig. 2: OVERVIEW OF DASHBOARD

Fig. 3 below shows the product brand page of the customer segmentation dashboard. This page explains on the brand preferred by the E-commerce users. The visualizations show the most popular product brand to the least popular product. The dashboard also shows the table that display the number of product brands each user viewed, purchased, or added to cart.

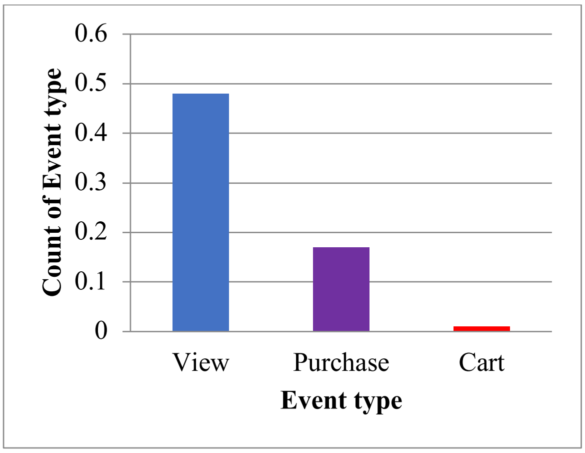

Fig. 3: EVENT TYPE DASHBOARD

In statistics, the mean squared error (MSE) of an estimator (of a procedure for estimating an unobserved quantity) measures the average of the squares of the errors that is, the average squared difference between the estimated values and the actual value. MSE is a risk function, corresponding to the expected value of the squared error loss. The fact that MSE is almost always strictly positive (and not zero) is because of randomness or because the estimator does not account for information that could produce a more accurate estimate.

The coefficient of determination (R^2) measures how well a statistical model predicts an outcome. The outcome is represented by the model's dependent variable. The lowest possible value of R^2 is 0 and the highest possible value is 1.

Comparisons are made to show the robustness of the proposed method in relation to the previous methods, Means-Multiple Linear Regression (MLR) and Support Vector Regression (SVR).

Table 1: COMPARATIVE PERFORMANCE ANALYSIS

Method	**Mean Squared Error (MSE)**	**coefficient of determination (R^2)**
Support Vector Regression (SVR)	0.16	0.86
Means-Multiple Linear Regression (MLR)	0.18	0.88
K-Means Clustering	0.09	0.98

The Fig. 4 and Fig. 5 show the graphical representation of Mean Squared Error (MSE) and coefficient of determination (R2) parameters for described K-Means Clustering approach, Means-Multiple Linear Regression (MLR) and Support Vector Regression (SVR) models.

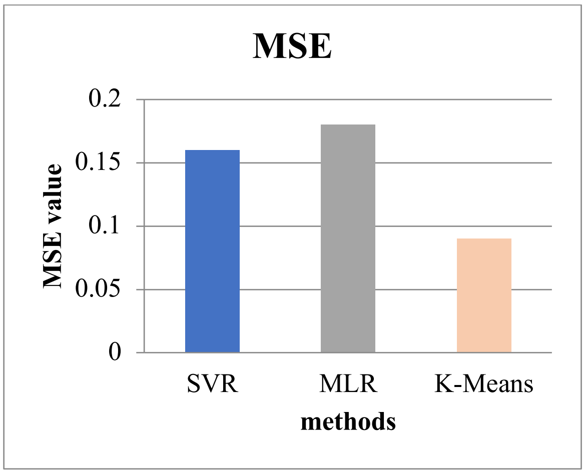

Fig. 4: COMPARATIVE ANALYSIS IN TERMS OF 'MSE'

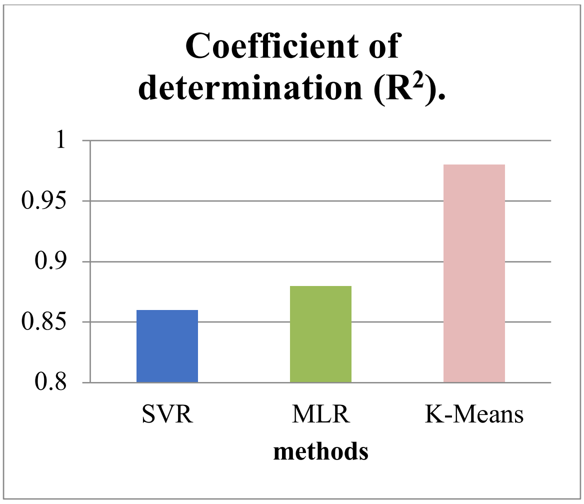

Fig. 5: COMPARATIVE ANALYSIS IN TERMS OF 'COEFFICIENT OF DETERMINATION (R^2)'

Results indicate that the use of K-Means Clustering has improved the performance of Intelligent Customer Segmentation with lower MSE and higher Coefficient of determination (R^2). The current research findings and interpretation inform marketing decision with e-commerce purchasing patterns to guide vendors toward effective engagement when confronted with volatility on data analysis.

V. CONCLUSION

In this paper, K-Means Clustering approach based Intelligent Customer Segmentation to increase sales using Customer Purchase Behavior Data is described. A customer segmentation strategy allows firms to target particular groups of consumers, resulting in more efficient marketing resource allocation and greater potential for cross and up-selling. The dataset used for this research is Malaysia's E-commerce dataset from MDEC repository for machine learning based on customers purchasing behavioral. To process the collected data and segment the customers, learning algorithm is used which is known as K-Means clustering. K-Means clustering is a process of classifying the dataset into certain number of clusters where each cluster will be defined with k centers. Data visualization and Dashboards are used for visualizing the results of customer segmentation to the end users. We evaluate the described model through a set of evaluation metrics, Mean Squared Error (MSE) and Coefficient of determination (R^2). Results indicate that the use of K-Means Clustering has improved the performance of Intelligent Customer Segmentation with lower MSE and higher Coefficient of determination (R^2). The current research findings and interpretation inform marketing decision with e-commerce purchasing patterns to guide vendors toward effective engagement when confronted with volatility on data analysis. We have done this project with as minimum flaws as possible and can further be enhanced by including major identification of statistics of poeple and improving the accuracy of the output. In this project we have implemented k-means algorithm, it can be further enhanced by using few complex algorithms such as conventional neural networks algorithms.

VI. REFERENCES

[1] Z. -H. Sun and X. Ming, "Multicriteria Decision-Making Framework for Supplier Selection: A Customer Community-Driven Approach," in IEEE Transactions on Engineering Management, vol. 70, no. 10, pp. 3434-3450, Oct. 2023, doi: 10.1109/TEM.2021.3089279.

[2] Z. Yang, Q. Li, V. Charles, B. Xu and S. Gupta, "Online Product Decision Support Using Sentiment Analysis and Fuzzy Cloud-Based Multi-Criteria Model Through Multiple E-Commerce Platforms," in IEEE Transactions on Fuzzy Systems, doi: 10.1109/TFUZZ.2023.3269741.

[3] M. Zavali, E. Lacka and J. de Smedt, "Shopping Hard or Hardly Shopping: Revealing Consumer Segments Using Clickstream Data," in IEEE Transactions on Engineering Management, vol. 70, no. 4, pp. 1353-1364, April 2023, doi: 10.1109/TEM.2021.3070069.

[4] Q. Jiang and Y. Jiang, "Analysis of e-commerce customer data mining based on Apriori optimization algorithm," 2022 2nd International Signal Processing, Communications and Engineering Management Conference (ISPCEM), Montreal, ON, Canada, 2022, pp. 155-160, doi: 10.1109/ISPCEM57418.2022.00037.

[5] M. Ghahramani, A. O'Hagan, M. Zhou and J. Sweeney, "Intelligent Geodemographic Clustering Based on Neural Network and Particle Swarm Optimization," in IEEE Transactions on Systems, Man, and Cybernetics: Systems, vol. 52, no. 6, pp. 3746-3756, June 2022, doi: 10.1109/TSMC.2021.3072357.

[6] A. Solichin and G. Wibowo, "Customer Segmentation Based on Recency Frequency Monetary (RFM) and User Event Tracking (UET) Using K-Means Algorithm," 2022 IEEE 8th Information Technology International Seminar (ITIS), Surabaya, Indonesia, 2022, pp. 257-262, doi: 10.1109/ITIS57155.2022.10009981.

[7] B. Melović, B. Rondović, S. Mitrović-Veljković, S. B. Očovaj and M. Dabić, "Electronic Customer Relationship Management Assimilation in Southeastern European Companies—Cluster Analysis," in IEEE Transactions on Engineering Management, vol. 69, no. 4, pp. 1081-1100, Aug. 2022, doi: 10.1109/TEM.2020.2972532.

[8] S. Miloudi, Y. Wang and W. Ding, "A Gradient-Based Clustering for Multi-Database Mining," in IEEE Access, vol. 9, pp. 11144-11172, 2021, doi: 10.1109/ACCESS.2021.3050404.

[9] P. Zhou, C. Lu, J. Feng, Z. Lin and S. Yan, "Tensor Low-Rank Representation for Data Recovery and Clustering," in IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 43, no. 5, pp. 1718-1732, 1 May 2021, doi: 10.1109/TPAMI.2019.2954874.

[10] Y. Yuan, K. Dehghanpour, F. Bu and Z. Wang, "A Data-Driven Customer Segmentation Strategy Based on Contribution to System Peak Demand," in IEEE Transactions on Power Systems, vol. 35, no. 5, pp. 4026-4035, Sept. 2020, doi: 10.1109/TPWRS.2020.2979943.

[11] A. Syaputra, Zulkarnain and E. Laoh, "Customer Segmentation on Returned Product Customers Using Time Series Clustering Analysis," 2020 International Conference on ICT for Smart Society (ICISS), Bandung, Indonesia, 2020, pp. 1-5, doi: 10.1109/ICISS50791.2020.9307575.

[12] Y. Feng, X. Wang and L. Li, "The Application Research of Customer Segmentation Model in Bank Financial Marketing," 2019 2nd International Conference on Safety Produce Informatization (IICSPI), Chongqing, China, 2019, pp. 564-569, doi: 10.1109/IICSPI48186.2019.9095900.

[13] X. Chen, W. Sun, B. Wang, Z. Li, X. Wang and Y. Ye, "Spectral Clustering of Customer Transaction Data With a Two-Level Subspace Weighting Method," in IEEE Transactions on Cybernetics, vol. 49, no. 9, pp. 3230-3241, Sept. 2019, doi: 10.1109/TCYB.2018.2836804.

[14] 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," in IEEE Transactions on Knowledge and Data Engineering, vol. 30, no. 3, pp. 559-572, 1 March 2018, doi: 10.1109/TKDE.2017.2763620.

[15] I. Maryani, D. Riana, R. D. Astuti, A. Ishaq, Sutrisno and E. A. Pratama, "Customer Segmentation based on RFM model and Clustering Techniques With K-Means Algorithm," 2018 Third International Conference on Informatics and Computing (ICIC), Palembang, Indonesia, 2018, pp. 1-6, doi: 10.1109/IAC.2018.8780570.

[16] R. Taniguchi, Y. Ohtaka and S. Morishita, "Prediction of Purchase Behavior of Customers in a Store by Cellular Automata," 2015 Third International Symposium on Computing and Networking (CANDAR), Sapporo, Japan, 2015, pp. 436-441, doi: 10.1109/CANDAR.2015.37.

[17] J. Zhu, H. Wang, M. Zhu, B. K. Tsou and M. Ma, "Aspect-Based Opinion Polling from Customer Reviews," in IEEE Transactions on Affective Computing, vol. 2, no. 1, pp. 37-49, Jan.-June 2011, doi: 10.1109/T-AFFC.2011.2.

[18] X. Zhang, G. Feng and H. Hui, "Customer-Churn Research Based on Customer Segmentation," 2009 International Conference on Electronic Commerce and Business Intelligence, Beijing, China, 2009, pp. 443-446, doi: 10.1109/ECBI.2009.86.

[19] T. Jiang and A. Tuzhilin, "Improving Personalization Solutions through Optimal Segmentation of Customer Bases," in IEEE Transactions on Knowledge and Data Engineering, vol. 21, no. 3, pp. 305-320, March 2009, doi: 10.1109/TKDE.2008.163.

[20] T. Iwata, K. Saito and T. Yamada, "Recommendation Method for Improving Customer Lifetime Value," in IEEE Transactions on Knowledge and Data Engineering, vol. 20, no. 9, pp. 1254-1263, Sept. 2008, doi: 10.1109/TKDE.2008.55.