Estimating Customer Segmentation based on Customer Lifetime Value Using Two-Stage Clustering Method

Pradnya Paramita Pramono
Department of Industrial Engineering
Faculty of Engineering, Universitas Indonesia
Depok, Indonesia
pradnya.paramita51@ui.ac.id

Isti Surjandari
Department of Industrial Engineering
Faculty of Engineering, Universitas Indonesia
Depok, Indonesia
isti@ie.ui.ac.id

Enrico Laoh
Department of Industrial Engineering
Faculty of Engineering, Universitas Indonesia
Depok, Indonesia
enricolaoh@ui.ac.id

***Abstract*— In order to cope with the competitive environment related to beauty industry sector in Indonesia, companies need to manage and evaluate customer interactions by enhancing Customer Relationship Management (CRM). This study aims to specify customer segment that has similar lifetime value with clustering method, hence company can conduct appropriate strategies to the right segment. Two-stage clustering method for segmenting customers is proposed in this study. Ward's method is used for choosing an initial number of cluster and K-Means method to perform clustering analysis. Two approaches using LRFM (Length, Recency, Frequency, Monetary) model and extended model called LRFM - Average Item (AI) variables in clustering process are compared by validity index to obtain the best result for customer segmentation. The result shows that adding new variable Average Item in LRFM model have no significant difference or better results in clustering. The ranking process based on Customer Lifetime Value (CLV) score is conducted using weighted LRFM model variables. Final weight score for all variables are obtained from Fuzzy AHP method. In summary, company also get several inferences such as customer characteristics of high and less potential customers. It can be a guideline for making the sale and marketing strategies.**

Keywords— Customer Segmentation, Customer Relationship Management, Customer Lifetime Value, Clustering, Fuzzy AHP

I. Introduction

In recent years, there has been massive increase in the competition among firms to sustaining their business field. One of the business sectors that exponentially growth in Indonesia is beauty industry. Ministry of Industry Republic of Indonesia in 2018 conveyed the growth of beauty industry increase 20% than national economic growth in 2017. The increase in growth was driven by large demand from the domestic and export markets as people began to pay attention in personal care. Thus, this study conducts at one of beauty clinic which is currently developing to increase market share.

In order to survive with these competitive industries, company needs to be more concerned with Customer Relationship Management. Customer Relationship Management (CRM) has an important role for supporting business strategies and build long-term interactions with customers [1]. CRM encourages companies to conduct analysis with all of the customer information, sales, marketing strategies, market trends by using the technology and resources to gain new insight related to customer behavior and customer value [2]. Previous research indicated that there are three main components that are interrelated with each other in CRM, namely CRM strategy, operational CRM and analytical CRM [3].

Customer Lifetime Value (CLV) concept is a part of Strategic Customer Relationship Management (CRM) that is widely used for predicting revenues that can be obtained from a customer by identifying their lifetime values [4]. Based on Pareto principle [5], 20% of customers have a large contribution for achieving 80% of company's sales. By knowing the customer value, company is able to implement marketing strategies that are more focused on the homogeneous group of customers. Moreover, the company can increase customer loyalty and market share. The marketing strategy that is appropriate for each group can help companies retain existing customers. According to [6] research, retaining customers is important and more profitable than bringing in new customers or acquisition.

Dividing heterogeneous customer into segments will be helpful for companies to execute their marketing strategies. Customer segmentation is one of an effective method to satisfy customer needs and preferences. Most of the studies show that customer lifetime values is often used as a basis in determining customer segmentation [7]. The common method used for segmenting customers is the clustering method. Clustering is also referred as data segmentation in several applications by grouping large data into groups that have similarities [8]. Various algorithms used in clustering are K-Means, Ward's method, DBSCAN (Density-Based Spatial Clustering of Application with Noise), SOM (Self Organizing Map), and others. One popular clustering method in the customer segmentation process that is widely used in various fields including data mining, statistical analysis, and several business applications is K-Means Clustering [9]. K-Means Clustering method is based on the means value for each data in the cluster [10]. The major problem in K-Means clustering analysis is termination of the algorithm or in other words determining the number of cluster. The ideal number of cluster is the level of minimum variation within cluster and the maximum variation between clusters. Recent studies find that combine several clustering methods can be used to identify optimal number of initial cluster [11] [12].

Based on transactional data, there are some CLV measurement variables that can be used to measure customer value segmentation [10]. RFM (Recency-Frequency-Monetary) model analysis which was developed by Hughes (1994), is widely used in several studies to evaluate customers based on their buying behavior [6]. RFM variables are selected to be important variables to identify value of

978-1-7281-1941-0/19/$31.00 ©2019 IEEE

customer based on their behavioral characteristics of transactions.

In addition, Chang and Tsay add one variable, namely Length [13]. The LRFM model is a development of the RFM model [13]. Calculation of CLV based on LRFM variables is used to predict the pattern of customer purchases in the future by considering the historical data of current and past customers. In this study, two approaches are proposed. The first approach is based on LRFM variables and the second approach, the extended of LRFM model with another variable besides LRFM model is considered. The new variable is Average Item (AI), which refers to the ratio of total item and the frequency purchased by customer in specific period of time.

Variables that are used as the basis for evaluating customer lifetime values have different weights according to the characteristics of the industry [9]. Therefore, the weight value of the variables is determined subjectively by some experts who have knowledge and experience in related business fields [14]. The Fuzzy Analytical Hierarchy Process (FAHP) method was used in this study to give weight for each variable and determine the CLV for each segment. Later, the results of CLV can be analyzed based on the ranking results of CLV [15]. Some variables might consist ambiguity, thus FAHP method is used to provide more accurate results and overcome the uncertainty of human judgments.

In summary, this study focuses on estimating customer segmentation based on customer lifetime value using two-stage clustering method, Ward's method is used for choosing an initial number of cluster and K-Means method to perform clustering analysis. Variables are chosen by comparing LRFM model and extended model of LRFM variable with Average Item in the clustering process. Thus, CLV is calculated according to the best results between two approaches. In the last section, the customer lifetime values are ranked to identify segment characteristics and recognize potential customer segment. The result of this study can be used as a guideline for making sale and marketing strategies according to customer characteristics.

II. METHODOLOGY

The structure in this study is based on CRISP methodology, such as data understanding, data preparing, data processing, and model evaluation. The brief method is shown in Fig. 1.

A. Data Preparation

Secondary data for customer online transaction in Beauty Clinic from January 2018 – December 2018 were collected in this paper. There was 3899 lists of customers are gained. For the first step of pre-processing data, this phase involves 3 steps: Data reduction, feature selection and data transformation.

Data reduction were done to identify unique value of customer, remove unreasonable records such as those of customer who have zero amount of monetary and transaction records who inconsistent and become outlier. In the selection stage, the attributes of the data are selected. The result of selection process is to extract the value based on selected variables for each customer.

Since the segmentation is on the basis of CLV and LRFM model, the selected features then transformed according to this method are included last purchase date, count purchase which is the frequency of customer purchases, total money expended by customer during one year which refers to monetary, the duration between the first transaction and the last transaction which refers to length and the last variable named Average Item (AI) which refers to the total average items purchased by customer in one year period. Thus, two approaches are used for segmenting customer: First, Model 1 consist of LRFM variables and Model 2 consist of extended LRFM variables with Average Item.

Data normalization were done by several steps, the data is normalized in a way that can be exploited by clustering and spatial analysis method. In this paper, data normalization was calculated through the following formulas:

$$v' = \frac{v - min_A}{max_A - min_A}\left(new_{max_A} - new_{min_A}\right) + new\_min_A \qquad (1)$$

This method performs a linear transformation on the original data. Suppose that v is a value that will be normalized; $minA$ and $maxA$ are the minimum and maximum values of an attribute A; new_{max_A} and new_{min_A} are the new interval number after normalization.

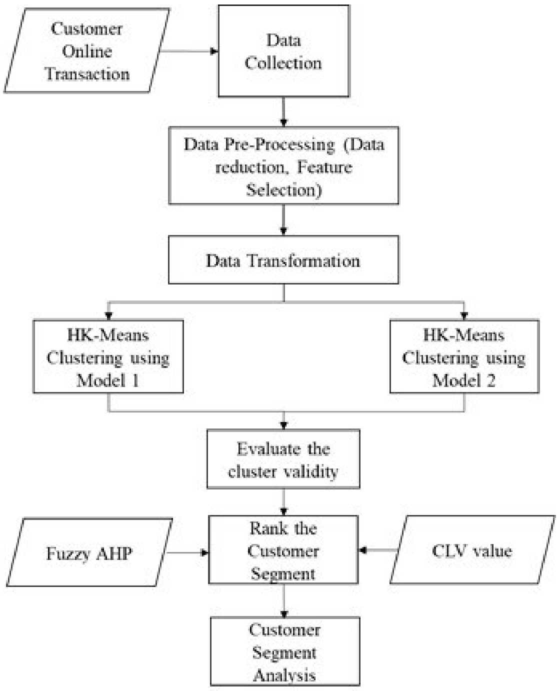
Customer Online Transaction
Data Collection
Data Pre-Processing (Data reduction, Feature Selection)
Data Transformation
HK-Means Clustering using Model 1
HK-Means Clustering using Model 2
Evaluate the cluster validity
Fuzzy AHP
Rank the Customer Segment
CLV value
Customer Segment Analysis

Fig. 1. Methodology Steps.

B. Data Clustering

Clustering algorithms can be classified as hierarchical and partitional. This study uses both hierarchical and partitional algorithms to divide the customer base. The hierarchical algorithms can be classified as agglomerative and divisive based on a bottom-up or top-down decomposition. One of the major problems in the cluster analysis is termination of the algorithm or in other words determining the number of cluster. The ideal number of cluster is the level of minimum variation within cluster and the maximum variation between clusters. This study applies two-stage clustering method, namely Hierarchical K-Means (HK-Means). First, Ward's method is used to do clustering group estimation and determine group number (k). Ward's method is used to minimize variance because this method keeps the minimum sum square of error while merge the data within cluster and maximize the sum square of error between clusters.

The second stage uses the K-means clustering operation to separate the data into k groups. K-Means clustering intends to partition n objects into k clusters in which each object belongs to the cluster with the nearest mean. The objective of K-Means clustering is to minimize total intra-cluster variance, or, the squared error function:

$$J = \sum_{j=1}^{k}\sum_{i=1}^{n}\left\|x_i^{(j)} - c_j\right\|^2 \qquad (2)$$

From the equation above J is the objective function of K-Means clustering; k is the number of cluster; n is the number of cases; $x_i^{(j)}$ is the case i; and c_j is the centroid for cluster j.

Furthermore, the clustering results will be evaluated by several validity methods. There are several internal validity index such as *Dunn Index, Davis Bouldin Index (DBI), Silhoutte, and Sum of squares within cluster (SSWC).* In this study, DBI is used as a validity index. DBI is a simple and the most widely used criterion to evaluate the validity of clustering results and determine the number of clusters. DBI is formulated as follows:

$$DBI = \frac{1}{n_c}\sum_{i-1}^{n_c} R_i\,, where \qquad (3)$$

$$R_i = \max_{j=1\ldots.nc,i\neq j}\left(R_{ij}\right), \qquad i = 1\ \ldots.n_c \qquad (4)$$

$$R_{ij} = \frac{S_i + S_j}{d_{ij}} \qquad (5)$$

$$d_{ij} = d\left(v_{i,}v_j\right),\ s_i = \frac{1}{\|c_i\|}\sum_{x\in c_i} d\ (x, v_i) \qquad (6)$$

Where $d\ (x, y)$ denotes the euclidean distance between x and y; $c_i$is the cluster i; v_i is the centroid of cluster i; and $\|c_i\|$ refers to the norm of c_i.

C. Fuzzy AHP

AHP technique weighs evaluation criteria based on a hierarchical structure. In AHP method, Decision maker's opinions are represented in a pair-wise comparison matrix in which each element shows the importance of one criterion over the others. Since experts' judgments are always ambiguous, multiple versions of fuzzy AHP method were introduced applying the theory of fuzzy. In order to solve these problems, fuzzy linguistic preference relation was applied in AHP method with two striking features of deriving consistent priorities with fewer computations [16]. Previous study presented by Chang (1996) finds that Triangular Fuzzy Number (TFN) is preferred for pairwise comparison scale of Fuzzy AHP and extent analysis method was used for the synthetic extent value of pairwise comparison [17]. Based on TFN as illustrated in Fig. 2. the parameters l, m, u, respectively denote the smallest, the modal value and the largest value of fuzzy number. In this study, Fuzzy AHP is used to clarify the final clustering results. By weighting the average cluster centroid with the final weight score of Fuzzy AHP, CLV value can be obtained. Thus, the ranking process based on CLV score for each segment will guide the company to identify high and less loyal customer.

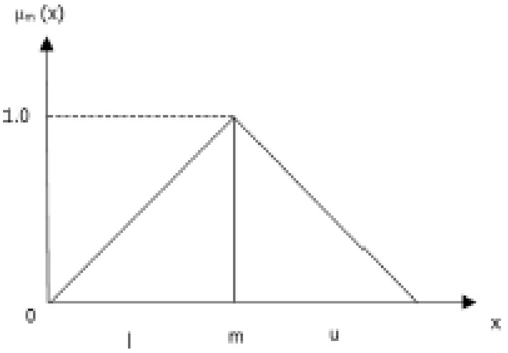

Fig. 2. Triangular Fuzzy Number.

III. RESULTS AND ANALYSIS

Historical data from customer transactions in this study are processed using a two-stage clustering approach to get the characteristics of a homogeneous customer group. The formed segment is measured by a variable measuring the customer lifetime value. Furthermore, the cluster results for model 1 and model 2 with extended variables are compared to get the best results. Giving weight for each variable using the Fuzzy AHP method is also done to identify the CLV and the ranking of the cluster formed.

TABLE I. SEQUENCE STEP CONCENTRATION COEFFICIENTS OF WARD'S METHOD FOR LRFM MODEL

Stage	Cluster	Coefficients	Coefficients Difference
3898	2	708.939	301.886
3897	3	407.053	181.181
3896	4	225.872	69.67
3895	5	156.202	31.095
3894	6	125.107	25.153
3893	7	99.954	

TABLE II. SEQUENCE STEP CONCENTRATION COEFFICIENTS OF WARD'S METHOD FOR EXTENDED LRFM MODEL

Stage	Cluster	Coefficients	Coefficients Difference
3898	2	776.072	314.172
3897	3	461.9	183.398
3896	4	278.502	60.411
3895	5	218.091	

A. Estimating number of cluster

Two-stage clustering approaches are used in this study. The first stage is applying hierarchical clustering namely Ward's method to determine number of clusters. Based on the sequence step agglomeration of Ward's method, the number of clusters can be estimated correctly. From Table 1, coefficient in stage 3897 shows that the difference is quite large compared to the results in stage 3896. Since this is a large value before a significant reduction, the number of cluster 3 is chosen to be processed as an initial number of cluster in k-means clustering.

For model 2 as an extended variable, the results of coefficients difference in Table 2 with great gaps were obtained at stage 3898 and 3897. Thus, 183.398 is a large amount of coefficient distance compared to the distance results in stage 3896. Then, the number of clusters of 3 was used in the initial cluster and proceeds to the next stage.

B. Segmenting the customer

The number of clusters generated based on Ward's method is processed in clustering analysis. In this study, DBI index is used to evaluate the cluster results as shown in Table 3. Clustering process for model 1, with 4 LRFM variables have DBI value of 0.707. Meanwhile, model 2 with an extended variable, namely LRFM-AI has a DBI of 0.859. DBI value on extended variables shows greater results. It indicates that the addition of Average Item variable does not give more significant clustering results. The significance of Length, Recency, Frequency, Monetary and Average Item variables are tested with ANOVA based on clustering results. F-score and p-value for model 1 in Table 4 reveal significant result, LRFM variables have F-score greater than 1.96 and probability score (sig.) less than 0.05. Meanwhile in Table 5, F-score and p-value are not significantly different compared with model 1. F-score for extended model 2 with new variable Average Item also determine low differences for each cluster, thus it makes the clustering results have no great differences.

Pearson correlation is also calculated to measure the correlation between new variable Average Item and LRFM model, the results indicate that new variable has no strong correlation with two existing variables Length and Recency. Since model 2 does not give either better or significant result, further analysis will be carried out based on model 1 with LRFM variables.

TABLE III. PERFORMANCE EVALUATION OF CLUSTERING METHODS

Parameter	**Davis Bouldin Index (DBI)**
LRFM	0.707
LRFM + Average Item	0.859

TABLE IV. ANOVA FOR LRFM MODEL

	Between Cluster		**Within Cluster**		**F**	**Sig.**
	Mean Square	***df***	***Mean Square***	***df***		
L	120.259	2	0.016	3896	7683.46	0.000
R	131.469	2	0.021	3896	6199.293	0.000
F	4.349	2	0.002	3896	1944.439	0.000
M	9.238	2	0.007	3896	1384.454	0.000

Final cluster centers generated for model 1 with LRFM variables is shown in Table 6. Total average value for each variable in the last row is used to compare with the LRFM cluster centers in the final analysis step. If the average LRFM value of a cluster exceeds the overall average value, then the up ↑ symbol appears. However, if the average LRFM value of a cluster does not exceed the overall average value, then the down ↓ symbol appears.

TABLE V. ANOVA FOR EXTENDED LRFM MODEL

	Between Cluster		**Within Cluster**		**F**	**Sig.**
	Mean Square	***df***	***Mean Square***	***df***		
L	120.079	2	0.016	3896	7626.954	0.000
R	131.366	2	0.021	3896	6178.988	0.000
F	4.379	2	0.002	3896	1972.061	0.000
M	9.276	2	0.007	3896	1394.359	0.000
AI	0.194	2	0.017	3896	11.348	0

TABLE VI. FINAL CLUSTER CENTRES

Cluster	**L**	**R**	**F**	**M**
Cluster 1	0.66	0.161	0.129	0.228
Cluster 2	0.055	0.719	0.014	0.059
Cluster 3	0.081	0.218	0.02	0.071
Average Value	0.265	0.366	0.054	0.119

C. Ranking the customer segments

The results of final weights for each variable with the Fuzzy AHP method presented in Table 7. The results indicate that the most important variable is Frequency and the least important variable is Recency.

TABLE VII. FINAL WEIGHTS OF THE CRITERIA

Criteria	**Weight**
Length	0.222
Recency	0.182
Frequency	0.305
Monetary	0.292

Next, to calculate CLV in each cluster, the normalized centroid score of each cluster is multiplied by the final weights. The simplest mathematical model for CLV is shown in the following equation:

$$CLV = L_{Ci}\ x\ WL_{Ci} + R_{Ci}\ x\ WR_{Ci} + F_{Ci}\ x\ WF_{Ci} + M_{Ci}\ x\ WM_{Ci} \tag{7}$$

Where L_{Ci} refers to normal centroid score variable Length of cluster c_i, WL_{ci} is weighted Length. R_{Ci} refers to normal centroid score variable Recency of cluster c_i, WR_{ci} is weighted Recency. F_{Ci} refers to normal centroid score variable Frequency of cluster c_i, WF_{Ci} is weighted Frequency. M_{Ci} refers to normal centroid score variable Monetary of cluster c_i, WM_{ci} is weighted Monetary.

TABLE VIII. CLV RESULTS FOR EACH CLUSTER

Cluster	**Total Customer**	**CLV**	**Rank**
Cluster 1	882	0.282	1
Cluster 2	1663	0.164	2
Cluster 3	1354	0.084	3

To determine the analysis of each cluster, the ranking process based on CLV is conducted. Firstly, the results in Table 8 reveal that Cluster 1 is the most important segment. Customer in cluster 1 are considered to the core customers and refers to high value loyal customers (LRFM, ↑↓↑↑). Cluster 1 consist of 882 loyal customers that have high average value of length and relationship with the company, often make purchases and spend a lot of money in the company. Customers in this class have a fairly low recency level and quite active in making purchases in the last few times. They tend to be gold customers.

Secondly, customers in cluster 2 may include the new customers group that refers to uncertain lost customers (LRFM, ↓↑↓↓). Cluster 2 consists of 1663 customers with low length of relationship, large recency value which means that customers very rarely made transactions in recent times and spent low monetary value. The number of customers included in cluster 2 is quite large compared to total customers in other segment groups. Company needs to develop closer relationship with this segment.

Finally, customers in cluster 3 are included to new customers group and refer to uncertain new customers (LRFM, ↓↓↓↓). Cluster 3 consists of 1354 customers who have only recently transactions to the company, so they have very low length of relationship with the company. They also spent low monetary value. However, customers in this class are new customers and may only make purchase during sales.

IV. CONCLUSION

This study proposed two approaches of CLV segmentation based on LRFM model and extended LRFM model with Average Item variable. The clustering was done using two-stage process of Ward's method to specify number of cluster and K-Means clustering to find the final results of clustering analysis. Based on Ward's method, the best initial number of cluster for K-Means clustering is three. Later, the optimum number of cluster generated by Ward's method is processed in K-Means clustering. The result shows that adding new variable in the extended LRFM model gives no difference in clustering results. For further analysis in K-Means clustering, segmentation process with LRFM model can be used as a parameter. Fuzzy AHP is proposed to define weight of each variable, the final score indicates that Frequency was the most significant variable in this study. The ranking process of CLV according to K-Means weighted centroid score for each cluster revealed that cluster 1 has the highest score, followed by cluster 2 and cluster 3, respectively. Customers in cluster 1 are potential and valuable to be gold customer. By having this CLV score, company can identify their customer characteristics and focus on those customers who bring the maximum benefit and loyalty. This study has some limitations that can direct to future researches. Since the segmentation is done based on LRFM variables, company can customize their marketing strategies only based on customer behaviors. Future work, study can use extended method to identify products which are bought frequently by the customer for each segment.

ACKNOWLEDGMENT

Authors would like to express gratitude and appreciation to Universitas Indonesia for funding this study through PIT-9 Research Grants Universitas Indonesia No: NKB-0061/UN2.R3.1/HKP.05.00/2019.

REFERENCES

[1] Z. Soltani and N. J. Navimipour, "Customer relationship management mechanisms: A systematic review of the state of the art literature and recommendations for future research," *Comput. Human Behav.*, vol. 61, pp. 667–688, 2016.

[2] Z. Zare-Hoseini, M. J. Tarokh, and H. J. Nooghabi, "Lifetime value model in the medical sector: A case study of a restoration and beauty clinic," *Int. J. Pharm. Healthc. Mark.*, vol. 5, no. 1, pp. 54–66, 2011.

[3] Khalid Rababah, Haslina Mohd, and Huda Ibrahim, "Customer Relationship Management (CRM) Processes from Theory to Practice: The Pre-implementation Plan of CRM System," *Int. J. e-Education, e-Business, e-Management e-Learning*, vol. 1, no. 1, 2011.

[4] C. H. Weng and T. C. K. Huang, "Knowledge acquisition of association rules from the customer-lifetime-value perspective," *Kybernetes*, vol. 47, no. 3, pp. 441–457, 2018.

[5] R. Srivastava, "Identification of Customer Clusters Using Rfm Model : a Case of Diverse Purchaser Classification," vol. 9, no. 4, pp. 201–209, 2017.

[6] A. J. Christy, A. Umamakeswari, L. Priyatharsini, and A. Neyaa, "RFM ranking – An effective approach to customer segmentation," *J. King Saud Univ. - Comput. Inf. Sci.*, 2018.

[7] F. Safari, N. Safari, and G. A. Montazer, "Customer lifetime value determination based on RFM model," *Mark. Intell. Plan.*, vol. 34, no. 4, pp. 446–461, 2016.

[8] J. Han, M. Kamber, and J. Pei, *Data Transformation by Normalization*. 2011.

[9] M. Khajvand, K. Zolfaghar, S. Ashoori, and S. Alizadeh, "Estimating customer lifetime value based on RFM analysis of customer purchase behavior: Case study," *Procedia Comput. Sci.*, vol. 3, pp. 57–63, 2011.

[10] C. F. Tsai, Y. H. hu, C. S. Hung, and Y. F. Hsu, "A comparative study of hybrid machine learning techniques for customer lifetime value prediction," *Kybernetes*, vol. 42, no. 3, pp. 357–370, 2013.

[11] H. Güçdemir and H. Selim, "Integrating multi-criteria decision making and clustering for business customer segmentation," *Ind. Manag. Data Syst.*, vol. 115, no. 6, pp. 1022–1040, 2015.

[12] R. Ait, A. Amine, B. Bouikhalene, and R. Lbibb, "Customer Segmentation Model in E-commerce Using Clustering Techniques and LRFM Model : The Case of Online Stores in Morocco," vol. 9, no. 8, pp. 2000–2010, 2015.

[13] D. A. Kandeil, A. A. Saad, and S. M. Youssef, "A two-phase clustering analysis for B2B customer segmentation," *Proc. - 2014 Int. Conf. Intell. Netw. Collab. Syst. IEEE INCoS 2014*, pp. 221–228, 2014.

[14] T. Hong and E. Kim, "Segmenting customers in online stores based on factors that affect the customer's intention to purchase," *Expert Syst. Appl.*, vol. 39, no. 2, pp. 2127–2131, 2012.

[15] D. R. Liu and Y. Y. Shih, "Hybrid approaches to product recommendation based on customer lifetime value and purchase preferences," *J. Syst. Softw.*, vol. 77, no. 2, pp. 181–191, 2005.

[16] T. C. Wang and Y. H. Chen, "Applying fuzzy linguistic preference relations to the improvement of consistency of fuzzy AHP," *Inf. Sci. (Ny).*, vol. 178, no. 19, pp. 3755–3765, 2008.

[17] D. Y. Chang, "Applications of the extent analysis method on fuzzy AHP," *Eur. J. Oper. Res.*, vol. 95, no. 3, pp. 649–655, 1996.