Clustering of potential customer in digital trade transaction

E Purwanto1,*, Z A Long2 and F Mohd3
Malaysian Institute of Information Technology, Kuala Lumpur University, Malaysia

*Corresponding author: purwanto.eko@s.unikl.edu.my

Abstract. The growth of digital trade transaction data in Indonesia is very rapid. Retail e-commerce in Indonesia is currently growing rapidly although it is still smaller than other countries, growing by around 30% every year for the last five years. Customer grouping is carried out based on different types of customer data, including customer value data, customer profile data, customer transaction data, and customer purchase order data. In the following sections, we briefly introduce each type. This study used data mining that was collected from potential customers by online questionnaires. The customer set data used were 309 respondents. The data collection that the authors do it by distributing questionnaires online through Google form. Clustering process used neural network algorithm. The competitive network used data clustering. The research result based on the results of data collection, the data grouping can be obtained with cluster 3 epoches 1000 data with kohonen parameter 0.01 and cluster 3 epoches 1000 data with kohonen parameter 0.01.

1. Introduction
Based on the results of the data collection by distributing questionnaires obtained data that the perpetrators of online transactions 65% aged between 19 to 25 years and 23.9% aged 26-35 years. Actors of online sales transactions based on the results of the data collection the questionnaire obtained data that 51.9% were as buyers, 45.5% as sellers and buyers and 2.6% as sellers. E-marketplace that are often used in online sales by online transactions are based on the results of questionnaire distribution showing 71.5% using Shopee, 41.5% using Tokopedia as a sales tool, and 37.2% using Bukalapak as an online sales tool. Based on the results of questionnaire distribution, it was found that the e-marketplace used by buyers in online transactions was 82.4% using Shopee, 43.8% using Tokopedia and 33% using Bukalapak. Knowledge of online transaction users is an important issue in business strategy, online transaction users are one of the sources of profit for online businesses. For this reason, a good understanding of the group of online transactions is needed. Understanding of online transaction data can be used by companies to invest in potential online transactions. Some problems need to be resolved such as sparse data and low recommended accuracy [1]. Social media sites can be used for e-commerce by inserting product recommendations to customers as a product marketing strategy [2]. Every SME in Indonesia has a variety of customers, but SME do not yet have a customer mapping method so they cannot know which customers are loyal or not [3]. The recommendation system is a tool for analyzing each group and for classifying the top-ranked items in each group [4].

Customer data grouping based on the type of customer data consisting of customer value data, customer profile data, customer transaction data, and customer purchase order data [5]. The criteria used in grouping, namely present value, potential value, and customer loyalty, are used to evaluate customers. The three criteria will be calculated, according to the value of the three criteria, customers can be grouped into different segments. The results of customer clustering were obtained using three
approaches, RFM, chi-square automatic interaction detection, and logistic regression [5]. Research on customer clustering is data that is usually collected starting from data about the background or demographic characteristics of customers. Customer data collected includes the order of goods purchased, where the goods can be real products or high-level conceptual goods. Grouping according to the order of items based on similarities. Data from two customers with the same order of items must be assigned to the same group. This grouping by order of items has the advantage of being able to provide an explanation of the dynamic changes in the order (that is the dynamic nature of customer behavior is included for consideration). The weakness in this grouping is that in fact, a customer usually buys more than one item in one transaction. The item order failed to represent actual customer buying behavior. So item sequence clustering is an unsuccessful approach to classifying customers based on their buying behavior [5].

The distribution of customer data includes the first, consumer information such as gender, age, marital status and education level are grouped into the category of individual consumer characteristics. Second is consumption characteristic data such as the payment method, consumption frequency grouped into the consumption characteristics category. The third is consumer attitude factors such as customer satisfaction, customer loyalty and consumer trust grouped into consumer attitude categories [6]. The resulting grouping of customers is user behavior data that provides businesses with a deeper understanding of the types of users that are on the system. Customer segmentation can provide benefits, namely for a better knowledge of the types of users in a system can produce a better business and marketing strategy then can increase the duration of use of the application by the user if he always receives relevant content [6]. To determine the performance of clusters, K-Means and Davies Bouldin Index methods can be used [7]. The K-Means algorithm is used for grouping customer data based on transaction data which aims to determine customer behavior [8]. The K-means clustering algorithm is used in the input classification with selected cross-validation testing and one of the data sets becomes a few percent for training and other tests by compiling the cluster files formed by the cluster model as the dependent variable and the remaining variables as independent variables [9]. Grouping sales with the criteria used in calculations with the K-means clustering algorithm obtained 3 clusters, namely sellers, resellers and dropships [10].

2. Material and Methods

This study used data mining that was collected from potential customers by online questionnaires. The customer set data used were 309 respondents. There are 9 indicators used for clustering attributes. These indicators are age, role online transaction, transaction duration, the number of marketplaces for sale, number of market transactions for transactions, Number of product categories, types of digital payment usage, and number of social media for transactions. The research uses methods like Figure 1 below:

![Figure 1. Research methods](image-url)
The data collection that the authors do is by distributing questionnaires online through Google form. A total of 309 respondents participated in filling out the questionnaire. Based on the results of the distribution of questionnaires obtained data that 65.2% of online transaction perpetrators aged 19-25 years, 23.9% of online transaction perpetrators aged 26-35 years and 10.9% aged under 18 years and over 36 years. The average online transaction done in one month based on the results of filling out the questionnaire is 33.1% answered once a month, 37% answered 2-4 times a month, 13% answered 5-10 times a month 16.9% answered more than 10 times in a month. Based on the education level of the perpetrators of online transactions, 62.4% had a bachelor's degree, 24.8% had a senior high school level and 12.8% had a bachelor and associate’s degree. Based on the type of transaction carried out the results obtained 51.8% as a buyer, 45.6% as a seller and 2.6% as a seller and buyer.

Pre-processing was the initial process before entering into the clustering process. Data with 9 variables were processed first. There was data that needs to be reduced. Input data also were be normalized. The number of variables is 9 variables data from 309 respondents were used as input data for clustering. The input data was in the form of a matrix 309x9. Clustering process used neural network algorithm. Competitive network used data clustering. It is unsupervised learning. The architecture of competitive network that used in this research, like figure below:

![Clustering model architecture](image)

**Figure 2.** Clustering model architecture for (a) 3 output and (b) 5 output

Based on the results of data collection by distributing questionnaires to the number of respondents 309 obtained preliminary data such as indicator 1 cluster 1 number 11 data, cluster 2 number 201 data, cluster 3 number 74 data, cluster 4 number 21 data and cluster 5 number 2 data. Indicator 2 cluster 1 number 8 data, cluster 2 number 160 data and cluster 3 number 141 data. Indicator 3 cluster 1 number 101 data, cluster 2 number 114 data, cluster 3 number 40 data and cluster 4 number 54 data. Indicator 4 cluster 1 number 56 data, cluster 2 number 128 data, cluster 3 number 56 data, cluster 4 number 27 data, cluster 5 number 19 data, cluster 6 number 13 data and cluster 7 number 10 data. Indicator 5 cluster 1 number 3 data, cluster 2 number 149 data, cluster 3 number 49 data, cluster 4 number 42 data, cluster 5 number 29 data, cluster 6 number 22 data and cluster 7 number 15 data. Indicator 6 cluster 1 number 32 data, cluster 2 number 200 data, cluster 3 number 51 data, cluster 4 number 16 data, cluster 5 number 8 data and cluster 6 number 2 data. Indicator 7 cluster 1 number 161 data, cluster 2 number 84 data, cluster 3 number 37 data, cluster 4 number 14 data, cluster 5 number 6 data, cluster 6 number 1 data and cluster 7 number 6 data. Indicator 8 cluster 1 number 44 data and cluster 2 number 265 data. Last indicator 9 cluster 1 number 20 data, cluster 2 number 118 data, cluster 3 number 82 data, cluster 4 number 75 data, cluster 5 number 11 data and cluster 6 number 3 data.

Based on the results of the data collection, data grouping can be obtained with cluster 3 epoches 1000 data with kohonen parameter 0.01 as follows:

| Cluster 1 | Cluster 2 | Cluster 3 |
|-----------|-----------|-----------|
| 120       | 101       | 88        |
While by grouping data with cluster 5 epoches 100 with kohonen parameter 0.01 as follows:

| Cluster 1 | Cluster 2 | Cluster 3 | Cluster 4 | Cluster 5 |
|-----------|-----------|-----------|-----------|-----------|
| 71        | 54        | 56        | 66        | 62        |

### Table 2. Cluster 5 epoches 1000

#### 3. Results and Discussion

Based on Table 1 it can be seen that clustering using 1000 epoches and output 3 Parameter of clustering cluster 3 epoches 100 kohonen parameter 0.01.

Cluster 1 : 2.2985 2.1803 1.3722 0.8728 1.1621 0.9400 1.1676 0.7816 1.1024  
Cluster 2 : 2.3362 2.5137 2.6008 1.0261 1.6830 1.3124 1.6309 0.9604 2.1471  
Cluster 3 : 2.4682 2.6216 2.4955 3.2899 4.0415 1.5875 2.8247 0.8434 2.3296

Our experiment use K-means clustering [7]. The results of clustering generated centroid value of the cluster. Centroid value was related to cluster division of input data. Input data that were close to the centroid value of the cluster will be entered into the cluster. Centroid value shown in Figure 3.

![Figure 3. (a) Value of cluster centroid (3 output) and (b) total distance (3 output)](image)

The clustering process was an iterative process. So that the optimal clustering results were determined by convergence. That was a condition where the clustering process has produced a fixed value even though epoches are added.

Based on Table 2 it can be seen that clustering using 1000 epoches and output 5 parameter of clustering cluster 5 epoches 1000 kohonen parameter : 0.01 the value of cluster centroid.

Cluster 1 : 2.4510 2.4055 2.5693 0.8442 1.0050 0.8794 1.2131 0.8311 1.0603  
Cluster 2 : 2.3397 2.6359 2.6064 4.1101 4.6332 1.6856 3.2078 0.8993 2.5229  
Cluster 3 : 2.2345 2.4661 2.0711 1.7266 1.4525 1.4760 1.5824 0.9260 2.6494  
Cluster 4 : 2.1874 2.1570 1.0255 0.7996 1.1331 0.9701 1.1161 0.8106 1.1669  
Cluster 5 : 2.5850 2.5111 2.4903 1.2607 3.3233 1.3761 2.2729 0.8338 1.8703

The results of clustering generated centroid value of the cluster. Centroid value was related to cluster division of input data. Input data that were close to the centroid value of the cluster will be entered into the cluster. The centroid value shown in Figure 4.
The clustering process was an iterative process. So that the optimal clustering results were determined by convergence. That was a condition where the clustering process has produced a fixed value even though epochs are added.

4. Conclusion
This study used data mining techniques that was collected from respondent of potential customers by online questionnaires. The customer set data used were 309 respondents. Based on the results of the distribution of questionnaires obtained data that 65.2% of online transaction perpetrators aged 19-25 years, 23.9% of online transaction perpetrators aged 26-35 years and 10.9% aged under 18 years and over 36 years. The average online transaction done in one month based on the results of filling out the questionnaire is 33.1% answered once a month, 37% answered 2-4 times a month, 13% answered 5-10 times a month 16.9% answered more than 10 times in a month. Based on the education level of the perpetrators of online transactions, 62.4% had a bachelor's degree, 24.8% had a senior high school level and 12.8% had a bachelor and associate’s degree. Based on the type of transaction carried out the results obtained 51.8% as a buyer, 45.6% as a seller and 2.6% as a seller and buyer. The clustering process used neural network algorithm. Based on the results of the data collection, data grouping can be obtained with cluster 3 epoches 1000 data with kohonen parameter 0.01 and cluster 5 epoches 1000 data with kohonen parameter 0.01.

References
[1] Jianrui C, Chunxia Z, Uliji and Lifang C 2019 Complex Intell. Syst. 6 147
[2] Ramalingam V V, Pandian A and Kirthiga M 2018 J. Phys. 1000 012029
[3] Syakur M A, Khotimah B K, Rochman E M S and Satoto B D 2018 Mater. Sci. Eng. 336 012017
[4] Kishore G K and Babu D S 2017 J. Comput. Eng. 19 3
[5] Liu Y C and Chen L Y 2017 J. Eng. Res. Appl. 7 1
[6] Schuller D and Pekarek J 2018 Acta Univ. Agric. Silvic. Mendel. Brun. 66 2
[7] Sari P K and Purwadinata A J 2019 Data SCI Appl. 2 1
[8] Mulyawan B, Christanti V M and Wenas R 2019 IOP Conf. Ser.: Mater. Sci. Eng. 508 012123
[9] Mohammed M, Yadwad S and Kassie A 2018 2nd Int. Conf. Trends Electron. Inform. (ICOEI). p 115
[10] Aria R R 2020 Int. J. Inf. Syst. Technol. 3 2