IOP Conference Series: Materials Science and Engineering

PAPER • OPEN ACCESS

Recommendation Product Based on Customer Categorization with K-Means Clustering Method

To cite this article: Bagus Mulyawan *et al* 2019 *IOP Conf. Ser.: Mater. Sci. Eng.* **508** 012123

View the article online for updates and enhancements.

Recommendation Product Based on Customer Categorization with K-Means Clustering Method

Bagus Mulyawan [*], **Viny Christanti M. and Riyan Wenas**
Faculty of Information Technology, Tarumanagara University
Jalan S.Parman No.1 Jakarta 11140 Indonesia

*bagus@fti.untar.ac.id

Abstract. Nowadays, web shopping is more than selling product. Many web shopping have basic analytics system to analyze customer data, transaction data including demographics, age and gender. They add many features in web shopping to maintain customer loyalty. In this research, we made a web shopping that can analyze customer shopping behavior. We used FM (Frequency and monetary) analysis based on a "transaction" data set. We categorize the customer based on how often they buy, how much they buy and how much the value of purchased item. We use K-Means algorithm to cluster the customer based on their transaction. Analyzing and understanding customers' buying behavior can help the store to know what they are looking for. Therefore, at every customer web page that is in the same cluster will appear recommended products accordance with the transaction that has been done. Recommendation Products are presented is the prediction of the type of goods that may be chosen by the customer. So that the recommendation products that appear on web pages between customers will be different. The data used for this test is "Istana Accessories" store data from January to June 2014. The results show that recommendation product from K-Means algorithm successfully obtained and displayed on the customer page.

1. Introduction

E-commerce or generally called electronic commerce is the distribution, purchase, sale, marketing of goods and services through the internet or computer network [1]. All components in the trade are applied to e-commerce such as product services, payment methods, shipping methods, and ways of promotion.

In general, in e-commerce based websites there are supporting features that aim to increase sales, such as top products and product recommendations. Based on the survey results from the Istana Accessories Shop, determining the top product is based on previous sales transactions. If many customers buy a product, it can be said that the product is a top product. Product recommendations are based on customer purchasing habits where looking for relationships between different items (products) in transaction data.

The accessories shop for Istana Accessories gadget is located in Roxy Square. This store provides various kinds of gadget accessories and serves sales to customers, most of which are retail stores. At this time, Istana Accessories does not have a website that is a place to facilitate customers in choosing products sold, to know detailed information about each product, and to order purchases. If the customer wants to do these three things, the customer must contact the shop owner through the chat application.

In order to create more efficient shopping activities, an online sales system was built using a website. Making a website for Istana Accessories presents several features, but its main features are focused on displaying product recommendations.

The recommendation product feature is created by applying K-means. Product recommendations are based on the relationship between a product and other products that are purchased together to form a pattern of product purchases.

The K-Means algorithm is used to determine top products and product recommendations that are based on customer purchasing characteristics. K-Means group customer based on attributes. For example, using the frequency shopping attribute (frequency) and the total monetary transaction at a certain time period. With the grouping of customers, transactions can be analyzed by a group of customers in a group. The products contained in the transaction can be made as top products and product recommendations. Top products are based on a product that has the highest sales quantity while other products are made product recommendations. Because of its personal nature of customers, top products and product recommendations displayed for each customer can be the same or different because it depends on the characteristics of the purchase and grouping.

Testing this application uses data from the Istana Accessories Shop. It is expected that through the existence of a website, it can expand sales marketing and increase customer loyalty.

Clustering is a data analysis method, which is often included as one of the data mining methods. The aim is to group data with the same characteristics into the same category and data with different characteristics into other categories [5]. The purpose of this clustering process is to group data into a category, so that objects in a category have a great resemblance to other objects in the same category, but differ from objects in other categories. This e-commerce system uses clustering with partition-based clustering. Partition approach is used because the number of existing categories has been determined first, then members are determined from each category.

2. Method

2.1 Algoritma K-Means

The K-Means method is one method of clustering data to partition existing data into one or more clusters / groups [2].

The K-Means method partitioned the data into clusters so that data that had the same characteristics was grouped into the same cluster and data that had different characteristics grouped into other groups. The purpose of clustering is to group objects until the distance of each object to the center of the group in a group is minimum.

The process of grouping using K-Means is generally carried out with the basic algorithm as follows [2]:

1. Determine the number of clusters.
2. Allocate data into clusters randomly.
3. Calculate the centroid (average) of the data in each cluster.
4. Allocate each data to the nearest centroid.
5. Return to stage 3, if there are still data that move clusters or if changes in the centroid value are above the specified threshold value.

To calculate the centroid cluster i, vi, use the following formula:

$$v_{ij} = \frac{\sum_{k=1}^{N_i} X_{kj}}{N_i} \qquad (1)$$

Where:

Ni: Amount of data that is a member of the i cluster.

The K-Means method has the following characteristics [2]:

1. K-Means are very fast in the clustering process and are very sensitive to random generation of early centroids.
2. Allows a cluster to have no members.
3. The results of clustering with K-Means are unique (always changing), sometimes good or bad

2.2 Euclidean Distance

The Euclidean Distance method is used to calculate the distance between data and centroid. This measurement is based on the value of the object in each dimension in learning. Euclidean Distance can calculate the distance between data as much as two dimensions and more. Euclidean Distance formula with 2 objects and 2 dimensions [5]:

$$D_{pq} = \sqrt{(p_2 - p_1)^2 + (q_2 - q_1)^2} \quad (2)$$

Information:
Dpq = distance of 2 objects
X1 = the first dimension of the first object
X2 = the first dimension of the second object

2.3 Silhouette Coefficient

One of the evaluation methods used in the K-Means clustering is the silhouette coefficient method. This method serves to test the quality of the resulting cluster. This method is a cluster validation method that combines the cohesion and separation method [3]. Cohesion measures how closely related objects in a cluster. Separation measures how different a cluster is with other clusters.

To calculate the silhoutte coefficient value, distance between objects is required by using the Euclidean Distance formula. After that the steps to calculate the silhoutte coefficient values are as follows:

1. For each object i, calculate the average distance from object i with all objects in one cluster. An average value is called a (i).
2. For each object i, calculate the average distance from object i with the object in the other cluster. Of all distances the average takes the smallest value. This value is called b (i).
3. After that, for objects i have a silhoutte coefficient value:

$$\mathbf{s(i) = (b(i) - a(i)) / max(a(i), b(i))} \quad (3)$$

The results of the calculation of the silhoutte coefficient value can vary between -1 to 1. The clustering results are said to be good if the silhoutte coefficient value is positive (a (i) <b (i)) and a (i) close to 0, so that the maximum silhoutte coefficient value will be generated that is 1 when a (i) = 0. So it can be said, if s (i) = 1 means that object i is already in the right cluster. If the value of s (i) = 0, the object i is between two clusters so that the object is not clear must be included in cluster A or cluster B. However, if s (i) = -1 means that the cluster structure generated is overlapping, so object i is more accurately included in another cluster. The average value of the silhoutte coefficient of each object in a cluster is a measure that shows how tight the data is grouped in the cluster. The following is the silhoutte value based on Kaufman and Rousseeuw.

0.7 < SC <= 1	Strong Structure
0.5 < SC <= 0.7	Medium Structure
0.25 < SC <= 0.5	*Weak Structure*
SC <= 0.25	*No structure*

3. Result and Discussion

Tests conducted on the application of determining the top product and product recommendation consist of testing the module and testing the data. Testing of the module aims to test whether each module in the application is running well. Testing of data is a test carried out to find out whether the system runs according to the concept and whether the top product and recommendation product produced form the correct pattern.

Testing of the data consists of top product testing and product recommendation for the application of K-Means and evaluation of K-Means clustering. The data used for this test is real data from January to June 2014.

The testing of the K-Means clustering was carried out using 6 months of data, namely from January 1, 2014 to June 30, 2014. Sales transactions that occurred for 6 amounted to 327 and 51 customers in the transaction.

In this test, customer groupings are carried out in several conditions, namely group division into 2 clusters (k = 2), 3 clusters (k = 3), 4 clusters (k = 4), 5 clusters (k = 5), 6 clusters (k = 6), and 7 clusters (k = 7). The group division aims to see the distribution of customers from the smallest grouping to larger groupings and to see the distribution of top products and recommendation products that result from different groupings.

Determination of the minimum number of product appearances is based on experiments conducted before testing. This test sets a minimum number of product appearances is 5. This value has been adjusted for transaction data for 6 months. Running clustering is done for k = 2, k = 3, k = 4, k = 5, k = 6, k = 7. The initial seed (centroid) is taken in certain rows of data

One of the test results, namely grouping 2 clusters is shown in Table 1.

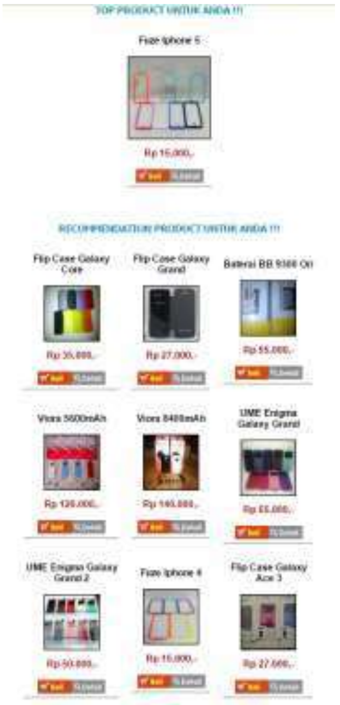
Tabel 1. *Clustering K-Means* 2 *Cluster*

Seed Awal		Cluster	Pelanggan	Top Product	Recommendation Product
Freq	Mone				
5	5	1	ANGELACC, TAKAPHONE, UNRI, GALLERYSMART, MUARACELL, LILY, HAPPYCELL, MEGAPHONE, PLANET, SURYACELL, KCELL, DJPONSEL, TWINS, BOYCELL, MAJESTY, SLICKSTONE, 88SELULARSHOP, PANCARASA, NOVEMBERPHONE, DNIZ, YUNICELL, KINGFORD, GOLDWIN, GALLERYPHONE, SIS, JOECELL, BLESS, ELLO, CHACHA, REVOLUTION	FUZE IPHONE 5	FLIP CASE GALAXY CORE, FLIP CASE GALAXY GRAND, BATERAI BB 9300 ORI, VIORA 5600MAH, VIORA 8400MAH, UME ENIGMA GALAXY GRAND, UME ENIGMA GALAXY GRAND 2, FUZE IPHONE 4, FLIP CASE GALAXY ACE 3
1	1	2	BBONE, SURYAPHONE, REDSTAR, GRAHAPHONE, JIITAPHONE, IACELL, MEGASHOP, RATUPONSEL, EZRA, PARADISESELULAR, PANCASELL, PRINCESS, NEYCELL, SPIRITPHONE, MULTICOM, STRAWBERRYCELL, MAKRO, MAGICCELL, ANDYCELL, ZENCOM, RUBYCELL	FUZE IPHONE 5	FLIP CASE GALAXY CORE, FLIP CASE UNIVERSAL 7", FLIP CASE GALAXY GRAND, FUZE IPHONE 4

Based on the test results in the table above, it can be seen that cluster 1 and cluster 2 are filled by a group of customers who have similar characteristics in terms of frequency and monetary. Cluster 1 is filled by 30 customers. This group of customers gets top products and product recommendations.

Cluster 1 is filled by 30 customers. This customer group gets the same top product and recommendation product. The top product shown is FUZE IPHONE 5. The product recommendations shown are FLIP CASE GALAXY CORE, FLIP CASE GALAXY GRAND, BATERAI BB 9300 ORI, VIORA 5600MAH, VIORA 8400MAH, UME ENIGMA GALAXY GRAND, UME ENIGMA GALAXY GRAND 2, FUZE IPHONE 4, FLIP CASE GALAXY ACE 3.

Cluster 2 is filled by 21 customers. This customer group gets the same top product and recommendation product. The top product shown is FUZE IPHONE 5. *The product recommendations shown are* FLIP CASE GALAXY CORE, FLIP CASE UNIVERSAL 7", FLIP CASE GALAXY GRAND, FUZE IPHONE 4.Based on the results of grouping on the table, it can be explained that cluster 1 is a group of customers with high frequency of shopping and total transactions. While cluster 2 is a group of customers with a low frequency of shopping and total transactions. Evidence of testing from clustering 2 clusters is shown in Figure 1.



TOP PRODUCT UNTUK ANDA !!!

Fuze Iphone 5

Rp 15.000,-

RECOMMENDATION PRODUCT UNTUK ANDA !!!

Flip Case Galaxy Core
Rp 35.000,-

Flip Case Galaxy Grand
Rp 27.000,-

Baterai BB 9300 Ori
Rp 55.000,-

Vizzo 5600mAh
Rp 120.000,-

Vizzo 8400mAh
Rp 145.000,-

UME Enigma Galaxy Grand
Rp 55.000,-

UME Enigma Galaxy Grand 2
Rp 50.000,-

Fuze Iphone 4
Rp 15.000,-

Flip Case Galaxy Ace 3
Rp 27.000,-

Figure 1. Top and Recommendation Product

Based on testing, it is known that the formation of 2 clusters and 3 clusters provides groupings of customers in each cluster. While in the formation of 4 clusters of up to 7 clusters, there are clusters that do not have members (customers).

3.1 Silhouette Coefficient Evaluation

K-Means evaluation using the Silhouette Coefficient method. The value of the silhouette coefficient is used to evaluate the cluster structure of k = 2 to k = 7. The average silhouette coefficient value and status structure of 2 clusters up to 7 clusters are shown in Table 2.

Tabel 2. Experiment result Average Silhouette Coefficient

Cluster	Average	Structure Status
2	0.55143678160487	Medium Structure
3	0.52291987994027	Medium Structure
4	0.3921899099552	Weak Structure
5	0.31375192796416	Weak Structure
6	0.26145993997013	Weak Structure
7	0.2241085199744	No Structure

Based on the table above, it is known that the two best average silhoutte coefficient values come from the evaluation of 2 clusters and 3 clusters that have the status of "Medium Structure". The status of "medium" indicates that in one cluster a group of customers is formed which has the same characteristics. Then between one cluster and another cluster has different customer characteristics.

Based on the evaluation results, candidates for the number of suitable clusters are 2 and 3 because they have the best structure status, namely medium. The value of the average silhouette coefficient 2 cluster is 0.55143678160487 while the 3 clusters are 0.52291987994027. This value does not show a significant difference. However, because the value of the average silhouette coefficient 2 cluster is higher, it is concluded that the clustering evaluation of data testing is 6 months (327 transactions from 51 customers) that the most suitable division of the number of clusters is 2.

4. Conclusion

The conclusion that can be obtained from the results of testing on the application of determining the top product and product recommendation are as follows:

1. Applications can display top products and product recommendations for each customer from the application of K-Means according to the specified settings.
2. The results of the silhouette coefficient evaluation of data testing show that the most suitable number of clusters is 2.

Suggestions for those who want to develop applications for determining top products and product recommendations on e-commerce with K-Means include:

1. Applications can be further developed by adding new features such as live chat between customers and operators and payment gateway systems to facilitate payment transactions.
2. Add another attribute for customer grouping.
3. The application of the K-Means method can be developed to create new features such as determining the best customer groups to be given shopping discount promotions.
4. Designing a product recommendation system for customers with other algorithms, such as the Collaborative Filtering algorithm.

5. References

[1] Irmawanti, Dewi, 2011, “Pemanfaatan E-Commerce Dalam Dunia Bisnis”, Jurnal Ilmiah Orasi Bisnis, Edisi VI.

[2] Han, Jiawei dan Kamber, Micheline. Data Mining Concepts and Techniques. 2nd Edition. San Francisco: Morgan Kaufmann, 2006.

[3] Ardha, “Metode Silhouette Coeffisien”, [Online] available : https://lookmylife.wordpress.com/2011/10/03/metode-silhoutte-coeffisien/ [6 June 2015].

[4] Alamsyah, 2008, “Pengertian E-Commerce”, Modul Introduction to E-Commerce, Universitas Gunadarma, Depok.

[5] J P Dias and H S Ferreira Automating the Extraction of Static Content and Dynamic Behaviour from e-Commerce Websites 297–304 ANT 2017 Procedia Computer Science 109C

[6] Luo Ya, 2012 The Comparison of Personalization Recommendation for Ecommerce 475 – 478 International Conference on Solid State Devices and Materials Science Physics Procedia 25

[7] Agusta, Yudi., 2007, “K-Means - Penerapan, Permasalahan dan Metode Terkait”, Jurnal Sistem dan Informatika, Volume 3.

[8] Ikhsan, M.; Dahria, M. dan Sulindawaty. “Penerapan Association Rule dengan Algortima Apriori pada Proses Pengelompokkan Barang di Perusahaan Retail”. Jurnal Sistem Komputer STMIK Triguna Dharma, Vol. 1, 2007