JIRAE, Vol. 3, No. 1, April 2018, 37–44
e-ISSN 2407-7259
DOI: 10.9744/JIRAE.3.1.37-44

The Application of Agglomerative Clustering in Customer Credit Receipt of Fashion and Shoe Retail

Michael Abadi Santoso[a], Budi Susanto[b], Gloria Virginia[c]

Faculty of Information Technology, Duta Wacana Christian University, Yogyakarta, Indonesia
[a]michael.abadi@ti.ukdw.ac.id, [b]budsus@ti.ukdw.ac.id, [c]virginia@ti.ukdw.ac.id

Abstract. Agglomerative Clustering is one of data mining methods to get a cluster in form of trees. In order to achieve these objectives, we used two agglomerative methods such as Single Linkage and Complete Linkage. Searching for nearest items to be clustered into one cluster also needs a similarity distance to be measured. We used Euclidean Distance and Cosine Similarity for measuring similarity distance between two points. The factors that promote high levels of accuracy depend on the pre-proceeding stage for clustering process and also affect the results obtained. Therefore, we conducted research through several stages: pre-processing such as ETL, normalization, and pivoting. The ETL process consisted of removing outliers using IQR method, data-cleaning and data-filtering processes. For normalization, we used Min-Max and Altman Z-Score methods to get the best normal value. The results of this research demonstrate that the highest accuracy occurs when using the Complete Linkage with Min-Max and the Euclidean method with the average purity of 0.4. The significant difference is observed when using the Z-Score and Cosine Similarity methods; the average purity is around 0.11. Besides, we found that the system also could not predict the customers' preferences in buying goods for the next period. Another result in the research is that transactional data in a company are not good enough to be clusterized.

Keywords: Agglomerative Clustering, Single Linkage, Complete Linkage, Data Warehouse, Data Mining.

1. Introduction

Nowadays in the business world, people constantly make efforts to develop their businesses by considering all aspects, which might affect decision-makings. One of the aspects, which contribute big impacts towards the stability of a company, is sales. In the retail business, customer preferences in purchasing products of a company in certain period of time can influence the balance of the company. By understanding customer preferences, a company is able to determine which goods are potentially sold to the consumers at a specific time.

One of the ways to analyze customer preferences is by making use of data mining. The data to be used in the research are big data from one of the big retail companies in Indonesia. The data itself consists of 17,927 transactional credit receipt data and 3,250 transactional cash receipt data from one branch. It is considered as big data because it has big volume and complex data sets. It also has 83 tables and multiple schemas, not only for determining customer preferences but also supporting the whole company's business process. They will be processed using data warehouse methods first before being mined to clear all of unnecessary tables and avoid slow query process in the future. The concept of the data warehouse is the ability to create query towards big data quickly. Sample of the case in real life is when thousands of transactional receipts are illustrated into their relational model of transaction, eventually it will create too many records which means it is not a good relational model to be applied in business intelligence. ETL (Extract, Transform, Load) is one of the ways to transform those data into data warehouse model. In the middle of the process, we need to make use of pivoting process towards those data to make sure we could read those data in the perfect way to be clusterized. One of the ETL processes is the pre-processing step such as clearing NULL field and setting the outlier for profiling the important data.

Clustering is one of the methods that will be used in this research. This method will identify objects that have similarity in some aspects, which then are called the centroid or characteristic vector. There are various methods which can be used in agglomerative such as Single Linkage, Complete Linkage, Average Linkage, Ward Method and Centroid Method. Clustering process will be conducted using Single Linkage and Complete Linkage. System will use data warehouse concept and business intelligence, starting from drawing the system's architecture, creating an ETL and dimensional model database, and pivoting the data. Then, the result will be issued in OLAP (Online Analytical Processing) form and the process of data clustering will begin.

2. Research Methods

2.1. Cluster Analysis

Cluster analysis is a method for grouping every object with a similar pattern into one or more groups [1]. The main purpose of cluster analysis is for grouping data with same characteristics. In clustering, data will be grouped depending on their similarity. If data are grouped into the same cluster, it means that they have short distance between each other. Hierarchical Agglomerative Clustering will be used in this research especially using Single Linkage and Complete Linkage.

2.2. Hierarchical Agglomerative Clustering

Hierarchical Agglomerative Clustering process starts with single clusters and progressively combines pairs of clusters, forming smaller numbers of clusters and it create a structural level like a tree [2]. The result of clustering can be visualized in the form of dendrogram. Dendrogram is a visual representation from every step in analyzing cluster and it is describing the relationships between all observations in a data set [2]. An object is connected to another by a single line in order to create one cluster. Dendrogram is very helpful for everyone because it helps people see cluster positions in a tree diagram, which shows us the relationship of objects in every cluster. Figure 1 is an example of dendrogram.

In Figure 1, the numbers on the bottom side represent objects, while the numbers on left indicate the distances between each other. Another version of dendrogram has a structure where the object is on the left and the distance is on the top.

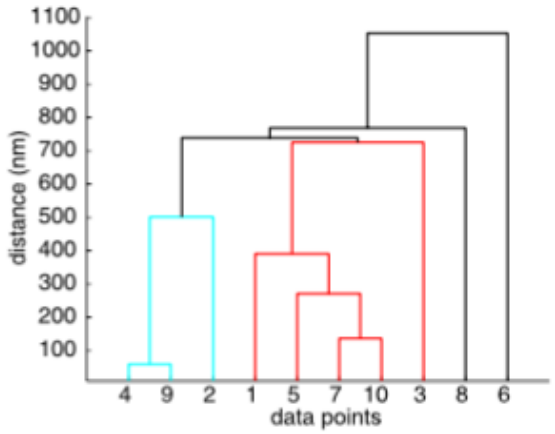

Figure 1. Example of dendrogram

Agglomerative method starts with the assumption that every object is a single cluster and then every two closest objects with the nearest distance will be grouped into one cluster. Following objects will be grouped as well with regard to the distance between the objects until all of the objects have been grouped. There are a lot of Agglomerative methods such as Single Linkage, Complete Linkage, Average Linkage, Centroid Method and Ward Method. Single Linkage and Complete Linkage will be used in this research.

2.3. Single Linkage

Single Linkage is a grouping method by finding the nearest distance between two clusters, while the distance is determined by the nearest pair of different clusters' data. Single Linkage method [3] is described in Eq. 1.

$$d(r,s) = \min\left(\text{dist}(x_{ri}, x_{sj})\right) \quad (1)$$
$$i \in \{1,\dots,i_r\} \text{ and } j \in \{1,\dots,j_s\}$$

where:

x = datum in any cluster
r, s = clusters being evaluated
i, j = i^{th} and j^{th} data of respective clusters r and s
dist = distance

Steps of Single Linkage are described as follows:

1. Decide k as the final number of clusters. User can set the final number of clusters that user wants and not limited to 1 cluster. Initially each datum is treated as an individual cluster.
2. Calculate distance for every two clusters by finding the minimum distance (i.e., iteratively calculated using Eq. 1).
3. Merge two clusters with the nearest distance. Given c = number of clusters after merging at some iteration.
4. If $c > k$, back to the third step.

2.4. Complete Linkage

Complete Linkage is a grouping method by finding the nearest distance between two clusters, while the distance is determined by the farthest pair of different clusters' data. Complete Linkage method [4] is described in Eq. 2.

$$d(r,s) = \max\left(\text{dist}(x_{ri}, x_{sj})\right) \quad (2)$$
$$i \in \{1,\dots,i_r\} \text{ and } j \in \{1,\dots,j_s\}$$

where:

x = datum in any cluster
r, s = clusters being evaluated
i, j = i^{th} and j^{th} data of respective clusters r and s
dist = distance

Steps of Complete Linkage are described as follows:

1. Decide k as the final number of clusters. User can set the final number of clusters that user wants and not limited to 1 cluster. Initially each datum is treated as an individual cluster.
2. Calculate distance for every two clusters by finding the maximum distance (i.e., iteratively calculated using Eq. 2).
3. Merge two clusters with the nearest distance. Given c = number of clusters after merging at some iteration.
4. If $c > k$, back to the third step.

2.5. Min-Max

The first step done before clustering the data is normalizing the data. Min-Max Normalization is one of the famous normalization methods. It is done by

deciding minimum and maximum boundaries of new normal data. Min-Max formula [5] is shown in Eq. 3.

$$X_i' = \frac{X_i - \text{min value of } A}{\text{max value of } A - \text{min value of } A} \times (D - C) + C \quad (3)$$

where:

X_i' = datum at i after being normalized
X_i = real datum at i
C = start range of normal data
D = end range of normal data
A = the dataset

2.6. Z-Score

Z-Score is another common normalization method. Normal value will be counted based on the mean value of each feature as well as each standard deviation. Z-Score formula [5] is shown in Eq. 4.

$$X_i' = \frac{X_i - \text{mean}(E)}{\text{std}(E)} \quad (4)$$

where:

X_i' = normalized datum at i
X_i = real datum at i
mean(E) = mean of dataset E
std(E) = standard deviation of dataset E
E = the dataset

2.7. Euclidean Distance

After normalizing the data, the next step is calculating similarity distance between all data pairs using one of similarity methods. Euclidean Distance is a similarity method which computes the square root of the sum of the square differences of data features. Euclidean Distance formula [6] is shown in Eq. 5.

$$\mathrm{D}(x_2, x_1) = \sqrt{\sum_{j=1}^{d} |x_{2_j} - x_{1_j}|^2} \quad (5)$$

where:

x_1, x_2 = the first and the second data
d = total number of dimensions
j = data dimension j

2.8. Cosine Similarity

Cosine Similarity is a method which measures similarity using the cosine of the angle between two data. In Cosine Similarity, dot multiplication between two vector data is computed before dividing it by the product of magnitudes of the two vectors. Cosine Similarity formula for two-dimensional data [6] is shown in Eq. 6.

$$\cos(C) = \frac{x_{1_j} x_{2_j} + x_{1_h} x_{2_h}}{\sqrt{{x_{1_j}}^2 + {x_{1_h}}^2} \times \sqrt{{x_{2_j}}^2 + {x_{2_h}}^2}} \quad (6)$$

where:

C = the angle between two data
x_1, x_2 = the first and the second data
j, h = features at j and h

2.9. Evaluation

Purity method will be used to measure the purity of the output. Purity is an accuracy evaluation of clustering. Purity can be measured by assigning a class for each cluster based on the most frequently occurring data and counting the correct data divided by the total number of data [7]. Purity formula is shown in Eq. 7.

$$\text{purity}(\Omega, \mathbb{C}) = \frac{1}{N} \sum_{k} \max_j |\omega_k \cap C_j| \quad (7)$$

where:

Ω = $\{\omega_1, \omega_2, ..., \omega_k\}$ is a set of cluster
$\mathbb{C}$ = $\{C_1, C_2, ..., C_j\}$ is a set of class
ω_k = a set of data in cluster ω_k
C_j = a set of objects in class C_j
N = total number of data

Binning is a method to obtain a data sample to calculate the purity. Binning is a smoothing and grouping method by looking at values around data and giving boundaries to each group [8]. Steps of binning method are described as follows:

1. Sort data from the smallest to the biggest.
2. Divide the data into a number of bins. There are two techniques to divide the data: equal-width (distance) partitioning and equal-depth (frequency) partitioning. The equal-width partitioning was used in this work.
3. Smoothing the data with three kinds of techniques, which are smoothing by bin-means, smoothing by bin-medians, and smoothing by bin-boundaries. In this work, smoothing by bin-boundaries was employed.

3. Result and Discussion

3.1. System Implementation

The system had been developed using Struts2 Java framework. Java Server Pages, JQuery and SemanticUI had been used for the front-end page. Besides that, Java had been used for the back-end.

3.2. Server Implementation

All of the Hadoop Ecosystem components had been implemented under Fedora 23 Linux Server 64-bit. Packets installed were Hadoop 2.72, Hive 2.0, Zookeeper 3.4.8, and Tez 0.9.2. The multinode cluster comprised two servers, with a namenode and two datanodes. The two servers were connected through a switch. One of the servers acted as the host server that managed all client requests. The server specification is as follows:

- Processor : Intel Core i3-2370M @ 2.40GHz
- Memory : 6GB RAM
- Harddisk : 500GB

Hive was used for SQL Operation and was implemented in the namenode. MapReduce is the data processing machine in Hadoop Ecosystem. The process

running under MR took much time because it was done after completing another process (i.e., writing to disc). Therefore, MR had been replaced with Apache Tez. Hiveserver2, instead of Hiveserver, was used for the Hive service process. Figure 2 is the system architecture in this work.

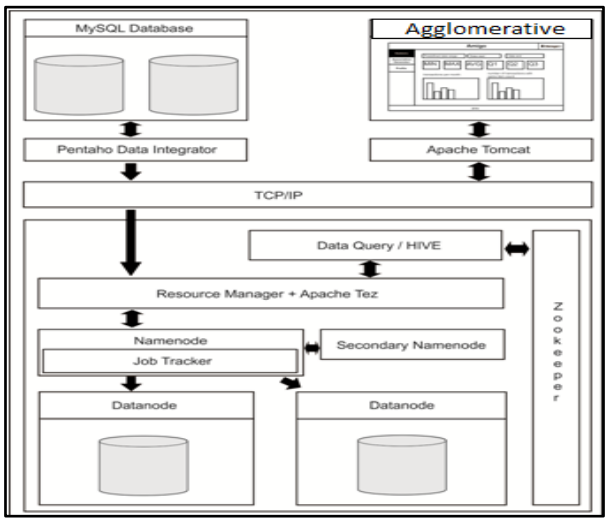

Figure 2. System architecture

The process flow itself is explained in Figure 3. MySQL database was the original database which stored all raw, original company data. All original data would go through filtering process which includes clearing the outlier before putting it into HDFS through Pentaho Data Integrator. All data in HDFS would be processed by Hive Map Reduce to filter the data based on conditional requirements that user can choose and determine in the application (i.e., choose the data period) to obtain the fact table. After all of the storing and mapping processes are done, user can normalize the data, calculate similarity distance, and create the clusters based on the requirements that the user desires. The output of the processes would be a scatter plot of the data that are already clusterized and a table that indicates the data itself.

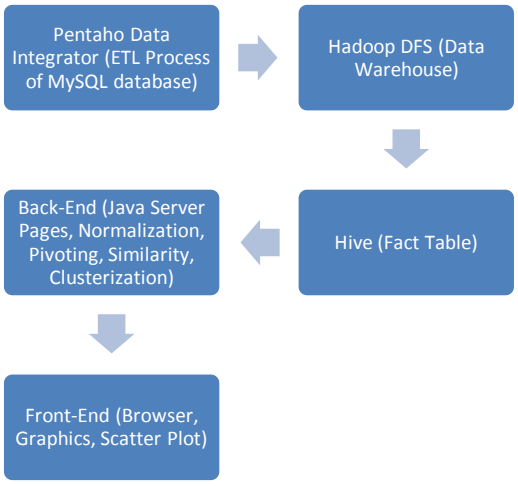

Figure 3. Process flow

3.3. Agglomerative Clustering Implementation

Implementation of the Agglomerative Clustering was in the package called mas.algorithm.com. It had some classes such as Cluster, AgglomerativeClustering, SuperCluster, and ClusteringItem. Steps of the Agglomerative clustering are as follows:

1. Putting similarity values and names of the data.
2. Creating Cluster objects which consist of ClusteringItem with the names and the data.
3. Executing an operation to create SuperCluster as long as the Cluster size is still above 1.
4. If SuperCluster consists of two Cluster items, system will look for the farthest/nearest distance between them depending on Single or Complete Linkage.
5. System will find the nearest distance from ClusterPair and create SuperCluster from the ClusterPair. SuperCluster will be added into the ArrayList and those two items from ClusterPair will be deleted from Cluster.

3.4. Choosing the Cluster Size

There are 7 data testers in the work. The data testers were collected from transactional credit receipts in the company that had been pivoted with total 5 attributes on each transaction to be compared to another transaction. Those attributes are age, total for Child items, total for Woman items, total for Man items, and total for Shoes. The data itself had passed ETL process such as profiling data (i.e., setting the outliers) and clearing the NULL field. All of the preprocessing steps needed to be done to avoid low query process and also to avoid any data that could made big differences in the result, such as outlier data. The data testers were divided into 7 data testers in different time durations, to check whether or not the purity changes a lot while the total number of data increases.

The first step in the work used 4 data testers (Tester I until Tester IV) to analyze purity with different cluster sizes (i.e., 12 and 16) and consequently, based on the result, choose the target cluster size. Next, the 7 data testers were used for the evaluation step with the target number of clusters. The selected cluster sizes of 12 and 16 are the fairest numbers to be used since, after grouping by binning methods, the data in the clusters do not have too small or too big differences between groups of ages. The 12 clusters can be a combination of 3 groups of ages and 4 types of items, whereas the 16 clusters can be a combination of 4 groups of ages and 4 types of items. Types of items are Shoes, Man Items, Women Items, and Children Items. In the work, Min-Max and Z-Score will be used for the normalization process while Euclidean Distance and Cosine Similarity will be used for the similarity measurement.

Method-I to Method-IV are combinations of normalization and similarity methods and they would be used for getting the most cohesive clusters based on the resulting purity. Method-I was a method with Z-

Score and Cosine Similarity, while Method-II used Z-Score and Euclidean Distance. Method-III used Min-Max and Cosine Similarity, whereas Method-IV used Min-Max and Euclidean Distance. Table 1 describes data testers used in this work, while Tables 2 and 3 show purity values from data testers. The data testers from Table 1 were ordered based on total data to make sure that even if the total data in a time period is larger than that in the other time period, they do not have big difference of purity values and instead produce a consistently similar value. Each time period was associated with a specific event which might affect the amount of purchases, such as New Year (December to January), Chinese New Year (February), Easter (March to May), Ramadan / Muslim Fasting Month (June to July), and no event (October to November). The 7 data testers were used for determining the consistent purity result of each method.

Table 1. Information about data testers (ordered by total data)

Name	Month	Total Data
Tester I	February 2012	213
Tester II	December 2012	259
Tester III	June 2012	267
Tester IV	October – November 2012	447
Tester V	March – May 2010	508
Tester VI	June – July 2012	622
Tester VII	March – May 2012	743

Table 2. Purity values from data tester I and II (M=Method, C=Cluster, SL=Single Linkage, CL=Complete Linkage)

Method	C	Tester I		Tester II	
		SL	CL	SL	CL
M-I	12	0.1830	0.1549	0.1621	0.1544
	16	0.1500	0.1079	0.1589	0.1081
M-II	12	0.1971	0.3333	0.2046	0.3590
	16	0.1830	0.2910	0.2046	0.2509
M-III	12	0.2394	0.0845	0.1891	0.0730
	16	0.2018	0.0840	0.1891	0.0960
M-IV	12	0.2523	0.4507	0.2123	0.3745
	16	0.2206	0.3568	0.1853	0.3050

Table 3. Purity values from data tester III and IV (M=Method, C=Cluster, SL=Single Linkage, CL=Complete Linkage)

Method	C	Tester III		Tester IV	
		SL	CL	SL	CL
M-I	12	0.1610	0.1198	0.1800	0.0707
	16	0.2022	0.0636	0.1559	0.0620
M-II	12	0.2209	0.2509	0.1897	0.2974
	16	0.2471	0.2696	0.1672	0.3247
M-III	12	0.2172	0.0900	0.2202	0.0530
	16	0.2134	0.1011	0.2491	0.0450
M-IV	12	0.2280	0.4419	0.2395	0.3601
	16	0.2546	0.4719	0.2652	0.3553

Purity values which had been computed are presented in Tables 2 and 3. In both tables, the underlined cells indicate the highest purity values of both cluster sizes of each data tester. All of them occur on M-IV. Based on M-IV's values, 75% of the highest results take place in the cluster size of 12 (3 underlined cells) and 25% in the cluster size of 16 (1 underlined cell). Based on the results, the cluster size of 12 was selected to determine the method's accuracy. Comparing the purity values of all methods (M-I to M-IV) in Tables 2 and 3, it can be concluded that the purity results are almost similar between Single Linkage and Complete Linkage.

3.5. Method Evaluation

Based on results on the previous step, the cluster size of 12 was used to carry out the clustering process. The resulting purity values after clustering are shown in Tables 4 and 5. The results are also drawn in Figure 4.

Table 4. Purity values with Single Linkage

Data	M-I	M-II	M-III	M-IV
Tester I	0.1830	0.1971	0.2394	0.2523
Tester II	0.1621	0.2046	0.1891	0.2123
Tester III	0.1610	0.2209	0.2172	0.2280
Tester IV	0.1901	0.2102	0.1812	0.2125
Tester V	0.1692	0.1732	0.1988	0.2106
Tester VI	0.1800	0.1897	0.2202	0.2395
Tester VII	0.1951	0.1965	0.2368	0.2355
AVG	**0.1772**	**0.1989**	**0.2118**	**0.2272**

Table 5. Purity values with Complete Linkage

Data	M-I	M-II	M-III	M-IV
Tester I	0.1549	0.3333	0.0845	0.4507
Tester II	0.1544	0.3590	0.0730	0.3745
Tester III	0.1198	0.2509	0.0900	0.4419
Tester IV	0.0805	0.3243	0.0980	0.4183
Tester V	0.1240	0.3543	0.0748	0.4173
Tester VI	0.0707	0.2974	0.0530	0.3601
Tester VII	0.0659	0.3109	0.0630	0.3647
AVG	**0.1100**	**0.3186**	**0.0766**	**0.4039**

In Figure 4, Method IV Complete Linkage is producing the best purity values from all data testers, followed by Method II Complete Linkage. Other methods give fluctuating results. Figure 5 is an example of plotted clusters using Method IV. Therefore, Method IV which is the combination of Min-Max (as the normalization method) and Euclidean Distance (as the similarity method) will be used later for analyzing customer preferences. However, those resulting purity values are still below 0.5, as shown in Tables 4 and 5. One of the reasons that cause the purity being below average is because the characteristics of data used were not good enough to be clusterized. The other reason is due to the sampling data for obtaining the purity. The sampling data used for data comparison made use of a binning method, and the binning method was not good

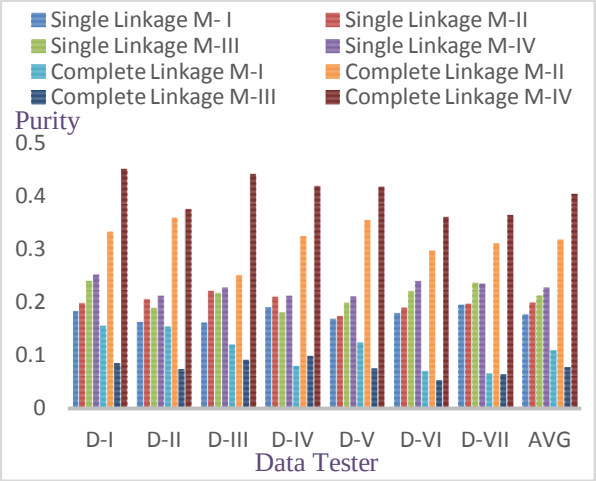

Figure 4. Comparison of purity values among data testers

enough to create the sampling data. The idea of using the binning method is grouping the data. By grouping the data, user could get real data to make the comparison to get the purity. The data being used was linear which is hard to be clusterized, and using binning method helps to make the linear data become more discrete. Sampling data is important for the work to get the purity result of the cluster made by the system. Despite that, the data used in this research is not categorized as good and suitable for clustering because the distribution of data values are still raw and have no special characteristics to describe similarity between each datum based on the features.

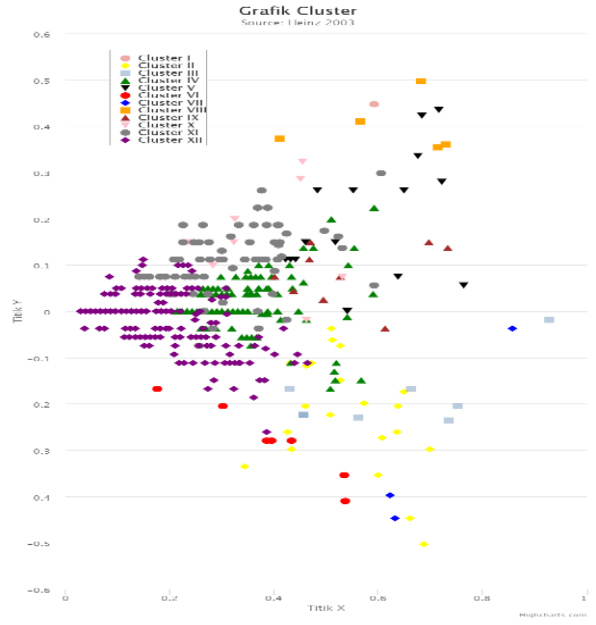

Figure 5. Scattered clusters of data tester V using Method IV

3.6. Customer Preferences on Buying Product Category Per Quarter Analysis

Seventy-two data testers, consisting of 36 credit receipt data and 36 cash receipt data, were used to analyze customer preferences. Each data tester has different amount of data, around 200 to 1200 per data tester. Credit receipt data are transactional sales data with credit card payment, whereas cash receipt data are transactional sales data with cash payment. All receipts were divided into 4 yearly quarters. Q1 was for the first quarter (January to March), Q2 the second quarter (April to June), Q3 the third quarter (July to September), and Q4 the fourth quarter (October to December). The first thing to do was to find the best pure cluster value. It was taken from the clusters with the minimum purity of 75% based on sampling data by the binning method and research data by the agglomerative method. The purity in this analysis was based on the data that really matched between manual clustering using the binning method and automatic clustering using the agglomerative method. Purity in this section was purity from the result of the binning method which was manual clustering by grouping the data based on ages and product types to get the sampling data. Tables 6 to 8 are the real products that matched or were correctly placed in the clusters produced by the binning method and the agglomerative calculation. The purity parameters in this section were coming from the binning method which had higher result than the purity in Section 3.5. Tables 6 to 8 are the results of customer preferences analysis for year 2010 to 2012 with the minimum purity of 75%. This purity resulted from comparison between the output clusters of the binning method and those of Section 3.5. For example, if there were 4 data in cluster A by the binning method and 30 data in the same cluster by the agglomerative method, and we found that only 2 data perfectly matched in cluster A, then the purity value was 50%. Type in Tables 6 to 8 is the cluster type based on data, which were grouped by the range of ages shown in Table 9 and 4 different types of retail products. The numbers in the tables are the total products that are in the correct cluster according to the clustering calculation with the minimum purity. The underlined cells are the dominant clusters in each quarter.

Table 6. Total products correctly placed in the cluster per product type per quarter in year 2010

TYPE	CREDIT				CASH			
	Q1	Q2	Q3	Q4	Q1	Q2	Q3	Q4
GOL-ICHILDREN	0	0	0	0	0	1	3	0
GOL-IWOMEN	0	0	10	12	1	0	0	0
GOL-IMEN	1	0	0	0	0	2	0	2
GOL-ISHOES	0	0	0	0	0	2	0	2
GOL-IICHILDREN	0	0	0	0	2	0	4	0
GOL-IIWOMEN	0	9	0	0	1	0	0	0
GOL-IIMEN	0	0	9	2	0	0	1	3
GOL-IISHOES	0	4	0	8	0	0	0	0
GOL-IIICHILDREN	0	0	0	0	3	1	0	1
GOL-IIIWOMEN	2	5	0	14	3	0	4	0
GOL-IIIMEN	2	0	0	0	1	3	0	1
GOL-IIISHOES	0	12	0	1	1	0	2	8

Table 7. Total products correctly placed in the cluster per product type per quarter in year 2011

TYPE	CREDIT				CASH			
	Q1	Q2	Q3	Q4	Q1	Q2	Q3	Q4
GOL-ICHILDREN	0	0	0	0	1	0	0	0
GOL-IWOMEN	0	0	0	0	0	3	1	7
GOL-IMEN	1	0	0	0	2	3	0	3
GOL-ISHOES	0	0	0	0	0	0	0	0
GOL-IICHILDREN	0	0	29	2	1	2	0	0
GOL-IIWOMEN	0	0	0	9	0	4	0	0
GOL-IIMEN	5	0	2	0	1	1	0	0
GOL-IISHOES	0	0	1	0	1	3	0	1
GOL-IIICHILDREN	0	6	8	0	0	0	1	0
GOL-IIIWOMEN	0	0	0	0	0	0	1	0
GOL-IIIMEN	0	0	0	0	0	1	0	0
GOL-IIISHOES	0	0	0	0	0	0	0	0

Table 8. Total products correctly placed in the cluster per product type per quarter in year 2012

TYPE	CREDIT				CASH			
	Q1	Q2	Q3	Q4	Q1	Q2	Q3	Q4
GOL-ICHILDREN	0	0	0	1	1	8	1	0
GOL-IWOMEN	50	0	53	0	0	2	5	6
GOL-IMEN	0	0	10	0	1	0	0	4
GOL-ISHOES	0	0	0	0	0	2	1	0
GOL-IICHILDREN	0	0	0	0	0	1	4	10
GOL-IIWOMEN	0	0	0	0	0	7	0	0
GOL-IIMEN	0	14	6	0	0	3	0	3
GOL-IISHOES	3	3	0	0	1	3	2	3
GOL-IIICHILDREN	0	5	0	3	16	1	0	0
GOL-IIIWOMEN	0	0	0	15	3	0	1	1
GOL-IIIMEN	0	4	3	0	0	0	0	0
GOL-IIISHOES	0	0	0	0	1	0	0	0

Table 9. Customer age group from 2010 – 2012 data

Customer Age Group	Range
GOL-I	10 – 22 years old
GOL-II	22 – 35 years old
GOL-III	> 35 years old

In the credit transaction Q1 year 2010, it has been found that the purchase of goods was dominated by the purchase of men and women type of goods, while in Q2 it was dominated by shoes type of goods, followed by Q3 and Q4 which were dominated by women type of goods. In credit transaction year 2011, Q1 was dominated by the purchase of men type of goods, followed by Q2 and Q3 which were dominated by children type of goods, and Q4 was dominated by women type of goods. In credit transaction year 2012, Q1 was dominated by the purchase of women type of goods, followed by Q2 which was dominated by men type of goods, while Q3 and Q4 were dominated by women type of goods. Between credit receipt and cash receipt transactions, similar domination of type of purchase goods in some quarters were located in Q1 year 2010 with the purchase of women type of goods, Q2 year 2010 with shoes type of goods, Q3 year 2010 with women type of goods, Q3 year 2011 with children type of goods, Q4 year 2011 with women type of goods, and Q3 year 2012 with women type of goods. Besides that, the same pattern in the yearly quarter can only be found in Q4 with women type of goods for credit transaction, while for cash transaction they can be found in Q1 with children type of goods and Q3 with women type of goods. In the work, it has also been found that age grouping shown in Table 9, based on the binning method, could not be predicted since age grouping was based on the value of minimum and maximum ages in each group. The age was available in the transactional data. Each transactional record had customer member information including age, address, and phone number. Data testers in the work were discovered to be unsuitable to help analyze customer preferences because the numbers of pure data in the clusters were too small.

4. Conclusion

Based on the research, it can be concluded that in order to do cluster analysis, preprocessing steps such as clearing all NULL fields and deciding the outliers of the data with IQR method need to be done to avoid low query and to avoid outlier data that could give big gap in the result. It has also been found that Complete Linkage, with average purity value of 0.4, has better purity value than Single Linkage. Based on the research result, Min-Max and Euclidean Distance are the combination of normalization and similarity methods which create better purity value. However, those purity values which were produced by the system were still below 0.5. The other conclusion is the system also could not predict the same dominant cluster for each yearly quarter since only 25% of the credit receipt data and 50% of the cash receipt data had the same pattern. The customer preferences between credit receipt and cash receipt data from a quarter to another quarter were different, as well. Since the data were not good enough, it can be concluded that the system also could not be used to help analyze customer credit receipt preferences in purchasing products in a certain period of time.

References

1. Wilk, J. and Pelka M., Cluster Analysis – Simbolic vs. Classical Data, *Acta Universitatis Lodziensis – Folia Oeconomica*, 286, 2013, pp. 205–213.
2. Sembiring, R.W., Zain, J.M., and Embong, A., A Comparative Agglomerative Hierarchical Clustering Method to Cluster Implemented Course, *Journal of Computing*, 2(12), Dec 2010, pp. 1–6.
3. Rashid, N.R, Sabri, A., and Safiek, M, A Comparison between Single Linkage and Complete Linkage in Agglomerative Hierarchical Cluster Analysis for Identifying Tourists Segments, *IIUM Engineering Journal*, 12(6), 2011, pp. 105–116.
4. Jena, A.P. and Paidi, A.N., Analysis of Complete-Link Clustering for Identifying and Visualizing

Multi-Attribute Transactional Data Using MATLAB, *IJCA Proc. of International Conference on Emergent Trends in Computing and Communication*, 2014, pp. 44–50.
5. Patro, S.G.K. and Sahu, K.K., Normalization: A Preprocessing Stage, *International Advanced Research Journal in Science, Engineering and Technology (IARJSET)*, 2(3), Mar 2015, pp. 20–22, doi: 10.17148/IARJSET.2015.2305.
6. Sasirekha, K. and Baby, P., Agglomerative Hierarchical Clustering Algorithm – A Review, *International Journal of Scientific and Research Publications (IJSRP)*, 3(3), Mar 2015.
7. Deepa, M. and Revathy, P., Validation of Document Clustering Based on Purity and Entropy Measures, *International Journal of Advanced Research in Computer and Communication Engineering (IJARCCE)*, 1(3), May 2012, pp. 147–152.
8. Patel, K., Bhojak, P., Shah, V., and Agrawal, V., Comparison of Various Data Cleaning Methods in Mining, *International Journal of Advance Research in Engineering, Science & Technology (IJAREST)*, 3(5), May 2016, pp. 918–923.