CLASIFICATION SYSTEM OF LIBRARY BOOK BASED ON SIMILARITY OF THE BOOK TITLE USING K-MEANS METHOD (CASE STUDY LIBRARY OF BHAYANGKARA SURABAYA)

[1]ARIF MARDI WALUYO, [2]EKO PRASETYO, [3]ARIF ARIZAL

[1,2,3]Informatics Department, Faculty of Engineering, Bhayangkara Surabaya University

Jl. Ahmad Yani No 114 Surabaya

E-mail: [1]rahulpamalia88@gmail.com, [2]eko@ubhara.ac.id, [3]qariff@ubhara.ac.id

ABSTRACT

In the grouping of book data in the library of Universitas Bhayangkara Surabaya at this time, the grouping is still based on the title and the existing field. So that resulted in the laying of some books whose title is not in accordance with the field of place. To facilitate the grouping of library books, in this research will provide a solution by doing the grouping of books based on the similarity of the title using K-Means method with the distance dissmilarity. The data are grouped a number of 500 titles in the library of Bhayangkara University Surabaya. The data will be processed through the Pre-processing process first of each book title by using the Information Retrieval System which results in the basic word. The basic word that will be used as a feature in the process of grouping so that can be known similarity. The result of the research is that it can be concluded that the application of Library Book Grouping System Based on Similarity of Book Title Using K-Means Method (Case Study of Bhayangkara Library Surabaya) is suitable for data that has been specified on each title. And some processes there are clusters that are always consistent in putting the book data in accordance with the similarity. Of all test results that have the best silhouette value is on using the value of K = 7, ie in the process to 1 with the value of silhouette = 0.2221

Keywords: *Information Retrieval System, K-Means, Library, Pre-processing, Silhouette.*

1. INTRODUCTION

Library is one place as a medium of learning through reading books that are provided in accordance with the field title of the book. The task of the library is to develop a collection of books, manage, and care for library materials, provide services as well as carry out the library administration. One example is like grouping books on every shelf that must be appropriate & neat, so it can facilitate visitors in the search book. If the library is well managed, it can be used as a place to read as well as learn comfortably for library visitors.

1.1. Backgroud

In the classification of book data in the library of Universitas Bhayangkara Surabaya at this time, the grouping is still based on the existing field. So that resulted in putting some books there whose title does not match with other titles. Things like this can make it difficult for visitors to find the book they are looking for. The previous researcher, Krisna Dwi Ananta, has been researching in the Final Project to create a grouping system of books in UBHARA library using K-Means method with Cosine Similarity distance. However, the results of the grouping system of the book is still not relevant or less suitable if applied. Because in the selection of attributes only based on the number & thickness of the existing book pages. This can cause multiple titles of books that have different genres mixed in one cluster.

Therefore, in this final project conducted a research entitled "Classification System of Library Book Based on Similarity of The Book Title Using K-Means Method (Case Study of Bhayangkara Library Surabaya)". In this research Pre-processing will be done in advance of each book title which will be used as a feature so it can be known similarity. It is expected that this research can produce a system for grouping books based on the similarity of the title of the book to facilitate library visitors in finding the desired book.

1.2. Formulation of the problem

Some of the main issues related to the research are as follows:

1) How do the process the Pre-processing of each book title to get the title resemblance using the InformationRetrieval System, which results will be used as a feature in the grouping process?
2) How to create library grouping system based on similarity of book title using K-Means method with Dissmilarity distance measurement?
3) How to measure cluster validation or grouping accuracy in each grouping result?

1.3. Limitations of the problem

Limitations of the problems in this study are as follows:

1) For grouping process use K-Means method & use Dissmilarity distance measurement.
2) The feature used for grouping is the basic word processed with InformationRetrieval System technique as Pre-processing of the title of the book.
3) The data used as a grouping of only the title of the book contained the Indonesian language.
4) Data taken only the title of the book from the library UBHARA.

1.4. Research Purposes

This research has the following objectives::

1) Create an application that can do the grouping book based on the similarity of the title of the book in UBHARA library.
2) To analyze whether it is appropriate to use InformationRetrieval System as a pre-processing in order to get a base word that will serve as a feature in K-Means grouping and use Dissmilarity distance measurement.

2. THEORETICAL BASIS

Theoretical basis contains the theories that support in making research and system.

2.1. Data Mining

Tan (2006) defines data mining as a process for obtaining useful information from large database warehouses. Data mining can also be interpreted as extracting new information derived from large data chunks that aid in decision making. The term data mining is sometimes called knowledge discovery. Some of the techniques that are often mentioned in the literature of data mining in its application include: clustering, classification, association rule mining, neual network, genetic algorithm and others. What distinguishes perceptions of data mining is the development of data mining techniques for applications on large-scale databases. Before the popularity of data mining, these techniques can only be used for small-scale data.

2.2. K-Means Algorithm

K-Means is one method of nonhierarchy data grouping (sekatan) that seeks to partition existing data into two or more groups. This method partitions the data into groups so that the same characteristic data is entered into the same group and the different characteristic data are grouped into the other group. Understanding K-Means in this Final Project is referenced from Data Mining Concept and Application book using Matlab (Eko Prasetyo). The purpose of this data grouping is to minimize the objective function set in the grouping process, which generally attempts to minimize variation within a group and maximize variation between groups.

Grouping of data by K-Means method is generally done with the following algorithm:

1. Determine the number of groups.
2. Allocate data into groups at random.
3. Calculate the center of the group (sentroid / average) of the data in each group.
4. Allocate each data to the nearest centroid / average.

5. Return to step 3, if there is still data moving group, or if there is a change of centroid value above the specified threshold value, or if the value change on the objective function used is still above the specified threshold value.

In step 3, the centroid location of each group taken from the mean (mean) of all data values on each feature must be recalculated. If M denotes the amount of data in a group, i denotes the i feature in a group, and p denotes the data dimension, to compute the i-feature feature centroid. The formula is done as much as p dimensions so that i starts from 1 to p, the formula is as follows.

$$C_i = \frac{1}{M}\sum_{j=1}^{M} x_j \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (1)$$

In step 4, the re-allocation of data into each group in the K-Means method is based on the comparison of the distance between the data with each group's centroid. The data is reallocated explicitly to the group that has the centroid with the closest distance from the data. This allocation can be formulated as follows (MacQueen, 1967).

$$a_{i1} = \begin{cases} 1 \ \arg\min\{d(\mathrm{x_i},\ C_1)\} \\ 0 \ \text{other} \end{cases} \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (2)$$

a_{i1} is the value of the membership of point x_i to the center of group C_1, d is the shortest distance from the x_i to K group data after being compared, and C_1 is the 1[st] centroid.

The objective function used for K-Means is determined by the distance and value of the data membership in the group. The objective function used is as follows (MacQueen, 1967).

$$J = \sum_{i=1}^{N}\sum_{l=1}^{K} a_{ic} D(x_i\,, C_l)^2 \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (3)$$

N is the amount of data, *K* is the number of groups, a_{il} s the membership value of the data point x_i to the center of the group C_l, C_l is the center of the l[st] group, and D (x_i, C_l) is the point distance x_i to the C_i group followed. a has a value of 0 or 1. If a data is a member of a group, the value a_{il} = 1. Otherwise, the value of a_{il} = 0.

2.3. Distance Measurement

2.3.1. *Cosine Similarity*

In this method the term weight contained in each document is presented in a vector. For example document a is presented by vector a = { $a_1,\ a_2,\ a_3, \ldots\ a_t$ } and document b is presented by vector b = { $b_1,\ b_2,\ b_3, \ldots\ b_t$ }. This correlation can be quantified by the cosine angle between two vectors in the equation below:

$$Similarity\ = \cos(\theta)\frac{A.B}{||A||.||B||} = \frac{\sum_{i=1}^{n} Ai\ x\ Bi}{\sqrt{\sum_{i=1}^{n}(Ai)^2}\ x\ \sqrt{\sum_{i=1}^{n}(Bi)^2}} \ldots\ldots\ldots\ldots (4)$$

With | a | and | b | is the norm a and b, the value of sim (a, b) varies from 0 to 1. This value indicates that the higher the value (a.b) the greater the similarity of the two vectors.

2.3.2. *Dissimilarity*

Dissimilarity is a numerical degree in which two objects are different, their range is 0 to 1, or even to ∞. This distance is used to calculate in k-means contained in equation (1) and equation (3). If similarity is a similarity measure then dissimilarity is a measure of unlike, and if the interval is [0,1], then dissimilarity can be formulated as follows.

$$Dissimilarity = 1 - \ S \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (5)$$

With S is the value of the similarity result, which is found in (4).

2.4. IRS (*Information Retreival System*)

InformationRetrieval System, or in Indonesian language Information Retrieval System is a field at the intersection of information science and computer science.

The steps that occur in the process of indexing subsystem are as follows:

1. Word Token / Parsing, is a process for dividing text that can be a sentence, paragraph or document, into tokens / certain parts. Typically, the token separator between the tokens is spaces and punctuation.
2. Stopword Removal / Filtering, is the process of deletion or disposal of words that do not provide important information, common words in a document and conjunctions, such as: and, or, but, which, whereas and so on.
3. *Stemming, is a process contained in InformationRetreival system that transforms the words contained in a document to the root word by using certain rules.*

2.5. Cluster Accuracy Size

There are several measures of accuracy to know the quality of a grouping. One measure of precision that can be used is the coefficient of Silhouette, can be formulated as:

$$s(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}}$$

Where a (i) is the average distance of data incorporated in the cluster. b (i) is the average of the data distance to the other cluster the smallest.

3. ANALYSIS AND DESIGN

3.1. System Analysis

In this system will be pre-procecing first to each title of the book will be in the feature so that it can be known similarity. The data of this book is taken from UBHARA library database in the form of excel format used as a guide in this research and processed in making clustering application.

Book data that the author receives still must be processed first to fit the application made. Because seen from the similarity of the title then it must be processed first through the IRS (InformationRetrieval System). The process is to take the title of the book and then in tokenisasi function to retrieve each word, then processed with a stopword is to delete the words or words that are not needed, the last is the process of stemming the process to restore the word mejadi rood word or basic word. This last result is later used as a feature in the grouping process. The following will be given a sample process for processing up to the process of grouping.

3.2. System Planning

provide an overview of system design to be built or developed, and to understand the flow of information and processes in the system.

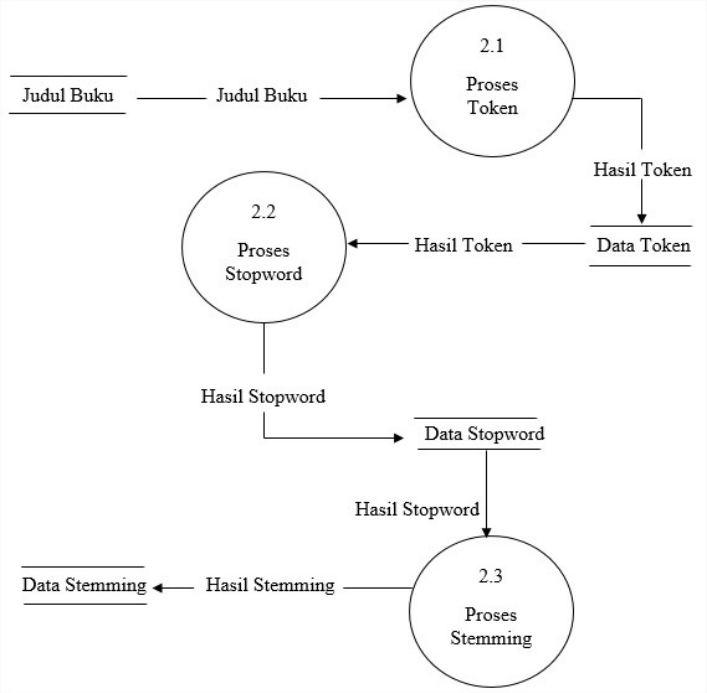

Image 1 Display DFD Preprocessing

Description of image 1 DFD levelPreprocessing using informationretrieval system & following its processing::

1. Token process, retrieve every word & delete unimportant characters.
2. Stopword process, delete the connect / word that often appears.

The stemming process, returning to the base word which will be used as a feature for grouping process.

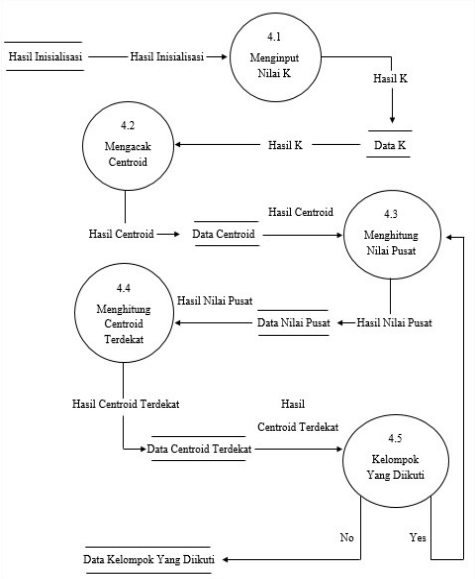

Image 2 Grouping DFD Views

Explanation of image 2 DFD level 2 Grouping is as follows:

1. Then the result of the initialization will be clustering, the process of grouping the book using k-means method.
   a. Determine the value of K, the number of groups.
   b. Scramble the initial centroid value.
   c. Count the center of the group (centroid / average) of the data in each group.
   d. Allocate each data to the nearest centroid / average.
   e. Return to step c if there is still data moving group or if there is a change in the value of the objective function used is still above the specified threshold value.

4. RESULTS & DISCUSSIONS

The experiment was done 3 times of testing with the value of K each 5, 6, and 7. Use data of 500 titles of books. After the pre-processing of 500 data obtained the number of basic words as much as 550 words that will be used as a feature in the grouping. From the data is already divided into 6 keywords, namely: Economics, Management, Accounting, Law, Engineering, Communication. This division of keywords that will later find out from the grouping results whether the title is categorized according to the same keywords.

4.1. Trial Results With 5 Clusters

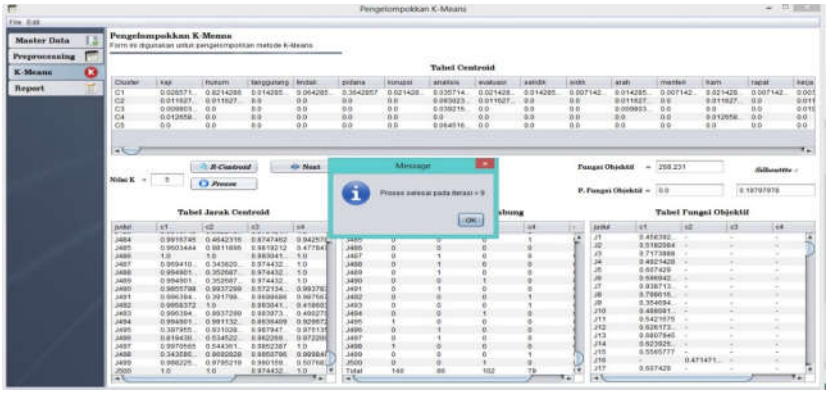

Image 3 Process 5 Clusters

From the grouping results in figure 3 can be analyzed for the keyword similarity is:

- C1 = With the amount of 140 data, Legal keywords as much as 120 data = 85.71% and Technique keywords as much as 20 data = 14.28%. So C1 into the most groups of Law..
- C2 = With the amount of 86 data, Economic keyword as much as 68 data = 79.06% and Technique keywords as many as 18 data = 20.93%. So C2 into the largest group of Economics.
- C3 = With the amount of 102 data, Keywords Communications as many as 66 data = 64.70% and Engineering keywords as much as 36 data = 35.29%. So C3 into the largest group of Communication.
- C4 = With the amount of 79 data, Accounting keywords as much as 75 data = 94.93% and Technique keywords as much as 4 data = 5.06%. So C4 into the most group of Accounting.
- C5 = With the amount of 93 data, Keyword Accounting as much as 76 data = 81.72% and Technique keywords as many as 17 data = 18.27%. So C5 entry into the most group of Accounting.
- From one process the precise value of silhouette = 0.1980 is obtained

4.2. Trial Results With 6 Clusters

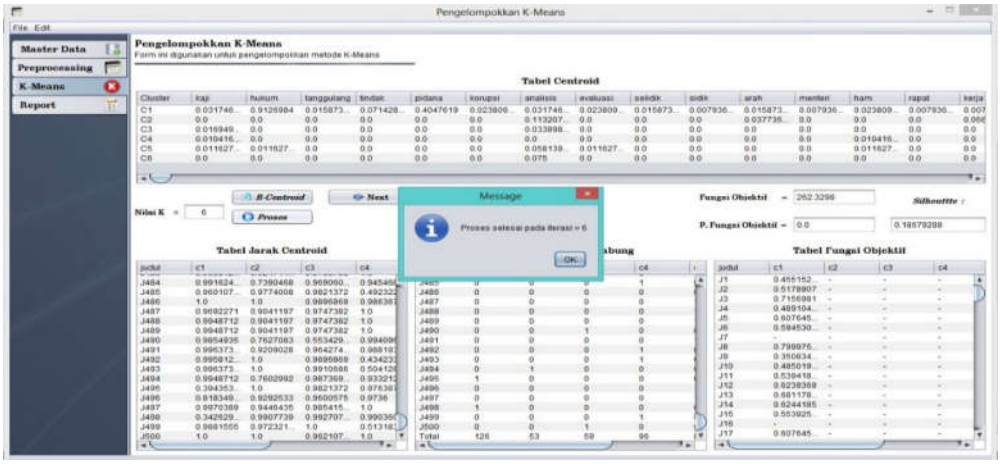

Image 4 Process 6 Clusters

From the result of grouping in the figure 4 can be analyzed for the keyword similarity is:

- C1 = With a total of 126 data, Legal keywords as many as 120 data = 95.23% and Technique keywords as much as 6 data = 4.76%. So C1 into the most groups of Law.
- C2 = With the amount of 53 data, Keywords Communications 27 data = 50.94% and Technique keyword as much as 26 data = 49.05%. So C2 into the largest group of Communication.
- C3 = With the amount of 59 data, the keyword Communication as much as 45 data = 76.27% and Technique keywords as much as 14 data = 23.72%. So C3 into the largest group of Communication.
- C4 = With the amount of 96 data, Accounting keywords as much as 77 data = 80.20% and Technique keywords as much 19 data = 19.79%. So C4 into the most group of Accounting.
- C5 = With the amount of 86 data, economic keyword as much as 62 data = 72.09% and Technique keyword as much as 24 data = 27.90%. So C5 into the largest group of Economics.
- C6 = With the amount of 80 data, Keyword Management as much as 74 data = 92.5% and Technique keywords as much as 6 data = 7.5%. So that C6 into the largest group of Management.
- From one process the precise value of silhouette = 0.1868 is obtained.

4.3. Trial Results With 7 Clusters

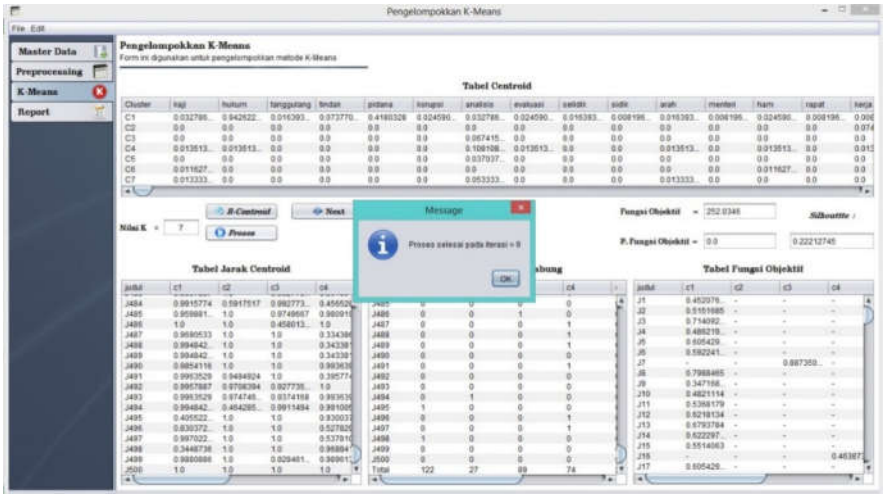

Image 5 Process 7 Clusters

From the results of the groupings in figure 5 can be analyzed for the keyword similarity is:

- C1 = With the amount of 122 data, Legal keyword as much as 119 data = 97.54% and Technique keyword as much as 3 data = 2.45%. So C1 into the most groups of Law.
- C2 = With 27 data, Technique keyword is 27 data = 100%. So C2 into the most group of Techniques.
- C3 = With the amount of 89 data, Keyword Management as much as 72 data = 80.89% and Technique keywords as many as 17 data = 19.10%. So C3 into the largest group of Management.
- C4 = With the amount of 74 data, Economic keyword of 19 data = 61.29% and Technique keywords as much as 6 data = 8.10%. So C4 into the largest group of Economics.
- C5 = With the amount of 27 data, the keyword Economy as much as 1 data = 3.70% and Technique keywords as much as 26 data = 96.29%. So C5 into the most group of Techniques.
- C6 = With the amount of 86 data, Accounting keywords as many as 77 data = 89.53% and Technique keywords as much as 9 data = 10.46%. So C6 into the most group of Accounting.
- C7 = With the amount of 75 data, Keywords Communications as much as 67 data = 89.33% and Technique keywords as many as 8 data = 10.66%. So C7 into the largest group of Communication.
- From one process is obtained preciseness value of silhouette = 0.2221.

5. CONCLUSIONS AND SUGGESTION

Based on the research that has been done can be drawn conclusion, among others:

1. Based on the test results, it is concluded that in the study of Library Books Grouping in UBHARA Based on Similarity of Book Title Using K-Means Method suitable for use on data that has been determined the key words to each title and some processes there are clusters that are always consistent in putting data books according to the similarity. Of all test results that have the best silhouette value is on the use of value K = 7, ie in the process to 1 with the value of silhouette = 0.2221.
2. If it produces a clustering of data in each cluster that does not have a resemblance title, it is because there are some titles that do not have similarities with other titles so that the data will be incorporated into clusters that have formed data that already has a resemblance title or will form a cluster new again.
3. Factors determining the value of K is very important in the process of groupingkarena if not in accordance with the data will be in the group will make the dissemination of data that does not have the same title will be incorporated in each cluster.

There are suggestions for further research including:

1. If the pre-processing of all book titles used as constraint features is present at processing time. Because if more and more data books will be more and more features are formed. This will make processing time in K-Means grouping longer.
2. For further research it is necessary to add pre-processing process to English so that the title of books using English as a whole can be grouped.

6. REFERENCES

[1] Agusta, Ledy. 2009. PerbandinganAlgoritma Stemming Porter DenganAlgoritmaNazief&AdrianiUntuk Stemming DokumenTeksBahasa Indonesia. KonferensiNasionalSistemdanInformatika 2009. Bali.
[2] Ananta, K. D. 2013. PengelompokkanBukuPerpustakaan UBHARA MenggunakanMetode K-Means DenganPengukuran Cosine Similarity. UBHARA. Surabaya.
[3] Hastie, T. et al. 2001. The Elements of Statistical Learning: data mining, interface, and prediction. New York: Springer- Vergal.
[4] Muhhammad, Ardiansyah. PenggunaanJarak Dynamic Time Warping (DTW) PadaAnalisis Cluster Data DeretWaktu(StudiKasusPada Dana PihakKetigaProvinsi Se-Indonesia). Malang.
[5] Ong, Johan Oscar. 2013. Implementasi Algoritma K-Means Clustering Untuk Menentukan Strategi Marketing President University.JurnalIlmiahTeknikIndustri. Bekasi.
[6] Prasetyo, Eko. 2012. DATA MINING – KonsepdanAplikasiMenggunakanMatlab, ANDI, Yogyakarta.
[7] Rismawan, T. danKusumadewi, S. 2008. Aplikasi K-Means UntukPengelompokkanMahasiswaBerdasarkanNilai Body Mass Index (BMI) &UkuranKerangka. SNATI. Yogyakarta.
[8] Rousseeuw, Peter J. 1987. Silhouttes : a graphical aid to the interpretation and validation of cluster analysis. North-Holland.
[9] Suprihatin. 2011. Klastering K-Means untukpenentuanNilaiUjian. Universitas Ahmad Dahlan. Yogyakarta.
[10] Sutarno NS. 2003. "PerpustakaandanMasyarakat". Jakarta: YayasanObor Indonesia, p.7.
[11] Syafrianto, Andri. 2012. PerancanganAplikasi K-Means UntukPengelompokkanMahasiswa STMIK ELRAHMA Yogyakarta BerdasarkanFrekuensiKunjungankePerpustakaandan IPK. Yogyakarta.
[12] Tala, Fadillah Z. 2003. A Study Of Stemming Effects On Information Retreival In Bahasa Indonesia. Master of Logic Project. Institute for Logic, Language and Computation Universiteit van Amsterdam. The Netherlands.