IT FOR SOCIETY, Vol. 03, No. 02
ISSN 2503-2224

Documents Clustering Using K-Means Algorithm

R.B. Wahyu*, Arnold Vito Ezra Panjaitan
President University
*Corresponding author: rbw0101@gmail.com

Abstract *- Nowadays in the digital era, people could easily access and stored a wide range of information through the Internet into documents. With the huge number of unstructured documents with various type of information in digital storage, people need an application that could help them organize and classify the documents automatically. Documents Clustering using K-Means Algorithm is a desktop-based documents clustering application which implement K-Means Algorithm to provides clustering output based on the documents content similarity up to 85% accuracy based on the user expectation.*

I. Introduction

With the ease of getting various kind of information in this digital era from the Internet, people start to live with paperless environment – not buying newspapers, magazines or books. Many of the information stored from the Internet will be useful for further occasion when it is needed such as for research, literature study, supporting ideas, or references. These various collections of information needed to be categorized to make them easily accessed based on their contents.

Although there have been several clustering applications developed, none has been clustering the documents and then grouped the documents into folders based on the documents category as accurate as possible. The existence of such application will greatly help people in classifying their scattered documents into a set of folders which contains similar information one to another internally, but substantially different with other folders.

The major problems are: 1. Currently there is no such application that implement the clustering results into grouping the documents into folders 2. Retrieve the information of each documents, 3. Implement K-Means Algorithm for clustering the documents, 4. The accuracy of clustered documents using K-Means Algorithm. These can be solved by having an application to clusters and categorized the documents using K-Means Algorithm.

The objectives of this research are:

- To create a desktop-based documents clustering application using K-Means Algorithm application
- To implement the clustering result from K-Means Algorithm into grouping the documents into folders based on their contents similarity

Limitations

This application uses an English stop words collection; thus, the application could only read documents which its content is in English.

All documents to be processed must be in text format such as .txt, .pdf and .doc documents.

The clustering result does not guarantee 100% accuracy based on the user expectation because the application will merely learn and clustered the documents based on the documents content similarity one to another.

Other limitations are: the processing time is heavily depending on the number of documents, the documents content length and number of clusters given.

II. Methods

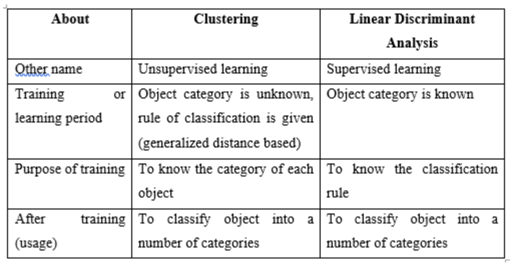
This section explains the methods used in the development of the application.

1.1 Clustering

Clustering is an unsupervised learning technique to group objects based on their distance and/or similarity. It is called as an unsupervised learning technique because clustering let the machine (computer or program) learns mere from objects given, then the machine will automatically categorize these objects into which group [19].

In clustering, we only specify the rule of clustering method of how the machine will compute the distance between objects and compute the distance between clusters. The purpose of performing clustering technique is to know how the objects will be categorized based on the rule given to the machine.

Table 1 Clustering vs Linear Analysis [15]



About	Clustering	Linear Discriminant Analysis
Other name	Unsupervised learning	Supervised learning
Training or learning period	Object category is unknown, rule of classification is given (generalized distance based)	Object category is known
Purpose of training	To know the category of each object	To know the classification rule
After training (usage)	To classify object into a number of categories	To classify object into a number of categories

1.2 Document Clustering

Documents clustering is an action of collecting the similarities of documents contents, then automatically categorized those documents with their contents similarities. Its goal is to learn automatically the category of every documents then create clusters that similar internally but substantially different with the other clusters [7].

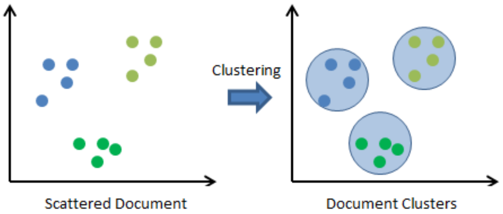


Figure 1 Documents Clustering Illustration [9]

1.3 K-Means Algorithm

K-Means algorithm is a type of unsupervised learning algorithm developed by J. Mac Queen (1967) to find groups for a set of data based on their attribute with the number of groups as a cluster represented by variable K [20]. Variable K representing a positive integer number which is used to determine how many cluster's centroids will be used to classify the data, in which the centroid represent the value of the center of each cluster.

Generally, k-means algorithm works by taking a unique attribute from each data given as a data point, then assign these data points to the nearest cluster centroid, after that the centroids will begin moving to better fit the clusters themselves. These data points and centroids will keep moving until the algorithm find the minimum fitting error between the data sets and centroids by updating the centroids to be the mean value of the clusters repeatedly in each loop until no movement needed anymore [23].

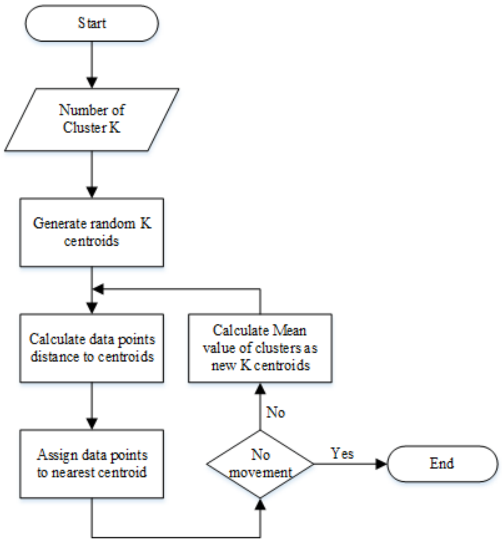

Start
Number of Cluster K
Generate random K centroids
Calculate data points distance to centroids
Calculate Mean value of clusters as new K centroids
No
Assign data points to nearest centroid
No movement
Yes
End

Figure 2 Basic K-Means Algorithm Flowchart [18]

1.4 Information Retrieval

Information retrieval is an activity of obtaining relevant information from a large amount of information databases [6]. Information retrieval help the algorithm to identify the relevant information that represent each document. By having the relevant information representation of each document, the document clustering algorithm will be able to provide a qualified clustering output of the documents.

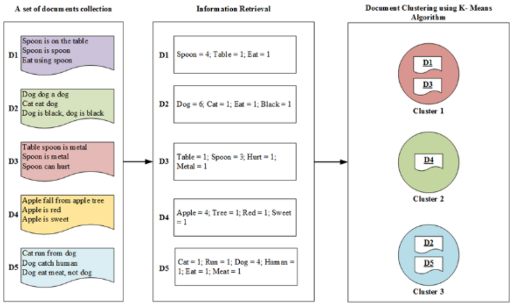

A set of documents collection
Information Retrieval
Document Clustering using K- Means Algorithm
D1
Spoon is on the table
Spoon is spoon
Eat using spoon
D2
Dog dog a dog
Cat eat dog
Dog is black, dog is black
D3
Table spoon is metal
Spoon is metal
Spoon can hurt
D4
Apple fall from apple tree
Apple is red
Apple is sweet
D5
Cat run from dog
Dog catch human
Dog eat meat, not dog
D1 Spoon = 4; Table = 1; Eat = 1
D2 Dog = 6; Cat = 1; Eat = 1; Black = 1
D3 Table = 1; Spoon = 3; Hurt = 1; Metal = 1
D4 Apple = 4; Tree = 1; Red = 1; Sweet = 1
D5 Cat = 1; Run = 1; Dog = 4; Human = 1; Eat = 1; Meat = 1
D1
D3
Cluster 1
D4
Cluster 2
D2
D5
Cluster 3

Figure 3 Information Retrieval role in Document Clustering

1.5 Terms Frequency (TF) – Inverse Document Frequency (IDF)

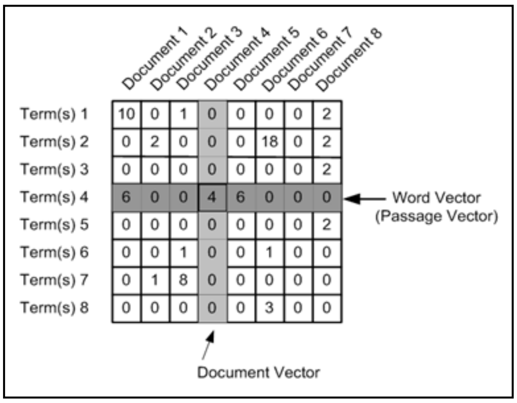
Terms Frequency (TF) is the numerical measurement of how frequent a word appears in a document [22], which will show how important a word to a document.

$$TF_{(t)} = log_{10}\frac{Number\ of\ t\ appearence\ in\ a\ document}{document\ length}$$

Inverse Document Frequency (IDF) is the statistical weight measurement of how important a word to a set of documents collection [22]. IDF diminishes the weight of words that appears very frequently in a set of documents collection, but it scales up the rare ones.

$$IDF_{(t)} = \frac{Total\ documents\ in\ the\ collection}{Number\ of\ documents\ with\ word\ t\ in\ it}$$

Basically, the ways to get the relevant information of each document using Information Retrieval method is by getting the TF-IDF value of each document information. After having the TF-IDF of each word among the documents, the documents will be represented as a mutual comparable vector with the intention to have the numerical model of each document. Each document vector will be represented using the TF-IDF result of each term (word) from each document [23].



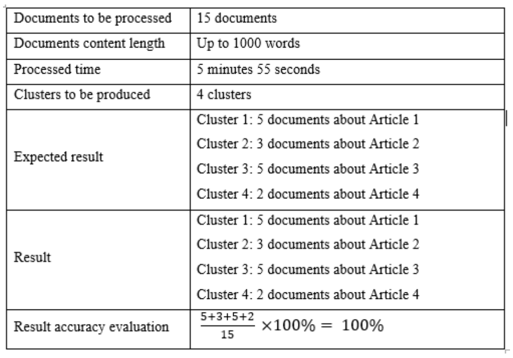
	Document 1	Document 2	Document 3	Document 4	Document 5	Document 6	Document 7	Document 8
Term(s) 1	10	0	1	0	0	0	0	2
Term(s) 2	0	2	0	0	0	18	0	2
Term(s) 3	0	0	0	0	0	0	0	2
Term(s) 4	6	0	0	4	6	0	0	0
Term(s) 5	0	0	0	0	0	0	0	2
Term(s) 6	0	0	1	0	0	1	0	0
Term(s) 7	0	1	8	0	0	0	0	0
Term(s) 8	0	0	0	0	0	3	0	0

Figure 4 TF-IDF vector space model [12]

III. Experimental Results

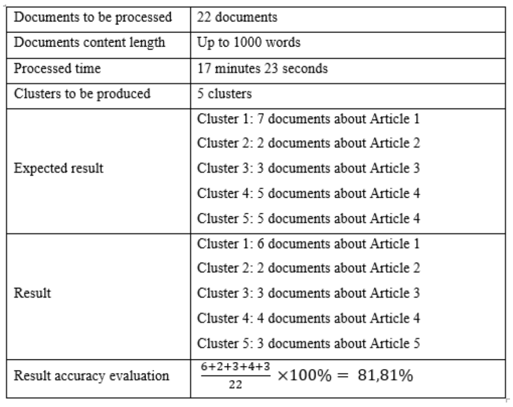
Tests are done in order to evaluate the effectiveness of the proposed methods in previous section. Test results are described in tables below:

Table 2 K-Means Performance 1st Evaluation



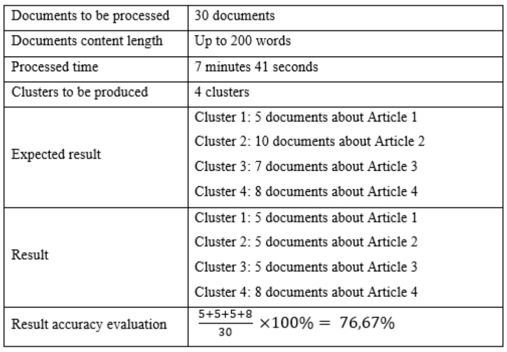
Documents to be processed	15 documents
Documents content length	Up to 1000 words
Processed time	5 minutes 55 seconds
Clusters to be produced	4 clusters
Expected result	Cluster 1: 5 documents about Article 1 Cluster 2: 3 documents about Article 2 Cluster 3: 5 documents about Article 3 Cluster 4: 2 documents about Article 4
Result	Cluster 1: 5 documents about Article 1 Cluster 2: 3 documents about Article 2 Cluster 3: 5 documents about Article 3 Cluster 4: 2 documents about Article 4
Result accuracy evaluation	$\frac{5+3+5+2}{15} \times 100\% = 100\%$

Table 3 K-Means Performance 2nd Evaluation



Documents to be processed	22 documents
Documents content length	Up to 1000 words
Processed time	17 minutes 23 seconds
Clusters to be produced	5 clusters
Expected result	Cluster 1: 7 documents about Article 1 Cluster 2: 2 documents about Article 2 Cluster 3: 3 documents about Article 3 Cluster 4: 5 documents about Article 4 Cluster 5: 5 documents about Article 4
Result	Cluster 1: 6 documents about Article 1 Cluster 2: 2 documents about Article 2 Cluster 3: 3 documents about Article 3 Cluster 4: 4 documents about Article 4 Cluster 5: 3 documents about Article 5
Result accuracy evaluation	$\frac{6+2+3+4+3}{22} \times 100\% = 81{,}81\%$

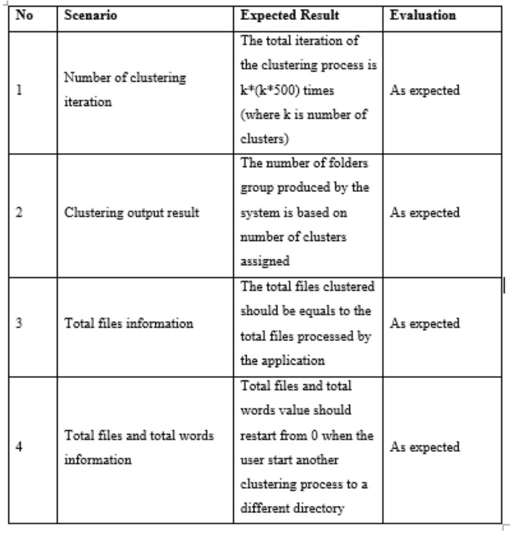
Table 4 K-Means Performance 3rd Evaluation



Documents to be processed	30 documents
Documents content length	Up to 200 words
Processed time	7 minutes 41 seconds
Clusters to be produced	4 clusters
Expected result	Cluster 1: 5 documents about Article 1 Cluster 2: 10 documents about Article 2 Cluster 3: 7 documents about Article 3 Cluster 4: 8 documents about Article 4
Result	Cluster 1: 5 documents about Article 1 Cluster 2: 5 documents about Article 2 Cluster 3: 5 documents about Article 3 Cluster 4: 8 documents about Article 4
Result accuracy evaluation	$\frac{5+5+5+8}{30} \times 100\% = 76{,}67\%$

Based on the Evaluation data of the K-Means Algorithm performance on Table 2 to Table 4, this application could produce $\frac{100+81{,}81+76{,}67}{3}\ x\ 100\% = 86{,}16\%$ accuracy.

Table 5 Documents Clustering Results Evaluation



No	Scenario	Expected Result	Evaluation
1	Number of clustering iteration	The total iteration of the clustering process is k*(k*500) times (where k is number of clusters)	As expected
2	Clustering output result	The number of folders group produced by the system is based on number of clusters assigned	As expected
3	Total files information	The total files clustered should be equals to the total files processed by the application	As expected
4	Total files and total words information	Total files and total words value should restart from 0 when the user start another clustering process to a different directory	As expected

IV. Discussion

This section presents discussion on why Documents Clustering using K-Means Algorithm uses the methods it uses.

- Clustering

Documents Clustering using K-Means Algorithm utilizes clustering method for classifying the objects rather than another classifying the objects method such as Linear Discriminant Analysis.

The advantage of this method is:
System could maximize the classification result without needed to have a specific training example.

- K-Means Algorithm

This application uses K-Means Algorithm to calculate the documents similarity one to another.

The advantage of this method is:
K-Means algorithm works by taking unique attribute from each data, then assigning these attributes to the nearest cluster to better fit the clusters themselves by finding the minimum fitting error between the data sets and centroids by updating the centroids to be the mean value of the clusters.

V. Conclusions

First, Document Clustering using K-Means Algorithm application is a desktop-based application developed to classify text documents purely based on its content similarity.
Second, the clustering result of the documents does not guarantee 100% accuracy compared to the user expectation. The accuracy of this application is up to 85% compared to user expectation because the classification process is really depending on the number of documents, number of clusters, the document contents, and the documents length.
In the future, Documents Clustering using K-Means Algorithm will have user interface enhancement and an advanced documents processing method and the results itself.

REFERENCES

[1] Ambler, S. W. (n.d.). *Agile Modeling*. Retrieved March 15 2017, from UML 2 Use Case Diagrams: AN Agile Introduction: http://www.agilemodeling.com/artifacts/useCaseDiagram.htm

[2] Erb, E. (n.d.). *Github*. Retrieved April 20, 2017, from Document Clustering Program in Java: https://github.com/ezraerb/DocumentCluster

[3] *File: K Means Example Step 1.svg*. (n.d.). Retrieved March 20, 2017, from Wikipedia: https://en.wikipedia.org/wiki/File:K_Means_Example_Step_1.svg

[4] *File: K Means Example Step 2.svg*. (n.d.). Retrieved March 20, 2017, from Wikipedia: https://en.wikipedia.org/wiki/File:K_Means_Example_Step_2.svg

[5] *File: K Means Example Step 3.svg*. (n.d.). Retrieved March 20, 2017, from Wikipedia: https://en.wikipedia.org/wiki/File:K_Means_Example_Step_3.svg

[6] *Information Retrieval*. (n.d.). Retrieved March 22, 2017, from http://www.doc.ic.ac.uk/~nd/surprise_97/journal/vol4/hks/inf_ret.html

[7] Jajoo, P. (2008). *Document Clustering.* Retrieved February 5, 2017

[8] *K-Means Clustering*. (n.d.). Retrieved February 4, 2017, from A Tutorial on Clustering Algorithms: https://home.deib.polimi.it/matteucc/Clustering/tutorial_html/kmeans.html

[9] Kunwar, S. (n.d.). *Text Documents Clustering using K-Means Algorithm*. Retrieved October 20, 2016, from Code Project: https://www.codeproject.com/Articles/439890/Text-Documents-Clustering-using-K-Means-Algorithm

[10] Osinski, S., & Weiss, D. (n.d.). *Carrot2*. Retrieved April 27, 2017, from Carrot2 Search: http://search.carrot2.org/stable/search

[11] Osinski, S., & Weiss, D. (n.d.). *Carrot2 User and Developer Manual*. Retrieved April 27, 2017, from Carrot2: http://download.carrot2.org/head/manual/index.html#chapter.introduction

[12] Rose, B. (n.d.). *Document Clustering with Python*. Retrieved March 19, 2017, from http://brandonrose.org/clustering

[13] Shah, N., & Mahajan, S. (2012). Document Clustering: A Detailed Review. *International Journal of Applied Information Systems (IJAIS)*. Retrieved February 6, 2017

[14] Sousa, S. d. (n.d.). *The Advantages and Disadvantages of RAD Software Development*. Retrieved October 4, 2016, from Susan de Sousa's My PM Expert: www.my-project-management-expert.com/the-advantages-and-disadvantages-of-rad-software-development.html

[15] Teknomo, K. (n.d.). *Difference of Cluster Analysis and Discriminant Analysis*. Retrieved February 2, 2017, from Revoledu: http://people.revoledu.com/kardi/tutorial/LDA/Cluster%20and%20discriminant%20analysis.html

[16] Teknomo, K. (n.d.). *Discriminant Analysis Tutorial*. Retrieved February 2, 2017, from Revoledu: http://people.revoledu.com/kardi/tutorial/LDA/

[17] Teknomo, K. (n.d.). *Euclidean Distance*. Retrieved March 23, 2017, from Revoledu: http://people.revoledu.com/kardi/tutorial/Similarity/EuclideanDistance.html

[18] Teknomo, K. (n.d.). *How the K-Mean Clustering algorithm works?* Retrieved March 1, 2017, from Revoledu: http://people.revoledu.com/kardi/tutorial/kMean/Algorithm.htm

[19] Teknomo, K. (n.d.). *What is Clustering?* Retrieved January 30, 2017, from Revoledu: http://people.revoledu.com/kardi/tutorial/Clustering/clustering.htm

[20] Teknomo, K. (n.d.). *What is K-Mean Clustering?* Retrieved January 31, 2017, from Revoledu: http://people.revoledu.com/kardi/tutorial/kMean/WhatIs.htm

[21] toletol, K. (n.d.). *Rapid Application Development (RAD) Model*. Retrieved October 6, 2016, from Wikipedia: https://en.wikipedia.org/wiki/File:RADModel.JPG

[22] *What does tf-idf mean?* (n.d.). Retrieved March 23, 2017, from http://www.tfidf.com/

[23] Zong, J. (n.d.). *K Means Clustering with Tf-idf Weights*. Retrieved March 10, 2017, from http://jonathanzong.com/blog/2013/02/02/k-means-clustering-with-tfidf-weights