Volume 6, Issue 8, August 2016 ISSN: 2277 128X

International Journal of Advanced Research in Computer Science and Software Engineering

Research Paper

Available online at: www.ijarcsse.com

# Web Document Clustering System Using K – Means Algorithm

**Irwan Bastian*, Rozaliyana, Metty Mustikasari**

Information System Departement, Gunadarma University,
Indonesia

***Abstract— Nowadays there are increasing volume of text documents flowing over the internet, huge collections of documents in digital libraries and repositories, and digitized personal information. Clustering is a useful technique that organizes a large quantity of unordered text documents into a small number of meaningful and coherent clusters. Document clustering is one of popular technique data analysis. This paper proposes documents Clustering using K-Means algorithm. The documents similarity is measured using Winnowing algorithm and Cosine algorithm. The application runs in web browser. System will cluster documents based on the number of cluster entered. System is evaluated by using precision and in terms of time execution.***

***Keywords— Clustering documents, K-Means, Winnowing, Cosine Similarity***

## I. INTRODUCTION

Document Clustering is one on text mining to construct a huge volume of documents into categories with similar content. Clustering on large text dataset can be effectively done using partitional clustering algorithms. The K-Means algorithm is the most suitable partitional clustering approach for handling large volume of data. K-Means clustering algorithm uses a similarity metric that determines the distance from a document to a point that represents a cluster head. This similarity metric plays a vital role in the process of cluster analysis. The usage of suitable similarity metric improves the clustering results.

This paper illustrates a study to develop a system that classify documents by using K-Means algorithm. This system is constructed to classify documents into clusters based on the number of clusters entered by user. The purpose of this research is to create a web system that classify a separate text document datasets. The similarity measurement of this approach using cosine similarity and based on the number of cluster input to be formed.

The source documents to be tested is a text document with plaintext format. The source documents are using Indonesian archive online news portal consisting of kompas.com, tempo.com, and detik.com. The maximum number of documents that were tested is 50 documents.

## II. LITERATURE REVIEW

Several studies related to the document clustering using K-Means clustering algorithm had been done by some previous studies. Steinbach M, et al (2007) conducted a study to evaluate two clustering algorithms namely Hierarchical Clustering and K-Means. These results indicated that the approach with Hierarchical Clustering better than K-Means. However, a derivative of K-Means such as bisecting K-Means delivered results close to the results of Hierarchical Clustering algorithms. This result was due to that approach by bisecting K-Means clustering produces significantly fairly consistent [6].

Huang (2008) proposed similarity measures for text document clustering. She compared and analyzed the effectiveness of these measures in partitional clustering for text document datasets. Her experiments utilized the standard K-Means algorithm and she reported results on seven text document datasets and five distance/similarity measures that had been most commonly used in text clustering [3].

Ravindran and Thanamani (2015) proposed K-Means Document Clustering using Vector Space Model. They used Cosine Similarity of Vector Space Model as the centroid for clustering. Using this approach, the documents could be clustered efficiently even when the dimension was high because it used vector space representation for documents which was suitable for high dimensions [4].

## III. MATERIALS AND METHOD

### *A. Clustering*

Cluster analysis is a process of grouping a set of physical or abstract objects into classes based on the class of similar objects [2]. A cluster is a collection of object data that are similar to each other to form a separate cluster that different from the others. Clustering method has been developed in various research areas including data mining, statistics, machine learning, spatial databases, biological, and marketing. Document clustering has been applied to improve the effectiveness of information retrieval. The application of this clustering relies on a hypothesis that relevant documents will tend to be in the same cluster.

There is a formal definition of the cluster analysis in the literature. Suppose S is the set of objects that have an N elements:

$$S : \{o_1, o_2, o_3, \dots ,o_N \} \qquad (II.1)$$

As defined in equation II.1, cluster analysis divides S into a number k set C1, C2, C3,..., Ck. The set is called a cluster. As shown in equation II.2, a cluster Ci is a subset of S.

$$Ci \subseteq S \qquad (II.2)$$

Solutions or the output of a cluster analysis is stated as a set of all clusters.

$$C = \{C1, C2, C3,..., C4 \mid Ci \subseteq S, \forall i \in 1 \dots k\} \qquad (II.3)$$

The system designed is supported by several algorithms for clustering documents. The algorithms used in supporting this development are Winnowing algorithm, K-Means algorithm, and cosine similarity function.

### B. K-Means

In this study K-Means algorithm was selected to perform clustering process [7]. The standard K-Means algorithm works as follows. Given a set of data objects D and a pre-specified number of clusters k, k data objects are randomly selected to initialize k clusters, each one being the centroid of a cluster. The remaining objects are then assigned to the cluster represented by the nearest or most similar centroid. Next, new centroid are re-computed for each cluster and in turn all documents are re-assigned based on the new centroid. This step iterates until a converged and fixed solution is reached, where all data objects remain in the same cluster after an update of centroids. The generated clustering solutions are locally optimal for the given data set and the initial seeds. Different choices of initial seed sets can result in very different final partitions. The K-Means algorithm works with distance measures which basically aims to minimize the within-cluster distances [6].

### C. Winnowing Algorithm

Winnowing algorithm is the algorithm used to perform the process of checking the similarity of the word. This algorithm is widely used to detect plagiarism. The input of the Winnowing algorithm is a text document. The result is a collection of hash value is formed from the numerical value calculation of each ASCII characters. The collecting of hash values are processed are called fingerprint [1] [5].

### D. Cosine Similarity

Cosine similarity is a measure of similarity between two vectors of an inner product space that measures the cosine of the angle between them. Cosine similarity is one of the most popular similarity measurements to compare of two documents [3].

Given two documents $\vec{t}_a$ and $\vec{t}_b$, cosine equation is:

$$SIMc\left(\vec{t}_a \cdot \vec{t}_b\right) = \frac{\vec{t}_a \cdot \vec{t}_b}{\left|\vec{t}_a\right| \cdot \left|\vec{t}_b\right|} \qquad (II.4)$$

Where $\vec{t}_a$ and $\vec{t}_b$ is m-dimensional vector of the term set T = $\{t_1, ..., t_m\}$. Each document represents a term of weight in the document. As a result, the Cosine similarity is a non-negative bounded between [0,1] [5].

## IV. PROPOSED METHOD

### A. Framework Method

The system's workflow in the process of grouping documents on this application is illustrated at figure 1. Primary data needed in developing this application is a text file that is stored in text format collected in a corpus.

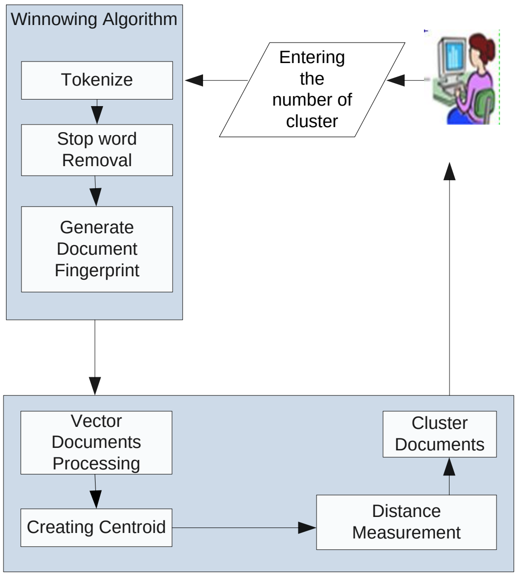


Fig. 1 Proposed Method

Figure 1 shows that after number of cluster is entered, the documents will be tokenized and all stop words were removed. After the document fingerprint is resulted, document vector is processed, centroid is created and similarity distance is measured between data and centroid and then the documents are clustered.

### *B. System Design*

This stage explains the use of activity diagrams. This diagram shows that if the user enter the number of clusters the document clustering is processes. If the user select the home page, the results of document clustering is displayed. If the user select clustering page, the results of similarity value of each group is displayed on the "clustering results" page. On the “info” page the user can see the details of usage of the application.

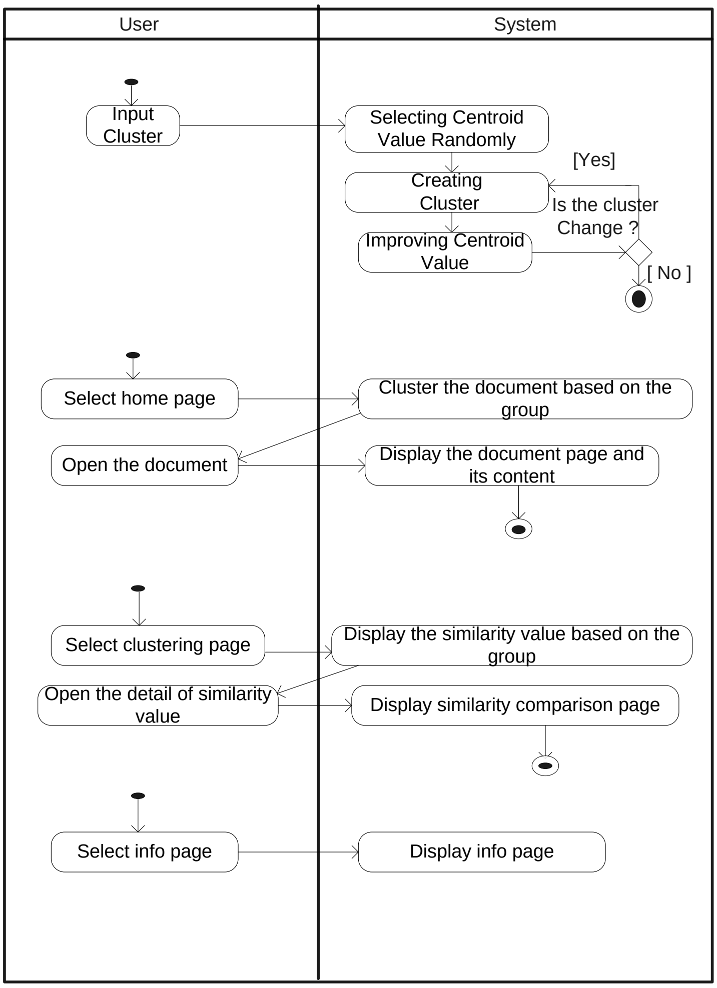
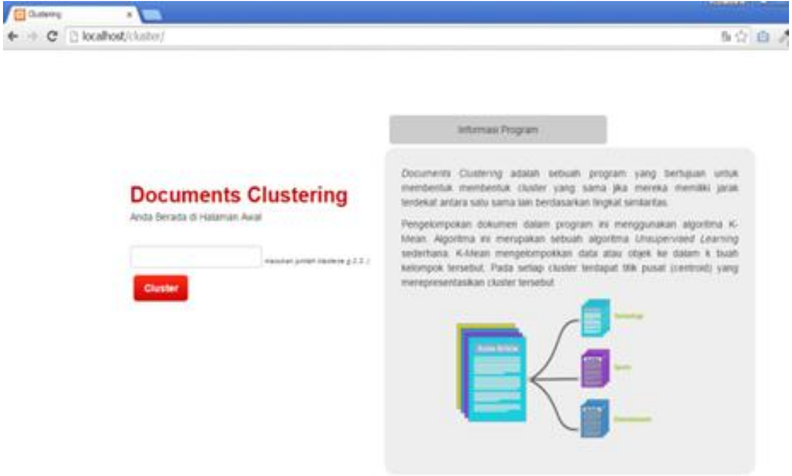

User
System
Input Cluster
Selecting Centroid Value Randomly
[Yes]
Creating Cluster
Is the cluster Change ?
Improving Centroid Value
[ No ]
Select home page
Cluster the document based on the group
Open the document
Display the document page and its content
Select clustering page
Display the similarity value based on the group
Open the detail of similarity value
Display similarity comparison page
Select info page
Display info page

Fig. 2 Activity Diagram

### *C. Web User Interface*

*1) Main Page*: The main page is the main web interface that user may enter number of clusters to be formed. This web page explains a brief information about the subject of document clustering.



Documents Clustering

Anda Berada di Halaman Awal

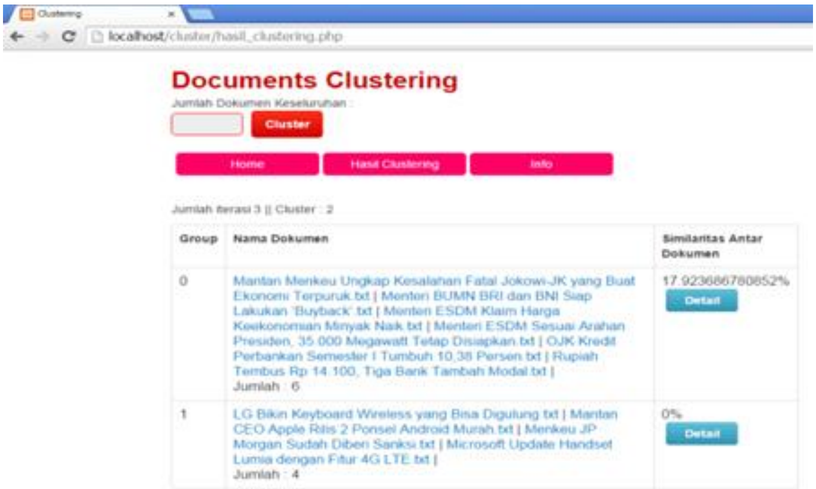
masukan jumlah klaster e.g 2,3..)

Documents Clustering adalah sebuah program yang bertujuan untuk membentuk membentuk cluster yang sama jika mereka memiliki jarak terdekat antara satu sama lain berdasarkan tingkat similaritas.

Pengelompokan dokumen dalam program ini menggunakan algoritma K-Mean. Algoritma ini merupakan sebuah algoritma *Unsupervised Learning* sederhana. K-Mean mengelompokkan data atau objek ke dalam k buah kelompok tersebut. Pada setiap cluster terdapat titik pusat (centroid) yang merepresentasikan cluster tersebut.

Fig. 3 Main page

*2) Clustering Result Page:* This page shows the documents that have the same degree of similarity. In order to see the similarity details users can choose a detail button.



Documents Clustering

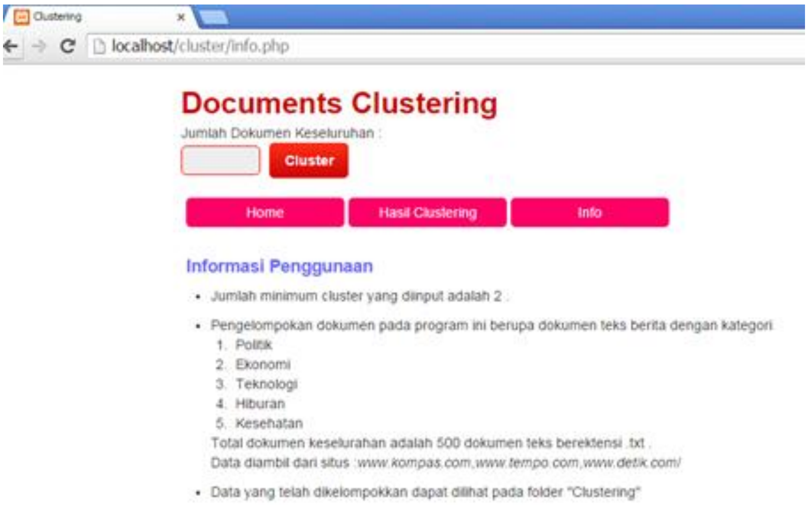
Jumlah Dokumen Keseluruhan :

Jumlah iterasi 3 || Cluster : 2

| Group | Nama Dokumen | Similaritas Antar Dokumen |
|---|---|---|
| 0 | Mantan Menkeu Ungkap Kesalahan Fatal Jokowi-JK yang Buat Ekonomi Terpuruk.txt \| Menteri BUMN BRI dan BNI Siap Lakukan 'Buyback'.txt \| Menteri ESDM Klaim Harga Keekonomian Minyak Naik.txt \| Menteri ESDM Sesuai Arahan Presiden, 35.000 Megawatt Tetap Disiapkan.txt \| OJK Kredit Perbankan Semester I Tumbuh 10,38 Persen.txt \| Rupiah Tembus Rp 14.100, Tiga Bank Tambah Modal.txt \|<br>Jumlah : 6 | 17.923686780852% |
| 1 | LG Bikin Keyboard Wireless yang Bisa Digulung.txt \| Mantan CEO Apple Rilis 2 Ponsel Android Murah.txt \| Menkeu JP Morgan Sudah Diberi Sanksi.txt \| Microsoft Update Handset Lumia dengan Fitur 4G LTE.txt \|<br>Jumlah : 4 | 0% |

Fig. 4 Clustering Results Page

*3) Info Page:* Info page is a web page that explains a brief information about application usage.



Documents Clustering

Jumlah Dokumen Keseluruhan :

Informasi Penggunaan

- Jumlah minimum cluster yang diinput adalah 2 .
- Pengelompokan dokumen pada program ini berupa dokumen teks berita dengan kategori
  1. Politik
  2. Ekonomi
  3. Teknologi
  4. Hiburan
  5. Kesehatan

  Total dokumen keselurahan adalah 500 dokumen teks berektensi .txt .
  Data diambil dari situs :www.kompas.com,www.tempo.com,www.detik.com/
- Data yang telah dikelompokkan dapat dilihat pada folder "Clustering"

Fig. 5 Info Page

## V. RESULT AND DISCUSSION

### *A. Experiment Result*

This Experiment evaluates the relation between number documents and execution time and number clusters and execution time. This study also evaluates relevant document that has been clustered. This study uses 50 documents. The document is extracted from Kompas.com, Tempo.com and Detik.com.

*1) Comparison Number Documents and Execution Time:* This experiment tested the number documents and execution time. In this experiment the cluster number entered is the three clusters. The comparison result can be seen in table 1.

Table I Comparison Of Number Of Documents And Execution Time

| No | Number of Documents | Time (s) |
|---|---|---|
| 1 | 10 | 0.415 |
| 2 | 20 | 1.916 |
| 3 | 30 | 4.401 |
| 4 | 40 | 8.628 |
| 5 | 50 | 14.305 |

Table 1 shows that the greater the number of documents, the greater the execution time required for clustering documents. This comparison can be seen in the fig 6.

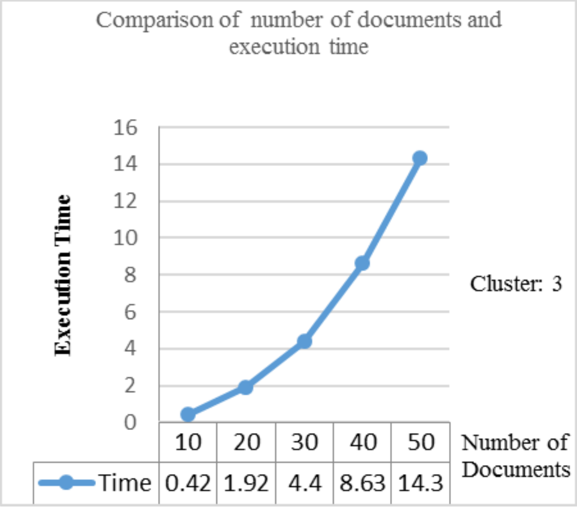


Fig. 6 Comparison of number of documents and execution time

*2) Comparison the Number of Cluster and Execution Time:* In this phase, the number of cluster and execution time is tested. This experiment compares the number of clusters that included in the execution time. The results of this experiment can be seen in Table 2.

Table II Comparison of the number clusters and execution time

| No | Cluster | Execution Time |
|---|---|---|
| 1 | 3 | 4.071 |
| 2 | 4 | 4.159 |
| 3 | 5 | 4.142 |
| 4 | 6 | 4.594 |
| 5 | 7 | 4.205 |

Table 2 shows that the number of clusters entered did not affect the execution time. This is due to the reduction vectors of documents processed by the system. The comparison chart can be seen in figure 7.

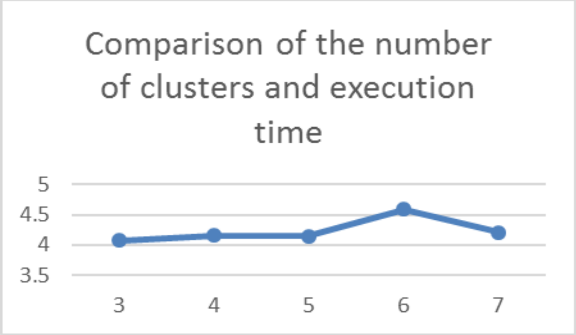


Fig. 7 Comparison of number of clusters and execution time

### *B. Relevant Documents Evaluation*

Based on the examination of the first test, the results of grouping category are shown at table 3. There are four categories: Entertainments category, Politics category, Technology category and Economics category. In this experiment, the input clusters formed is k = 5. Execution time results is 28.13 seconds. To evaluate relevant document of clustered documents using precision term. From Table 3, It is shown that the precision value of group 0 is 0.99, group 1 is 0.99, group 2 is 0.67, and group 3 is 0.6. As the precision average is near to 1, therefore it can be stated that clustering result is quite relevant.

Table III The Precision of Documents clustered

| Groups | Categories | Precision |
|---|---|---|
| 0 | Entertainments | 0.99 |
| 1 | Technologies | 0.99 |
| 2 | Politics | 0.67 |
| 3 | Economies | 0.6 |

## VI. CONCLUSION

The web application of K-Means algorithm using Cosine similarity measurement has been successfully implemented. The data needed in developing this application is a text file that is stored in text format collected in a corpus. In this study shows that the more documents are processed the longer execution time required. The execution time is not influenced by the number of cluster as it is due to the reduction vectors of document processed by the system.

Using cosine similarity measurement proved that it is producing significant clustering. The precision values show that this algorithm is quite accurate. In order to improve the effectiveness of document processing and classification, the algorithm should be extended. In addition to enrich with terms weighting, it is better using combination of similarity measurement.

## REFERENCES

[1] Elbegbayan, Norzima. Winnowing, *A Document Fingerprinting Algorithm*. TDDC03 Project, Spring 2005.

[2] Han, Jiawei., Kamber, Micheline. *Data Mining: concept and Techniques*. Elsevier : UK., 2006.

[3] Huang, Anna. *Similarity Measures for Text Document Clustering*. Hamilton-New Zealand: Proceedings of The New Zealand Computer Science Research Student Conference. 2008.

[4] Ravindran R.M, Thanamani A.S., *K–Means Document Clustering using Vector Space Model*, Bonfring International Journal of data mining , Vol 5, No.2 July 2015.

[5] Sceilmer, Saul, Daniel S. Wilkerson, dan Alex Aiken. *Winnowing: Local Algorithms for Document Fingerprinting. San diego: In Proceedings Of The ACM SIGMOD International Conference On Management Of Data*. 2003 (Online).

[6] Steinbach, Michael., Karypis, George., Vipin Kumar. *A comparison of Document Clustering Techniques*. Technical Report #00-034, Department of Computer Science and Engineering, University of Minnesota, 2007.

[7] S Jaiganesh, P. Jaganathan. *An Appropriate Similarity Measure for K-Means Algorithm in Clustering Web Documents*. International Journal for Scientific Research & Development. Vol. 3, Issue 02, 2015. ISSN (online): 2321-0613.