K-Means and K-Medoids for Indonesian Text Summarization

K K Purnamasari
Faculty of Engineering and Computer Science, Universitas Komputer Indonesia, Indonesia
Email : ken.kinanti@email.unikom.ac.id

Abstract. The purpose of this study is to build automated summation tools, especially in grouping methods such as K-Means and K-Medoids. Finding the best method between the two algorithms, this study focuses on comparing the two methods to summarize thesis report documents. This system is divided into Filtering, Tokenization, TF-IDF, Cosine Similarity, and Clustering. Based on 50 test documents, the average accuracy rate is 51.16% for K-Means and 63.35% for K-Medoids. K-Means has a smaller accuracy value than K-Medoids. The accuracy of the resulting K-Means also depends on the size and center of the initial cluster chosen. So, as the next stage of development, research needs to be done that compares the results of the combination of initial size and center cluster values for K-Means and continue with several other classifications.

1. Introduction
Development of online technology has caused enormous information needs. Various sources of information like online-news articles are generally real-time and unlimited. It leads to generate very large number of sentences in the online world [1]. Thorough manually reading process in obtaining core information from an article takes a long time. This condition triggered the beginning of research in automatic summarization system development. Automatic summarization process will identify the most important things in a text and produce less number of sentences [2] [3] [4]. The reduced-size text will help people to find relevant information faster and easier.

Several studies on automatic summarization system in a number of clustering methods have been carried out [5] [1]. One of the most popular clustering methods is K-Means. Summarization system in Indonesian text with K-Means algorithm show 61% f-measure in news articles [6] and 51% accuracy in thesis report documents [7]. In other cases, K-Medoids as K-Means modification usually used to increase the performance. In the purpose of finding the best method between K-Means and K-Medoids in the case of text summarization, this study compares both methods to summarize thesis report documents [8].

The purpose of this study is to build automated summation tools, especially in grouping methods such as K-Means and K-Medoids. Finding the best method between the two algorithms, this study focuses on comparing the two methods to summarize thesis report documents. This system is divided into Filtering, Tokenization, TF-IDF, Cosine Similarity, and Clustering. Based on 50 test documents, the average accuracy rate is 51.16% for K-Means and 63.35% for K-Medoids.

2. Method
The entire process built is divided into five main parts: Filtering, Tokenization, TF-IDF, Cosine Similarity, and Clustering in K-Means or K-Medoids (see Figure 1). Filtering and Tokenization stages are used to prepare text to be calculated. TF-IDF and Cosine Similarity are used to determine the weight of each generated token. In this work, K-Medoids use TF-IDF vectors, and K-Means use Cosine Similarity results.
2.1 Filtering
Filtering removes various characters that are not used. In this study, deleted characters are all characters except space, letters, period, question mark, and exclamation.

2.2 Tokenization
A sentence in a text consists of a series of words that begin with a capital letter and ends with a period, question mark, or exclamation [9]. Based on this condition, the text can be processed into a collection of tokens. In this work, a token can be a single character or word.

2.3 TF-IDF
TF-IDF is a method that can be used to give weight to each word in the text. This method gives weight to each word in the text by calculating the frequency of term’s occurrence in a document (Term Frequency / TF) and the possibility of a document having a word (Inverse Document Frequency / IDF). Several variations of the TF-IDF weighting scheme are often used in the case of information retrieval. The possibility of a document having a word can be calculated using the formula in Equation 1 [9].

\[
IDF = \log\left(\frac{N}{df}\right)
\]  

(1)

2.4 Cosine Similarity
Cosine Similarity models term into vectors and calculate the cosine value of two tokens. The token used can be in the form of words, paragraphs, or even the entire source-text. Sorted cosine value will indicate the similarity among terms. High scores will indicate a large degree of similarity. In this study, the similarity level is calculated using Equation 2, where t shows the index indices of the sentence term, Wt.b1 shows the weight of the token t in blocks b1, and Wt.b2 shows the weight of the token t in block b2 [10].

\[
CS(b_1, b_2) = \frac{\sum_{t=1}^{n} W_{t.b1} W_{t.b2}}{\sqrt{\sum_{t=1}^{n} W_{t.b1}^2} \sqrt{\sum_{t=1}^{n} W_{t.b2}^2}}
\]  

(2)

2.5 Summarization (K-Means)
K-Means method is included in the category of non-hierarchical clustering that is easy and simple, quickly finds convergence, and adaptable to the distribution of data. This method creates a cluster by calculating distances between data and the center of each cluster (centroid) [11]. The process of K-Means used in this study shown in Figure 2.
Determination of the number of clusters in this study is adjusted to the summarization goal. Clustering could be done with discard sentences that are not important enough, based on their relevances to the overall contents. This condition supports the use of two clusters. Determination of centroid is randomly done at the beginning of every iteration, using average score of the cluster as a centroid value. Calculation of the distance between data and centroid is done using Euclidean Distance. This method calculates the distance between data with several initial cluster centers. Based on this calculation, the shortest distance that determines the closeness of the data with a cluster can be determined, using Equation 3 below.

\[ Ed(x_i, y_j) = \sqrt{(x_i - y_j)^2} \]  \hspace{1cm} (3)

2.6 Summarization (K-Medoids)
K-Medoids is a clustering algorithm reminiscent of the k-means algorithm. Both K-Means and K-Medoids algorithms are partitional (breaking dataset up into groups) and both try to minimize the distance between data in a cluster. One of the data values will be chosen as the center of the cluster. K-Medoids chooses data as center (medoids) and can use arbitrary distances, while K-Means only minimizes Euclidean distances [12] (see Figure 3).
3. Results and Discussion

Automatic Summarization in this research is evaluated by using 50 pieces of test data (806 sentences) from thesis report documents. K-Means and K-Medoids accuracy percentage results for each document are obtained as shown in Table 1. In almost all documents, K-Medoids is better than the K-Means algorithm.

The result shows that K-Medoids better than K-Means. However, in general, there are no big differences between both of them. It may be related to the number of clusters (k parameter) selection at the beginning of the process [13].
Table 1. Result.

| D[i] | Number of Sentences | Accuracy of K-MEANS (%) | Accuracy of K-MEDOIDS (%) |
|------|---------------------|-------------------------|--------------------------|
| 1    | 17                  | 52.94                   | 64.71                    |
| 2    | 21                  | 52.38                   | 57.14                    |
| 3    | 13                  | 53.85                   | 61.54                    |
| 4    | 16                  | 50.00                   | 62.50                    |
| 5    | 19                  | 52.63                   | 57.89                    |
| 6    | 14                  | 50.00                   | 71.43                    |
| 7    | 10                  | 50.00                   | 70.00                    |
| 8    | 30                  | 50.00                   | 53.33                    |
| 9    | 12                  | 50.00                   | 58.33                    |
| 10   | 26                  | 50.00                   | 57.69                    |
| 11   | 19                  | 52.63                   | 52.63                    |
| 12   | 12                  | 50.00                   | 66.67                    |
| 13   | 13                  | 46.15                   | 61.54                    |
| 14   | 15                  | 53.33                   | 66.67                    |
| 15   | 14                  | 57.14                   | 78.57                    |
| 16   | 14                  | 42.86                   | 57.14                    |
| 17   | 28                  | 53.57                   | 60.71                    |
| 18   | 23                  | 52.17                   | 56.52                    |
| 19   | 23                  | 52.17                   | 60.87                    |
| 20   | 16                  | 50.00                   | 68.75                    |
| 21   | 16                  | 50.00                   | 62.50                    |
| 22   | 11                  | 54.55                   | 63.64                    |
| 23   | 20                  | 50.00                   | 60.00                    |
| 24   | 18                  | 50.00                   | 55.56                    |
| 25   | 8                   | 50.00                   | 62.50                    |
| 26   | 17                  | 52.94                   | 52.94                    |
| 27   | 11                  | 54.55                   | 72.73                    |
| 28   | 13                  | 53.85                   | 76.92                    |
| 29   | 13                  | 53.85                   | 61.54                    |
| 30   | 10                  | 50.00                   | 70.00                    |
| 31   | 11                  | 54.55                   | 63.64                    |
| 32   | 26                  | 50.00                   | 61.54                    |
| 33   | 11                  | 54.55                   | 90.91                    |
| 34   | 20                  | 50.00                   | 60.00                    |
| 35   | 19                  | 52.63                   | 68.42                    |
| 36   | 14                  | 50.00                   | 57.14                    |
| 37   | 22                  | 50.00                   | 50.00                    |
| 38   | 12                  | 50.00                   | 58.33                    |
| 39   | 14                  | 50.00                   | 64.29                    |
| 40   | 21                  | 52.38                   | 66.67                    |
| 41   | 8                   | 37.50                   | 62.50                    |
| 42   | 18                  | 61.11                   | 72.22                    |
| 43   | 16                  | 50.00                   | 56.25                    |
| 44   | 10                  | 60.00                   | 60.00                    |
| 45   | 19                  | 47.37                   | 63.16                    |
| 46   | 18                  | 50.00                   | 50.00                    |
| 47   | 14                  | 50.00                   | 64.29                    |
4. Conclusion

Based on the research that has been done, it is obtained the results of summarization accuracy of 51.16% for K-Means and 62.35% for K-Medoids. K-Means has a smaller accuracy value than K-Medoids. The accuracy of the K-Means produced also depends on the size and center of the initial cluster selected. So, as the next stage of development, it is necessary to do a study that compares the results of a combination of size values and the initial cluster center for K-Means and continue with several other classifications.

References

[1] S. Saraswathi and R. Arti, "Multi-Document Text Summarization Using Clustering Techniques and Lexical Chaining," ICTACT Journal on Soft Computing, 1(1), p. 29, 2010.
[2] Kiyoumarsi, F. (2015). Evaluation of automatic text summarizations based on human summaries. Procedia-Social and Behavioral Sciences, 192, 83-91.
[3] Pal, A. R., Maiti, P. K., and Saha, D. (2013). An approach to automatic text summarization using simplified leks algorim and wordnet. International Journal of Control Theory and Computer Modeling (IJCTCM), 3(4/5), 15-23.
[4] Al-Hashemi, R. (2010). Text Summarization Extraction System (TSES) Using Extracted Keywords. Int. Arab J. e-Technol., 1(4), 164-168.
[5] Cai, X., Li, W., Ouyang, Y., and Yan, H. (2010, August). Simultaneous ranking and clustering of sentences: a reinforcement approach to multi-document summarization. In Proceedings of the 23rd International Conference on Computational Linguistics (pp. 134-142). Association for Computational Linguistics.
[6] Pratama, F. (2014). Rancang Bangun Aplikasi Peringkas Teks Otomatis artikel Berbahasa Indonesia Menggunakan Metode Term Frequency Inverse Document Frequency (Tf-Idf) Dan K-Mean Clustering (Doctoral dissertation, Universitas Islam Negeri Sultan Syarif Kasim Riau), 30-33.
[7] Purnamasari, K. K., and Widiastuti, N. I. (2017). Jurnal Ilmiah Komputer dan Informatika (KOMPUTA), 6(2).
[8] Velmurugan, T. (2012). Efficiency of k-means and k-medoids algorithms for clustering arbitrary data points. Int. J. Computer Technology and Applications, 3(5), 1758-1764.
[9] Manning, C. D., Raghavan, P., and Schütze, H. (2009). Probabilistic information retrieval. Introduction to Information Retrieval, 220-235.
[10] Rahutomo, F., Kitasuka, T., and Aritsugi, M. (2012, October). Semantic cosine similarity. In The 7th International Student Conference on Advanced Science and Technology ICAST, 4(1), 1.
[11] Singh, G. (2013). A study of encryption algorithms (RSA, DES, 3DES and AES) for information security. International Journal of Computer Applications, 67(19).
[12] Shinde, S. K., and Kulkarni, U. V. (2011). Hybrid Personalized Recommender System Using Fast K-medoids Clustering Algorithm. Journal of Advances in Information Technology, 2(3), 152-158.
[13] Pham, D. T., Dimov, S. S., and Nguyen, C. D. (2005). Selection of K in K-means clustering. Proceedings of the Institution of Mechanical Engineers, Part C: Journal of Mechanical Engineering Science, 219(1), 103-119.