Journal of Engineering and Applied Sciences 13 (Special Issue 3): 3188-3192, 2018
ISSN: 1816-949X

Document Clustering Using Combination of K-means and Single Linkage Clustering Algorithm

Anthon Roberto Tampubolon, Novita Sijabat, Ester Tambunan and Sanny Simarmata
Del Institute of Technology, North of Sumatera, 22381 Sumatera Utara, Indonesia

Abstract: Document clustering is a technique for classifying documents based on similarity levels of objects within documents. Document clustering is also can be applied to Retrieve Information (IR) based on calculation of term frequency-inverse document frequency and the vector space model-cosine similarity. K-means is one method of grouping One powerful partitioning technique but K-means may be trapped in a local optimum because of centroid random selection. In this study, we build an application document clustering and conduct experiments on the final project document of student in Del Institute of Technology. The experimental results showed that, the K-means clustering which is a partition can be optimized using one of the techniques that single linkage hierarchical clustering based cluster variance (variance within and variance between).

Key words: Document clustering, K-means, single linkag, trapped, frequency, technique

INTRODUCTION

Clustering is a technique to grouping the data or object based on similarity. Grouping with clustering technique is based on the similarity of documents, where objects in a cluster have similar high internal homogeneity and external heterogeneity or high distinction with objects in other clusters. Document clustering can be applied to various fields such as data mining and information retrieval (Kaur and Kaur, 2013; Gupta and Rajavat, 2014). one of the techniques document clustering for information retrieval is to provide weighting word (term) that is based on calculation of the Term Frequency-Inverse Document Frequency (TF-IDF) and the vector space model-cosine similarity (Bafna *et al*., 2016).

There are two methods of clustering those are partitioning clustering and hierarchical clustering. In this study, we will combine the two method. K-means method is a powerful method by partitioning clustering techniques. K-means has the advantage that, it can group the data in large numbers with effective compute (Kaur and Kaur, 2013; Wu *et al*., 2015). However, K-means can be trapped into local optimal because the initial centroid is randomly determined. Therefore, to solve the problem, this study will employ optimized model using single linkage algorithm method in order to determine the initial centroid in the process of grouping data.

Single linkage clustering is hierarchical clustering method that classify documents gradually. This method performs clustering with accurate results (Kaur and Kaur, 2013). Thus, we will combine the K-means and single linkage and will compare it with the K-means method. To analyze the performance of both methods, we will use cluster variance method.

There are two kinds of cluster variance, the Variance within clusters (V_w) and the Variance between clusters (V_b). Vw value is used to view the results of the variance of the spread of objects that exist in a cluster (internal homogeneity) and V_b value is the value that is used to view the variance for distributing data among clusters (external heterogeneity) (Rutterford *et al*., 2015).

MATERIALS AND METHODS

In this research, we will build a prototype application to implement clustering documents in the field of Information retrieval, i.e., grouping of documents by a search based on keywords entered by the user (Pawar *et al*., 2016). The data employed in this research is the final project document of students in Del Institute of Technology. Final project document is a document of research done by the students to complete their studies at the Del Institute of Technology. The document totaled 130 documents and can be accessed through the academic website in http://akademik.del.ac.id.

System architecture: The first step from the architecture is preprocessing and transformation. An initial phase is preprocessing documents, where the output is a collection of terms or words. Text preprocessing consists of four general stages: stage case folding, tokenization, filtering and stemming (Gupta and Rajavat, 2014). First step in preprocessing and transformation is case folding is to change the whole character to lowercase. The entire contents of the abstract and keywords final documents will be transformed into lowercase. Second step is Tokenization. Tokenization is the process of separating the query documents into tokens and ignoring

Corresponding Author: Anthon Roberto Tampubolon, Del Institute of Technology, North of Sumatera, 22381 Sumatera Utara, Indonesia

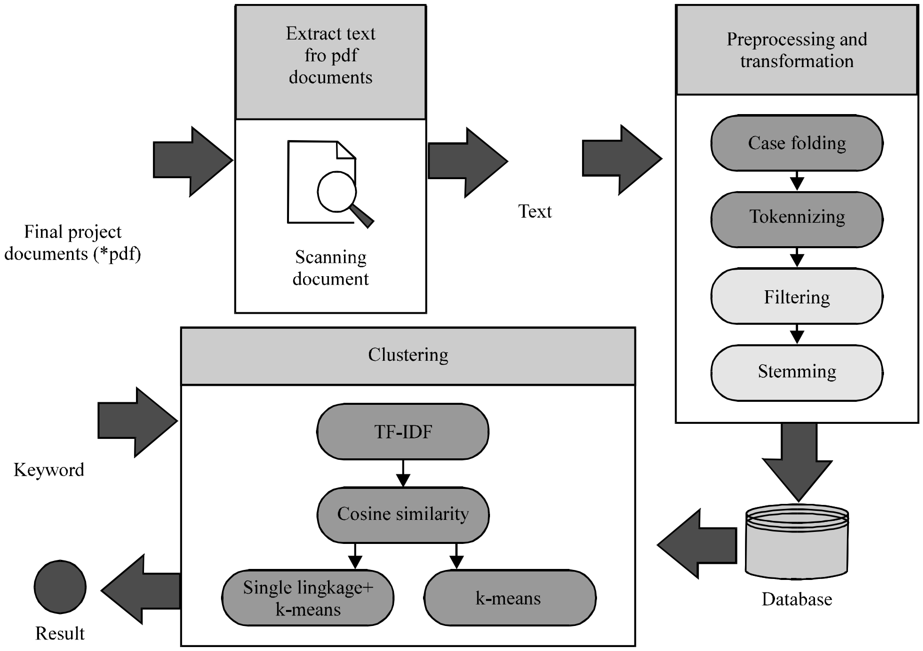


Fig. 1: System architecture for document clustering

punctuation and numbers. The entire string final project document will be separated into tokens. Third step is filtering, in filtering the stopper will delete the words in the text included stop list. Words are included in the stop list are words which do not contain important information. The next stage is the stage stemming preprocessing, i.e., getting basic words of remaining term from the previous process. Stemming algorithm used is algorithm Nazief and Adriani for Indonesian language documents.

The second step from the architecture is clustering process. Clustering process consists of four general stage: term weighting-TF-IDF, vector space model-cosine similarity, K-means algorithm and the combination of K-means and single linkage clustering.

The first stage is term weighting-TF-IDF to improve the ability to distinguish the different words (term) or to examine the similarity of certain words (term), term weighting is important in this process. Term weighting conducted by comparing multiple documents to a keyword by giving weight values to the document that will be compared. One of the methods of weighting the document is the Term Frequency-Inverse Document Frequency (TF-IDF) (Bafna *et al*., 2016). Tf-idf illustrates how important the word (term) in a document and the corpus. Term frequency is the frequency of occurrence of a term in a document and also in the corpus. This is the formula of term weighting-TF-IDF:

$$\text{IDF(query)} = 1+\log_e\left(\frac{\text{Total number of document}}{\text{Number of document}}\right) \quad (1)$$

Second stage is Vector Space Model-Cosine similarity. Cluster is a collection of objects where the data objects in a cluster are similar to each other and different from the objects in other clusters. Clustering uses a distance function in classifying the elements that are most similar to each other (Wu *et al*., 2015). The elements are represented in the form of data matrix where the columns associated with the condition and rows associated with a distance matrix. The higher the value, the more the differences on the objects (Bafna *et al*., 2016).

Third stage is K-means. K-means clustering method is based on the distance that divides the data into a number of clusters K-means clustering method is one method of non-hierarchical cluster analysis classified as a method of grouping which are unsupervised because the data analyzed do not have a class label. It means that the process of grouping does not have a definite cluster members. Objects that have been entered into a particular cluster can still be moved to another cluster. Object displacement on K-means will stop when all the data residing on a given cluster do not move to another cluster (Gupta and Rajavat, 2014; Steinbach *et al*., 2000). Basic K-means algorithm for finding k clusters:

- Select K points as the initial centroids
- Assign all points to the closest centroid
- Recomputed the centroid of each cluster
- Repeat 3 and 4 steps until centroids don't change

The last stage of clustering process is the combination of K-means and single linkage clustering. K-means clustering method is one method partitional clustering algorithm that is simple and most commonly used in grouping the data by clustering (Gupta and Rajavat, 2014). Quality of K-means method will be better when using a large data set. However, K-means can be trapped into local optimal because the initial centroid is randomly determined. K-means weaknesses can be addressed by changing the process of centroid. Search initial centroid at the k-means can be replaced by other methods, that is single linkage clustering method that will be applied in this research. Clustering process of final project document employs the K-means the K-means optimized single linkage clustering method. Initiation of the initial centroid on the method of K-means will be determined by the method of single linkage clustering. K-means method Started by determining the number K clusters, K-means method will produce the determination of the initial centroid randomly. This method is a combination of single linkage clustering and K-means, initial centroid is determined by finding the average of the data residing on a cluster generated in single linkage clustering method. At this stage of single linkage clustering each data is assumed to be a cluster. If there are n data, the number of initial clusters is k, where k = n. Calculated the distance between the cluster using the cosine similarity. The cluster with the most minimal distance incorporated into the new cluster so, that the number of clusters becomes k = k-1. In the single linkage method:

$$k \approx \sqrt{n/2} \qquad (2)$$

Where:

k = The number of clusters
n = The number of documents

Thus, the process of merging clusters on a single linkage clustering method will stop when the number of clusters that formed it meets $k \approx \sqrt{n/2}$ cluster. In state $k \approx \sqrt{n/2}$, then proceed with the K-means method.

In K-means clustering, cluster members put in a cluster that has a distance that is closest to its centroid, proses search iteration cluster center and grouping objects into clusters formed continue until there are no more objects to move.

Figure 3 flowchart combination of single linkage clustering and K-means show explanations workmanship sequence of research using single linkage method and K-means clustering.

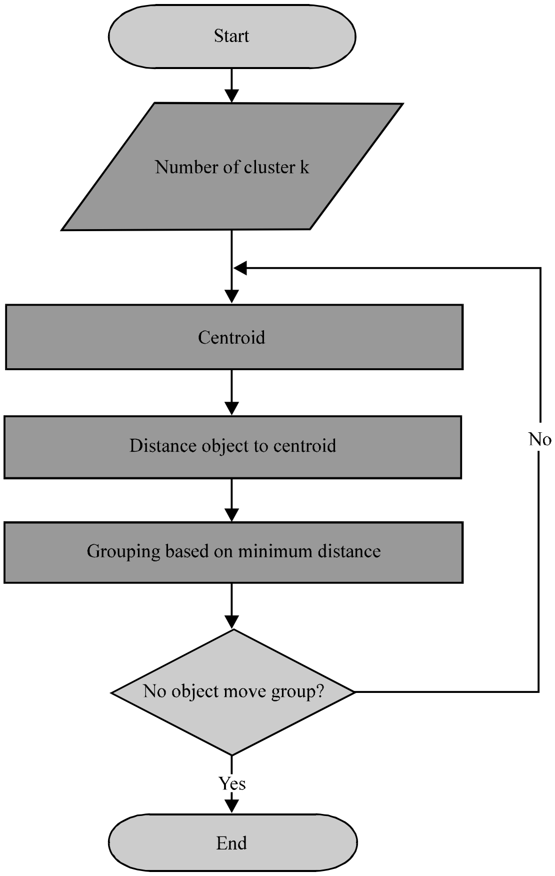

Start
Number of cluster k
Centroid
Distance object to centroid
Grouping based on minimum distance
No object move group?
No
Yes
End

Fig. 2: K-means flowchart

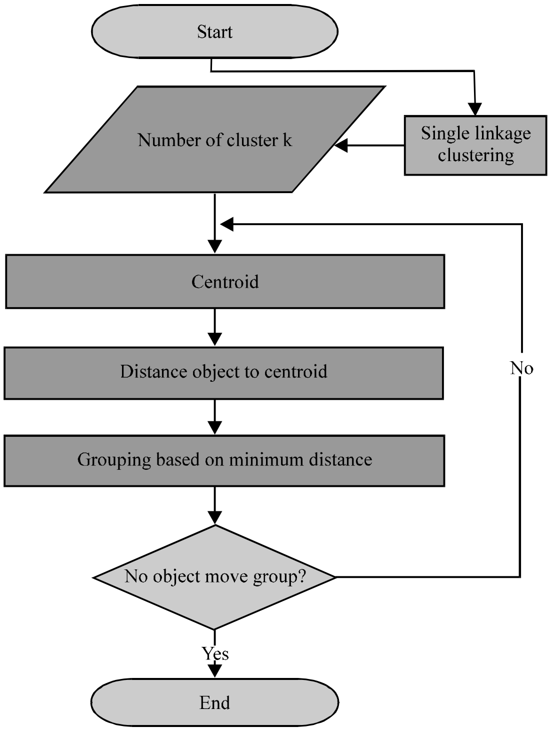

Start
Number of cluster k
Single linkage clustering
Centroid
Distance object to centroid
Grouping based on minimum distance
No object move group?
No
Yes
End

Fig. 3: Combination of single linkage and k-means flowchart

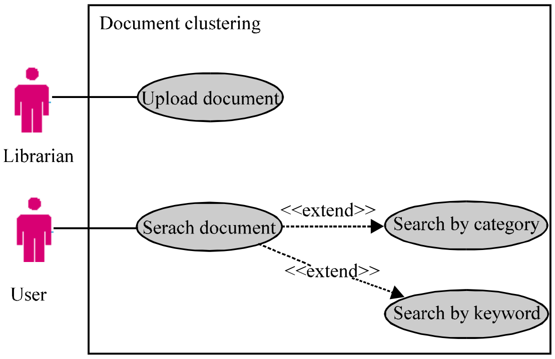
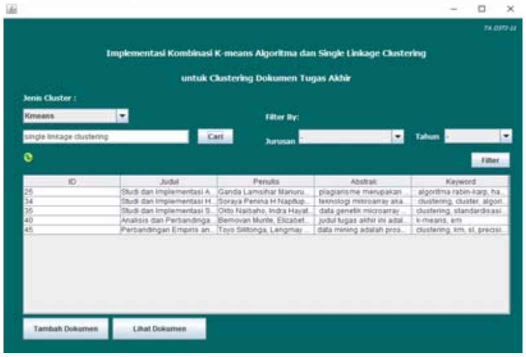


Fig. 4: Use case diagram of document clustering system



Implementasi Kombinasi K-means Algoritma dan Single Linkage Clustering

untuk Clustering Dokumen Tugas Akhir

Jenis Cluster : Kmeans

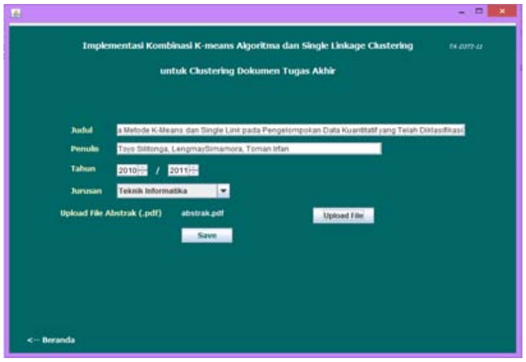
single linkage clustering

Jurusan -

Tahun -

ID	Judul	Penulis	Abstrak	Keyword
25	Studi dan Implementasi A...	Ganda Lamsihar Manuru...	plagiarisme merupakan ...	algoritma rabin-karp, ha...
34	Studi dan Implementasi H...	Soraya Penina H Napitup...	teknologi mikroarray ala...	clustering, cluster, algori...
35	Studi dan Implementasi S...	Otto Naibaho, Indra Hayat...	data genetik microarray ...	clustering, standardisasi...
40	Analisis dan Perbandinga...	Bernovan Munte, Elizabet...	judul tugas akhir ini adal...	k-means, em
45	Perbandingan Empiris an...	Toyo Silitonga, Lengmay ...	data mining adalah pros...	clustering, km, sl, precisi...

Fig. 5: Home view user interface



Implementasi Kombinasi K-means Algoritma dan Single Linkage Clustering

untuk Clustering Dokumen Tugas Akhir

Judul: a Metode K-Means dan Single Link pada Pengelompokan Data Kuantitatif yang Telah Diklasifikasi

Penulis: Toyo Silitonga, LengmaySimamora, Toman Irfan

Tahun: 2010 / 2011

Jurusan: Teknik Informatika

Upload File Abstrak (.pdf): abstrak.pdf

Fig. 6: Upload document view user interface

Implementation: This system developed in java application, Fig. 4 show use case diagram of the system, Fig. 5 and 6 show the user interface of the system.

RESULTS AND DISCUSSION

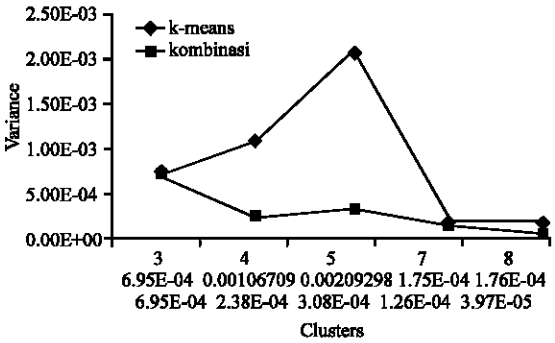


Fig. 7: Comparison of Variance within cluster (V_w)

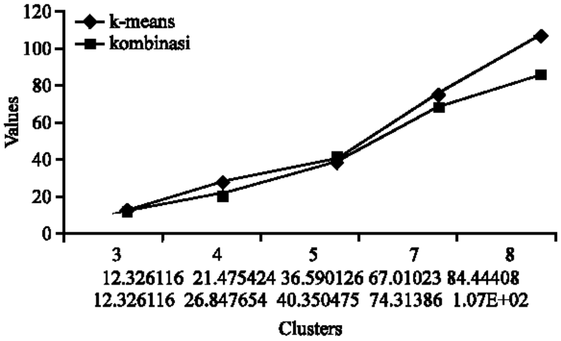


Fig. 8: Comparison of Variance between cluster (V_b)

Experimental result: We have tested each method on 130 final project document of Del Institute of Technology. For analysis, we get the cluster variance (V_w and V_b) of 20, 35, 50, 100 and 130 documents for each method. For value of Variance within cluster (V_w) will be better if the value is getting smaller but for value of Variance between (V_b) will be better if value is getting larger.

Figure 7 shows the comparison Variance within cluster (V_w) from each method referred to different scenario based on different cluster such as 20, 35, 50, 100 and 130 documents. We can see that the quality of the combination of single linkage and K-means cluster produces better quality than the K-means non-optimization.

The performance of the K-means non-optimization and combination of K-means and single linkage clustering is almost similar when clustering 7. However, the process iteration is much higher in non-optimization compared to other method. It means that K-means and single linkage clustering provide better result for big dataset. Figure 8 shows comparison value of variances between clusters based on every scenario referred to different cluster such as 20, 35, 50, 100 and 130 documents. We can see that the quality of the combination of single linkage and K-means cluster produces better quality than the K-means non-optimization. It means that K-means and single linkage clustering provide better result for big dataset.

CONCLUSION

There are two conclusions of this study: first is cluster quality produced by a combination of single linkage and K-means is better than K-means without optimization. And second is based on analysis of variance variance between and within cluster cluster, cluster quality produced by a combination of SL with the K-means getting better for large data.

REFERENCES

Bafna, P., D. Pramod and A. Vaidya, 2016. Document clustering: TF-IDF approach. Proceedings of the International Conference on Electrical, Electronics and Optimization Techniques (ICEEOT), March 3-5, 2016, IEEE, Maharastra, India, ISBN:978-1-4673-9940-1, pp: 61-66.

Gupta, M. and A. Rajavat, 2014. Comparison of algorithms for document clustering. Proceedings of the 2014 International Conference on Computational Intelligence and Communication Networks (CICN), November 14-16, 2014, IEEE, Indore, India, ISBN:978-1-4799-6930-2, pp: 541-545.

Kaur, M. and U. Kaur, 2013. Comparison between K-mean and hierarchical algorithm using query redirection. Int. J. Adv. Res. Comput. Sci. Software Eng., 3: 1454-1459.

Pawar, S.S., A. Manepatil, A. Kadam and P. Jagtap, 2016. Keyword search in information retrieval and relational database system: Two class view. Proceedings of the International Conference on Electrical, Electronics and Optimization Techniques (ICEEOT), March 3-5, 2016, IEEE, Lavale, Pune India, ISBN:978-1-4673-9940-1, pp: 4534-4540.

Rutterford, C., A. Copas and S. Eldridge, 2015. Methods for sample size determination in cluster randomized trials. Intl. J. Epidemiol., 4: 1051-1067.

Steinbach, M., G. Karypis and V. Kumar, 2000. A comparison of document clustering techniques. Proceedings of the 6th ACM SIGKDD World Text Mining Conference, Volume 400, August 20-23, 2000, Boston, pp: 1-2.

Wu, G., H. Lin, E. Fu and L. Wang, 2015. An improved K-means algorithm for document clustering. Proceedings of the 2015 International Conference on Computer Science and Mechanical Automation (CSMA), October 23-25, 2015, IEEE, Hangzhou, China, ISBN:978-1-4673-9167-2, pp: 65-69.