Computing Group News Documents Using K-Means and K-Nearest Neighbor

Sitti Arni[1], Syaharullah Disa[2]
[1]Department of Information Systems, Sekolah Tinggi Manajemen Informatika Komputer Profesional Makassar, Indonesia. 90233
[2]Department of Information Engineering, Sekolah Tinggi Manajemen Informatika Komputer AKBA Makassar, Indonesia. 90245

{arnist1903@gmail.com[1], syahruldisa181073@gmail.com[2]}

Abstract. The aim of the study was to reduce the computational time of the document search by using the K-Means and K-Nearest Neighbor algorithm. The search result through search engines with display the documents according to the keywords entered. The number of documents that were displayed will make the user difficult to fine the required documents and it a long time for the display process of the documents. K-Means algorithm is used for the clustering documents obtained through online with search engine whereas algorithm K-Nearst Neighbor is used for grouping the clustering result document that are done with offline. The clustering with K-Means can reduce computational time on the news grouping by using K-Nearst Neighbor. The combination of two methods result an average time of 0.5011 seconds. Whereas the grouping that uses pure K-Nearst Neighbor requires the computational time 2.4841 seconds. The test result indicated that the combination of the two algorithms resulted accurate classification with the faster computation time.

Keywords: *Classification, Clustering, Information Retrieval.*

1 Introduction

Browsing and searching activities carried out by the user to make it easier to get information quickly, relevant and as needed, are supported by an increase in information flow that is very fast and the development of Information Retrieval (IR) technology which is a document search system.

The trend of changing document search patterns through digital technology is not new anymore at this time. The rapid increase in the number of documents, causing manual searches is very difficult. To find documents, an information retrieval system is needed that is able to classify documents.

Research on text document retrieval including the Implementation of theMethod *K-Nearest Neigbour* by Weighting *Term Frequency* (TF). *Inverse Document Frequency* (IDF). *Inverse Class Frequency* (ICF) for Creative Idea Categorization on Companies (Herlambang, 2017). The accuracy results obtained have the highest similarity average of 90% with k = 1 for creative idea categorization. The research did not measure the computational time of thealgorithm *K-Nearest Neighbor*. Cluster-based information retrieval for the Indonesian language text information retrieval system uses partitional clustering methods with Bisecting K-Mean and

ICOST 2019, May 02-03, Makassar, Indonesia

DOI 10.4108/eai.2-5-2019.2284616

Buckshot algorithms, and hierarchical agglomerative methods with UPGMA cluster clusters and Complete Link calculation algorithms (Hamzah A, 2009). Multithreading implementation to improve Information Retrieval performance with the GVSM method, where researchers conduct testing by searching documents using threads and without threads in various document collections (Pardede J, 2014).

K-Nearest Neighbor is an effective method for classifying but has high computation time. *K-Nearest Neighbor is* less efficiently used in cases with large amounts of data (Jiang et al., 2012). Calculation of the distance between *testing* data anddata *training* must be done on the wholedata *training*. This study aims to develop *K-Nearest Neighbor* so that it can be used to reduce computing time. The proposed development method usestechniques *clustering* with *K-Means*. K-Means will divide the training data into several *clusters* with the same radius.calculation *K-Nearest Neighbor* will only be done on clusters that have the closest distance to the testing data (Jiang *et al*., 2012).

2 Material and Methods

This study uses data obtained in the form of a news article from kompas.com consisting of 5 categories. These categories include techno, science, automotive, economics and education. Each category represented 20 news articles so that the amount of data samples 100. The news article Kompasberita stored in a folder in the form of files that berekstensi.txt.

Using K-Means Clustering

Clustering Method using K-Means algorithm has the ability to group data in large enough quantities to computation time quickly and efficiently. The stages of K-Means algorithm as follows:

a. Specify the number of clusters
b. Determine the value of the centroid. Centroid value for the initial iteration is done randomly. Centroid value which is the stage of iteration, then use the formula in Equation.
   1.

$$\overline{V}_{ij} = \frac{1}{N_i}\sum_{k=0}^{N_i} X_{kj} \quad , \qquad (1)$$

   V_{ij} : the average cluster centroid to-i to variable j.
   N_i : the amount of data that is a member of the cluster to-i
   i,k : the index of the cluster
   j : the index of variables
   X_{kj} : k-th data values that exist in the cluster to the variable j.
c. Calculate the distance between the centroid point at each data point. To calculate the cosine similarity distance formula used in equation.
   2.

$$\cos\alpha = \frac{AxB}{|A||B|} = \frac{\sum_{i=1}^{n} A_i x B_i}{\sqrt{\sum_{i=1}^{n}(A_i)^2}\, x \sqrt{\sum_{i=1}^{n}(B_i)^2}}$$

(2)

cos α : *cosine similarity*
A : vektor A, yang akan dibandingkan kemiripannya
B : vektor B, yang akan dibandingkan kemiripannya
A x B : dot product antara vektor A dan vektor B
|A| : panjang vektor A
|B| : panjang vektor B
|A||B| : cross product antara |A| dan |B|

d. Group data by specifying the cluster members to calculate the cosine similarity value most in the cluster.
e. Back to stage 2, repeat until the value of the resulting fixed centroid and the cluster members do not move to another cluster.

Data Classification Using the K-Nearest Neighbor (KNN)

One algorithm used in the classification or prediction of new data is a K-Nearest Neighbor. The working principle of the K-Nearest Neighbor is to find the shortest distance between the data to be evaluated by the K nearest neighbors in the training data. The stages of K-Nearest Neighbor algorithm is as follows:

a. Determine the value of K
b. Calculate the distance of each object on the data given sample using the formula cosine similarity to the equation
3.

$$\cos \alpha = \frac{AxB}{|A||B|} = \frac{\sum_{i=1}^{n} A_i x B_i}{\sqrt{\sum_{i=1}^{n} (A_i)^2} x \sqrt{\sum_{i=1}^{n} (B_i)^2}}$$

(3)

cos α : *cosine similarity*
A : vektor A, yang akan dibandingkan kemiripannya
B : vektor B, yang akan dibandingkan kemiripannya
A x B : dot product antara vektor A dan vektor B
|A| : panjang vektor A
|B| : panjang vektor B
|A||B| : cross product antara |A| dan |B|

c. Sort objects in a group that has the largest cosine similarity
d. If the set value K = 5, then taken 5 cosine similarity value of the largest and most the majority of which will be used to take a decision.

3 Results and Discussion

Some menus contained in this document perangkingan Applications include: Dataset menu will display data news article. Perangkingan app featuring news documents dataset as in figure 1.

Figure 1. *Data Set*

Screening menu, featuring dataset screening results, in the form of data is already done preprocessing. The results of the screening process as in figure 2.

No	Term	Doc Id	Jumlah	Bobot
1	jakarta	1	2	1.99934
2	com	1	1	0.653926
3	usaha	1	3	4.28135
4	jasa	1	1	2.95651
5	moody	1	3	16.5644
6	investors	1	2	11.0429
7	service	1	2	5.62682
8	depan	1	4	6.28087
9	prediksi	1	2	4.95388
10	capai	1	1	1.30195

Figure 2. Screening Data

K-Means menu, display a form to fill in the number of clusters to be created, iterative process of clustering and each cluster member of each iteration. Results Clustering based K-Mean stages can be seen in Figure 3.

Figure 3. iterations on K-Mean Process for Total Cluster = 5

Menu K-Means and K-Nearest Neighbor, display form data input of testing and the number of clusters, the distance in the data table testing to all members of the data cluster testing, the order of closest to farthest distance data and the results of decisions according to the value of k is determined. KNN classification process with K-Mean clustering uses data testing id-data = 5 in the data screening by the number of clusters = 5. Layout-id system to display the data 5 can be seen in Figure 4.

Data Testing

ID Data	Kategori
5	Ekonomi

Figure 4. Data Testing with id = 5

Calculate the distance of each sample data in the cluster to 1 for all data other than the data on the testing of 1, while the layout of the system as in Figure 5.

Tabel Jarak dengan semua data pada Cluster yang dekat dengan data testing:

No	Id Data	Jarak	Kategori
1	1	0.048	Ekonomi
2	2	0.007	Ekonomi
3	3	0.154	Ekonomi
4	4	1.005	Ekonomi
5	6	0.014	Ekonomi
6	8	0.084	Ekonomi
7	9	0.205	Ekonomi
8	11	3.928	Ekonomi

Figure 5. Process Calculations Distance All Your Data with Data Testing on Cluster 1

Data sorted by the largest value of cosine similarity, as in Figure 6.

Tabel Jarak dengan semua data pada Cluster yang dekat dengan data testing yang sudah diurutkan jarak terdekat:

No	Id Data	Jarak	Kategori
1	19	20.093	Ekonomi
2	20	18.120	Ekonomi
3	21	15.239	Ekonomi
4	49	14.836	Ekonomi
5	18	11.563	Ekonomi
6	28	10.753	Ekonomi
7	17	9.540	Ekonomi
8	11	3.928	Ekonomi
9	40	2.431	Ekonomi

Figure 6. Sequence Data Based on Distance Nearest Testing Data in Cluster 1

If it is determined the value of K at KNN = 5, then the cosine value similatiry taken five largest and most majority that would be used to make decisions, such as in Figure 7. The result of the decision at K = 5 is the economy, together with the data testing on id = 5, namely economy.

Jika K =1 ,maka Hasil = Ekonomi, Proses KNN membutuhkan waktu 0.50633597374 seconds

Jika K =3 ,maka Hasil = Ekonomi, Proses KNN membutuhkan waktu 0.468783140182 seconds

Jika K =5 ,maka Hasil = Ekonomi, Proses KNN membutuhkan waktu 0.465476989746 seconds

Jika K =7 ,maka Hasil = Ekonomi, Proses KNN membutuhkan waktu 0.473922967911 seconds

Jika K =9 ,maka Hasil = Ekonomi, Proses KNN membutuhkan waktu 0.483791828156 seconds

Figure 7. The process of KNN with K-Mean and Calculation Time Computing

Menu K-Nearest Neighbor, display form data input testing, testing distance table data to all cluster members testing the data, the data sequence closest distance to the farthest and the decision of the appropriate values of k determined. KNN classification process with the data testing the data id = 5 in the data screening, as in Figure 8.

Data Testing

ID Data	Kategori
5	Ekonomi

Figure 8. Testing Data with Data Id = 5

Calculate the distance of each data screening by testing the data examined, ie id = 5 which is the id of existing data in the test 1. The same calculation is done for all the data in the data screening in addition to the existing data on testing 1 as in Figure 9.

Tabel Jarak dengan semua data pada Cluster yang dekat dengan data testing:

No	Id Data	Jarak	Kategori
1	1	0.048	Ekonomi
2	2	0.007	Ekonomi
3	3	0.154	Ekonomi
4	4	1.005	Ekonomi
5	6	0.014	Ekonomi
6	7	0.097	Ekonomi
7	8	0.084	Ekonomi
8	9	0.205	Ekonomi
9	11	3.928	Ekonomi
10	12	0.340	Ekonomi

Figure 9. Distance Data Screening with Data Testing

Sort the data into groups that have the greatest similarity cosine values, as in Figure 10.

Tabel Jarak dengan semua data pada Cluster yang dekat dengan data testing yang sudah diurutkan jarak terdekat:

No	Id Data	Jarak	Kategori
1	19	20.093	Ekonomi
2	20	18.120	Ekonomi
3	21	15.239	Ekonomi
4	49	14.836	Ekonomi
5	147	13.946	Sains
6	18	11.563	Ekonomi
7	28	10.753	Ekonomi
8	17	9.540	Ekonomi
9	149	6.328	Sains
10	103	5.898	Sains

Figure 10. Ordering Process Based Screening Data Distance Nearest Testing Data

If it is determined the value of K at KNN = 5, then taken 5 cosine similarity most majority to take decisions, as in Figure 11. The result of the decision at K = 5 is the economy, with data testing on id = 5 is economy.

Jika K =1 ,maka Hasil = Ekonomi, Proses KNN membutuhkan waktu 2.60434293747 seconds

Jika K =3 ,maka Hasil = Ekonomi, Proses KNN membutuhkan waktu 2.59490704536 seconds

Jika K =5 ,maka Hasil = Ekonomi, Proses KNN membutuhkan waktu 2.535187006 seconds

Jika K =7 ,maka Hasil = Ekonomi, Proses KNN membutuhkan waktu 2.58087420464 seconds

Jika K =9 ,maka Hasil = Ekonomi, Proses KNN membutuhkan waktu 2.62053108215 seconds

Figure 11. The process of KNN and Calculation Time Computing

The Results of the K-Mean Process for Determining Cluster

The results of purity calculations to determine the number of clusters to be selected in the next process can be seen in Figure 12.

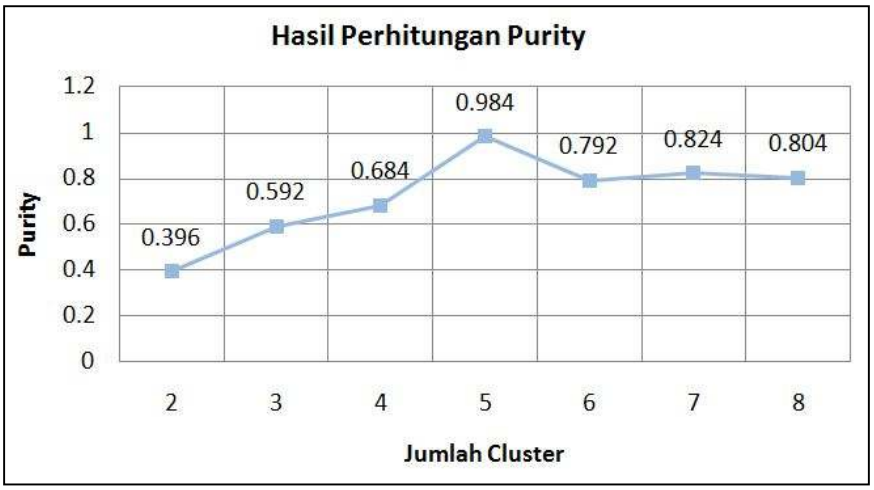

Figure 12. K-Mean Process for Determining Cluster

Figure 12 shows that the highest purity is obtained on the number of clusters = 5 at 0.984. Testing selected number of clusters is {2,3,4,5,6,7,8}, assuming that the number of categories of news = 5, then ideally the best cluster number is 5, so have the number of clusters = 5 and the number of clusters around 5 (3 down and 3 above 5).

KNN process results with K-Mean for Total cluster = 5

The result of the calculation accuracy of KNN with K-Mean for the number of clusters = 5 which has the highest purity grades can be seen in Figure 13.

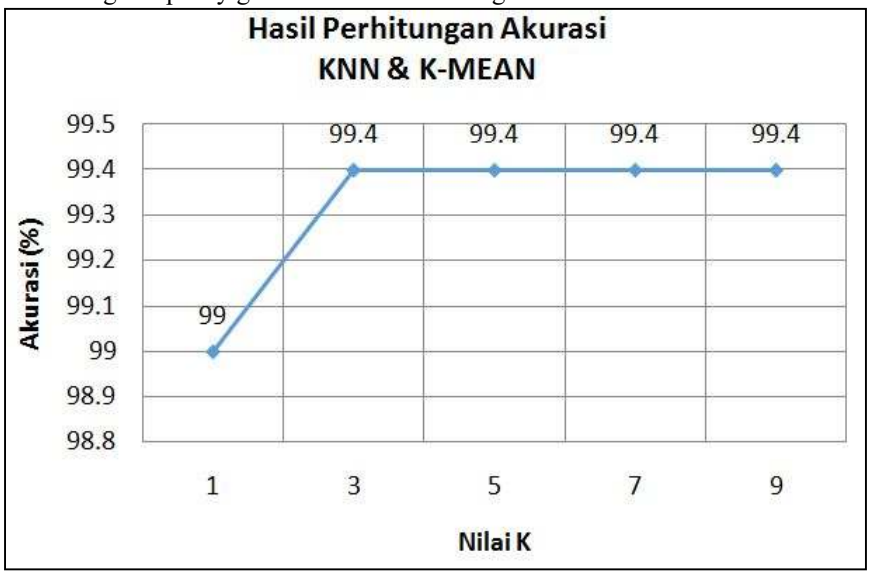

Figure 13. The process of KNN with K-Mean for Total Cluster = 5

Figure 13 shows the highest accuracy is obtained at K = 3, 5, 7 and 9 was 99.4% and the lowest accuracy

Pure KNN Process Results Without K-Mean

Pure KNN process accuracy results without K-Mean can be seen in Figure 14.

at K = 1 at 99%. At K = 1, 3, 5, 7 and 9 consistent 99%, resulting in the decision to classify the news could use K = 1, 3, 5, 7 or 9.

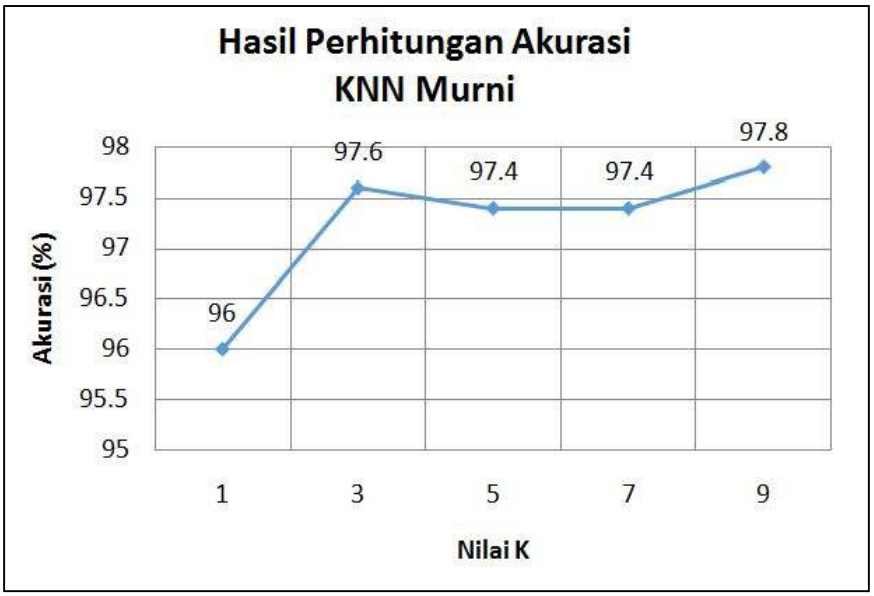

Figure 14. Process Without Pure KNN K-Mean

Figure 14 shows the highest accuracy was obtained for pure KNN without K-Mean at K = 9 97.8% and the lowest accuracy at K = 1 is 96%.

Results Comparison of Time Computing KNN with K-Mean and KNN Pure

On average computing in pure KNN and KNN with K-Mean can be seen in Figure 15.

Figure 15. Comparison of computing time with K-Mean KNN and KNN Pure

Figure 15 shows the average computation time for KNN with K-Mean is 0.5011 seconds while the computing time with pure KNN is 2.4841 seconds. It can be concluded that a pure KNN KNN longer than using a K-Means. It is caused by the amount of data processed by pure KNN more that all data contained in the data screening compared to the amount of data processed by KNN with K-mean that only processes data on the cluster that has the shortest distance with the data testing

Comparison of Computing Time and Accuracy Using KNN and K-Mean

The average computation time and accuracy on any number of clusters in the K-Mean KNN can be seen in Table 1.

Table 1. Average Time Computing and Accuracy in KNN with K-Mean

Total Cluster	Average	
	Time Computing	Accuracy
2	1,2661	97,82
3	0,8486	98,32
4	0,6337	97,82
5	0,5011	99,32
6	0,4243	98,70
7	0,3687	98,06
8	0,3344	98,60

Table 1 shows that the highest accuracy is obtained on the number of clusters = 5 is 99.32 and the lowest accuracy obtained on the number of clusters = 2 and 4, namely 97.82. Such accuracy is higher than the average accuracy on pure KNN at 97.24. This study shows that, in addition to the computing time can be reduced there is increased accuracy of KNN classification.

4 Conclusion

Merger KNN method for classifying dokemen news and clustering using K-Nearest Neighbor can reduce computation time search news document in Indonesian language. The test results showed the average computation is required 0.5011 seconds. Compared with the use of the method of K-Nearest Neighbor KNN without requiring computation time 2,4841detik.

References

[1] Hamzah A. *Temu Kembali Informasi Berbasis Kluster untuk Sistem Temu Kembali Informasi Teks Bahasa Indonesia*, Jurnal Teknologi, 2(1), 1-7. (2009)

[2] Jiang, Shengyi. Dkk : *An Improved K-Nearest-Neighbor Algorithm For Text Categorization*. School of Informatics, Guangdong University of Foreign Studies, 510420 Guangzhou, China. (2012)

[3] Oktaviani, Diar Puji : *Menjadi Programmer Jempolan Menggunakan PHP*. Yogyakarta : Mediakom, Indonesia (2010)

[4] Pardede, J. *Implementasi Multithreading untuk Meningkatkan Kinerja Information Retrieval dengan Metode GSVM,* Jurnal Sistem Komputer, 4(1), 2252-3456 (2014)

[5] Prilianti, K. R. danHendra Wijaya : *Aplikasi Text Mining untuk Automasi Penentuan Tren Topik Skripsi dengan Metode K-Means Clustering*. Jurnal Cybermatika (2014)

[6] Purwanti, Endah : *Klasifikasi Dokumen Temu Kembali Informasi dengan K-Nearest Neghbour.* Surabaya, Fakultas Sains dan Teknologi Universitas Airlangga, Indonesia (2015)

[7] Raharjo, Budi : *Mudah Belajar PHP*. Bandung : Informatika Bandung, Indonesia (2015)

[8] Rismawan, Tedy.dkk : *Sistem Pendukung Keputusan Berbasis Pocket Pc Sebagai Penentu Status Gizi Menggunakan Metode Knn (K-Nearest Neighbor)*. Laboratorium Komputasi dan Sistem Cerdas, Jurusan Teknik Informatika, Fakultas Teknologi Industri, Universitas Islam Indonesia (2008)