Document Similarity Detection using K-Means and Cosine Distance

Wendi Usino[1], Anton Satria Prabuwono[2], Khalid Hamed S. Allehaibi[3], Arif Bramantoro[4], Hasniaty A[5], Wahyu Amaldi[6]
Faculty of Information Technology, Universitas Budi Luhur, Jakarta, Indonesia[1,2,4,6]
Faculty of Computing and Information Technology Rabigh King Abdulaziz University, Rabigh, Saudi Arabia[2,4]
Faculty of Computing and Information Technology, King Abdulaziz University, Jeddah, Saudi Arabia[3]
Institute of Visual Informatics, Universiti Kebangsaan Malaysia, Bangi, Selangor, Malaysia[5]
Faculty of Engineering, Universitas Hasanuddin, Makassar, Indonesia[5]

***Abstract*—A two-year study by the Ministry of Research, Technology and Education in Indonesia presented the evaluation of most universities in Indonesia. The findings of the evaluation are the peculiarities of various dissertation softcopies of doctoral students which are similar to any texts available on internet. The suspected plagiarism behavior has a negative effect on both students and faculty members. The main reason behind this behavior is the lack of standardized awareness among faculty members with regard to plagiarism. Therefore, this study proposes a computerized system that is able to detect plagiarism information by using K-means and cosine distance algorithm. The process starts from preprocessing process that includes a novel step of checking Indonesian big dictionary, vector space model design, and the combined calculation of K-means and cosine distance from 17 documents as test data. The result of this study generally shows that the documents have detection accuracy of 93.33%.**

***Keywords*—K-means; cosine distance; cluster; document similarity; document frequency; inverse document frequency; preprocessing; vector space model**

I. Introduction

There is a two-year study conducted by the Academic Performance Evaluation team in the Ministry of Research, Technology and Higher Education that found an indication of plagiarism in various universities in Indonesia. The finding starts from several anomalies in doctoral dissertations which are published in electronic format. It is explicitly said in the report that there are several irregularities contained in the dissertation document footprint and a number of dissertations similar to open source texts on the internet.

Plagiarism behavior has been identified for both students and faculty members. One of the main reasons of this behavior problem found is due to the different standard of plagiarism issue between faculty members graduated from local universities and those graduated overseas. There is a perception amongst academics in Indonesian institution that overseas graduates are stricter in plagiarism issue [1]. This perception triggers reluctance amongst faculty members to raise the plagiarism issue due to the fear of being harshly judged. There should be a further concern about standards and consistency in preventing plagiarism in higher education institutions, especially in law enforcement. The detection of plagiarism is non-trivial given the facts that there is an increasing amount of information generated from an easy access of various websites, large databases and social media that pose serious problems for publishers, researchers and educational institutions.

Document grouping [2] is a technique to organize a large number of documents. This technique is usually the unsupervised learning which has no identification of any classes. Unlike classifications technique, data are grouped into groups according to their similarity. The obtained cluster indicates a meaningful category. The result is used as a basis for documents classification. Document categorization is also useful for fast information retrieval and data mining. Any documents have possibility of having word similarities within the same topic. The cluster contains very similar documents.

One of existing non-hierarchical cluster methods is K-means [3] that partitions existing data into one or more clusters. It is important to note that K-means algorithm is considerably sensitive to outliers. Outliers are data far from the majority of other data, and thus inapplicable when inserted into a cluster. This kind of data can distort the cluster mean value excessively. Due to the time constraint, this research assumes the outliers are insignificant to the result of the research.

This research aims to combine two different measurements for detecting Indonesian documents similarity by utilizing Indonesian big dictionary. In detail, the research objective is to implement the document similarity detection system in order to observe the performance of the proposed technique.

II. Literature Review

A. Theoretical Background

Data analysis is currently the heart of most computing applications, especially during the design phase. The data analysis process is practically categorized as simplification and exploration, which is based on the availability of a model that appropriately represents the real data source. The main procedures in both types of procedures during hypothesis formation and decision making are clustering and classification based on postulated model and analysis results.

Clustering is the arrangement technique of patterns (usually described as points in multidimensional space or measurement vectors) into groups based on similarity or dissimilarity.

Clustering is a non-trivial method of analyzing particular data. It is a method of creating a collection of items that are somewhat related in one or more features. The purpose of grouping is to provide similar data groups. Clustering is often misinterpreted as classification. The simplest difference is regarding to the measure. The measure in clustering is based on the intra-cluster distance that should be reduced in order to get the best clustering results. The term of clustering is commonly used to refer a grouping method of data that are not yet labeled. Clustering and classification [4] have different terminology and assumptions for the grouping process components, as well as the context in which grouping is being used.

A comprehensive survey is very important step due to the large amount of literatures regarding to grouping issue. Accessibility to the survey is another issue to reconcile different vocabulary and assumptions regarding to grouping in various research. Authors in [5] present that typical pattern grouping activities involve the modeling of data point, the calculation of data point relatedness, clustering or grouping, the abstraction of processed data, and eventually the evaluation of result. These activities are accommodated in our research to fulfill the objective.

K-means is one of existing cluster partition algorithms. In this algorithm, the partition that has data considered as cluster *k*. Other clustering algorithms are also proposed to handle document grouping tasks for automatic grouping and enhanced partition of K-means algorithm, such as a method for initializing centroid [6], [7], oncology-based K-means algorithm, domain ontological grouping [8], and dataset based analysis to increase the efficiency of the K-means algorithm in case that the false document is given as input [9].

Cosine distance is a measure of the similarity between two vectors based on the cosine angle between them. This study proposes a document similarity detection system by clustering and calculating the cosine angle between the examined documents.

In a combined algorithm of K-means and Cosine distance, there are *n* data points that are divided into *k* clusters based on several similarity measurement criteria. The K-means algorithm is relatively agile and thus considered as a common clustering algorithm. Vector quantization, cluster assessment, feature discovery are several examples of K-means utilization as surveyed in [3].

K-means algorithm starts from selecting the number of *k* clusters, assigning each data point to the nearest cluster center, and moving each cluster center to the average and last data points. These steps are repeated several times to achieve the convergence. The final result of the K-means algorithm is the suitable number of clusters. Creating the number of clusters before implementing the algorithm is considered as impractical. It also requires in-depth knowledge of the field of clustering. Before applying vector space models to the text documents, an information retrieval is performed through a preprocessing. Preprocessing input is plain text documents and its output is a set of tokens utilized in the vector model.

B. Related Works

There are several related works that are reviewed to observe state of the art in the research domain. Rajeswari et al. [2] implements K-means algorithm to a group of news articles with 20 categories that requires predefined cluster names. The result shows that K-means algorithm is unable to cluster the article documents automatically without considering the feature as a cluster label. Moreover, there is an issue of over clustering. It means that there is a need of clustering repetition until all documents are correctly clustered. Hence, it can be inferred that the combination with another algorithm is required in several domains.

Bhattacharjee et al. [10] proposes the use of cosine similarity measure to cluster sentiment analysis between -2 (very negative) and +2 (very positive) for 8000 comments on telecommunication domain. The result shows 82.09% accuracy for two classes of negative and positive. It outweighs previous works have 71.5% accuracy in average. It inspires us to include cosine technique for clustering our documents.

Bafna et al. [11] proposes the combination of K-means and hierarchical algorithms. It initially starts from small dataset and advances to an extended one while different clusters are being created. In total, there are 10,000 documents for the whole classification process. The result shows that the combined algorithm is able to classify two classes of positive-negative and three classes of positive-negative-neutral.

Shirkhorshidi et al. [12] compares several similarity measurements based on distance. He utilizes 15 generic datasets to reproduce the clustering results. Hence, the distance measurement performance can be measured based on its category and dimension.

Rani and Sahu [13] compare several clustering techniques in measuring textual contents and articles. The searching keywords are identified based on the most relevant content. Matrix of keywords is built to compare the different algorithms in Matlab. However, it only chooses news article that triggers a further research question for other article types.

III. System Design

The research of detecting document similarity by using K-means algorithm and Cosine distance method is considered as applied research. The result of applied research can be directly incorporated to solve the problems. Moreover, the combination between K-means algorithm and Cosine distance method requires a specific process schema to fulfil the objective of the research.

Fig. 1 illustrates the process flow of document similarity detection system that emphasizes on the preprocessing. The document requires preprocessing in order to calculate the similarity between documents. More specifically, the purpose of preprocessing is to extract special features on documents for information retrieval. The first step in preprocessing is filtering that eliminates any punctuations and special characters.

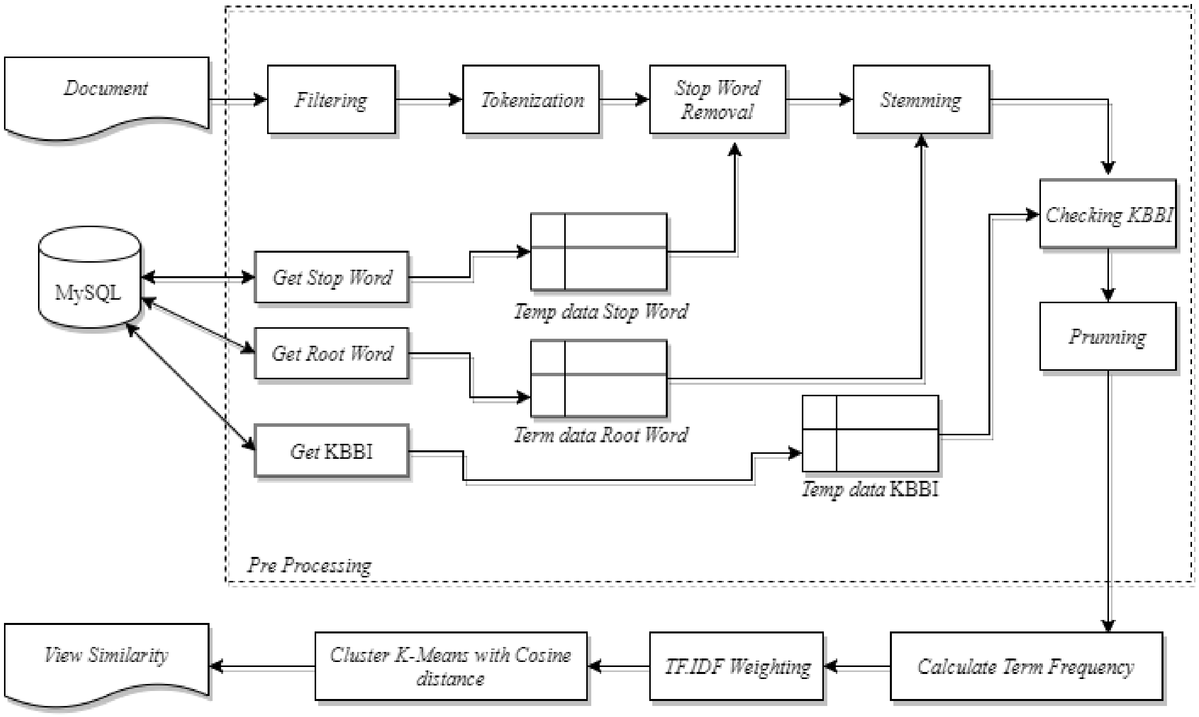


Fig. 1. Process Scheme to Detect Document Similarity.

The second step is tokenization. It is a function to convert a sentence into a list of words. The third step is stop word removal. Any words that have no meaning in the vector space are removed in this step. The fourth step is stemming which uses Indonesian language stemming to convert any words into their basic forms. The last step in preprocessing is pruning that sorts all words and removes any words with low frequency.

Although the common preprocessing stage consists of filtering, tokenization, stop word removal, stemming and pruning; this study proposes to add one process before the stemming process. It is a process of validating the term whether it is registered in the Indonesian big dictionary (KBBI) [14] or not. This process ensures that the term has a real concept in Indonesian language and, therefore, is applicable for any documents in Indonesian language.

Object oriented design with unified modeling language is utilized to design the system. The design starts from the results of data collection and literature study to obtain the system requirements specifications. It includes program specification design, system process design, system flow design and system application interface design.

Once the system has been developed, the testing is conducted by including the preprocessing, vector space model, K-means clustering, cosine similarity, and accuracy testing.

It is important to note that this research uses purposive sampling technique which is one of the common sampling techniques in other research works. This sampling technique deliberately takes samples based on predetermined criteria. Hence, the size of the dataset is considerably not big since we are focusing on the design of the system.

In order to acquire a better accuracy and performance of the detection model, the document is preprocessed to prepare the terms readable by K-means & cosine distance algorithm. The weighting through the vector space model is also required by the algorithm. This research uses dataset which consists of three article categories, namely sports, news and finance. Each category has 20 articles which are proportionally distributed in 17 similarity detection scenarios as provided in Table 1.

TABLE I. GROUPING THE DATASETS

ID	Description
1	20 sport articles
2	17 sport articles
3	14 sport articles
4	11 sport articles
5	8 sport articles
6	20 news articles
7	17 news articles
8	14 news articles
9	11 news articles
10	8 news articles
11	20 finance articles
12	17 finance articles
13	14 finance articles
14	11 finance articles
15	8 finance articles
16	3 sport, news, finance articles
17	6 sport, news, finance articles

Class diagram is generally used by the system developers to obtain the glimpse of the system structure before the code is written. It is also useful to ensure that the system is implemented based on the most optimized design. Fig. 2 shows the class diagram for similarity detection system using K-means and cosine distance. In this diagram, there are four classes designed to accommodate the system requirement: Document, Distance, Tokenization and Term Frequency-Inverse Document Frequency. Each class has its properties and methods to seamlessly develop the application. Document class behaves as main class and uses three other similarity information classes due to the nature of document concept in the real world.

The similarity detection system starts from the user input. It advances all the processes until the system outputs the similarity distance value in matrix view as illustrated in Fig. 3.

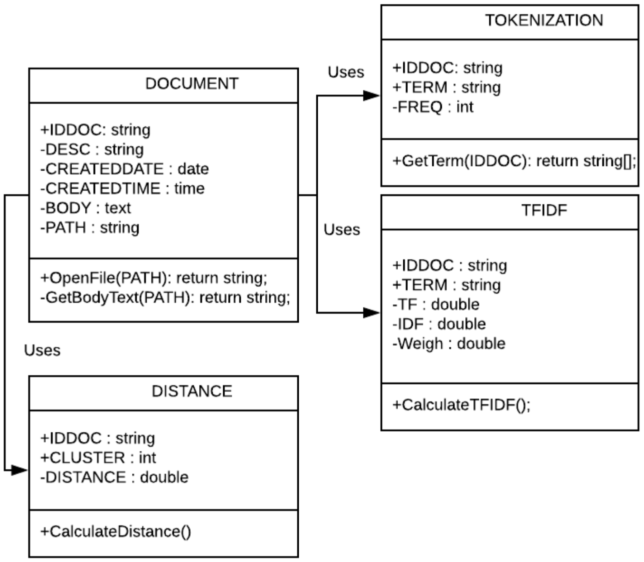


Fig. 2. Similarity Detection Class Diagram.

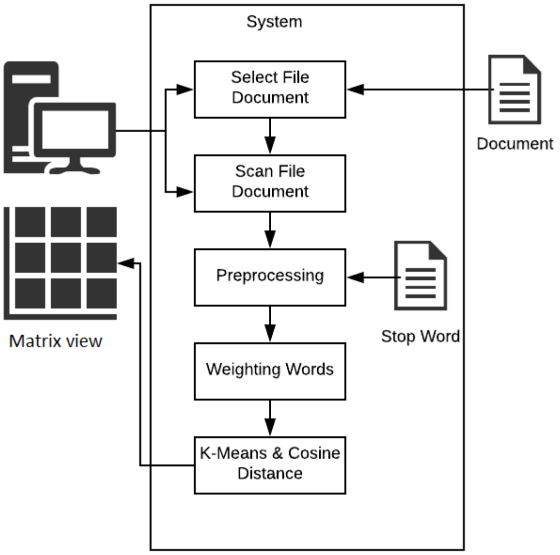


Fig. 3. System Process Design.

IV. Evaluation

The evaluation is required to examine the reliability of the proposed approach. Preprocessing is the initial stage for processing input data before being processed in the main stages of the vector space model. Preprocessing is required to establish the uniformity and ease of reading during the subsequent processes. In this study, the proposed preprocessing steps are tokenization, stop word removal, stemming, Indonesian big dictionary based checking and pruning. The result of the preprocessing is represented in Table 2.

It can be inferred that the preprocessing is required to process 17 documents with 52,805 words. It produces 30,969 words which equals to 58.64% of the total words in the document being tested. It creates a vector of 1,500 unique words in 7.997 milliseconds. In other words, it is around 0.47 milliseconds per document. This preprocessing time is considered as reliable in this research.

Vector space model is tested to determine the time that is required to process data provided during preprocessing. There are three calculations during vector space model testing, namely term frequency, document frequency & inverse document frequency as normally used as an evaluation technique in clustering [11]. Term weight is added in this research to increase the reliability.

Each of the evaluation is represented in different figure to show the clarity and interdependency of the calculation. Fig. 4 illustrates the testing of calculating term frequency for all documents. The calculation of term frequency that is produced by the system will be used as a reference for the calculation process and word weighting process.

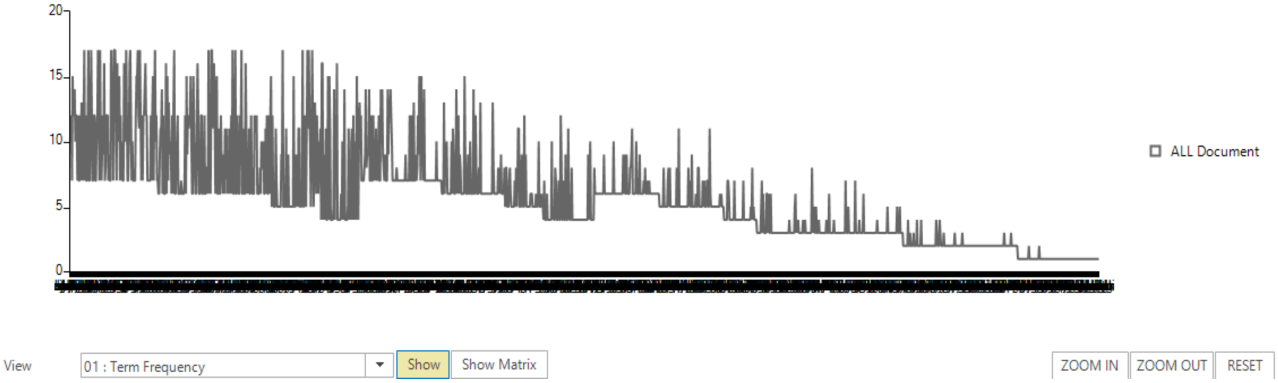
Fig. 5 is a graph of the application test in calculating the document frequency. This graphs shows that the higher the value in each document, the more terms contained in the document.

Fig. 6 represents a graph of an application test in calculating inverse frequency document. The more terms that the document has, the smaller inverse value the document implies. Accordingly, the fewer terms appear on a document, the more inverse value is resulted from the document.

The application test of generating vector model space is divided into several tests on 17 preprocessed articles. Each article has a feature vector of 1,500 words. Hence, the matrix dimension created during the term frequency process is 17 x 1500.

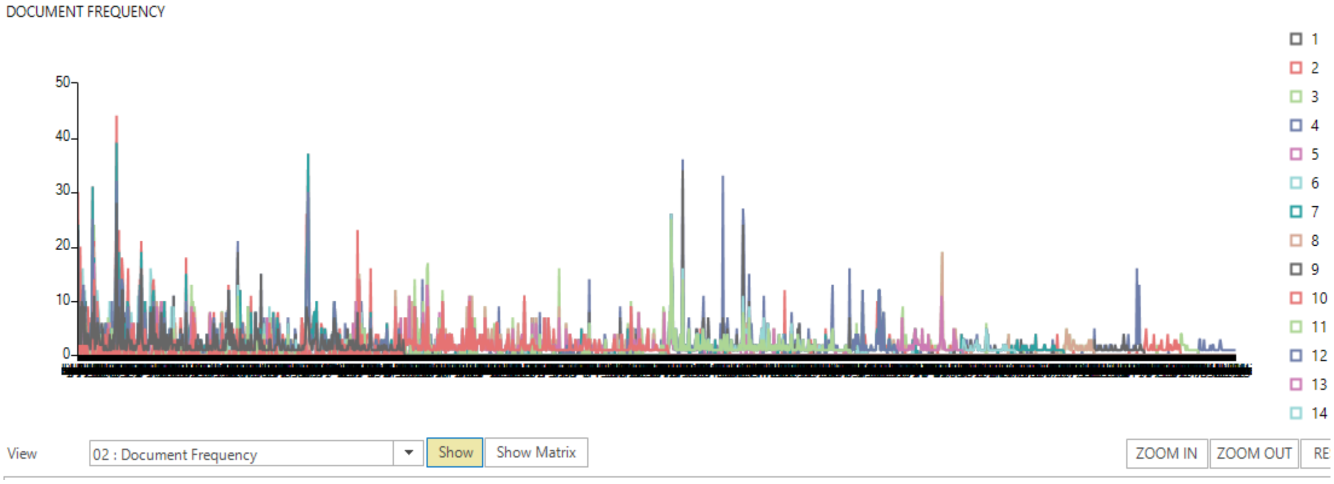
TABLE II. Result of Preprocessing

	Number of Terms		**Duration (ms)**
	Before	**After**	
Total	52,805	30,969	7.997
Average	3,106.17	1,821.7	0.4704
Average Percentage		58.64%	



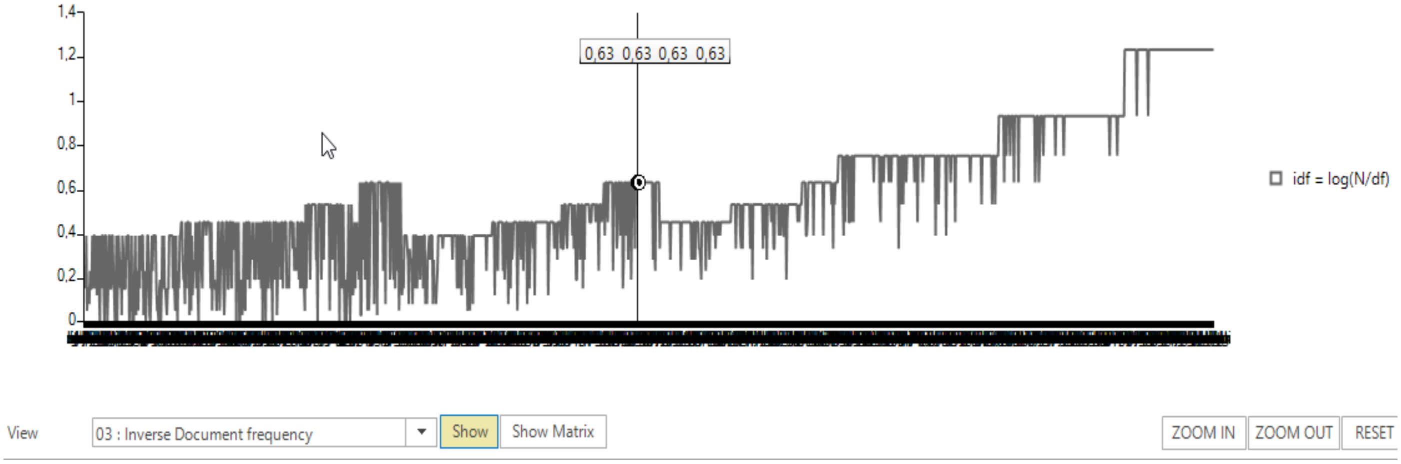
View: 01 : Term Frequency

Fig. 4. Application Testing on Term Frequency.



View: 02 : Document Frequency

Fig. 5. Application Testing on Document Frequency.



View: 03 : Inverse Document frequency

Fig. 6. Application Testing on Inverse Document Frequency.

Table 3 presents the duration of the vector space model generation. The total duration of Term Frequency (TF), Document Frequency-Inverse Document Frequency (DF-IDF), and weighting calculation has the average of 119.2969 milliseconds. For K-means and cosine distance testing, the data are taken from the preprocessing and vector space model.

In Table 4, the cluster value represents similar documents, while cosine value is the angular distance to other documents. It can be concluded that the K-means algorithm and cosine distance are able to detect the similarity of documents. Out of 15 total documents, there are 14 correct documents, and one wrong document. It means that the arbitrary accuracy is 93.33%.

TABLE III. THE DURATION FOR GENERATING VECTOR SPACE MODEL (IN MILLISECONDS)

	TF	DF-IDF	Weighting
Test 1	27.0380	31.0140	58.5830
Test 2	26.5502	32.4380	59.3690
Test 3	27.9404	34.9580	60.0000
Average	27.1762	32.8033	59.3173

TABLE IV. RESULTS OF K-MEANS TESTING AND COSINE DISTANCE

ID	Short Text	Cosine	Cluster
1	11 finance articles.docx	0.839715	2
2	11 news articles.docx	0.804417	0
3	11 sport articles.docx	0.908695	1
4	14 finance articles.docx	0.961371	2
5	14 news articles.docx	0.924069	0
6	14 sport articles.docx	0.931172	1
7	17 finance articles.docx	0,966877	2
8	17 news articles.docx	0.929583	0
9	17 sport articles.docx	0.95024	1
10	20 finance articles.docx	0.948105	2
11	20 news articles.docx	0,908088	0
12	20 sport articles.docx	0.894802	1
13	3 sport, news and finance articles.docx	0.42862	1
14	6 sport news and finance articles.docx	0.507354	0
15	8 finance articles.docx	0.726018	2
16	8 news articles.docx	0.731459	0
17	8 sport articles.docx	0.857179	1

V. CONCLUSION

The documents are initially passed in the similarity detection system by preprocessing them to get the vector. In preprocessing, this research validated the terms in documents to Indonesian big dictionary. Vector Space Model is used to calculate the document similarity by combining the K-means and cosine distance algorithms. The simple accuracy measurement formula is applied to identify the results of document similarity detection. In the test result, the processing time of 17 document schemes at the preprocessing stage is 7.997 milliseconds, while the processing time of vector space model process is 119,296 milliseconds. The system delivers the document similarity detection accuracy of 93.33%. In the future, it is expected to apply this research in a bigger dataset by including online articles in order to improve its reliability.

REFERENCES

[1] T. S. Adiningrum, “Reviewing plagiarism: an input for indonesian higher education,” Journal of Academic Ethics, pp. 107-120, 2015.

[2] F. Rozi, and F. Sukmana, “Document grouping by using meronyms and type-2 fuzzy association rule mining,” Journal of ICT Research and Applications, 11(3), pp. 268-283, 2017.

[3] K. Rajeswari, O. Acharya, M. Sharma, M. Kopnar, and K. Karandikar, "Improvement in k-means clustering algorithm using data clustering," in Proc. International Conference on Computing Communication Control and Automation, pp. 367-269, 2015.

[4] P. Arabie, and G. De Soete, Clustering and classification, World Scientific, 1996.

[5] D. Pal, A. Jain, A. Saxena, and V. Agarwal, “Comparing various classifier techniques for efficient mining of data,” in Proc. of the International Congress on Information and Communication Technology, Computer Science Engineering, India, pp. 191-202, 2016.

[6] K. B. Aljanabi and A. H. Aliwy, “An efficient algorithm for initializing centroids in k-means clustering,” Journal of Kufa for Mathematics and Computer, 3(2), pp. 18-24, 2016.

[7] L. H. Patil, and M. Atique, “A novel approach for feature selection method TF-IDF in document clustering,” In Proc. of Advance Computing Conference (IACC), 2013 IEEE 3rd International, pp. 858-862, 2013.

[8] Y. Cheng, Y. Qiao, and J. Yang, “An improved markov method for prediction of user mobility,” in Proc. of 12th International Conference on Network and Service Management and Workshops (CNSM 2016), pp. 394–399, 2016.

[9] T. Kanungo, D. M. Mount, N. S. Netanyahu, C. D. Piatko, R. Silverman, and A. Y. Wu, “An efficient k-means clustering algorithm: Analysis and implementation,” IEEE Transactions on Pattern Analysis & Machine Intelligence, (7), pp. 881-892, 2002.

[10] S. Bhattacharjee, A. Das, U., Bhattacharya, S. K. Parui, and S. Roy, “Sentiment analysis using cosine similarity measure,” in Proc. of IEEE 2nd International Conference on Recent Trends in Information Systems (ReTIS), pp. 27-32, 2015.

[11] P. Bafna, D. Pramod, and A. Vaidya, “Document clustering: TF-IDF approach,” in Proc. of Electrical, Electronics, and Optimization Techniques (ICEEOT), International Conference on, pp. 61-66, 2016.

[12] A. S. Shirkhorshidi, S. Aghabozorgi, and T. Y. Wah, “A comparison study on similarity and dissimilarity measures in clustering continuous data,” PloS one, 10(12), pp. 1-20, 2015.

[13] U. Rani, and S. Sahu, “Comparison of clustering techniques for measuring similarity in articles,” in Proc. of The 3rd International Conference on Computational Intelligence & Communication Technology (CICT), pp. 1-7, IEEE, 2017.

[14] D. Moeljadi, I. Kamajaya, and D. Amalia, “Building the kamus besar bahasa indonesia (kbbi) database and its applications,” in Proc. of The 11th International Conference of the Asian Association for Lexicography, pp. 64-80, 2017.