DESIGN OF A FEASIBLE DOCUMENT CLUSTERING STRATEGY FOR PLAGIARISM DETECTION OF A NATIONAL RESEARCH, REPORT AND MANAGEMENT PORTAL SYSTEM IN KOREA

Kwangho Song [a], Joonwoo Jeon [b], Yoo-Sung Kim [c]
[abc] Inha University, Incheon, Korea
Corresponding email: yskim@inha.ac.kr

Abstract

A feasible document clustering strategy is proposed to improve the execution time of plagiarism detections in national research report management portal system in Korea. The proposed document clustering scheme uses both a dimension reducing technique and a synonym dictionary to produce the appropriate clusters even with large number of documents. Hence, since the actual plagiarism detection is needed to be applied to only the related clusters not entire documents in the database, the execution time efficiency of the plagiarism detection can be improved.

***Keywords*:** Plagiarism Detection, Document Clustering, Dimension Reduction, Synonym Dictionary.

1. Introduction

Nowadays, document plagiarisms have become a big problem, in not only academic communities, but also social organizations throughout the world (Velásquez et al., 2016). In general, the process of the previous plagiarism detection systems such as 'turn-it-in' (iParadigms, 1997) have been effective to a point. It aims to treat massive amount of documents consists of three steps; the first step is to store the meta data of documents and document themselves into the database, the second step is to select the similar documents from the database to the input target document, and the last is to locate exactly plagiarized areas in the detected similar documents.

One of the popular methods for the second step, selecting similar documents from the database is filtering (Song et al., 2015). Filtering is a kind of sifting method that can remove the unrelated documents from the candidate list and finally collect only the related documents from the database through similarity checking against the input target document. Hence, the number of documents that should be scanned to detect plagiarized areas in the last step can be reduced. To reduce the number of documents to be scanned in the last step, as another alternative way, clustering can be used in the first step of storing documents. Clustering is a kind of grouping method that classifies documents into subgroups according to their similarity between themselves (Kang, Joo & Lee, 2003). Hence, if the documents are stored according to the clustering results in the database, only the documents in the similar clusters will need to be scanned to detect the plagiarized areas instead of whole documents in the database in the last step.

A document plagiarism detecting system that can be used in a Web environment, in dealing with a huge amount of documents, has been developed by our research team (Song et al., 2014). In this system, the filtering method using Jaccard-coefficient algorithm and Cosine-similarity algorithm for the second step of selecting similar documents is adapted. However, this system seems to have some weaknesses. The most critical problem is time delay caused by filtering process in the second step. As such, the filtering process needs more running time than that for the last step of actual detecting plagiarized areas in the Web environments.

In this paper, as a solution to the problem, the clustering method is used before the first step to reduce the execution time of the second step of selecting similar documents. Also, to get higher accuracy of the document clustering results, a feasible document clustering strategy which uses a dimension reduction technique and synonym dictionary for plagiarism detection is proposed.

2. Related work

To manage the copyrights of the national research reports in Korea, we developed a national research report management portal system that is working in the Web environment, in which a document plagiarism detection system is developed and employed (Song et al., 2014; Song, Min & Kim, 2015). This plagiarism detection system is able to detect not only the typical cloning types of plagiarisms, but also the synonym substitutions and spacing manipulation types (Song et al., 2014). [Figure 1] shows the process flow of the system. In that system, the two-phase filtering scheme, which is developed to efficiently select only similar documents from the database to the input target document by the Jaccard-coefficient algorithm and the Cosine distance algorithm, is used for the second step. In addition, for the last step of actual detecting plagiarized areas, the Euclidean distance algorithm is also used to compute the similarity between target document and one of the returned results from the filtering process.

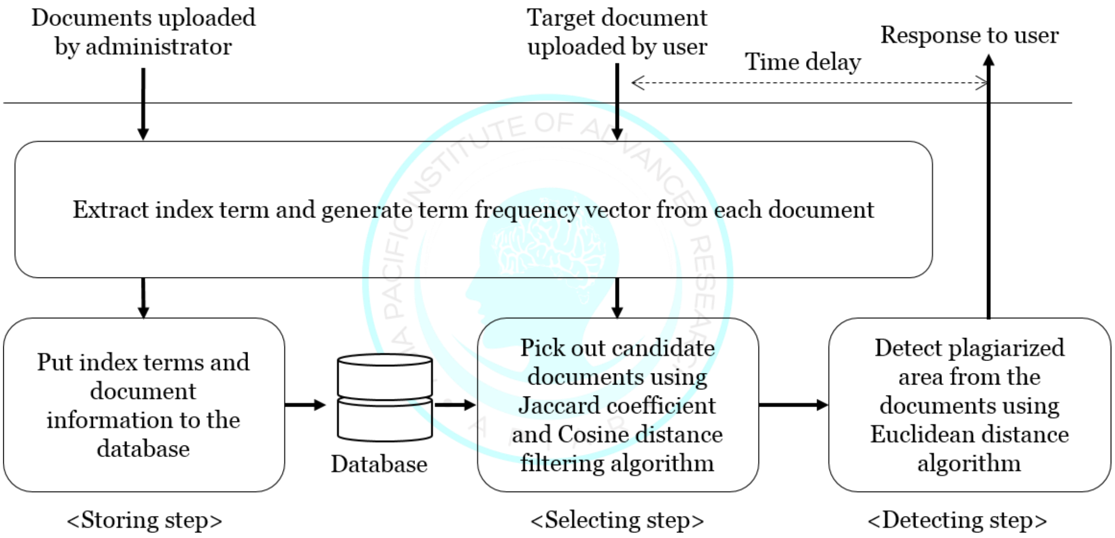
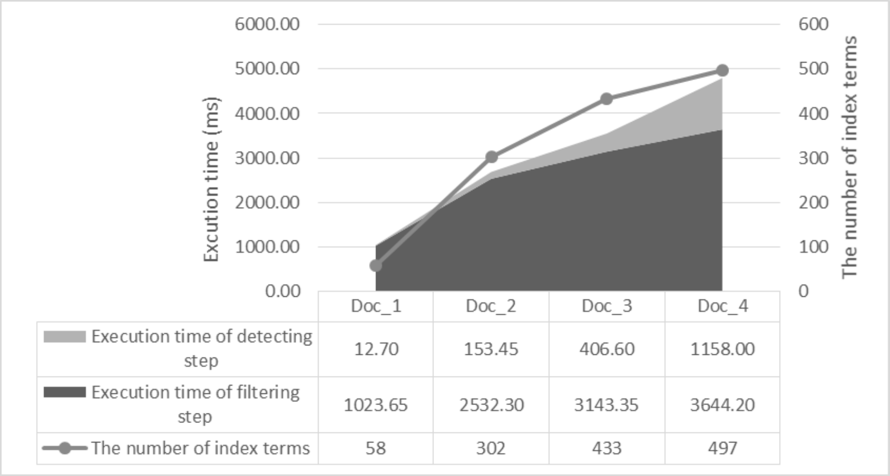

Documents uploaded by administrator
Target document uploaded by user
Time delay
Response to user
Extract index term and generate term frequency vector from each document
Put index terms and document information to the database
Database
Pick out candidate documents using Jaccard coefficient and Cosine distance filtering algorithm
Detect plagiarized area from the documents using Euclidean distance algorithm
<Storing step>
<Selecting step>
<Detecting step>

Figure 1: Process Flow of Plagiarism Detection System

However, although the two-phase filtering scheme can reduce the number of candidate documents to be scanned to detect plagiarisms in the last step, the more documents are added, the total execution time is required. Especially the execution time for filtering step comprises predominantly the total execution time irrespective of the numbers of index terms in documents of as shown in [Figure 2].



	Doc_1	Doc_2	Doc_3	Doc_4
▨ Execution time of detecting step	12.70	153.45	406.60	1158.00
■ Execution time of filtering step	1023.65	2532.30	3143.35	3644.20
●— The number of index terms	58	302	433	497

Figure 2: Execution Time Constitution in Plagiarism Detection System

3. A feasible document clustering strategy for plagiarism detection system

It is very important to reduce the filtering execution time for providing efficient plagiarism detection services to users in the web environments. So, as a time reducing way, applying a document clustering strategy as the first step before storing document into the database is proposed as shown in [Figure 3].

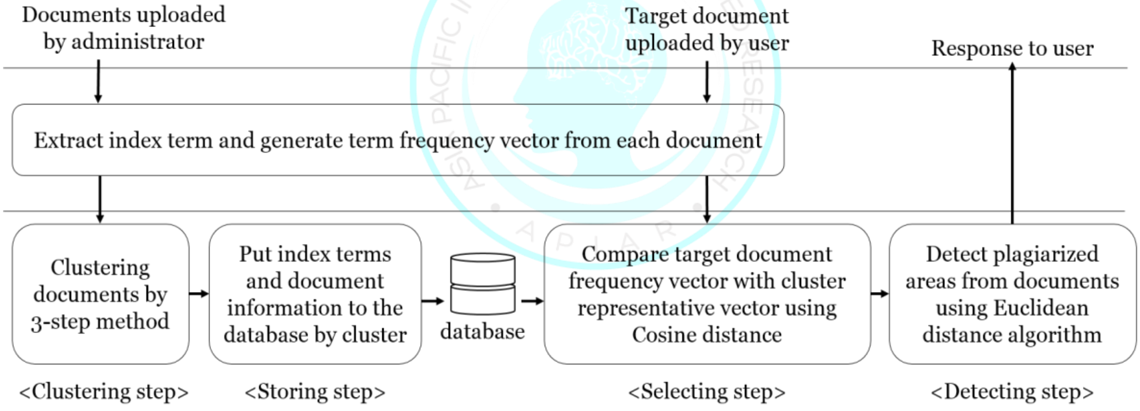

Documents uploaded by administrator
Target document uploaded by user
Response to user
Extract index term and generate term frequency vector from each document
Clustering documents by 3-step method
Put index terms and document information to the database by cluster
database
Compare target document frequency vector with cluster representative vector using Cosine distance
Detect plagiarized areas from documents using Euclidean distance algorithm
<Clustering step>
<Storing step>
<Selecting step>
<Detecting step>

Figure 3: Process Flow of Enhanced Plagiarism Detection System

However, inappropriate document-clustering may not reduce the execution time of the filtering step but rather increase the execution time from time to time. Thus, a feasible document clustering strategy in which a dimensional reduction technique and synonym dictionary is used is proposed to improve the execution time of plagiarism detections in the national research report management portal system in Korea.

According to the service process flow in [Figure 3], when the administrator uploads a bunch of documents, the system extracts index terms and computes their appearance frequencies in each document to generate the term frequency vector for each document as same in the previous system. Then, the system begins the document clustering shown in [Figure 4].

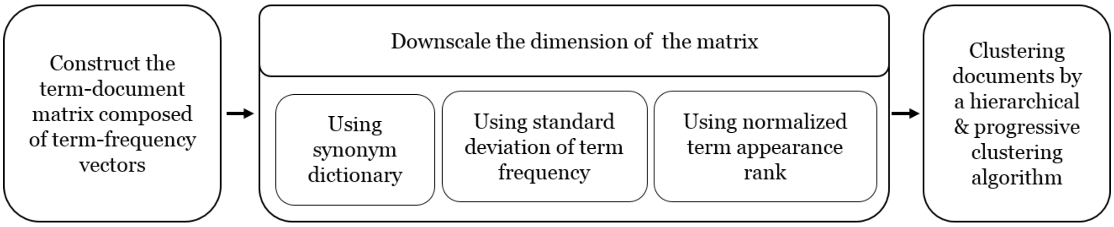

Construct the term-document matrix composed of term-frequency vectors
Downscale the dimension of the matrix
Using synonym dictionary
Using standard deviation of term frequency
Using normalized term appearance rank
Clustering documents by a hierarchical & progressive clustering algorithm

Figure 4: Process Flow of Document Clustering

For document clustering, the system constructs the term-document matrix from the term-frequency vectors for all documents. In the term-document frequency matrix, the frequency and dimension of analogous words among the index terms in the matrix are merged by using the synonym dictionary, so the matrix can be transformed into a smaller matrix that consists of normalized appearance ranking vectors of index terms and standard deviation respecting appearance frequencies of index terms in each document. Through this series of processes, we can get the dimension-reduced matrix to enhance the accuracy of clustering result and to compensate the drawback which caused by ‘curse of dimensionality’. In succession, the clustering step, which uses the modified term-document matrix to make the document groups, is carried out before the storing step. In this step, the clustering algorithm is a hierarchical agglomerative clustering algorithm such like ‘Single Link’, ‘Complete Link’ or ‘Group Average Link’ (Fagin et al., 2000) which is known as producing more accurate results than flat clustering methods. Therefore, the structure of each cluster looks like tree shape. Leaf node represents a document and the root is the representative feature vector of the cluster. Based on this clustering algorithm, the system makes the document cluster progressively to handle incrementally uploaded documents and to improve the accuracy of clustering results. Lastly, the system puts the index terms and the meta data of document into the database by cluster in the storing step as shown in [Figure 3].

When user who wants to check whether his or her target document has the plagiarized areas or not requests and submits the target document to the plagiarism detection system. The system extracts the index terms and makes term frequency vector of target document through the same process mentioned above. In succession, as the selecting step, the system selects the proper cluster, which has the similar representative feature vector with the term frequency vector of the target document through the similarity check using Cosine similarity algorithm. Unlike previously, the system does not check the similarity between the target document and all documents stored in the database, the number of documents that should be checked are significantly reduced. Therefore, the time performance of the service can be improved naturally.

After that, in the detecting step, the plagiarism detection system detects plagiarized areas from the documents passed from the filtering process through the similarity check of each sentence between the original documents in cluster and the target document based on Euclidean distance algorithm that was used before in (Song et al., 2015). At close of play, the user gets a response from the system whether the target document has plagiarized areas or not.

Conclusion

In this paper, we proposed a feasible document clustering strategy to improve the time performance of the plagiarism detections in the national research, report and management portal system in Korea. By applying the document clustering strategy as the first step of the plagiarism detection system, the enhanced plagiarism detection service is performed through four main steps; clustering, storing, selecting, and detecting. Hence, the documents uploaded into the national research report management portal system are grouped based on their similarities into clusters and stored by clusters. Hence, since the plagiarism detection is required against only the cluster which has the similar representative vector to that of the input target document, we can expect an improvement of the execution time of plagiarism detections against the national research reports.

Acknowledgement

This research project was supported by Ministry of Culture, Sports and Tourism(MCST) and from Korea Copyright Commission in 2015.

References

i. Fagin, R., Maarek, Y., Ben-Shaul, I. & Pelleg, D., 2000. *Ephemeral Document Clustering for Web Applications*. IBM Research Report RJ 10186.

ii. Kang, D. H., Joo, K. H. & Lee, W. S., 2003. An Effective Incremental Text Clustering Method for the Large Document Database. *The KIPS Transactions*, Part D, 10(1), pp. 57-66. (in Korean)

iii. iParadigms. 1997. *Turnitin*. Available at: http://turnitin.com/ko/

iv. Song, K. H., Min, J. H. & Kim, Y. S., 2015. A Copyright Management Service Model for National Research Reports in Korea. *International Journal of Advances in Computer Science & Its Applications*, 5(2), pp. 267-271.

v. Song, K. H., Min, J. H., Lee, G. Y. & Kim, Y. S., 2014. Development of an Online Document Plagiarism Detection System. *Database Research*, 30(3), pp. 13-23. (in Korean)

vi. Song, K. H., Min, J. H., Lee, G. Y., Shin, S. C. & Kim, Y. S., 2015. An Improvement of Plagiarized Area Detection System Using Jaccard Correlation Coefficient Distance Algorithm. *Computer Science and Information Technology*, 3(3), pp. 76-80.

vii. Velásquez, J. D., Covacevich, Y., Molina, F., Marrese-Taylor, E., Rodríguez, C. & Bravo-Marquez, F., 2016. DOCODE 3.0 (DOcument COpy DEtector): A System for Plagiarism Detection by Applying An Information Fusion Process From Multiple Documental Data Sources. *Information Fusion,* 27, pp. 64-75.