Computer Science and Information Technology 3(3): 76-80, 2015
DOI: 10.13189/csit.2015.030304 http://www.hrpub.org

# An Improvement of Plagiarized Area Detection System Using Jaccard Correlation Coefficient Distance Algorithm

**Kwangho Song[1], Jihong Min[1], Gayoung Lee[1], Sang Chul Shin[2] ,Yoo-Sung Kim[1,*]**

[1]Department of Information and Communication Engineering, Inha University, South Korea
[2]Daolsoft Co., Ltd. 2F Won Seong Building, South Korea



**Abstract** In this paper, a plagiarized area detection system is proposed in which Jaccard correlation coefficient is used for filtering to improve the processing time against huge volume of documents. Hence, the proposed system does filter to efficiently detect plagiarized area against huge volume of original documents by two algorithms; Jaccard coefficient distance algorithm and Cosine distance algorithm. Since Jaccard coefficient distance algorithm computes the distance between two document based only on the existence of words while Cosine distance algorithm uses word's frequency also, Jaccard coefficient distance algorithm is faster than Cosine one. Hence, for the efficiency, we use Jaccard coefficient distance algorithm as the first filter. According to the experiment result of the performance comparison between the proposed system and the previous our system, the newly proposed system outperforms the previous one with about 30% reduced processing time.

**Keyword** Plagiarism, Jaccard Correlation Coefficient Distance, Filtering

## 1. Introduction

Plagiarism which is defined as using another person's words or ideas without giving credit to that person becomes a big social problem worldwide [1]. According to a study [2], plagiarism is rife also in USA. Furthermore suspect of plagiarism is not restricted to only student in University. According to [2], up to 91% of undergraduate students, 94% of grad students, and 99% of faculty have experience of copying or paraphrasing some parts from a written resource without citation. Also recently in Korea, according to [3], large portion of research reports written by national institutes is seemed as the results of plagiarism and overlapping publications.

The types of plagiarism has Cloning, Paraphrasing with no citation, Summarizing with no citation, Montage with no citation, Recycling with no citation, Manipulation, etc. [4][5]. The criterion of plagiarism is different from country to country. And there are no clear criteria in Korea or other country.

But common criteria are that quoted sentences must have a citation. And a guideline in [6] provides an interesting value as following, some document that contains less than 15% of words from other sources would probably indicate that plagiarism has not occurred. However, if the matching text is one continuous block or document in which more than 25% of words are used from other sources would probably be considered as plagiarism.

Previous studies related to plagiarism detection may be classified into two-folds; comprehensive study for entire system [7] and limited study on plagiarism detection algorithm [8][9]. Most of the previously proposed systems had good performance to detect plagiarisms of cloning type. However, it couldn't detect other plagiarism types. So we developed a system for detecting other plagiarism types [10]. But our system has a weakness at filtering performance against huge volume of original documents. Therefore in this study, we propose a new filtering scheme that can improve the processing performance of the previous system.

This paper is organized as follows. Chapter 2 introduces the previous related systems. And Chapter 3 describes the system proposed to overcome the weaknesses of the previous system. Chapter 4 describes the performance evaluation. And, in Chapter 5 we conclude this study.

## 2. Related Work

### 2.1. Previous System

The proposed system in [7] makes the posting file in which the index terms and their occurrences in the original documents are included, and compares original documents against target document in the unit of sentence by using the posting file. [7] proposes a conceptual design of plagiarism detection system, but it has no implementation details.

Proposed system in patent [8] and [9] parses an original document into list of index terms and makes a multiple keyword index. And target document is processed in the

same way for the original documents to make inquiry keyword. After then, the system compares index keywords in the index with inquiry keywords to detect plagiarized area. The system works well to detect original cloning type. However, it is difficult to detect paraphrasing type such as synonym substitution and manipulation type such as spacing manipulation.

So based on these previous studies, we developed a plagiarism detection system [10]. Using that system, we can detect not only typical cloning type but also synonym substitution and spacing manipulation types. For these functions, we used the Cosine distance algorithm for filtering and Euclidean distance algorithm to compute the similarity between target document and one of the returned results from the first filtering step, respectively. The process flow of filtering in [10] is in [Figure 1] and [Figure 2] is the filtering code with Cosine distance algorithm.

But, that system needs long processing time generally. Especially against large document which has huge number of distinct index terms that system requires long execution time to normalize two vectors and to compute the Cosine distance between them. Because, as shown in [Figure 2], the Cosine distance algorithm uses the frequency of words in document to generate vector, we need to normalize the vector in order to not be affected by the size of vector. Also, since to compute Cosine distance between two vectors needs a lot of complex operation, computing Cosine distance between two vectors is time-consuming task also. So against huge volume of original documents, our system has low performance which may become a big obstacle to employ the system for real service environments.

## 3. Efficient Filtering Scheme for Plagiarism Detection System

In this paper, we suggest the advanced filtering scheme in which Jaccard coefficient distance algorithm is firstly used to fast reduce the candidate documents which should be checked in more detail. Since, only against the returned results from the first filtering, the expensive computations are performed for computing Cosine distance between two vectors, the processing time to detect plagiarism can be reduced from that of the previous system especially against the huge volume of original documents.

The advanced filtering scheme adopts Jaccard coefficient distance algorithm into the existing system's filtering phase. Since the Jaccard coefficient distance algorithm doesn't take account of vector's magnitude, it doesn't need to calculate normalization step. That's reason why we choose the algorithm to reduce the execution time. So in newly proposed system, filtering phase consists of two steps; one with Jaccard distance, the other with Cosine distance. As seen at [Figure 3], the new step that based on Jaccard correlation coefficient algorithm, [Formula 1], are inserted between the preprocessing step and the filtering step of the existing system.

$$J(A,B) = \frac{|A \cap B|}{|A \cup B|} \qquad \text{[Formula 1]}$$

After preprocessing step, we set two vectors to indicate index term's presences in original document and in target document, respectively. And then, equalize each other. After then, calculate Jaccard correlation coefficient distance between two vectors for filtering returned documents. We determined that the threshold for filtering is 25% on the basis of [6]. As a result of the first filtering step, we can obtain a plagiarism candidate document set. Using this set we proceed to the second filtering step with Cosine distance. The process flow of the new system is in [Figure 3], and pseudo code for the filtering is shown in [Figure 4].

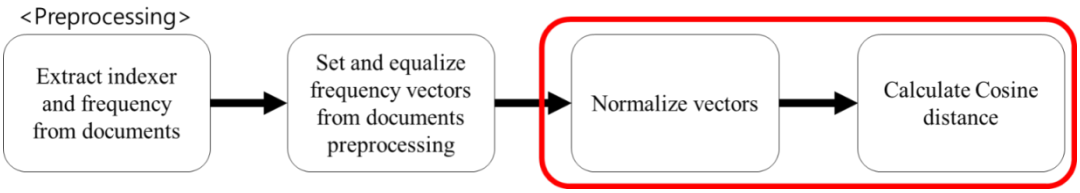


**Figure 1.** Process Flow for Filtering

| **Algorithm** | **Filtering with Cosine Distance Algorithm** |
|---|---|
| 1. | Create vector A with index terms of the target document |
| 2. | Create vector B with index terms of each original document |
| 3. | Compare the index terms between two vector A and B |
| 4. | For the index terms of vector A and B which are not matched to each other, the frequencies of those terms are set to zero in vector A and B, respectively. |
| 5. | Normalize vector A and B |
| 6. | Calculate Cosine distance between vector A and B |
| 7. | If the Cosine distance value is smaller than the threshold, the original document is not candidate for next processing step. |

**Figure 2.** Pseudo Code for Filtering

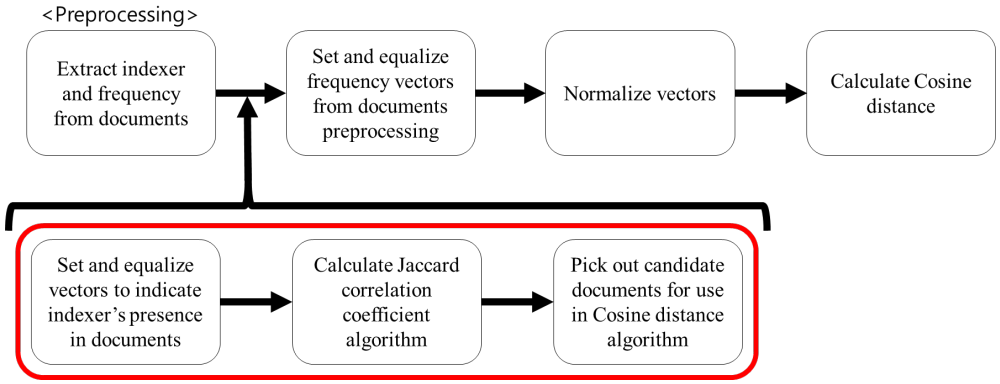


**Figure 3.** Advanced Filtering Step

**Algorithm** **Advanced Filtering with Jaccard Coefficient and Cosine distance**

1. Create vector A with index terms of the target document
2. Create vector B with index terms of each original document
3. Compare the index terms between two vector A and B
4. For the index terms of vector A and B which are not matched to each other, the frequencies of those terms are set to zero in vector A and B, respectively
5. Count the number of the same words between two vectors
6. Calculate Jaccard coefficient and pick out documents above 25% threshold
7. Process the remaining filtering phase with Cosine distance algorithm

**Figure 4.** Pseudo Code of Advanced Filtering with Jaccard Coefficient

Original Doc. : 사람은 누구나 자기를 알아주는 사람을 위해 헌신한다. Target Doc. : 사람이 사람이라고 다 사람이 아니고 사람다워야 사람이다.

Pre-Processing

| Sentence | 사람은 누구나 자기를 알아주는 사람을 위해 헌신한다. | | | | | | |
|---|---|---|---|---|---|---|---|
| Words | 사람 | 누구 | 자기 | 알다 | 사람 | 위하다 | 헌신 |
| Synonym | 사람 | 아무 | 나 | 알다 | 사람 | 위하다 | 헌신 |
| Location in Sentence | 0 | 4 | 8 | 12 | 17 | 21 | 23 |

Pre-Processing

| Sentence | 사람이 사람이라고 다 사람이 아니고 사람다워야 사람이다. | | | | |
|---|---|---|---|---|---|
| Words | 사람 | 사람 | 사람 | 사람 | 사람 |
| Synonym | 사람 | 사람 | 사람 | 사람 | 사람 |
| Location in Sentence | 0 | 4 | 12 | 20 | 26 |

**Figure 5.** Example of Preprocessing and Results for Original and Target Documents

For example, let us assume that original document and target document are written in Korean as in [Figure 5]. Two sentences in this example have different meaning even though a common word '사람' is used frequently in two sentences. That is, the target document is not likely plagiarized from the original document. But with the filtering step in which Cosine distance is used, since they have higher similarity value of 63% the target document should be passed to check whether the target document is plagiarized or not. That is, the target document is regarded as plagiarism suspicious even though the target document has different meaning from the original one. On the other hand, with the newly proposed filtering step, since the similarity between two example documents is 17%, the target document is not passed to the next plagiarism checking phase. Given this example, if the filtering step based on Jarccard coefficient distance algorithm is added, the number of documents that should be checked for the plagiarism detection must be decreased. And then, the total execution time is naturally shorter than before.

Also, because the amount for calculation of Jaccard coefficient distance algorithm is less than the amount for calculation of Cosine distance algorithm, we can reduce the overall computation time to check the plagiarism. Furthermore, it gives us the stability of performance for total filtering phase because of pre-selecting for calculation of Cosine distance algorithm.

# 4. Performance Evaluation

Before the performance evaluation of the new filtering step, we needed to verify the accuracy of the new filtering step. So, we compare the precision and recall using [Formula 2] and [Formula 3] to evaluate the filtering accuracy between the previous system's filtering step in which only Cosine distance algorithm is used(marked as 'Only Cosine' hereafter) and the newly proposed filtering

step in which Jacaard correlation coefficient distance algorithm is also used in addition to Cosine distance algorithm(marked as 'Both Jaccard-Cosine' hereafter).

$$Precision = \frac{Relevant\ documents\ among\ retrived\ document}{Retrived\ documents}$$

[Formula 2]

$$Recall = \frac{Relevant\ documents\ among\ retrived\ document}{Relevant\ documents}$$

[Formula 3]

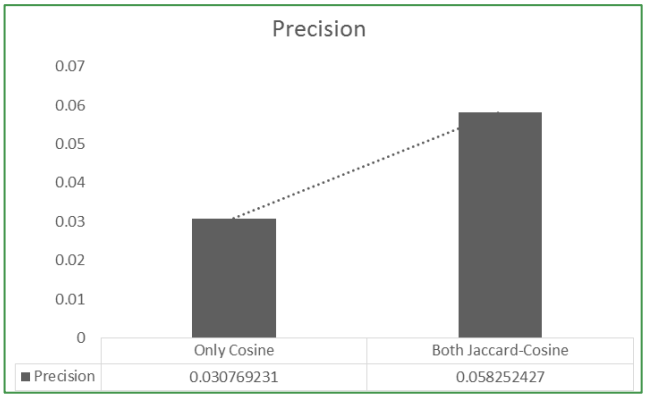


**Figure 6.** Filtering Step Precision Comparison between Cosine Only and Jaccard-Cosine

As a result, recall is the same as '1' for both cases. But precision is different. As shown in [Figure 6], the precision value of 'Both Jaccard-Cosine' filtering step is almost twice than the precision of 'Only Cosine' filtering step. This results can show us that the new filtering step based on the Jaccard coefficient distance algorithm can passing well the suspicious one of plagiarism as much as the existing system's filtering step. At the same time, the new one is effective to filter-out well the non-plagiarized one than the existing filtering step.

To evaluate the performance of the new filtering step, we compare the processing time with that of the existing system against the same original document set. The filtering results of the existing system and the new one are in [Figure 7] and [Figure 8], respectively.

From comparing the running times marked with underlines in [Figure 7] and [Figure 8], we can see that new filtering scheme outperforms the previous system in terms of processing time to detect plagiarism of the target document against the original documents.

In order to obtain a statistical result, we have conducted the similar test 200 times. The result of 200 times test is [Figure 9]. We can see in [Figure 9] that our new filtering scheme in which Jaccard coefficient algorithm is used in addition to Cosine distance algorithm has less average, min, max execution times than using only cosine distance algorithm. We can save processing time up to 30% by new filtering system. Also, performance stability is better than the previous system because new one's error sticks are shorter than ones of the previous system using only Cosine distance.

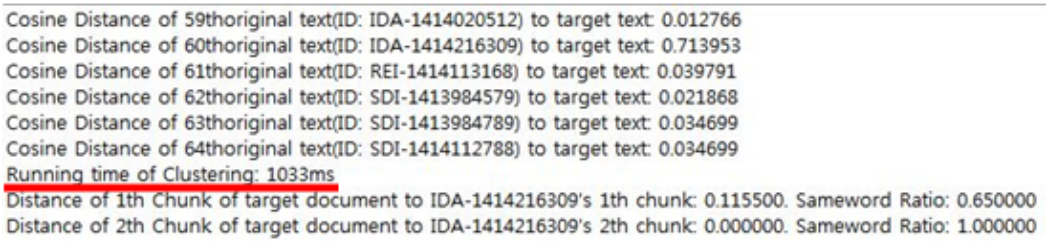


**Figure 7.** Filtering Result of the Existing System

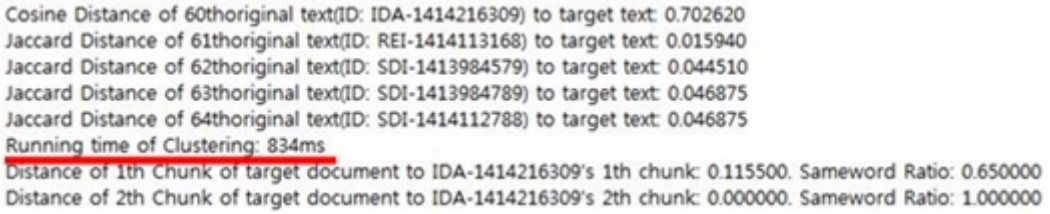


**Figure 8.** Filtering Result of the New System

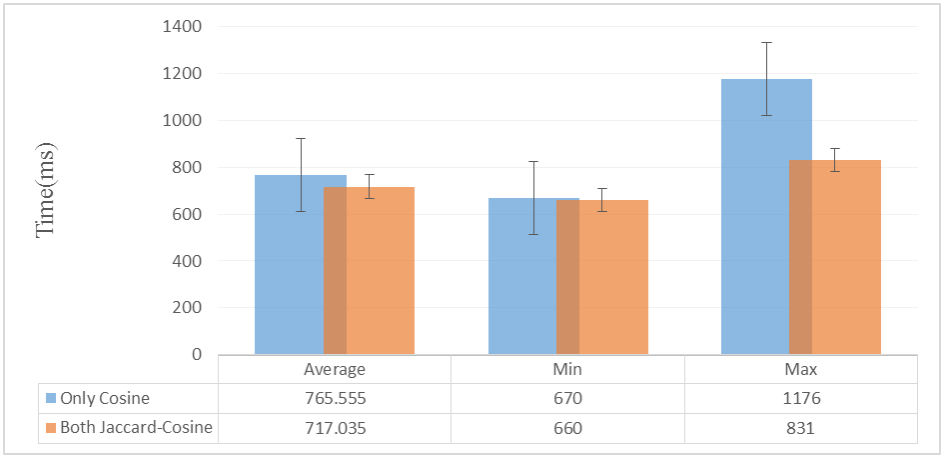


**Figure 9**. Comparison of Processing Times of the Entire System

## 5. Conclusions

In this study, we have studied the ways to solve the shortcomings that consuming very long time in the filtering step of existing plagiarism detection system. In order to reduce the processing time, we put pre-filtering step with Jaccard coefficient algorithm before the filtering step with Cosine distance algorithm. Since Jaccard coefficient algorithms can filter unrelated documents out more quickly than Cosine distance algorithm, the pre-filtering step is very helpful to decrease the number of candidate original documents for detecting plagiarism against target document. As a result, we can save the processing time up to 30% compared with previous one. So in massive document circumstance, our plagiarism detection system with new filtering method is more efficient than previous one.

## Acknowledgements

This research project was supported by Ministry of Culture, Sports and Tourism(MCST) and from Korea Copyright Commission in 2014.

## REFERENCES

[1] www.merriam-webster.com/dictionary/plagiarism, "Merriam-Webster Online Dictionary"

[2] Donald L. McCabe, 2005, "Cheating among college and university students: A North American perspective", *International Journal for Educational Integrity*, 1(1).

[3] National Research Council for Economics, Humanities and Social Sciences, "2012 research report ethic appraisal", *National research Council for Economics, Humanities and Social Sciences,* 2013

[4] http://www.plagiarism.org/plagiarism-101/types-of-plagiarism/, "Type of plagiarism"

[5] Kwak, D. C., 2007."A Study on the Types of Plagiarism and Appropriate Citation Practices of Writing Research Papers", *Journal of the Korean Society for Library and Information Science*, 41(3), 103-126.

[6] The University of the West Indies, "Guidelines for staff and students on plagiarism", http://sta.uwi.edu/

[7] Park, D. H., Park, M. S., Park, J. H., Kwon, H. C., 2000, "Phrase search using posting file in Korean information retrieval system", *Spring Scholarship Conference on The Korean Institute of Information Scientists and Engineers*, Korea, Spring, 27(1), 384-386.

[8] Muhayu Corp., 2013, "APPARATUS AND METHOD FOR CALCULATING DOCUMENT PLAGIARISM AND RECORD MEDIA RECORDED PROGRAM FOR REALIZING THE SAME METHOD", *Republic of Korea patent publication*, 10-1264151.

[9] Wisenut Corp., 2014, "DOCUMENT PRE-WORKING APPARATUS FOR HIGH SPEED COPY SECTION DETECTION AND METHOD OF THE SAME", *Republic of Korea patent publication*, 10-14538660000.

[10] Song, K. H., Min, J. H., Lee, G. Y., Kim, Y. S., 2014, Development of Plagiarized Area Detection system Using Synonym Dictionary, Proceedings of the 41th Winter Scholarship Conference on The Korean Institute of Information Scientists and Engineers, Korea, December 18-20.