A Method for Plagiarism Detection over Academic Citation Networks

Sidik Soleman and Atsushi Fujii
Department of Computer Science, Tokyo Institute of Technology, Tokyo, Japan

Keywords: Plagiarism Detection, Citation Behavior, Information Retrieval, Content Analysis.

Abstract: Whereas in the academic publication, citation has been used for a long time to borrow ideas from another document and show the credit to the authors of that document, plagiarism, which does not indicate the appropriate credit for a borrowed idea, has of late become problematic. Because plagiarism detection has been formulated as finding partial near-duplicate in response to a document for a suspected case of plagiarism, in this paper we propose a method to improve the similarity computation between text fragments. Our contribution is to formulate three document similarities based on citation and content analysis, and to combine them in our method. We also show the effectiveness of our method experimentally and discuss its advantages and limitation.

1 INTRODUCTION

Reflecting the rapid progress of science, technology, and culture, an increasing number of academic publications have recently been available by means of digital libraries or general-purpose search engines on the Web. Whereas an academic publication should include the novel ideas proposed by the authors, most of the residue include known facts or knowledge in a large body of literature, for which citation provides a practical solution to easily indicate the source of each idea and also credit to the authors of each source.

This customary has resulted in a huge network in which each academic publication (i.e., document) and citation is represented as a node and a directed link between two nodes, which we shall call "academic citation network (ACN)". In practice, the entire ACN can be divided into more than one subnetwork, each of which roughly corresponds to a different discipline. However, we use only "ACN" to refer to both the entire and a partial ACN, without loss of generality.

Whereas in principle the authors who wish to borrow specific content from other documents are responsible for citing appropriate documents, in practice misconduct associated with missing or deceptive citations has of late become a crucial problem. Such conducts are generally termed "plagiarism" and is defined, for example, in the Merriam-Webster dictionary[1] as "the act of using another person's words or ideas without giving credit to that person".

[1]https://www.merriam-webster.com/dictionary/plagiarism

Plagiarism has a significantly negative impact on our society in terms of the following perspectives. First, it discourages the spirit of the invention and creativeness because the credit is not given to the right people. Second, the evaluation of each publication can purposefully be manipulated given that the frequency of a document being cited has been used to measure the achievement for a research project and intellectual contribution of individual researchers. Finally, plagiarism can decrease the trust of the academia. The above background has motivated us to explore plagiarism detection over the ACN.

A single case of plagiarism can generally be represented as "party X plagiarized document Q using one or more documents S_1,...,S_i,...,S_n, where X, Q, and S_i are variables representing a plagiarist, plagiarized document, and source document, respectively". Plagiarized documents, which refer to a resultant document, should not be confused with the source documents. In contrast, a task of plagiarism detection can be different depending on the purpose of a user. In the following, (a)-(c) are example scenarios for plagiarism detection associated with different resolution of analysis.

(a) To determine if a document in question is a plagiarized one, in which the input can potentially be non-plagiarized one.

(b) To find one or more source documents for each of the plagiarized ones as an evidence of plagiarism, in addition to (a).

(c) To identify how the fragment in a source docu-

Soleman, S. and Fujii, A.
A Method for Plagiarism Detection over Academic Citation Networks.
DOI: 10.5220/0006938402740281
In *Proceedings of the 10th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management (IC3K 2018) - Volume 1: KDIR*, pages 274-281
ISBN: 978-989-758-330-8
Copyright © 2018 by SCITEPRESS – Science and Technology Publications, Lda. All rights reserved

ment has been modified in the plagiarized one, in addition to (a) and (b).

Although it may also be important to determine whether a plagiarism is due to a deliberate intention or an innocent mistake, in this paper we focus only on intentional cases.

Because as in the general representation for plagiarism above, document Q usually consists of fragments of S_i ($1 \leq i \leq n$) with optional modification, plagiarism detection has often been recast as detection for partial near-duplicate text in a document collection. Thus, a system for plagiarism detection can be realized with a straightforward application of information retrieval (IR), and more precisely the purpose is to search the document collection for one or more fragments resembling those in a document in question. Finally, the candidate documents whose similarity score or whose ranking in descending order of the similarity score is above a predetermined threshold are presented to the user. Systems for plagiarism detection that follows the IR approach generally rely on the similarity between the plagiarized document and a candidate for its source document.

In this paper, we propose a method for plagiarism detection, focusing mainly on the computation for the similarity score between two documents. Our contribution is, unlike existing methods for plagiarism detection relying only on a single type of document similarity, to formulate three document similarities based on citation and content analysis and to combine them in our method.

Section 2 surveys past research on plagiarism detection to clarify our focus and approach. Sections 3 and 4 elaborates our method for plagiarism detection and our experiment to evaluate its effectiveness, respectively. In Section 5, we conclude our work.

2 RELATED WORK

We can categorize existing work into three categories according to type of information that is compared to calculate document similarity, i.e. method based on textual content, citation, and combination of both.

2.1 Method based on Textual Content

The methods in this type calculate the document similarity by comparing the textual contents between two documents. Several methods have been proposed to compare the textual contents, e.g. bag-of-word model, word n-gram model, and fingerprinting.

In the bag-of-word model, a document is represented as a set of words or terms. Each term is usually assigned with a certain weight. Since some fragments of document are more important than the other parts, e.g. the methodology section is more important than the introduction one, (Alzahrani et al., 2012) considered word distribution in each document section to calculate its weight.

(HaCohen-Kerner et al., 2010) found that comparing contents in abstract section of documents is promising despite its short size. While (Soleman and Fujii, 2017a) discovered that the distinction between citing and non-citing sentences in documents is important when calculating the document similarity. They categorize a sentence as citing one if it contains at least one citation anchor. Citation anchor is a symbol or characters in body text referring to a document in reference list.

Unlike the bag-of-word model, the word n-gram model preserves word order since it transforms a document into a set of substrings consisting of a sequence of n words. According to the experiment conducted by (Barrón-Cedeño and Rosso, 2009), they found that using word 2 and 3-gram are effective.

The fingerprinting methods transform a document into a collection of substrings and/or often apply a mathematical function to transform the substrings into unique fixed-size strings. For example, (Kasprzak and Brandejs, 2010) used word 5-gram to generate substrings and used a hash function to transform the substrings. While (Grozea et al., 2009) used character 16-gram and (Sánchez-Vega et al., 2017) proposed several types of character n-gram to generate the substrings.

As the content from source document may have been modified by plagiarist in plagiarized one, some methods proposed to utilize lexical dictionaries to handle the word substitution by means of synonym, such as the method proposed by (Chong and Specia, 2011), and (Chen et al., 2010). Besides using the contents of candidate source documents to be compared with the content of a document in question, (Soleman and Fujii, 2017b) demonstrated that using sentences citing the candidate source documents as additional information for them is also useful.

2.2 Method based on Citation

In the field of citation analysis, bibliographic coupling (Kessler, 1963) is a well-known method to measure the similarity between documents with respect to citation link. Two document are likely to be similar or related if they cite the same documents.

Motivated by the above study, (Gipp and Meuschke, 2011) compared pattern of citation anchors between two documents for calculating the doc-

ument similarity. While (HaCohen-Kerner et al., 2010) compared the document title in reference list to calculate the document similarity. However, they found that several false-positives are produced by their method. It means that innocent and non-source documents are identified as plagiarized and source ones, respectively.

Since the methods in this type only work when document contains citation information or reference list, they should be combined with the other types of method.

2.3 Combination of Textual Content and Citation

The methods in this type combined more than one type of document similarity to improve their effectiveness, since the document similarities should complement each other. For example, (HaCohen-Kerner et al., 2010) found that the combination of the content-based method, i.e. the content similarity in the abstract section, and citation-based method is promising. While (Pertile et al., 2016) successfully improved the effectiveness of their method by combining content-based method, i.e. the content similarity in document level and several citation-based methods.

Although (HaCohen-Kerner et al., 2010) and (Pertile et al., 2016) combined the citation and content-based methods, they considered the citation information and the textual contents as two independent entities. However, some text fragments in documents and the citation information are closely related.

To address this problem, (Soleman and Fujii, 2017a) proposed a citation-based method by calculating the content similarity of citing sentences, and they linearly combined it with a content-based method, i.e. the content similarity in non-citing sentences. They, however, did not consider the cited documents or the citation directions when calculating the content similarity of citing sentences.

3 PROPOSED METHODS

As we have mentioned earlier, one of our main contribution is to formulate three document similarities based on citation and content analysis and to combine them in our method. We use the information from the ACN to extract citing sentences, to perform content analysis, and to formulate the document similarity calculations.

Suppose we have ACN consisting three documents q, y, and z where q and y cite z. One or more sentences $s_{q,1}$, $s_{q,2}$, $s_{q,3}$, ... , $s_{q,n}$ in q and $s_{y,1}$, $s_{y,2}$, $s_{y,3}$, ... , $s_{y,n}$ in y contain citation anchor referring to z. We describe our hypotheses to formulate the document similarities as follows:

1. $s_{q,i}$ and $s_{y,i}$ are likely to contain novel contents borrowed from z. Thus, they can be used for additional contents for z to emphasis novel content described in z. Any document having similar content significantly to the additional content of z, that document is likely to be a plagiarized one and z is the source document.
2. Since citing sentences contain borrowed contents from the cited document, the non-citing sentences in q, y, and z should contain novel content described in each document. Hence, when non-citing sentences in a plagiarized and its source document are compared, they should have a significant content similarity.
3. If $s_{q,i}$ and $s_{y,i}$ have a significant content similarity, q and y have the same citation behavior towards z. Thus, q and y are similar. Typically, the more documents are cited by two documents with the same citation behavior, the more similar they are.

Now, let's say we want to detect plagiarism for document in question q given a document collection D. We elaborate our document similarities between q and $d \in D$ by the following text. However, we first describe the method that we use to calculate the content similarity between two text fragments.

Given text fragment s_q from q and s_d from d, we calculate the content similarity between the text fragments by means of bag-of-word model and TF·IDF[2] term weighting. We also perform several preprocessing[3], i.e. text lowercasing, stopword removal, and stemming. Thus, we calculate the weight of term t in s_q as follow:

$$TF(t, s_q) = |\{t' \in terms(s_q) : t' = t\}| \quad (1)$$

$$IDF(t) = log \frac{|D|}{\{d \in D : t \in terms(d)\}} \quad (2)$$

$$w(t, s_q) = TF(t, s_q) \cdot IDF(t) \quad (3)$$

After calculating the weight for each term in s_q and s_d, we transform them into vector representations, i.e. $\vec{s_q}$ and $\vec{s_d}$, we calculate their content similarity by means of cosine similarity:

$$sim_{cont}(s_q, s_d) = cos(\vec{s_q}, \vec{s_d}) = \frac{\vec{s_q} \cdot \vec{s_d}}{||\vec{s_q}|| \, ||\vec{s_d}||} \quad (4)$$

[2]Term frequency·inverted document frequency

[3]https://www.nltk.org/

Next, we describe our three proposed document similarity calculations that are formulated based on our hypotheses above.

According to our first hypothesis, we use citing sentences as additional contents for the cited document. Thus, we use sentences citing d as its additional content. We consider a sentence as citing one if it contains at least one citation anchor. Since the additional content contain the novel content described in d, we should compare it with the part of q containing the novel content, i.e. its non-citing sentences (see the second hypothesis). Thus, our first document similarity $sim_{add}(q,d)$ is calculated as:

$$cited(d) = \{d' \in D : d' \; cites \; d\} \tag{5}$$

$$ncs(q) = concat(\{s \in sentence(q) : s \; not \; citing\}) \tag{6}$$

$$cs(d',d) = concat(\{s \in sentences(d') : s \; citing \; d\}) \tag{7}$$

$$sim_{add}(q,d) = sim_{cont}(ncs(q), concat(\sum_{d' \in cited(d)} cs(d',d))) \tag{8}$$

Based on the second hypothesis, we compare the novel contents described in q and d by comparing their non-citing sentences. Therefore, our second document similarity $sim_{nc}(q,d)$ is calculated as:

$$sim_{nc}(q,d) = sim_{cont}(ncs(q), ncs(d)) \tag{9}$$

Based on the third hypothesis, we calculate the similarity between q and d by comparing their citation behavior. Hence, our third document similarity $sim_{cb}(q,d)$ is calculated as follow:

$$citing(d) = \{d' \in D : d \; cites \; d'\} \tag{10}$$

$$cb(q,d) = \sum_{d' \in \{citing(q) \cap citing(d)\}} sim_{cont}(cs(q,d'), cs(d,d')) \tag{11}$$

$$cb_{smth}(q,d) = sim_{cont}(concat(\sum_{d' \in \{citing(q) - citing(d)\}} cs(q,d')), concat(\sum_{d' \in \{citing(d) - citing(q)\}} cs(d,d'))) \tag{12}$$

$$sim_{cb}(q,d) = \frac{cb(q,d) + cb_{smth}(q,d)}{min(|citing(q)|, |citing(d)|)} \tag{13}$$

We use $cb_{smth}(q,d)$ as smoothing score for the similarity of citation behavior, since (Soleman and Fujii, 2017a) found that comparing the content of citing sentences regardless of the cited document is also useful. This may also alleviate the problem when plagiarist modifies some citation anchors, i.e. replacing them with the other ones. In the similarity of citation behavior, we also use $min(|citing(q)|, |citing(d)|)$ to anticipate when plagiarist reduces or adds the number of cited documents.

Finally, we combine our three document similarities by a linear combination in our proposed method. Hence, the final document similarity score between q and d is:

$$ds(q,d) = \alpha sim_{cb}(q,d) + (1-\alpha) sim_{nc}(q,d) + \beta sim_{add}(q,d) \tag{14}$$

where $\alpha, \beta \in [0,1]$. We use α to prioritize between the similarity of citation behavior and non-citing sentences. While β is used to determine how much the similarity of additional content should be considered. We can also use a machine learning algorithm to combine our document similarities. Hence, α, β are automatically determined by the algorithm.

4 EVALUATION

4.1 Evaluation Scenarios

In this paper, our task is to identify the candidate source documents in a collection, given a document in question. We can perform this task either by ranking or classifying the documents in collection according to their document similarity scores.

In the ranking task, the document in question is a plagiarized document, while in the classification task, it is either a plagiarized or an innocent one, and the objective is to classify whether a pair of document in question and candidate source document is the pair of plagiarized and source document. We use both ranking and classification scenario in the evaluation.

4.2 Dataset

We evaluate the proposed method by the dataset developed by (Pertile et al., 2016). It is constructed by exhaustive investigation of two document collections, i.e. ACL[4] and PubMed[5]. For this evaluation, however, we only used dataset from the ACL since these documents have more consistent citation format.

(Pertile et al., 2016) created the dataset by performing pairwise content comparison between documents in the collection by means of several document

[4]http://aclanthology.info/

[5]https://www.ncbi.nlm.nih.gov/pubmed/

similarities to select top-n pairs. They asked 10 annotators to judge whether a pair in the top-n pairs is suspected as a plagiarism case. If the pair is suspected to be a plagiarism case, they labeled it as positive pair, otherwise it is labeled as negative one. However, the positive pairs in this dataset might be better to be addressed as suspected self-plagiarism since the documents in a pair share one or more authors.

In the evaluation, we use the positive and negative pairs for the classification scenario. While we use the positive pairs and the document collection for the ranking scenario. One of the documents in a positive pair is used as the document in question and the other one is used as the source document.

Since the documents in the dataset are in the PDF format, we use Grobid (Lopez, 2009) to extract their contents, to parse their reference lists, to extract and link citing sentences to the cited documents. The complete information about this dataset is shown in Table 1.

Table 1: The detail information of the dataset.

Type	Detail
Topic	computational linguistics
Positive pair	41
Negative pair	52
Target document	4685
Plagiarized document	40
Source/plagiarized document	1.025
Avg. word (target)	2557.7
Avg. word (plagiarized)	2797
Kappa	0.675 *(substantial)*
Agreement rate	84%

4.3 Evaluation Methods

We measured the performance of the methods by using precision (P), recall (R), and $F1$, which are calculated as:

$$P = TP/(TP+FP) \tag{15}$$

$$R = TP/(TP+FN) \tag{16}$$

$$F1 = 2 \times P \times R/(P+R) \tag{17}$$

where TP (true positive) is the number of retrieved source documents in the ranking scenario, or the number of correctly predicted positive pairs in the classification one. FP (false positive) is the number of retrieved non-source documents in the ranking scenario, or the number of negative pairs predicted as positive ones in the classification scenario. FN (false negative) is the number of source documents that are not retrieved in the ranking scenario, or the number of positive pairs predicted as negative ones in the classification scenario.

We also calculate Mean Average Precision (or MAP) to measure ranking quality in the ranking scenario:

$$AP(q,n) = {}^{1}/_{|S_q|} \sum_{i=0}^{n} P(i) \times sourceDoc(i) \tag{18}$$

$$MAP(n) = {}^{1}/_{|Q|} \sum_{q \in Q} AP(q,n) \tag{19}$$

where $AP(q,n), S_q, P(i), sourceDoc(i)$, and Q are the average precision of input document q at the cut-off n, the source documents of q, the precision at rank i, a function that returns 1 if document in rank i is one of the source documents of q, otherwise it returns 0, and the set of documents in question, respectively.

In addition, we calculate the percentage of document pairs predicted correctly or accuracy in the classification scenario:

$$A = (TP+TN)/(TP+FP+FN+TN) \tag{20}$$

where TN (true negative) is the number of correctly predicted negative pairs in the classification scenario.

4.4 Baseline Methods

In the evaluation, first, we compare the proposed method with the method that compares the content of two documents by means of bag-of-word model. Suppose we have a document in question q and a collection D. Thus, we calculate the document similarity score between q and $d \in D$ in this method as:

$$bowm(q,d) = sim_{cont}(q,d) \tag{21}$$

Second, we compare our method with the method proposed by (Soleman and Fujii, 2017a). Their method $cnc(q,d)$ calculate document similarity by comparing the content similarities between citing and non-citing sentences.

$$\begin{aligned} sim_{cs}(q,d) = sim_{cont}(&concat(\sum_{d' \in citing(q)} cs(q,d')), \\ &concat(\sum_{d' \in citing(d)} cs(d,d'))) \end{aligned} \tag{22}$$

$$cnc(q,d) = \lambda sim_{cs}(q,d) + (1-\lambda) sim_{nc}(q,d) \tag{23}$$

where λ is a weighting variable between 0 and 1.

Lastly, we compare our method with a citation-based one calculated by counting the same cited documents:

$$cbm(q,d) = |citing(q) \cap citing(d)| \quad (24)$$

4.5 Evaluation Results

In this section, we discuss the results from the ranking and the classification scenario. We also discuss the error and successful cases that happened in the classification scenario.

Table 2 shows the results from the ranking scenario for R, P, and F1. According to the R scores, the baseline method *bowm*, *cbm*, and *cnc* achieved a significant high score. These happened because during the dataset creation, (Pertile et al., 2016) pooled the document pairs that have a significant content similarity, i.e. the top-30 document pairs to be annotated. Hence, improving their performances in this scenario is quite difficult.

Among the baselines, the method *cnc* achieved the best performance in R for the cut-off=30. While our method achieved the same performance as *cnc* in R, P, and F for every cut-off. At the cut-off=30, our method and *cnc* improved the R scores for one and three plagiarized documents when we compared them with *bowm* and *cbm*, respectively.

According to the MAP scores shown on Table 3 at cut-off=30, *cnc* is the best method among the baseline methods, while ours achieves the best performance of all the methods. Our method improved the ranking position of one source document compared with the baseline method *bowm* and *cnc*. The rank of this source document in *bowm*, *cnc*, and our method is 32, 30, 24, respectively.

In this evaluation, we found the best λ for *cnc* was .2. We also found the optimal α and β for our method were between .1 and .3, and between .2 and .4, respectively. These results suggest that the content similarity of non-citing sentences should be prioritized, but the similarity of citation behavior should not be ignored. Additionally, using citing sentences as additional content for the cited document is also useful.

Table 4 shows the MAP scores for each document similarity method that we proposed. Among them, sim_{nc} achieves the best MAP score, while sim_{add} is the lowest one. The reason for sim_{add} to have the lowest MAP scores is because some source documents (20 of 40) do not have any sentences citing them, or the citing sentences are not extracted or identified.

In the classification scenario, we used SVM (Support Vector Machine) algorithm[6] to perform this task. Thus, α, β, and λ are decided automatically by the SVM. We did stratified 10-fold cross-validation and also searched for the optimum parameters in the SVM, i.e. type of kernel, C, and γ.

Since the method (Lopez, 2009) failed to extract some citing sentences and/or identify the cited documents in the ranking scenario, we manually performed these tasks on both positive and negative pairs for the classification scenario. Thus, we could give the ideal situation for all the document similarity methods except for sim_{add} since it was not possible to do these tasks manually on all documents in the collection.

Table 5 shows the evaluation results in the classification scenario. Our similarity of citation behavior (sim_{cb}) achieved .3466, .17, .4228 and .338 higher than *cbm* for P, R, F1 and A, respectively. These results indicate that citing sentences should be considered when comparing citations or reference lists.

Since we could not give ideal situation for sim_{add}, i.e. only 31 of 93 document pairs that the candidate source documents have sentences citing them, its performance is the lowest among our three document similarities in P, R, F1, and A. Despite its performance, sim_{add} is still useful when we consider the situation where the content of candidate source documents are not available in the document collection.

The combination of sim_{cb} and sim_{nc} also scored .0433, .045, .0446, and .0424 higher than *cnc* in the terms of P, R, F1, and A, respectively. These results suggest that citation anchors should also be considered when comparing citing sentences. Additionally, this combination performed .0655, .0501, and .0433 higher than *bowm* for P, F1, and A, respectively. It indicates that citing and non-citing sentences should be distinguished when comparing documents.

According to F1 and A scores, our method (sim_{cb}, sim_{nc}, sim_{add}) is the best one. It performed .0501, .4228, and .0585 higher than the baseline method *bowm*, *cbm*, and *cnc* for F1, respectively. This indicates that our document similarity methods complement each others. In addition, our method also made the least FN and FP compare with the baseline methods according to Table 6.

Our method produced three FN and six FP according to the table 6. Three FN happened because the similar text fragment between these pairs are short. They share less than three citing sentences, and a few non-citing ones. Thus, to detect plagiarism for small textual overlap remains difficult when documents are long.

While the reason for the FP is because a few shared citing sentences containing multiple citation anchors. Typically, these citing sentences only list

[6] http://scikit-learn.org/

Table 2: The recall (R), precision (P), and F1 scores of the baselines and proposed method in the ranking scenario.

Cut-off	*bowm*			*cbm*			*cnc*			ours		
	P	R	F1	P	R	F1	P	R	F1	P	R	F1
10	.1025	.975	.1855	.095	.9	.1719	.1025	.975	.1855	.1025	.975	.1855
30	.0342	.975	.0661	.0325	.925	.0628	.035	1	.0676	.035	1	.0676
100	.0105	1	.0208	.0098	.925	.0194	.0105	1	.0208	.0105	1	.0208

Table 3: The MAP scores of the baselines and proposed method in the ranking scenario.

Cut-off	*bowm*	*cbm*	*cnc*	ours
10	.9625	.8467	.9625	.9625
30	.9625	.8478	.9633	.9635
100	.9633	.8478	.9633	.9635

Table 4: The MAP scores of each proposed document similarity method in the ranking scenario.

Cut-off	sim_{cb}	sim_{nc}	sim_{add}
10	.7099	.95	.2622
30	.7109	.9508	.266
100	.7114	.9508	.266

several existing studies. Although their number is a few, they contribute to a significant scores when calculating the similarity of citation behavior. There are two FP associated with this error. Hence, in the future, it is also important to consider the type or function of citation when calculating the similarity of citation behavior. For the rest of the FP, we could not find the reason for them.

In the classification, we also compare the prediction results between our method (sim_{cb}, sim_{nc}, sim_{add}) and *bowm*. Our method and *bowm* make the same correct and incorrect prediction for 78 and 7 document pairs, respectively. Our method incorrectly classifies one document pair where *bowm* correctly classifies it. Our method, however, correctly classifies seven document pairs (two are positive pairs and five are negative pairs) that are incorrectly classified by *bowm*. This indicates that our method could predict correctly topically similar document-pair, which could be considered as difficult case according to *bowm*. Hence, it could increase TP and decrease FP at the same time.

We also investigate the reason for these correctly predicted document pairs by only our method. Two positive pairs are correctly classified because they cite many the same documents with similar citing sentences. While five negative pairs are correctly predicted because they use different citing sentences for the same cited documents. Therefore, sim_{cb} could successfully predict these pairs. Additionally, four of five negative pairs are also correctly classified because they have different content in the non-citing sentences. Hence, sim_{nc} could correctly predict these pairs.

5 CONCLUSIONS

In this paper, we address the problem of detecting source document in plagiarism detection by means of ranking or classifying the candidate source documents. A few existing methods combine more than one document similarity score. In this paper, we formulate three document similarities based on citation and content analysis, i.e. the similarity of citation behavior, non-citing sentences, and the similarity between non-citing sentence in a document in question and sentences citing candidate source documents, and combine them in our proposed method.

In the ranking scenario, our method is slightly better than the baselines according to their MAP scores. In the classification scenario, our method achieves the best performance in F1 and A. Our method performs .0501, .4228, and .0585 higher than the baseline method *bowm*, *cbm*, and *cnc* for F1, respectively. Our method also produces the least FN and FP.

The evaluation results suggest that the document similarity calculations combined in our method complement each other. When comparing citation anchors or reference lists, we should not ignore the citing sentences, and we also should consider the citation anchors when comparing the citing sentences.

The evaluation results also imply that we should distinguish between citing and non-citing sentences when calculating document similarity. Also, using citing sentences as the additional content for the cited document document is useful in plagiarism detection.

In the future, since the evaluation was conducted on suspected plagiarism case, it is also important to evaluate our method in real case of plagiarism. Additionally, the type or function of citation should also be considered in the similarity of citation behavior.

Table 5: The evaluation scores in the classification scenario.

Features (kernel, C,γ)	P	R	F1	A
All $(rbf, 5, 1)$	.865	.925	.8874	.8933
Ours $(sim_{cb}, sim_{nc}, sim_{add})$ $(rbf, 10^3, 10^{-4})$	**.89**	**.925**	**.8981**	**.9044**
sim_{cb}, sim_{nc} $(rbf, 10^3, 10^{-3})$	.885	.9	.8842	.8933
sim_{add} $(rbf, 10^3, 1)$	.45	.25	.3091	.5823
sim_{nc} $(rbf, 10^3, 10^{-3})$	.855	.855	.8485	.862
sim_{cb} $(rbf, 10^3, 1)$	.8933	.66	.7213	.8063
cnc $(linear, 1, -)$	.8417	.855	.8396	.8509
cbm, *bowm* $(rbf, .1, 10)$	.8195	.9	.848	.85
cbm $(rbf, 10, 10)$	.5467	.49	.4753	.5664
bowm $(linear, 5, -)$	.8195	.9	.848	.85

Table 6: The number of TP, FN, TN, and FP of our method and the baselines.

Method	TP	FN	TN	FP
Ours	**38**	**3**	**46**	**6**
cnc	35	6	44	8
sbc	20	21	32	20
scont	37	4	42	10

REFERENCES

Alzahrani, S., Palade, V., Salim, N., and Abraham, A. (2012). Using structural information and citation evidence to detect significant plagiarism cases in scientific publications. *J. Am. Soc. Inf. Sci.*, 63(2):286–312.

Barrón-Cedeño, A. and Rosso, P. (2009). On automatic plagiarism detection based on n-grams comparison. In *Proceedings of the 31th European Conference on IR Research on Advances in Information Retrieval*, pages 696–700. Springer.

Chen, C. Y., Yeh, J. Y., and Ke, H. R. (2010). Plagiarism detection using rouge and wordnet. *Journal of Computing*, 2(3):34–44.

Chong, M. and Specia, L. (2011). Lexical generalisation for word-level matching in plagiarism detection. In *Proceedings of International Conference Recent Advances in Natural Language Processing*, pages 704–709. Association for Computational Linguistics.

Gipp, B. and Meuschke, N. (2011). Citation pattern matching algorithms for citation-based plagiarism detection: greedy citation tiling, citation chunking and longest common citation sequence. In *Proceedings of the 11th ACM Symposium on Document Engineering*. ACM.

Grozea, C., Gehl, C., and Popescu, M. (2009). Encoplot: Pairwise sequence matching in linear time applied to plagiarism detection. In *Proceedings of the SEPLN 2009 Workshop on Uncovering Plagiarism, Authorship, and Social Software Misuse*, pages 10–18. ceur-ws.org.

HaCohen-Kerner, Y., Tayeb, A., and Ben-Dror, N. (2010). Detection of simple plagiarism in computer science papers. In *Proceedings of the 23rd International Conference on Computational Linguistics*, pages 421–429. Association for Computational Linguistics.

Kasprzak, J. and Brandejs, M. (2010). Improving the reliability of the plagiarism detection system: lab report for pan at clef 2010. In *Notebook Papers of CLEF 2010 Labs and Workshops*. ceur-ws.org.

Kessler, M. M. (1963). Bibliographic coupling between scientific papers. *American Documentation*, 14(1):10–25.

Lopez, P. (2009). Grobid: Combining automatic bibliographic data recognition and term extraction for scholarship publications. In *Proceedings of 13th European Conference on Digital Libraries*, pages 473–474. Springer.

Pertile, S. D. L., Moreira, V. P., and Rosso, P. (2016). Comparing and combining content-and citation-based approaches for plagiarism detection. *J. Assn. Inf. Sci. Tec.*, 67(10):2511–2526.

Sánchez-Vega, F., Villatoro-Tello, E., Montes-y-Gómez, M., Rosso, P., Stamatatos, E., and Villaseñor-Pineda, L. (2017). Paraphrase plagiarism identification with character-level features. *Pattern Analysis and Applications*, pages 1–13.

Soleman, S. and Fujii, A. (2017a). Plagiarism detection based on citing sentences. In *Proceedings of 21st International Conference on Theory and Practice of Digital Libraries*, pages 485–497. Springer.

Soleman, S. and Fujii, A. (2017b). Toward plagiarism detection using citation networks. In *Proceedings of 12th International Conference on Digital Information Management*, pages 202–208. IEEE.