Plagiarism Detection Based on Citing Sentences

Sidik Soleman(✉) and Atsushi Fujii

Tokyo Institute of Technology, Tokyo, Japan
soleman.s.aa@m.titech.ac.jp, fujii@cs.titech.ac.jp

Abstract. Plagiarism, which is one of the forms of academic misconducts, is problematic. It results in discouraging innovation, and losing trust in the academic community. We modeled the plagiarism for academic publications, by means of the similarity between textual contents, and citation relations. Furthermore, we adopted the model in our proposed method for plagiarism detection. We evaluate our method using two types of dataset, namely auto-simulated and manually judged dataset. Our experiment shows that our method outperforms the baseline, which only uses the similarity between textual contents, on the auto-simulated dataset and the manually judged one for the ACL sub-dataset.

Keywords: Plagiarism detection · Information retrieval · Citation analysis

1 Introduction

Digital archives for academic publications have enabled us to efficiently access a large volume of scientific information. However, its misuse and misconduct have of late become a crucial problem. Plagiarism is "the act of using another person's words or ideas without giving credit to that person"[1], which results in discouraging innovation and losing trust in the scientific research community. To alleviate this problem, a number of methods for detecting plagiarisms specifically for academic publications have been proposed.

In a broad sense, plagiarism detection (PD) is a task to identify whether a document in question is produced by means of plagiarism, and is often requested to present one or more source documents as evidences for the plagiarism. However, in this paper we consider only cases where an input document is a plagiarized one and focus only on identifying one or more source documents for the input document.

As with an adversarial information processing like filtering spam e-mails, a person who conducts plagiarism, or a plagiarist for short, usually intends to hide the plagiarism, for example, by means of editing and summarizing source documents. As a result, PD is a cat-and-mouse game between plagiarists and people who develop PD systems.

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

© Springer International Publishing AG 2017
J. Kamps et al. (Eds.): TPDL 2017, LNCS 10450, pp. 485–497, 2017.
DOI: 10.1007/978-3-319-67008-9_38

Whereas the above scenario is associated with intentional plagiarism, detecting unintentional plagiarism is also important to avoid innocent mistakes. Fang et al. [1] investigated approximately 2 000 papers that were once indexed by PubMed[2] but retracted later and found that 9.8% of them were retracted due to being judged as a plagiarized paper. Irrespective whether those papers are associated with intentional or unintentional plagiarism, effective methods for plagiarism detection will have a significant impact on our society.

One of the crucial steps in PD is to measure the similarity between two documents. In the field of citation analysis, it is well-known that the number of same citations between two documents can be a good indicator whether they are related/similar or not, i.e. bibliographic coupling [2]. The more same citations two documents have, the more related they are. In this paper, we proposed a model for plagiarism that combines the similarity between textual contents and citation relations. More precisely, our model combines the similarity between textual contents in citing and non-citing sentences. We further applied this model to our PD system, which identifies source documents.

In this paper, our contribution is twofold. First, we modeled plagiarism by means of the similarity between textual contents and citation relations, and applied this model to PD system. Second, we evaluated the effectiveness of our PD system.

2 Related Work

Generally, the existing PD systems that focus on identifying source documents can be classified into two categories as shown in Fig. 1. These categories are search engine-based and direct comparison-based PD system.

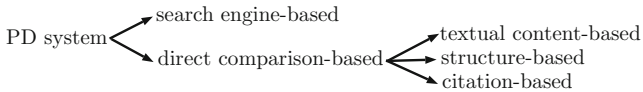

Fig. 1. The categories of PD systems

The search engine-based PD system, which was introduced in PAN workshop[3], utilizes a search engine to identify source documents, because plagiarists are likely to use a search engine to find source documents when plagiarizing document in the Web [3]. The PD systems generate a number of queries from input document, and submit to a search engine for retrieving source documents. Therefore, the system should produce queries that represent the source documents in order to be able to retrieve them. However, the performances of the

[2] https://www.ncbi.nlm.nih.gov/pubmed.
[3] a competition for plagiarism detection.

systems are often limited due to the capabilities of the search engine, e.g. query length, and document-query weighting scheme.

Unlike the previous category, the direct comparison-based PD systems compare input and target documents[4] directly, one by one. In this category, the PD systems can be divided into three types based on the aspects that the systems use for comparing documents, namely textual content, structure, and citation-based.

In the textual content-based type, the PD systems compare textual contents of input and target documents whether they have significant similarity. The systems use various textual comparison strategies, e.g. word n-gram [4].

In the structure-based type, the PD systems take the structure of document into consideration when comparing input and target documents since there are some parts of the document that may be less important than the others. For instance, Alzahrani et al. [5] used section-based component to represent the structure of document, such as *introduction, method,* and *conclusion section* as the components. They put different weight for each component, thus the important components have heavier weight than the less important ones, e.g. *method section* has heavier weight than *introduction section* has. They used these weights to re-weight terms in input and target documents when comparing them.

In the citation-based type, the PD systems consider citation relations when comparing input and target documents. There are two kind of citation considerations in the existing PD systems. First, the existence of citations is the sign of innocent case, such as in the system developed by Alzahrani et al. [5]. Thus, input document is not a plagiarized one and target documents that are similar to the input one are not source documents, as long as the input one cites them.

The second consideration is that the existence of citation relations are used to measure the similarity between input and target documents, which is motivated by the phenomena in citation, i.e. bibliographic coupling [2]. HaCohen-Kerner et al. [6] compared reference lists between input and target documents whether they have a significant degree of reference overlap. However, their system resulted in producing many false positives. It means that the innocent documents are labeled as plagiarized ones. One possible explanation is that these innocent ones cite the same documents with others, but their contents may be different.

Different from HaCohen-Kerner et al. [6], Gipp et al. [7] used the pattern of citation anchors[5] in input and target documents. They generated a number of chunks of citation anchors from the input and the target documents to compare whether the documents have a significant degree of chunk overlap or not. Since their system is likely to fail when there is no citation anchor, recently this work was extended by Pertile et al. [8], where they combined the similarity of textual content in document level, the similarity of reference list, and the pattern of citation anchors.

[4] Target documents are a set of documents in a collection where source documents exist.

[5] Citation anchors refer to characters in citing sentences that point to documents in reference list.

In summary, the recent works in PD consider citation relations to measure the similarity between input and target documents. However, these works may fail when there is no citation relations, or produce false positives. Thus, comparing citation relations alone is not sufficient.

To alleviate this problem, we proposed a model for plagiarism that combines the similarity between textual contents and citation relations, and adopted this model to our PD system. More precisely, we combined the similarity between textual contents in citing and non-citing sentences. Hence, a document is likely to be a plagiarized document, when it has a significant amount of citing and/or non-citing sentences that are similar to the other documents.

3 Proposed Approach

3.1 Model for Plagiarism

As mentioned previously, we model plagiarism by means of the similarities between textual contents in citing and non-citing sentences. Thus, given input (X) and target document (Y), their similarity score is calculated as follows:

$$\begin{aligned} Score(X,Y) = \alpha\ & Sim(Cite(X), Cite(Y)) + \\ (1-\alpha)\ & Sim(NCite(X), NCite(Y)) \end{aligned} \tag{1}$$

with

- $Cite$: a function that returns citing sentences from a document.
- $NCite$: a function that returns non-citing sentences from a document.
- α: a weighting parameter with value [0,1]. Thus, by tuning this value, we are able to prioritize between the similarity of citing and non-citing sentences.
- Sim: a function that measures the similarity of textual content.

Next, given $\boldsymbol{d_1}$ and $\boldsymbol{d_2}$ as vectorized text fragments generated by using bag-of-word method (i.e. word as the dimension of the vector), we define Sim, which calculates the similarity of textual content, by the following equation:

$$Sim(\boldsymbol{d_1}, \boldsymbol{d_2}) = \frac{\boldsymbol{d_1} \cdot \boldsymbol{d_2}}{\|\boldsymbol{d_1}\|\,\|\boldsymbol{d_2}\|} \tag{2}$$

In order to transform a text fragment to its vector representation, we calculate a weight for each word in the text fragment based on the frequency of that word in the text fragment, and inverted document frequency of that word in a document collection, by the following equation:

$$w_t = f_t\ log\frac{N}{n_t} \tag{3}$$

with

- f_t: total number of word t that appears in the text fragment.
- N: total number of documents in document collection.
- n_t: total number of documents in document collection that contain word t.

Unlike the PD systems that only consider citation anchors/reference lists, our model is still able to perform PD when citation relations are not available since the model compares non-citing sentences. In addition, when citation relations are available, our model considers them by means of the similarity between the textual contents in citing sentences. Therefore, our model is different from the textual content-based PD system, which does not consider the citation relations.

Regarding our task of PD that identifies source documents, the similarity score in our model is used to rank the target documents. Thus, the source documents ideally should be located at the top of the target document list.

3.2 PD System

Here, we describe our PD system, given an input document and a set of target documents in a collection. The system outputs a ranked document list, which in ideal situation, the source documents should be located at the top of the document list. Our PD system consists of three components as described in Fig. 2, namely sentence classification, preprocessing, and document comparison.

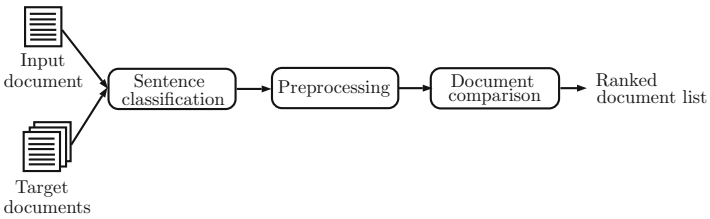

Fig. 2. The components in PD system

Sentence Classification. Since our model combines the similarity between the textual content in citing and non-citing sentences, all sentences in a document should be classified into two classes, i.e. citing and non-citing sentence. This component performs the classification based on the condition whether a sentence contains citation anchor or not. Thus, a sentence containing citation anchor is classified as citing sentence, otherwise it is non-citing sentence.

We employed regular expression to recognize citation anchors for the following formats:

- Combination of author name and publication year, e.g.: *(name, 2010), (name, 2010; name, 2010a), name (2010), [name, 2010b],* and *[name, 2010; name, 2010b].*

- Combination of author name, publication year, and page/paragraph number, e.g.: *(name, 2010, p.1), (name, 2010, para.1),* and *(name, 2010, p.i).*
- Citation anchor is a sequence of characters that refers to a document in reference list, e.g.: *[1], [LIZ2],* and *(1).*
- Combination of author name, publication year, and a document identification in reference list, e.g.: *[name, 2010 (1)]*

Preprocessing. This component performs some modifications to a text fragment, which is its input. First, the text fragment is lowercased, and any numerical character is removed. The next step is to remove any word that is considered as stopwords[6], and lastly words are stemmed using stemmer[7] for English language.

Document Comparison. This component measures similarity between input and target documents by applying our model for plagiarism, which is described in Sect. 3.1. An input document is compared with target documents one by one, and the target documents are sorted in descending order according to their similarity scores. The list of ordered target document is the output of this component, which is also the output of our system (i.e. the ranked document list).

4 Experiment

4.1 Dataset

To evaluate our system, we need dataset that is suitable for our PD task. Since identifying source documents is our goal, the dataset should consist of input documents and document collection containing their source documents. As we have mentioned earlier, we only use plagiarized documents as the input of our system. Additionally, because our model of plagiarism combines the textual similarity between content in citing and non-citing sentences, the dataset should contain citation relations. In this experiment, we used two types of dataset, namely auto-simulated and manually judged dataset.

The auto-simulated dataset was produced by Alzahrani et al. [5] by constructing plagiarized documents automatically since it is difficult to obtain verified plagiarized documents. In this dataset, they controlled the length and the obfuscation level of the plagiarized text fragment. They performed obfuscation by using several text modification techniques, such as verbatim copy-paste, word shuffling, synonym replacement, back-translation, and auto-summarization.

To construct the plagiarized documents, Alzahrani et al. [5] used document collection from Directory of Open Access[8]. First, they divided documents in the collection into two groups, namely plagiarized and target group. Second, they inserted text fragments from any document in the target group to any

[6] http://snowball.tartarus.org/algorithms/english/stop.txt.
[7] https://opennlp.apache.org/.
[8] http://doaj.org.

document in the plagiarized group after the text fragment is obfuscated. Thus, documents in the plagiarized group that are inserted with the text fragments are the plagiarized documents, and the ones where these text fragments come from are the source documents.

The manually judged dataset was created by Pertile et al. [8] by identifying documents that are suspected as the result of plagiarism in two document collections, namely ACL anthology[9] and PubMed[10]. First, they compared all documents in each document collection by using some similarity methods. Second, they pooled pairs of document from the top 30 ranked pairs for each similarity method. Lastly, they asked 10 annotators to judge these pairs of document by using the definition of plagiarism from ACM[11] and IEEE[12]. Thus, the identified pair of document is the pair of input and source document. The complete information about these datasets is described in Table 1.

Table 1. The statistics of the datasets

Type	Manually judged		Auto-simulated
	ACL	PubMed	
Topic	Computation linguistics	Biomedical and life science	Science and technology
Target document	4 685	1 440	8 657
Input document	40	60	3 950
Avg. word (target)	2 557.7	2 868.8	4 417
Avg. word (input)	2 797	3 732	5 263
Source/input document	1.025	1.05	2.5
Kappa	.675	.524	—
Agreement rate	84%	80%	—

4.2 Evaluation Method

Since our system outputted a ranked list of document, we used Mean Average Precision (MAP), which measured the ranking quality of a document list for evaluation method. In addition, we evaluated our system by measuring recall (R), precision (P), and F1. We calculated those methods by the following equations:

[9] http://aclanthology.info/.
[10] https://www.ncbi.nlm.nih.gov/pubmed/.
[11] http://www.acm.org/publications/policies/plagiarism_policy.
[12] http://www.ieee.org/publications_standards/publications/rights/plagiarism_FAQ.html.

$$MAP(n) = \frac{1}{|D|} \sum_{d=1}^{|D|} \frac{1}{|src_d|} \sum_{i=1}^{n} P(L_{d,i}) \tag{4}$$

$$P(L_{d,i}) = \frac{|\{s \in src_d \cap L_{d,i}\}|}{i} \tag{5}$$

$$R(L_{d,i}) = \frac{|\{s \in src_d \cap L_{d,i}\}|}{|src_d|} \tag{6}$$

$$F1(L_{d,i}) = \frac{2 \times P(L_{d,i}) \times R(L_{d,i})}{P(L_{d,i}) + R(L_{d,i})} \tag{7}$$

with

- n: cut-off value for ranked document list.
- $L_{d,i}$: top i documents of ranked document list for input document d.
- src_d: set of source documents for input document d.
- D: set of input documents.

MAP, R, P, and F1 produce score between [0,1]. When MAP score is 1, it means that all source documents for a given input document are located at the top of the ranked document list, consecutively. Thus, the higher the MAP score, the better the system performs.

When R score is equal to 1, it means that all source documents are contained in the ranked document list. While P score is 1, it suggests that all documents in the ranked document list are source document. Both R and P are combined as F1, thus the higher the F1 score, the better the system performs.

4.3 Experiment Result

Baseline. In our experiment, we compared our system with baseline that measured the similarity between the textual content of input and target documents without distinguishing citing and non-citing sentences. Thus, the baseline belongs to the category of textual content-based PD system. The processes in this baseline are similar to our system, except it does not perform sentence classification, and it uses Eq. 2 to compute the similarity scores.

Citation-Based PD. We also compared our system with the citation-based PD methods in Pertile et al. [8]. Generally, the methods compared list of reference and citation anchors between two documents, i.e. the number of similar references (BC), the overlap of references divided by its union (JR), the co-occurrence of citation anchors (CC), and the summation of a weighted similar reference, thus a reference cited by fewer documents has heavier weight (CF).

Result and Discussion. First, we present and discuss the results when our system only use the similarity of textual content in citing sentences (CS) or in non-citing ones (NS). Thus, we can identify which factor is the best and should

be prioritized in this experiment. In addition, we compare our system with the methods in Pertile et al. [8]. Second, we show and discuss the results of our system with the best weight (α) and its improvement, which is achieved. Lastly, we discuss the errors that happen in this experiment.

Table 2 presents the results when we only use the similarity of textual content in citing sentences (CS) or non-citing ones (NS) on the auto-simulated dataset. While the results of similar experiment on the manually judged dataset are shown in Tables 3 and 4. In Table 2, CS outperforms the baseline and NS at every cut-off based on their MAP scores. On average, the MAP scores of CS and NS are about .067 and .005 higher than the baseline, respectively. Moreover, we conducted 2-tailed paired t-test among them using their MAP scores (cut-off = 100), we found that their differences are significant at level 1%. Thus, these results suggest that CS should be prioritized on this dataset.

Table 2. The performance of baseline, NS, and CS on the auto-simulated dataset

Cut-off	Baseline				NS				CS			
	MAP	F1	R	P	MAP	F1	R	P	MAP	F1	R	P
10	.308	.100	.361	.058	.313	.107	.375	.063	**.379**	**.136**	**.435**	**.080**
30	.314	.052	.442	.028	.320	.054	.452	.029	**.384**	**.061**	**.488**	**.032**
100	.318	.024	.574	.012	.324	**.025**	**.582**	**.013**	**.386**	.023	.553	.012

Based on F1, R, and P scores in Table 2, CS outperforms the baseline and NS at cut-off 10 and 30, while NS outperforms CS and the baseline at cut-off 100. These results indicate that CS may be unable to identify some source documents that have a significant amount of similar non-citing sentences but not citing ones with input documents, and they are identified by NS. Thus, combining CS and NS is likely better, since both of them may complement each other.

In Table 3, which shows the MAP scores on the manually judged dataset, the performance of the baseline is pretty good. Since the ratio of input and source documents is approx. 1 (Table 1), and the R scores in Table 4 are close or equal to 1, the MAP score about .9 means the majority of the source documents are located at the first position in the ranked document lists for each input document. We found 7 out of 100 input documents (1 in the PubMed and 6 in ACL sub-dataset), which their source documents are not located at the first position in their ranked document lists. Consequently, to outperform the baseline on this dataset may be difficult. This happened probably due to the limitation of this dataset since Pertile et al. [8] only focused on pairs of document that have large amount of similar textual contents to be annotated. Thus, the source documents are mostly located at the top of the ranked document lists in our baseline.

The above reason may also explain why the performance of our system is different in both datasets. Since Alzahrani et al. [5] controlled the length of plagiarized text fragments in the auto-simulated dataset from the short to the

Table 3. The MAP scores of baseline, NS, CS, and methods in Pertile et al. [8] on the manually judged dataset

Sub-dataset	Cut-off	Baseline	NS	CS	CF	BC	JR	CC
PubMed	10 or more	**.993**	.978	.978	.900	.900	.900	.430
ACL	10	.908	.908	**.963**	—	—	—	—
	30	.908	.910	**.964**	—	—	—	—
	100	.910	.911	**.964**	.800	.810	.810	.620

long one, thus it is more difficult to identify source documents in this dataset. Moreover, they controlled the obfuscation level of the plagiarized text fragments from the light one to the heavy one, unlike the manually judged dataset.

On the PubMed sub-dataset, the MAP scores of CS and NS are about .015 lower than the baseline in Table 3, although their F1, R, and P scores in Table 4 are the same. While on the ACL sub-dataset, on average, the MAP scores of CS are .054 higher than the baseline and NS as shown in Table 3. In addition, CS outperforms the baseline and NS on this sub-dataset at cut-off 10 and 30 according to their F1, R, and P scores in Table 4. These results suggest that CS may be better if it is given heavier weight in the manually judged dataset.

In Table 3, CS also outperforms all the methods in Pertile et al. [8] based on their MAP scores on the manually judged dataset. These results indicate that comparing list of reference and citation anchors alone is not sufficient for PD.

Tables 5 and 6 present the experiment results of our system with the best weight (α) on the auto-simulated and manually judged dataset, respectively. As shown in Table 5, our system with $\alpha = .9$ achieves the best MAP scores at every cut-off on the auto-simulated dataset, which is about .07 higher than the baseline on average. We also conducted 2-tailed paired t-test between our system ($\alpha = .9$) and the baseline by using their MAP scores (cut-off = 100), we found that their difference is significant at level 1%.

Table 4. The F1, P, and R of the baseline, NS, and CS on the manually judged dataset

Sub-dataset	Cut-off	Baseline			NS			CS		
		F1	R	P	F1	R	P	F1	R	P
PubMed	10 or more	.220	1.000	.123	.220	1.000	.123	.220	1.000	.123
ACL	10	.181	.950	.100	.181	.950	.100	**.190**	**.975**	**.105**
	30	.064	.950	.033	.068	.988	.035	**.069**	**1.000**	**.036**
	100	.021	1.000	.011	.021	1.000	.011	.021	1.000	.011

Table 5. The best performance of our system on the auto-simulated dataset

Cut-off	Baseline				$\alpha = .5$				$\alpha = .9$			
	MAP	F1	R	P	MAP	F1	R	P	MAP	F1	R	P
10	.308	.100	.361	.058	.360	.135	.435	.080	**.383**	**.137**	**.440**	**.081**
30	.314	.052	.442	.028	.368	**.065**	**.518**	**.035**	**.388**	.062	.495	.033
100	.318	.024	.574	.012	.372	**.028**	**.633**	**.014**	**.390**	.024	.567	.012

In addition, our system ($\alpha = .9$) achieves the best F1, R, and P on the auto-simulated dataset in Table 5 at cut-off 10. While at cut-off 30 and 100, our system with $\alpha = .5$ achieves the best F, R, and P on this dataset, which is .007 higher than our system with $\alpha = .9$ based on their F1 scores.

In Table 6, our system does not outperform the baseline on the PubMed sub-dataset, although its best MAP score ($\alpha = .9$) is about .007 lower than the baseline and their F1, R, and P scores are the same. While on the ACL sub-dataset, we find that our system ($\alpha = .9$) outperforms the baseline about .067 higher on average based on their MAP scores, and also achieves the best F1, R, and P scores at cut-off 10 and 30.

Based on the MAP scores, our system achieves its best performance when $\alpha = .9$ on both datasets. Thus, it confirms our previous finding that it is better to give heavier weight on the similarity between textual content in citing sentences in this experiment. The results also suggest that the similarity between textual content in citing and non-citing sentences complement each other, since the MAP scores become worse when only using one of them. Moreover, we conducted 2-tailed paired t-test among the MAP scores (cut-off = 100) on the auto-simulated dataset when our system uses $\alpha = 0$ (NS), $\alpha = .9$, and $\alpha = 1$ (CS). We found that their differences are significant at level 1%.

Table 6. The best performance of our system on the manually judged dataset

Sub-dataset	Cut-off	Baseline				$\alpha = .9$			
		MAP	F1	R	P	MAP	F1	R	P
PubMed	10 or more	**.993**	.220	1.000	.123	.986	.220	1.000	.123
ACL	10	.908	.181	.950	.100	**.975**	**.190**	**.975**	**.105**
	30	.908	.064	.950	.033	**.976**	**.069**	**1.000**	**.036**
	100	.910	.021	1.000	.011	**.976**	.021	1.000	.011

Error Analysis. In this experiment, we conducted error analysis on the manually judged dataset using the best results of our system ($\alpha = .9$). We found 4 input documents (3 in the PubMed and 1 in the ACL sub-dataset), which their MAP scores are not as high as the others. We suspected that there are three reasons why these errors happened.

First, some of the target documents that are not annotated as source document have similarity scores higher than the source documents in our system. Moreover, we observed a significant similarity of textual contents between them and the input document. This happened probably because Pertile et al. [8] used cut-off (i.e. the top 30) instead of similarity threshold on their method to pool document pairs from the document collections to be annotated. Since their method might rank these target documents lower than the cut-off, these target documents are ignored and not annotated. We found two input documents from the PubMed sub-dataset associated with this error. Thus, determining similarity threshold to decide whether an input and a target document are a plagiarized and source document, respectively is another crucial issue in PD.

Second, the target documents mentioned above have similar topic with the input document. For instance, they discuss the same research problem, use the same learning algorithm, and evaluate their methods using the same dataset. However, their proposed methods are different. We also observed that they and the input document cite some documents together, and also use similar terminologies and descriptions. We identified one input document from each sub-dataset associated with this error.

Lastly, we suspect that our similarity method (see Eq. 2) may be sensitive to the length of text fragments when one of them is longer. We identified one input document from the ACL sub-dataset may associate with this error.

5 Conclusion

Plagiarism, which is one of the forms of academic misconducts, is problematic. It results in discouraging innovation and losing trust in the scientific research community. We proposed a method for plagiarism detection (PD) based on our model of plagiarism, which combines the similarity between textual contents in citing and non-citing sentences.

Given a plagiarized document as the input, our system identifies its source documents. We evaluated our system using two types of dataset, namely auto-simulated and manually judged dataset. In the evaluation, we compares our system with the baseline, which measures the similarity of textual contents between two documents without distinguishing citing and non-citing sentences.

According to the experiment results, our system does not outperform the baseline for the PubMed sub-dataset on the manually judged dataset, although the difference of their MAP (Mean Average Precision) scores is about .007. However, our system outperforms the baseline on the auto-simulated and the manually judge dataset for the ACL sub-dataset about .07 and .067 higher (MAP) than the baseline, respectively.

As for future work, we may extract more features from citation relations, and integrates them with the current system. Additionally, to reduce the vector sparsity of text fragments when measuring their similarities, we may use an algorithm to learn their vector representations.

References

1. Fang, F.C., Steen, R.G., Casadevall, A.: Misconduct accounts for the majority of retracted scientific publications. Proc. Nat. Acad. Sci. **109**(42), 17028–17033 (2012). doi:10.1073/pnas.1212247109. NAS
2. Kessler, M.M.: Bibliographic coupling between scientific papers. Am. Documentation **14**(1), 10–25 (1963). doi:10.1002/asi.5090140103. Wiley
3. Potthast, M., Gollub, T., Hagen, M., Graßegger, J., Kiesel, J., Michel, M., Oberländer, A., Tippmann, M., Barrón-Cedeño, A., Gupta, P., Rosso, P., Stein, B.: Overview of the 4th international competition on plagiarism detection. In: Forner, P., Karlgren, J., Womser-Hacker, C. (eds.) Working Notes Papers of the CLEF 2012 Evaluation Labs (2012)
4. Gupta, P., Rosso, P.: Text reuse with ACL: (upward) trends. In: Proceedings of the ACL-2012 Special Workshop on Rediscovering 50 Years of Discoveries, pp. 76–82. ACL (2012)
5. Alzahrani, S., Palade, V., Salim, N., Abraham, A.: Using structural information and citation evidence to detect significant plagiarism cases in scientific publications. J. Am. Soc. Inf. Sci. **63**(2), 286–312 (2012). doi:10.1002/asi.21651. Wiley
6. HaCohen-Kerner, Y., Tayeb, A., Ben-Dror, N.: Detection of simple plagiarism in computer science papers. In: Proceedings of the 23rd International Conference on Computational Linguistics, pp. 421–429. ACL (2010)
7. Gipp, B., Meuschke, N.: 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, pp. 249–258. ACM (2011). doi:10.1145/2034691.2034741
8. Pertile, S.D.L., Moreira, V.P., Rosso, P.: Comparing and combining content-and citation-based approaches for plagiarism detection. J. Assn. Inf. Sci. Tec. **67**(10), 2511–2526 (2016). doi:10.1002/asi.23593. Wiley