Vol.2017-IFAT-126 No.4
2017/3/30

Modeling Plagiarism in Citation Networks for Academic Publications

Sidik Soleman[1,a)] Atsushi Fujii[1]

Abstract: Digital archives of academic publications have enabled us to efficiently access a large volume of information. However, its misuses have become a crucial problem lately. In this paper, we model typical misuses of documents for academic publications, which enabled to propose a new model for automatic plagiarism detection.

1. Introduction

Digital archives for academic publications have enabled us to efficiently access a large volume of information. However, its misuses have of late become a crucial problem. Here, misuses, we refer to, are plagiarism and inappropriate citation.

Plagiarism is "the act of using another person's words or ideas without giving credit to that person"[*1], while inappropriate citation is a citing text that its assertion is not supported by the cited document. In the case of inappropriate citation, thus the person who should receive the credit is being denied. Therefore, inappropriate citation may also be a form of plagiarism. The problem of plagiarism results in discouraging innovation and losing trust in academic 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 is a task to identify whether a document in question was produced by means of plagiarism or not, and is often requested to present one or more source documents as evidences for the plagiarism. 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, plagiarism detection is a cat-and-mouse game between plagiarists and people who develop plagiarism detection systems.

To speed up the development of plagiarism detection systems, it is common to model and simulate the plagiarism because finding document created by means of plagiarism, or plagiarized document for short, is costly. For instance, Alzahrani et al. [1] modeled plagiarized document in academic publications as a document that has similar fragments with other documents without citing them, and simulated plagiarism by automatically combining and editing one or more documents to create a plagiarized document for evaluating their model.

In this paper, we modeled plagiarism for academic publication in citation networks. Citation is used when one borrows ideas or words from another person, and consequently links his/her document to the cited one. Therefore, ideas or words in citing texts can be associated with the cited document, and serve as complement to it when detecting plagiarism. For instance, it may be more reliable to also compare plagiarized document with fragments of text that cite (or *"citing texts to"*) its source document, instead of comparing the source document alone, because the citing texts may contain ideas that are used in the plagiarized one. Thus, by modeling plagiarism in citation networks, we may be able to improve the capabilities of plagiarism detection systems.

Whereas the above scenario is associated with intentional plagiarisms, detecting unintentional plagiarisms are also important to avoid innocent mistakes. Fang et al. [4] investigated approximately 2 000 papers that were once indexed by PubMed 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.

2. Related Work

As plagiarism detection is a cat-and-mouse game between plagiarists and people who develop systems of plagiarism detection, and collecting a large number of acutal plagiarized documents are costly, modeling and simulation strategy are commonly used to speed up the development of the systems.

In the early plagiarism detection, since plagiarist creates document by means of other documents, people modeled plagiarized document as the one that has similar fragments of text with the others. For example, Potthast et al. [6] [7] adopted this model and subsequently simulated it to create plagiarized document for PAN workshop, a competition on plagiarism detection. In their first simulation [6], plagiarized document was automatically generated by combining and editing one or more documents randomly from a document collection, while in the second simu-

[1] Tokyo Institute of Technology, Meguro–ku, Tokyo 152–8550, Japan
[a)] soleman.s.aa(at)m.titech.ac.jp

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

lation [7], the plagiarized one was manually written by means of one or more documents that were retrieved using a search engine.

In the PAN workshop, people used these datasets to test their proposed methods, which were mostly adopting the above model of plagiarized document, for example, Grozea et al. [3] applied character n-gram comparison to measure the degree of similarity between the plagiarized one and its source documents. However, this model ignores the fact of citation, which may be an inappropriate one, or an innocent case because citing text may be similar to the cited document.

To consider the fact of citation, people then modeled plagiarized document as the one that has similar fragments of text with the others and does not cite them, for instance, Alzahrani et al. [1] created plagiarized document, and proposed a method to detect plagiarism based on this model for academic publications. They created plagiarized document by means of one or more document without citing them, and used the existence of citation to filter the innocent case in their proposed method. However, this model underestimates the fact of inappropriate citations, which may go undetected.

More recently, a model of plagiarism was proposed by Gipp et al. [5], which was motivated by the action of plagiarists that target content containing citations, such as literature review section of a document. They modeled plagiarized document in their proposed method as the one that has similar fragments of text, structure of citation anchors [*2], and list of document in reference with others. However, this model also underestimates the inappropriate citation and ignores the potential of citation network especially citing text, as it focuses only on citation anchor and reference list.

To summarize, three models had been proposed to define plagiarized document as:

(1) The one that has similar fragments of text with others.
(2) The one that has similar fragments of text with others without citing them.
(3) The one that has similar fragments of text with others, similar structures/patterns of citation anchors, and similar documents in reference list [5].

However all of them ignore the problem of inappropriate citation, which is also form of plagiarism and the potential of citation networks to improve plagiarism detection systems, especially that use citing texts and citation links to model plagiarism.

The usefulness of citation network has been recognized in many applications, for example, Fujii et al. [2] re-ranked patents in patent retrieval using count of how many times patents are cited by the others because patent cited by many others is important (e.g. the patent become a basis of its citing patents).

Whereas above work uses citation link, Ritchie et al. [9] applied citing texts extracted from citation networks to increase terms of cited document for information retrieval, because citing texts contain description of some aspects of the cited document, consequently their terms should be good for index of the cited document. This reason also motivated Qavzinian et al. [8] to use citing texts for creating summary of the cited document in the task of automatic summarization.

[*2] citation anchor refers to the alphanumberic code that points to a document in reference list

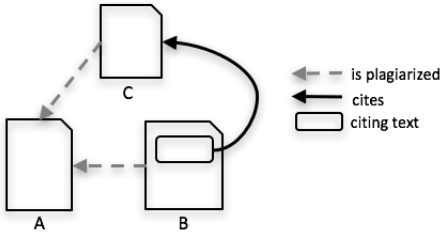

Fig. 1 Illustration of basic idea

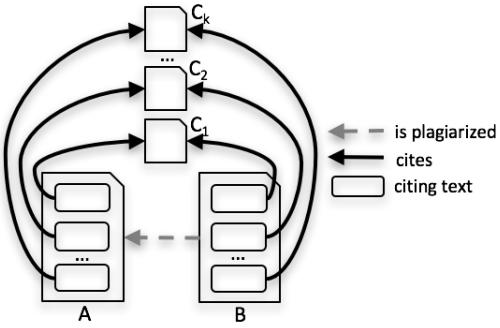

Fig. 2 Situation when a related work is plagiarized

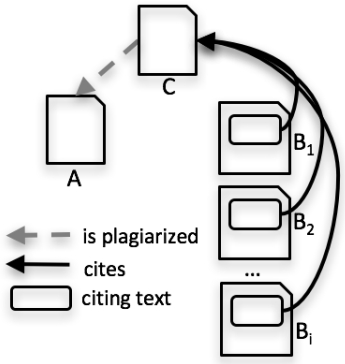

Fig. 3 Situation when large part of or novelty is plagiarized

Since there is a few work that model plagiarism detection that use citation network, despite its potential that has been shown in many applications, in this paper, we modeled it for academic publications.

3. Proposed Model of Plagiarism Detection

3.1 Basic Idea

The basic idea of our model of plagiarism detection is to use citation network of candidate source documents (e.i. citing text and citing-cited relations). For the sake of simplicity, we only assume two documents B and C, and also assume B cites C. B and C can be expanded to a set of documents, namely B_i's and C_k's, respectively. Document A is plagiarized by means of either B or C. Figure 1 illustrates this situation.

3.2 Detecting Plagiarism in Citation Networks

Here, we elaborate our ideas for modeling our plagiarism detection using citation networks. In short, we explain how to detect plagiarism using documents that are cited by the source document, using documents that cites source documents, and by de-

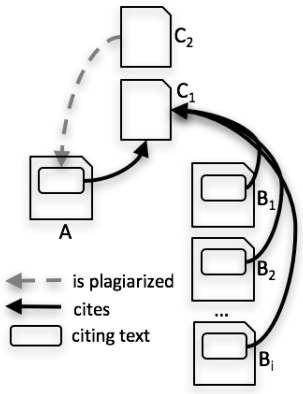

Fig. 4 Situation when inappropriate citation exists

tecting inappropriate citation. We also explain the situation when there is no plagiarism taking place. The details of them are explained as follows:

(1) **Using C to detect plagiarism for B**
We assume a situation where A is plagiarized by means of B. A possible scenario is to plagiarize a literature review in B regarding C. In such a case, A must satisfy the following conditions:
(a) A contains a specific amount of fragment that is similar to one in B, and
(b) Citation information, such as citation anchor and bibliography for C, is maintained.
In summary, a high similarity between A and B with respect to fragments associated with C and the existence of the appropriate citation information for C in A suggests the possibility that A has been plagiarized by means of B. Figure 2 illustrates this situation.

(2) **Using B to detect content plagiarism for C**
In this case, A must satisfy the following conditions:
(a) A contain a specific amount of fragment that is similar to one in C and/or B, and
(b) Citation information, such as citation anchor and bibliography for C, is not contained.
The reason why we use B instead of C is that using Bi's will be more reliable than a single document. In summary, a high similarity between A and B with respect to fragments associated with C and the absence of the appropriate citation information for C in A suggests the possibility that A has been plagiarized by means of C. Therefore, a possible scenario here is to plagiarize large part of text of C or novelty of C, which B agrees that the novelty belongs to C, as illustrated in Figure 3.

(3) **Detecting inappropriate citation**
This is the opposite case of the above (2), where the citation to C exists but the similarity between A and B, and/or between A and C is small. As in the above (2), increasing the number of Bis will be more reliable than relying only a single B. The situation of inappropriate citation is illustrated in Figure 4.

(4) **No plagiarism**
No similarity between A and B/C and no citation link between A and C.

3.3 Formal Definition

Based on previously explained conditions for plagiarism and inappropriate citation, here we formalized them using the following conditions:

- **Condition (I):**

$$Sim(A, B_i) = max_k\ Sim(A, B_i, C_k) >= \theta \qquad (1)$$

The equation above represents the similarity between A and B_i is the maximum similarity between A and B_i with respect to C_k, and also that similarity is equal or greater than a predetermined threshold theta. We can also replace max_k with sum_k, depending on the definition for the similarity between two documents.

- **Condition (II):**

$$Cite(A, C_k) = True \qquad (2)$$

The above condition represents A cites C_k with a citation anchor.

Based on two conditions above, we can perform classification whether A is plagiarized document and B/C is its source document. The rules for the classification are:

(1) Condition (I) is True AND Condition (II) is True:
A is a plagiarized document.
B is a source document.
(2) Condition (I) is True AND Condition (II) is False:
A is a plagiarized document.
C is a source document.
(3) Condition (I) is False AND Condition (II) is True:
A is a plagiarized document.
C is a victim of inappropriate citation
(4) Condition (I) is False AND Condition (II) is False:
No plagiarism

3.4 Implementation

To implement our model, first, we have to define the similarity function in the condition I or the equation 1, but before that, we have to understand the process of plagiarism detection. Generally, detecting plagiarism needs to compare an input document, in this time we assume that the input one is plagiarized document, with every document in a collection one by one, and ranks these documents according to their comparison scores in order to put the source documents on the top of the list, thus, the source documents can be identified easily. Therefore, we may ignore θ in equation 1 for this time. However, we still need to define the similarity function in equation 1 to compute document scores based on the classification rules.

In the equation 1, we measure similarity between A and B_i with respect to C_k. B_i represents a document in collection, hence, the number of B_i's is equal to the number of documents in the collection and *i* iterates from the first document to the last one in the collection. While C_k represents document that is cited by B_i, hence, the number of C_k's is equal to the number of documents cited by B_i. Therefore, the equation 1 means that it compare A and B_i to compute scores for C_k's, as a consequence, this equation function returns some scores for each C_k *(Score(C_k))*.

According to the rule (2) and (3) in Section 3.3, we define the equation 1 as:

$$Sim(A, B_i) = Score(C_k) = NPSim(A, B_i, C_k) + ICSim(A, B_i, C_k) \quad (3)$$

with:

- NPSim(A, B_i, C_k) represents the score of C_k when larger part or novelty of C_k is plagiarized in A, which is calculated using B_i, and
- ICSim(A, B_i, C_k) represents the score of C_k when it becomes the victim of inappropriate citation of A, which is also calculated using B_i.

We then define NPSim and ICSim based on the rule (2) and (3) in Section 3.3 as:

$$NPSim(A, B_i, C_k) = max_k\ (1 - Cite(A, C_k)) \times Sim(A, CiteText(B_i, C_k)) \quad (4)$$

$$ICSim(A, B_i, C_k) = max_k\ Cite(A, C_k) \times (1 - Sim(CiteText(A, C_k), CiteText(B_i, C_k))) \quad (5)$$

with:

- Cite(A, C_k) represents a condition whether A cites C_k with a citation anchor, if the condition is true, the value is equal to 1, or otherwise 0,
- CiteText(A, C_k) and CiteText(B_i, C_k) represent fragments of A and B_i that cites C_k,
- *Sim* represents a function that measures the similarity between two texts (e.g. cosine similarity function)
- max_k may also be replace with sum_k on the application, because we have B_i's.

Since C_k's are also contained in the collection, at some point, C_k is B_i. However, not all B_i's may become C_k, because it depends on whether they are cited by other documents or not. If a document is not cited by any document in collection, there is no chance this document to become C_k. In other words, the Score(C_k) is 0. This situation is not good for plagiarism detection system, because plagiarist may take advantage of this.

To overcome this situation, we add the score of C_k by defining another function based on the rules in Section 3.3. Since we also have not considered the rule (1), we define a function that represents this rule as well.

$$FinalScore(C_k) = Score(C_k) + DocSim(A, C_k) + CTSim(A, C_k) \quad (6)$$

$$DocSim(A, C_k) = Cite(A, C_k)(1 - Sim(A, C_k)) + (1 - Cite(A, C_k))(Sim(A, C_k)) \quad (7)$$

$$CTSim(A, C_k) = \frac{1}{k'} \sum_{k'} Sim(CiteText(A, C'_{k'}), CiteText(C_k, C'_{k'})) \quad (8)$$

with:

- FinalScore(C_k) is the final score for C_k that is used to rank document by the system of plagiarism detection,
- DocSim(A, C_k) is the document similarity between A and C_k
- *Sim* represents a function that measures the similarity between two texts
- CTSim(A, C_k) represents the normalized similarity score of citing texts to between A and C_k, and
- k' is the number of documents cited by C_k.

3.5 Advantage of Our Model

Since we use citation networks to model our plagiarism detection, it enabled us to use many documents to detect plagiarism for a document, which is more reliable than using that document alone. For example, when plagiarized document contains an original idea of the source document, our model compares the plagiarized one with the others that cite the source document by means of their citing texts. Since ideas or words in citing texts can be associated or original to the cited document, in this case the source document, to identify the original idea that is plagiarized is probably better to compare the plagiarized one with citing texts rather than comparing the plagiarized one alone.

We may see citation as vote, because many people vote an ideas belongs to a person, when there is somebody wanting to claim the idea, we can reject the claim without doubt because many people know who is the owner of it.

Because in our model we also address the problem of inappropriate citation, which is one of the forms of plagiarism, and the problem when people plagiarize a literature review, it may limits any modification related to citation. For example, if a citation is replaced its anchor or changed its content by plagiarists, our model may idenfity this citation as inappropriate one, or if they re-order its presentation in their documents, our model may still identify this plagiarized one, because our model compares citing texts with the respect to the cited documents. As a result, it may be difficult for plagiarists to do their actions in our model.

3.6 Simulation and Future Work

Since simulation is also common strategy in developing the system of plagiarism detection, in order to speed up the process, we plan to use this strategy to test our model of plagiarism.

We plan to test our model using the existing dataset from Alzahrani et al [1]. This dataset was constructed by automatically simulating plagiarism by means of one or more documents. Given a random document, they added fragment of text from other documents and also applied some text modification methods (e.g. auto-paraphrase, auto-summarization, or double-back translation). However, this dataset is still lack of inappropriate citation cases. To overcome this problem, we may do the following strategies:

- Annotating the case of inappropriate citation manually in this dataset, and/or
- Automatically creating instances of inappropriate citation cases using this dataset by randomly assigning citation anchors to fragments of text or replacing citation anchors of citations by other documents.

4. Conclusion

In this paper, we proposed a new model of plagiarism detection using citation networks. We model this by modeling the misuse

of documents, namely plagiarism and inappropriate citation. Inappropriate citation is also one of the forms of plagirism, because the person who should receive attribution of that citation is being denied.

We created rules to classify whether a document is plagiarized one, a source document, or a victiom of inappropriate citation. We also defined a scoring method for retrieving source documents.

By modeling plagiarism in citation networks, it enables us to use many documents to detect plagiarism for a document. As a result, it may be difficult for plagiarists to perform their actions in our model.

For future work, we plan to implement our model and evaluate its effectiveness by using existing dataset. Since inappropriate citation case is not available in the dataset, we may develop this as well.

References

[1] Alzahrani, S., Palade, V., Salim, N., and Abraham, A.: Using Structural Information and Citation Evidence to Detect Significant Plagiarism Cases in Scientific Publications, *Journal of the American Society for Information Science and Technology. Wiley Subscription Services, Inc., A Wiley Company*, Vol.63, No.2, pp.286–312 (2012).

[2] Fujii, A.: Enhancing Patent Retrieval by Citation Analysis, *Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval. ACM*, Amsterdam, Netherlands, pp.793–794 (2007).

[3] Grozea, C., Gehl, C., and Popescu, M.: ENCOPLOT: Pairwise sequence matching in linear time applied to plagiarism detection, *SEPLN 09 Workshop on Uncovering Plagiarism, Authorship, and Social Software Misuse (PAN 09). CEUR-WS.org*, San Sebastian (Donostia), Spain, pp.10–18 (2009).

[4] Fang, F.C., Steen, R.G., and Casadevall, A.: Misconduct Accounts for the Majority of Retracted Scientific Publications, *Proceedings of the National Academy of Science. National Academy Sciences*, Vol.109, No.42, pp.17028–17033 (2012).

[5] Gipp, B. and Meuschke, N.: Citation Pattern Matching Algorithms for Citation-based Plagiarism Detection: Greedy Citation Tiling, Citation Chunking and Longest Common Citation Sequence, *Proceedings of the 11th ACM Symposium on Document Engineering (DocEng'11). ACM*, Mountain View, California, USA, pp.249–258 (2011).

[6] Potthast, M., Stein, B., Eiselt, A., Barrón-Cedeño, A., and Rosso, P.: Overview of the 1st International Competition on Plagiarism Detection, *SEPLN 09 Workshop on Uncovering Plagiarism, Authorship, and Social Software Misuse (PAN 09). CEUR-WS.org*, San Sebastian (Donostia), Spain, pp.1–9 (2009).

[7] Potthast, M., Gollub, T., Hagen, M., Graßegger, J., Kiesel, J., Michel, M., Oberländer, A., Tippmann, M., Barrón-Cedeño, A., Gupta, P., Rosso, P., and Stein, B.: Overview of the 4th International Competition on Plagiarism Detection, *Working Notes Papers of the CLEF 2012 Evaluation Labs*, Rome, Italy, available from ⟨http://www.clef-initiative.eu/publication/working-notes⟩ (2012).

[8] Qazvinian, V. and Radev, D.R.: Scientific Paper Summarization Using Citation Summary Networks, *Proceedings of the 22nd International Conference on Computational Linguistics - Volume 1. ACL*, Manchester, United Kingdom, pp.689–696 (2008).

[9] Ritchie, A., Robertson, S., and Teufel, S.: Comparing Citation Contexts for Information Retrieval, *Proceedings of the 17th ACM Conference on Information and Knowledge Management. ACM*, Napa Valley, California, USA, pp.213–222 (2008).