Citation Based Plagiarism Detection - A New Approach to Identify Plagiarized Work Language Independently

Bela Gipp
UC Berkeley / OvGU
102 South Hall, Berkeley
+1 (510) 859-3860
gipp@berkeley.edu

Jöran Beel
UC Berkeley / OvGU
102 South Hall, Berkeley
+1 (510) 859-3860
beel@berkeley.edu

ABSTRACT
This paper describes a new approach towards detecting plagiarism and scientific documents that have been read but not cited. In contrast to existing approaches, which analyze documents' words but ignore their citations, this approach is based on citation analysis and allows duplicate and plagiarism detection even if a document has been paraphrased or translated, since the relative position of citations remains similar. Although this approach allows in many cases the detection of plagiarized work that could not be detected automatically with the traditional approaches, it should be considered as an extension rather than a substitute. Whereas the known text analysis methods can detect copied or, to a certain degree, modified passages, the proposed approach requires longer passages with at least two citations in order to create a digital fingerprint.

Categories and Subject Descriptors
H.3.3 [**Clustering**]: INFORMATION STORAGE AND RETRIEVAL – *Information Search and Retrieval*.

General Terms
Algorithms, Measurement, Languages

Keywords
Plagiarism Detection, Duplicate Detection, Citation Analysis, Citation Order Analysis, Language Independent

1. INTRODUCTION
Plagiarism is defined as the 'use or close imitation of the language and thoughts of another author and the representation of them as one's own original work.'[1]

Plenty of websites addressing students and scholars give advice on how to ensure that plagiarized text cannot be identified by a plagiarism detection system such as copyscape.com. The most common advice given is to paraphrase and use synonyms, or even copy from sources that were written in another language. Plagiarism detection services responded by integrating dictionaries and sophisticated data analysis methods. However, these systems still have unsatisfying detection rates if text is paraphrased or translated as shown at the *International Competition on Plagiarism Detection* in 2009 [6].

[1] Random House Compact Unabridged Dictionary, 1996

HT'10, June 13–16, 2010, Toronto, Ontario, Canada.
ACM 978-1-4503-0041-4/10/06.

2. RELATED WORK
Hundreds of papers have been published covering sophisticated approaches to detect plagiarism, and dozens of applications were developed. All of them use more or less sophisticated approaches to analyze the text, but ignore the used citations [3], [6]. These approaches deliver excellent results in detecting copied text passages, but fail if text has been paraphrased or translated—for example, from German to English. Instead of analyzing the words of a document, this paper suggests analyzing the used citations.

To our knowledge, applying citation analysis approaches to detect plagiarism has not yet been attempted. Several citation analysis approaches, however, have been developed as a measure of subject relatedness. In 1963, Kessler introduced [2] the concept of bibliographic coupling. Document A and Document B are bibliographically coupled if they cite one or more documents in common. Figure 1 illustrates this approach: Documents A and B are related because they both cite Documents 1, 2 and 3.

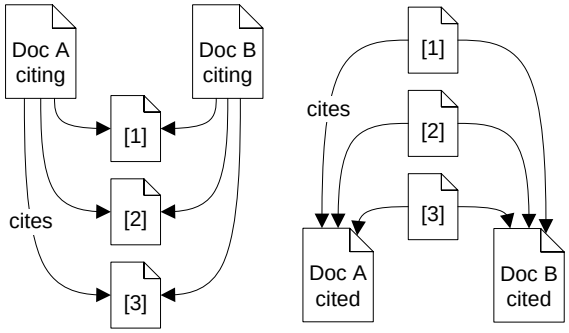

Figure 1: Bibliographic coupling (left) and co-citation (right)

A variation of this, called co-citation, was proposed by Marshakova [4] and Small [5]. Two documents are "co-cited" when at least one document cites both. This approach is illustrated on the right in Figure 1: Documents A and B are related because both are cited by Documents 1, 2 and 3. The more co-citations two documents receive, the more related they are. A further development of this approach is Citation Proximity Analysis, which identifies related documents by their co-occurrence of citations under consideration of their proximity to each other [1]. All approaches allow the calculation of the coupling strength and

are used to identify related articles by academic search engines such as *SciPlore.org* and *CiteSeer*.

3. THE NEW APPROACH – CITATION ORDER ANALYSIS

The new approach, which we call *Citation Order Analysis* (COA), is similar to bibliographic coupling, but also analyses the order of citations within the document. This allows the creation of a citation-based digital fingerprint. By using tolerant sequence analysis algorithms, such as the Levenshtein distance, plagiarized text can also be detected if the order of citations has been slightly changed, as shown in Figure 2.

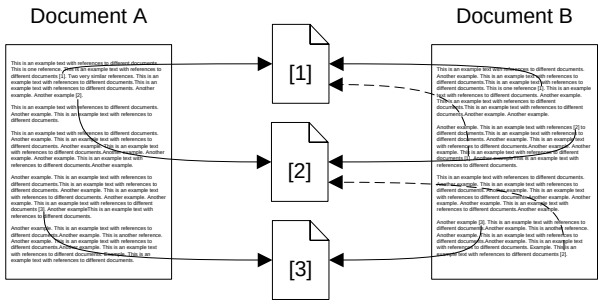

Figure 2: Example Citation Order Analysis (COA)

These steps are performed in our plagiarism detection system:

1. The document is parsed and a series of heuristics applied to process the citations, including their position within the document[2].

2. Citations are matched with their entries in the bibliography.

3. The citation-based similarity of the documents is calculated. In the basic version, only the order is considered; in the more advanced version, the distance between two citations is evaluated as well. Even if a document is translated, the order of citations within sentences or paragraphs might change due to different sentence structures or writing styles.

4. COMPARISON OF RESULTS

A comparison with the existing approaches is problematic, as both approaches have their own strengths. Whereas text-based approaches detect local similarity, like copied sentences, this citation-based approach analyzes global similarity. The interpretation, for instance, of a precision and recall value only makes sense when compared to other approaches. Since no other approaches exist for paraphrased and translated scientific text, such a comparison is not feasible. The test sets, like the PAN-PC-09 that was used at the first International Competition on Plagiarism Detection in 2009, are tailored to compare the performance of classical plagiarism detection systems, but are unsuitable to test this new approach, as citations were ignored.

To evaluate our approach, we ran a test on 0.8 million scientific publications from open access repositories and hid among them 20 specially-designed plagiarized documents. To create a more realistic test scenario, we deleted some citations, added new ones, changed the order slightly, and changed the citation style. The outlined approach identified 19 of the test documents, along with hundreds that contained at least some plagiarized sections. One very short document was not identified; it cited five sources, of which we deleted two. Precision and recall could be improved by considering the overall citation counts and the expected probability that they are co-cited. Rarely cited documents form a better digital fingerprint than frequently cited documents.

By lowering the threshold, not only can plagiarism be detected, but also documents which have not been cited, that were involved in the creation process. For example, Tom reads paper A, which cites paper B and C. Later he writes his own paper and remembers the interesting ideas in papers B and C. He cites them, but does not cite paper A, which had originally brought his attention to papers B and C in the first place. This is not usually considered plagiarism, but knowledge concerning which papers were involved in the creation process can be of interest.

Converting the pdf files to xml and the extraction of the necessary citation information took on average two seconds per publication on a 4x2.33 GHz Linux server. Approximately 96% of the citations could be identified and matched with the entries in the bibliography. The time for the similarity computation using an optimized version of the Levenshtein distance is negligible.

5. CONCLUSION

This paper describes a new approach to detect plagiarism and scientific documents that have been read but not cited. The main advantage of this approach is its language independency and its immunity to paraphrasing. The computational complexity is minimal if compared to text-based approaches. The main disadvantages are the dependence of (correct) citations, and that short passages with few citations cannot be detected.

However, the weaknesses of this approach are, to a large extent, the strengths of the text-based approaches. To achieve the best possible results, combining the classical text-based method with this citation order-based approach is recommended.

6. REFERENCES

[1] B. Gipp and J. Beel. Citation Proximity Analysis (CPA) - A new approach for identifying related work based on Co-Citation Analysis. *Proceedings of the 12th International Conference on Scientometrics and Informetrics (ISSI'09)*, volume 2, pages 571–575, 2009. Download from: www.SciPlore.org

[2] M. M. Kessler. Bibliographic coupling between scientific papers. *American Documentation*, 14:10–25, 1963.

[3] R. Lukashenko, V. Graudina, and J. Grundspenkis. Computer-based plagiarism detection methods and tools: An overview. In *Proceedings of the 2007 international conference on Computer systems and technologies*, page 40. ACM, 2007.

[4] I.V. Marshakova. System of document connections based on references. *Scientific and Technical Information Serial of VINITI*, 6(2):3–8, 1973.

[5] H. Small. Co-citation in the scientific literature: a new measure of the relationship between two documents. *Journal of the American Society for Information Science*, 24:265–269, 1973.

[6] B. Stein, editor. *PAN-09 3rd Workshop on Uncovering Plagiarism, Authorship and Social Software Misuse and 1st International Competition on Plagiarism Detection*, 2009.

[2] The citations were parsed using a modified version of parsCit (http://wing.comp.nus.edu.sg/parsCit) in combination with the authors' self-developed software, which is available upon request.