Counting Co-occurrences in Citations to Identify Plagiarised Text Fragments

Solange de L. Pertile[1], Paolo Rosso[2], and Viviane P. Moreira[1]

[1] Instituto de Informática - UFRGS – Brazil
{slpertile,viviane}@inf.ufrgs.br
[2] Natural Language Engineering Lab. - ELiRF
Department of Information Systems and Computation
Universitat Politècnica de València, Spain
prosso@dsic.upv.es

Abstract. Research in external plagiarism detection is mainly concerned with the comparison of the textual contents of a suspicious document against the contents of a collection of original documents. More recently, methods that try to detect plagiarism based on citation patterns have been proposed. These methods are particularly useful for detecting plagiarism in scientific publications. In this work, we assess the value of identifying co-occurrences in citations by checking whether this method can identify cases of plagiarism in a dataset of scientific papers. Our results show that most the cases in which co-occurrences were found indeed correspond to plagiarised passages.

1 Introduction

Plagiarism is one of the most serious forms of academic misconduct. It is defined as the act of the appropriation of another person's ideas, words, or works without giving credit to the original source. With the growing popularity of the Internet, many documents are freely available enabling students and researchers to reuse words from other authors without crediting them.

According to a study by Mccabe [10] 36% of undergraduate students and 24% of graduate students, admitted having copied or paraphrased sentences from the Internet without referencing them. More recently, the Journal of Zhejiang University-Science (China) [2]used the CrossCheck tool [1]to analyse the papers submitted to their revision process. They found that 22.8% (692 out of 2,233) of the papers presented unreasonable levels of copying or self-plagiairsm [14]. High levels of text reuse have also been found by Gupta & Rosso [9] who analysed papers accepted by the ACL.

The interest in plagiarism detection has been rising in the last few years. The PAN benchmarks [3] has been running for four years with an increasing number of participants all over the world [12]. PAN's evaluations aim at detecting different forms of plagiarism providing a standardised evaluation framework.

Usually, plagiarism detection relies on content analysis. The idea is to identify text fragments in common between a suspicious document and possible sources.

Automatic detection techniques have been proposed to deal with the various forms of plagiarism. Content analysis is more difficult in the presence of paraphrasing [6] and even more so when more than one language is involved, *i.e.* in cross-language plagiarism [11].

More recently, Gipp et al. [8] propose methods for plagiarism detection based on citation analysis. In their work, two documents which cite the same references are considered as having a high degree of similarity. The ideas are interesting since citation-based plagiarism detection could potentially be used in cases which content-based retrieval is typically ineffective. However, experimental evidence of the effectiveness of citation-based methods is limited to the application of the method in a prominent case of plagiarism concerning the doctoral thesis of a German politician. In another study, Alzahrani et al. [5] use the citations within a scientific paper in a different way. Cases in which the original source has been properly referenced are ignored by the content analysis phase. Thus, citations are used as a filter and not as an evidence of similarity across papers.

In this paper, we aim to bridge the gap between these two aforementioned works. We compute citation co-occurrences on the dataset used in [5] and assess whether they are effective in pointing out cases of plagiarism.

2 Identifying Co-Occurrences in Citations

Throughout this paper, we use the term *citation* to refer to the strings in the body of a scientific paper which point to where the original text was extracted from. The term *reference* is used to denote an entry in the Bibliography (or References) section of the paper.

Our aim is to compare the similarity of scientific papers based on the analysis of co-occurrences in citations. If two documents share at least a pair of citations within a text fragment, this is computed as an inter-document co-occurrence. Our assumption is that a high rate of inter-document co-occurrences is an indication of plagiarism. Given a pair of documents, these are represented as the co-occurrences of their citations. These intra-document co-occurrences are computed sliding a window of size s through the document. The inter-document co-occurrences are then calculated as the Jaccard similarity coefficient (or *overlap*) of these co-occurrences: $sim(w_i, w_j) = \frac{w_i \cap w_j}{w_i \cup w_j} = \frac{n_{i,j}}{n_i + n_j - n_{i,j}}$
where w_i and w_j are the windows i and j, respectively, $n_{i,j}$ is the number of shared co-occurrences between windows i and j, n_i and n_j are the number of co-occurrences in windows i and j, respectively.

More specifically, the steps involved in our process are the following:

– **Pre-processing:** Identify citations within the text of the document and link them to their corresponding entry in the list of references.

–**Computing co-occurrences within a document:** Slide a window of size s through the document and compute co-occurrences within this window.

– **Computing co-occurrences across documents:** For each pair of co-occurrences between a window in a suspicious document and a window in a source document, check whether they match using approximate string matching. References with a similarity score higher than a given threshold t are considered as being referring to the same paper.

Table 1. Results of Co-occurrence Analysis

	$s = 5$	$s = 15$	$s = 30$
Co-occurrences in Citations	90	160	161
Plagiarism with Co-occurrences	51	76	64
Precision	0.5667	0.4750	0.3975
Recall	0.0123	0.0183	0.0154
F1	0.0241	0.0353	0.0297

3 Experiments

The ideal dataset to analyse in our experiments would be a real collection of scientific papers with some cases of plagiarism. However, such a collection does not exist. Thus, we resorted to an artificial dataset originally described in [5] and available from [4], which is composed of scientific papers. There are 8,657 original and 6,755 suspicious papers containing verbatim and obfuscated cases of plagiarism. Annotation files revealing which fragments were plagiarised enable checking whether co-occurrences in citations are good indicators of plagiarism. At the moment, we can only handle papers which cite references using the numbered style. Thus, in our work, 4149 suspicious papers were compared against the 6035 source documents which adopt the numbered style.

In order to segment the references, we relied on the Ondux tool [7], which represents the state-of-the-art in information extraction by text segmentation. An extension of Levenshtein's Edit distance, called Carla [13], which accounts for inversions of substrings, was used to compare references across documents. The similarity threshold used was t=0.86, based on empirical observations. Window sizes (s) were 5, 15, and 30 through the documents.

The results are shown in Table 1. The smallest window (i.e., 5 lines) yielded the highest precision (56.67%), which means that in most cases in which co-occurrences were found, indeed correspond to plagiarism. The remaining cases with co-occurrences that were not considered plagiarism were due to three main reasons: (i) the suspicious document had cited the source from which text and references had been copied; (ii) two similar references were wrongfully treated as the same by our method, (iii) papers by the same authors and about the same topic had a high level of citation co-occurrence, but were not considered as plagiarism. In some cases, the paragraphs in the suspicious and source documents have identical contents and still were not annotated as plagiarism.

On the other hand, only a very small fraction of the cases of plagiarism have been identified. The main reason for the low recall is that, in most of the cases of plagiarism in this collection, the text fragment copied from the original did not include any references. Also, in some cases, the plagiarised fragment had been extracted from an article from a totally different area (e.g. the source from a plagiarised text in economy was a paper on veterinary). In such cases, it is very unlikely that the source and suspicious paper would share any references.

4 Conclusion

This work presented a study on the validity of using co-occurrences in citations to detect plagiarism in scientific documents. We carried out experiments on a dataset of scientific papers with cases of plagiarism simulated artificially. Our results have shown that most of the cases with co-occurrences in citations correspond to plagiarism. Moreover, nearly all of these cases were within paraphrased text fragments. On the other hand, only a small fraction of plagiarism cases involved text fragments with citations. This suggests that a hybrid approach which combines content similarity and citation analysis can potentially yield better detection quality. As future work, we plan to test whether citation co-occurrences help identify portions of text reuse within a real collection of scientific papers by comparing with the results of [9] on the ACL corpus.

Acknowledgements: This work was partially funded by CNPq (478979/2012-6). Solange Pertile's 5-month internship at NLE Lab of Universitat Politècnica de València was funded by CAPES. P.Rosso's work was carried out in the framework of the the VLC/CAMPUS Microcluster on Multimodal Interaction in Intelligent Systems and the European Commission WIQ-EI IRSES (no. 269180) and DIANA-APPLICATIONS-Finding Hidden Knowledge in Texts: Applications (TIN2012-38603-C02-01) research projects. We thank the authors of [5] for sharing their dataset with us and Enrique Flores for the preliminary brainstorming on how to identify co-occurrences in citations.

References

1. CrossCheck http://www.crossref.org/crosscheck/.
2. Journal of Zhejiang University-Science http://www.zju.edu.cn/jzus/.
3. PAN http://www.pan.webis.de.
4. Plagiarism corpus http://www.c2learn.com/plagiarism/corpus/v1/.
5. Alzahrani, S., Palade, V., Salim, N., and Abraham, A. Using structural information and citation evidence to detect significant plagiarism cases in scientific publications. *JASIST 63*, 2 (2012), 286–312.
6. Barrón-Cedeño, A., Vila, M., Marti, A., and Rosso, P. Plagiarism meets paraphrasing: Insights for the next generation in automatic plagiarism detection. *Computational Linguistics 39*, 4 (2013).
7. Cortez, E., da Silva, A. S., Gonçalves, M. A., and de Moura, E. S. Ondux: on-demand unsupervised learning for information extraction. In *SIGMOD* (2010), pp. 807–818.
8. Gipp, B., and Meuschke, N. Citation pattern matching algorithms for citation-based plagiarism detection: greedy citation tiling, citation chunking and longest common citation sequence. In *DocEng* (2011), pp. 249–258.
9. Gupta, P., and Rosso, P. Text reuse with acl: (upward) trends. In *Special Workshop on Rediscovering 50 Years of Discoveries* (2012), ACL '12, pp. 76–82.
10. Mccabe, D. L. Cheating among college and university students : A north american perspective. *International Journal for Educational Integrity 1* (2005).
11. Potthast, M., Barrón-Cedeño, A., Stein, B., and Rosso, P. Cross-language plagiarism detection. *Language Resources and Evaluation 45*, 1 (2011), 45–62.
12. Potthast, M., Gollub, T., Hagen, M., Tippmann, M., Kiesel, J., Stamatatos, E., Rosso, P., and Stein, B. Overview of the 5th International Competition on Plagiarism Detection. In *CLEF 2013 - Working Notes* (Sept 2013).
13. Ritt, M., Costa, A. M., Mergen, S., and Orengo, V. M. An integer linear programming approach for approximate string comparison. *European Journal of Operational Research 198*, 3 (2009), 706 – 714.
14. Zhang, Y. Crosscheck: an effective tool for detecting plagiarism. *Learned Publishing 23* (2010), 9–14.