Erschienen in: SIGIR'13 : proceedings of the 36th International ACM SIGIR Conference on Research and Development in Information Retrieval ; July 28 - August 1, 2013, Dublin, Ireland / Gurrin, Cathal et al. (Hrsg.). - New York : ACM, 2013. - S. 1119-1120. - ISBN 978-1-4503-2034-4

Demonstration of Citation Pattern Analysis for Plagiarism Detection

Bela Gipp[1,2], Norman Meuschke[1], Corinna Breitinger[1], Mario Lipinski[1], Andreas Nürnberger[2]

[1] Department of Statistics
University of California, Berkeley

[2] Department of Computer Science
Otto-von-Guericke-University Magdeburg

{gipp, meuschke, lipinski, breitinger}@berkeley.edu, andreas.nuernberger@ovgu.de

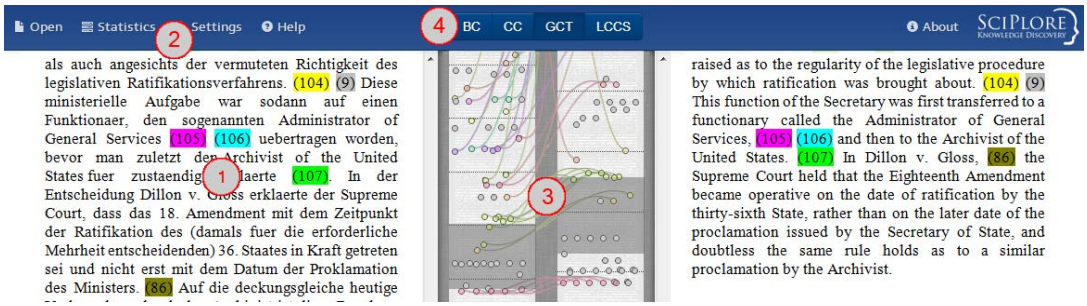

Figure 1: Screenshot of CbPD prototype (left plagiarized translation, right source document)

Categories and Subject Descriptors

H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval – Retrieval models.

General Terms

Algorithms, Experimentation

Keywords

Plagiarism Detection, Plagiarism Detection Systems, Citation-based Plagiarism Detection, Cross Language, Text Reuse

Limitations of Plagiarism Detection Systems

State-of-the-art plagiarism detection approaches capably identify copy & paste and to some extent slightly modified plagiarism. However, they cannot reliably identify strongly disguised plagiarism forms, including paraphrases, translated plagiarism, and idea plagiarism, which are forms of plagiarism more commonly found in scientific texts. This weakness of current systems results in a large fraction of today's scientific plagiarism going undetected.

Prof. Weber-Wulff, the organizer of a regular performance evaluation of Plagiarism Detection Systems (PDS), gives a disillusioning summary regarding available systems:

> *"[…] Plagiarism Detection Systems find copies, not plagiarism."*[5],
> and *"[…] for translations or heavily edited material, the systems are powerless […]"* [3].

While undisguised copy & paste-type plagiarism is software-recognizable and typically occurs in student essays where it has no serious consequences for society, a different story holds for disguised scientific plagiarism, which is currently non-recognizable. Plagiarized medical and pharmaceutical studies can jeopardize the safe treatment of patients when not based on original research.

Citation-based Plagiarism Detection

To address this problem, we proposed Citation-based Plagiarism Detection (CbPD) [2]. Compared to existing approaches, CbPD does not consider textual similarity alone, but uses the citation patterns within scientific documents as a unique, language-independent fingerprint to identify semantic similarity. Evaluations of real-world plagiarism cases have shown that plagiarists commonly disguise academic misconduct by paraphrasing copied text, but usually do not substitute or rearrange the citations copied from the source document. Motivated by these findings, we developed several CbPD algorithms; each tailored to a specific form of disguised plagiarism. We implemented the algorithms in a first citation-based plagiarism detection prototype capable of detecting strongly disguised plagiarism, even when no textual similarity remains.

We first demonstrated the advantages of the developed CbPD approach in evaluations using the plagiarized thesis of former German defense minister Karl-Theodor zu Guttenberg. While conventional detection approaches could not identify a single instance of translated plagiarism in the thesis, the novel approach detected 13 of the 16 translated plagiarisms present [2]. We used the bioscience full-text collection PubMed Central Open Access Subset (PMC OAS), which includes 200,000 publications, to demonstrate the practicability of the CbPD approach on a large corpus [4]. We machine parsed the citations (in text) and references (in bibliography) of documents in the PMC OAS and computed the citation-based similarities between all possible document pairs using each of the CbPD algorithms. The result is a database, which contains about 7 million references, 11 million citations and 750 million citation patterns. The database is the backbone of the presented prototype, which is available under www.citeplag.org. Besides the 200,000 publications from the PMC OAS, the prototype also allows inspection of the plagiarized thesis of former German defense minister Karl-Theodor zu Guttenberg.

The screenshot in Figure 1: Screenshot of CbPD prototype (left plagiarized translation, right source document) shows an excerpt from this thesis on the left and the corresponding text from the English source document on the right. The source document is an interpretation of the U.S. Constitution, which is translated nearly word-for-word, creating a plagiarism spanning 21 pages. Despite its length, no textual similarity is retained in the translated plagiarism, aside from four correctly cited longer quotes. Only upon visualizing citation patterns using CbPD does the disguised plagiarism become obvious (Figure 1).

Algorithms

As a raw measure of global document similarity, we include the Bibliographic Coupling (BC) score, a simple measure of the absolute number of shared references in the bibliographies of two documents, in our assessment. However, since BC does not consider the placement of citations in a document's full-text, we developed three "fine-tuned" algorithms which make up the core of the CbPD approach: Citation Chunking (CC), Greedy Citation Tiling (GCT) and Longest Common Citation Sequence (LCCS). Details on these detection algorithms can be found in this publication [1].

The *Citation Chunking* set of algorithms identify individual matching citation patterns, termed "citation chunks". The Citation Chunking algorithms are resilient to slight transpositions of citation order and inserted or deleted citations.

The *Greedy Citation Tiling* algorithms identify sets of matching citations in identical order, which we call "citation tiles". Longest citation pattern matches are identified first, subsequently the remaining shorter citation pattern matches are considered for citation tile formation.

The *Longest Common Citation Sequence* represents the maximum number of citations two documents share in identical order, when skipping over the non-matching citations. A document pair has either exactly one LCCS or no shared citations.

Prototype

The *CitePlag* prototype features a customizable side-by-side document visualization, see marker 1 Figure 1, to efficiently browse scientific documents for text and citation similarities and aid the user in identifying plagiarism. In this case, the suspicious document is displayed on the left and the source document is displayed on the right. When selecting the highlighted text or citation similarity in either document, the respective section in the other document is retrieved. The visualization of text and citation similarities is customizable to the user's preferences in the menu bar under 'settings', see marker 2 Figure 1.

A scrollable central document browser, see marker 3 Figure 1, schematically represents the documents to be compared and allows for quick interactive document navigation. The document browser visualizes citation-based similarities according the CbPD algorithm selected, see marker 4 Figure 1.

By highlighting matching citations in identical colors and connecting them in the document browser, see Figure 2, the CbPD approach visualizes not only the easy to spot global or copy & paste plagiarism instances, but also local and heavily disguised plagiarism instances, which remain invisible to conventional PDS.

The menu bar features document 'statistics', including graphs showing the sum of character-matches and bibliographic coupling strength per page. In a collapsible side tab, additional documents with high CbPD similarity scores are listed for comparison. The side tab also allows the user to set weighting coefficients for each of the CbPD algorithms, thus creating a hybrid CbPD algorithm with custom emphasis. The source-code is published as open-source under a creative commons license 3.0.

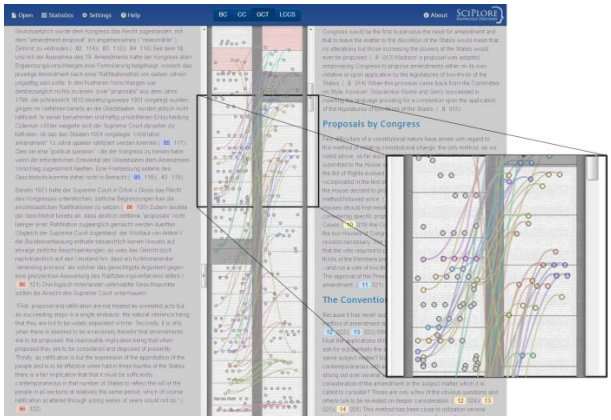

Figure 2: Citation patterns visualized

Using the CbPD approach allowed the identification of previously non-machine-detectable plagiarism cases. As a result, several publications, including a fraudulent medical study have already been retracted. The uncovered plagiarism can be browsed using the prototype. Nevertheless, the CbPD approach should be viewed as a supplement to existing software-based plagiarism detection methods and not a stand-alone replacement, since the individual strengths of the text-based and citation-based approaches to plagiarism detection complement each other. The strength of text-based approaches lies in their ability to detect even short instances of literal plagiarism, e.g. copy & paste. The citation-based approach excels when little or no textual similarity is present, as in the case of translated and heavily disguised plagiarism.

Acknowledgements

We wish to acknowledge André Gernandt, Leif Timm, Markus Bruns, Markus Föllmer and Rebecca Böttche for their contributions.

References

[1] Bela Gipp and Norman Meuschke. 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*, pages 249–258, Mountain View, CA, USA, 2011. ACM.

[2] Bela Gipp, Norman Meuschke, and Joeran Beel. Comparative Evaluation of Text- and Citation-based Plagiarism Detection Approaches using GuttenPlag. In *Proceedings of 11th ACM/IEEE-CS Joint Conference on Digital Libraries (JCDL'11)*, pages 255–258, Ottawa, Canada, 2011.

[3] HTW Berlin. Portal Plagiat - Softwaretest 2012. Online Source, 2012. Retrieved November 27, 2012 from: http://-plagiat.htw-berlin.de/collusion-test-2012/.

[4] U.S. National Center for Biotechnology Information. PubMed Central. Online Source, 2011. Retrieved September 27, 2011 from: http://www.ncbi.nlm.nih.gov/pmc/.

[5] Debora Weber-Wulff. Test Cases for Plagiarism Detection Software. In *Proceedings of the 4th International Plagiarism Conference*, Newcastle upon Tyne, UK, 2010.