Web-based Demonstration of Semantic Similarity Detection Using Citation Pattern Visualization for a Cross Language Plagiarism Case

Bela Gipp[1,2], Norman Meuschke[1,2] Corinna Breitinger[1], Jim Pitman[1] and Andreas Nürnberger[2]
[1]*University of California Berkeley, Berkeley, U.S.A.*
[2]*Otto-von-Guericke University, Magdeburg, Germany*

Keywords: Plagiarism Detection, Citation Analysis, Similarity Visualization.

Abstract: In a previous paper, we showed that analyzing citation patterns in the well-known plagiarized thesis by K. T. zu Guttenberg clearly outperformed current detection methods in identifying cross-language plagiarism. However, the experiment was a proof of concept and we did not provide a prototype. This paper presents a fully functional, web-based visualization of citation patterns for this verified cross-language plagiarism case, allowing the user to interactively experience the benefits of citation pattern analysis for plagiarism detection. Using examples from the Guttenberg plagiarism case, we demonstrate that the citation pattern visualization reduces the required examiner effort to verify the extent of plagiarism.

1 INTRODUCTION

Detecting academic plagiarism is an important task that remains tedious, especially for disguised plagiarism forms, such as paraphrases or cross-language plagiarism, because these forms exhibit little or no literal text overlap. Plagiarism detection systems (PDS) facilitate the identification of plagiarism, yet their detection and visualization capabilities are limited.

Today's PDS are typically web-based and allow users to upload documents for which the system retrieves suspiciously similar documents from a large reference collection, which often includes a subset of the Web (Stein et al. 2011). PDS in practical use rely exclusively on literal text string comparisons. The systems examine the percentage of lexical overlap among documents and treat overlap above a pre-defined threshold as an indicator for potential plagiarism. Subsequent to retrieving similar sources, systems rank the sources by "similarity score" and highlight the sections with the highest lexical similarity for user inspection. Due to their literal detection approach, current PDS capably identify copies, but fail to detect disguised plagiarism, including paraphrases or cross-language plagiarism (Weber-Wulff 2012).

2 RELATED WORK

In previous work, we introduced Citation-based Plagiarism Detection (CbPD) and showed that this approach can significantly increase the detection rate for disguised plagiarism (Gipp et al. 2011). CbPD examines order, proximity, and distinctiveness of in-text citations in academic literature to identify citation patterns that can serve as a language-independent similarity characteristic (Gipp and Meuschke 2011). We examined the prominent plagiarism case of Germany's former Minister of Defense, Karl Theodor zu Guttenberg, as a preliminary assessment of CbPD's ability to detect disguised plagiarism (Gipp et al. 2011).

Guttenberg's thesis is one of the most thoroughly investigated plagiarism cases to date. Volunteers of the crowd-sourced GuttenPlag Wiki project (GuttenPlag Wiki 2011) manually examined the thesis and revealed approx. 64 % of all lines of text to be plagiarized.

The barcode representation of the thesis in Figure 1 illustrates this finding.

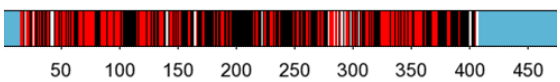

Figure 1: Plagiarized pages in Guttenberg's thesis (GuttenPlag Wiki 2011).

Gipp B., Meuschke N., Breitinger C., Pitman J. and Nürnberger A..
Web-based Demonstration of Semantic Similarity Detection Using Citation Pattern Visualization for a Cross Language Plagiarism Case.
DOI: 10.5220/0004985406770683
In *Proceedings of the 16th International Conference on Enterprise Information Systems* (ISS-2014), pages 677-683
ISBN: 978-989-758-028-4

Red bars in Figure 1 represent pages with plagiarism from multiple sources, while black bars indicate pages with plagiarism from a single source. White bars represent pages on which no plagiarism was found and blue sections represent the table of contents and bibliography.

We applied the CbPD algorithms to the verified cases of cross-language plagiarism in Guttenberg's thesis and compared the performance of our algorithms to state-of-the-art PDS. The CbPD algorithms identified 13 of the 16 instances of cross-language plagiarism in the thesis, while the other tested PDS found none (Gipp et al. 2011). This initial examination indicated that citation patterns are language independent and capable of identifying suspiciously similar documents for disguised plagiarism instances, which are undetectable by today's PDS.

However, it remained an open question to what degree a citation-based PDS could provide meaningful visual cues that allow a user to recognize the detected suspicious similarities. To answer this question, this paper uses the prominent cross-language plagiarism case of K.T. zu Guttenberg to demonstrate how an interactive, web-based visualization of citation patterns can facilitate the analysis of suspicious non-lexical similarities.

3 WEB-CLIENT

We visualize the longest instance of cross-language plagiarism in Guttenberg's thesis using CitePlag, a prototype of a PDS that integrates citation pattern analysis with traditional character-based detection methods (Gipp et al. 2013). The rationale is to offer an intuitive, highly interactive side-by-side document comparison, which the user can enhance with customizable highlights of citation-based and character-based similarity information.

Figure 2 shows the system architecture of CitePlag. The prototype's backend consists of a MySQL database and Java-based components for data disambiguation, document parsing, similarity detection and generating the output document format visualized in the fronted.

The document parser extracts metadata, citations, and references from the input documents and stores the data in the database. The data disambiguation component uses heuristics to consolidate all data stored in the database, e.g. to match references and to augment incomplete reference records with data from matching records with additional information.

The detector implements the citation-based and character-based detection algorithms. The detector reads the necessary data from the database to run the different algorithms and writes the identified citation and text patterns back to the database. The component for output conversion retrieves all information visualized in the fronted from the database and generates XML-based output documents.

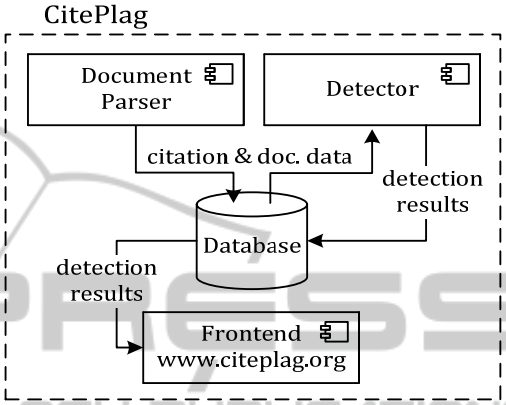

Figure 2: System architecture of the CitePlag prototype.

The frontend visualization, which is the focus of this paper, is implemented in HTML 5. Figure 3 shows two screenshots of the CitePlag user interface each displaying the Guttenberg thesis in German on the left and the retrieved source document, an English analysis of the U.S. constitution by the Congressional Research Service, on the right. The texts are individually scrollable. Footnotes in the documents are embedded in the full-text display and can be expanded or collapsed when clicking on a footnote marker.

In the left screenshot in Figure 3, citation pattern visualization is deactivated and only matching text is highlighted in identical colors. In this mode, the CitePlag user interface resembles the interfaces of most PDS in use today. In Figure 3, the only lexical match visible happens to be a legitimate quote. Identifying quotes as potential plagiarism is a common cause of false positives for plagiarism detection systems (Stein et al. 2007).

In the right screenshot in Figure 3, the interactive visualization of citation patterns is activated. In this mode, identical citation patterns are highlighted in addition to matching text in the full-text displays. Hovering over citations in the text opens a pop-up displaying metadata of the cited source. Additionally, a scrollable central document browser schematically displays the documents, while highlighting and connecting the matching patterns.

Lexical similarity among sections is shaded in

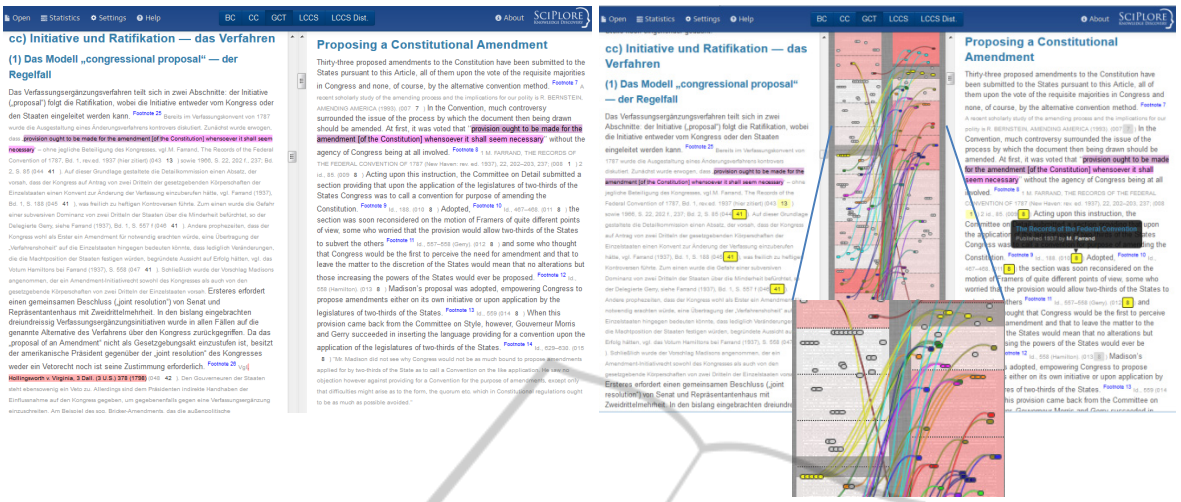

Figure 3: Interactive web-based citation pattern visualization for cross-language plagiarism (source: http://citeplag.org).

different intensities of red depending on severity. Clicking on citations or text highlights in either the full-text or the central browser aligns the matches, thus enabling a quick inspection of similarities.

The user can select among citation pattern visualization algorithms above the document browser.

4 DEMONSTRATION OF ALGORITHMS

This section presents the visualization algorithms implemented in the CitePlag prototype and demonstrates their benefits using examples from the excerpt of the Guttenberg cross-language plagiarism case. We published a detailed description of the algorithms in (Gipp and Meuschke 2011).

The reader is invited to visit the prototype at http://citeplag.org/ to interactively explore the Guttenberg plagiarism case and other plagiarism cases. CitePlag's source code is freely available under a MIT License.

4.1 Bibliographic Coupling

Bibliographic Coupling (BC) is a traditional and widespread citation-based similarity measure that denotes the number of identical bibliographic references in two academic documents (Fano 1956). A Bibliographic Coupling relationship is denoted in terms of a single value, the so called Bibliographic Coupling strength. Bibliographic Coupling strength is a raw measure of global document similarity, since it considers only the reference lists found in academic texts, but does not take into account the position or order of the citations within the full texts of the documents.

To visualize Bibliographic Coupling relations in CitePlag, we extended the original approach, which solely considers matching entries in the reference lists, to take into account matching citations in the full-texts instead. CitePlag highlights all citations of a source that both documents reference in the same color and connects each matching citation in the first document with every matching citation in the second document in the central document browser.

The leftmost image in Figure 4 visualizes the citation patterns formed by the Bibliographic Coupling approach for the longest instance of cross-language plagiarism in the Guttenberg thesis. In this particular case, the BC visualization results in a network of lines that are very dense and can thus become hard to trace.

The algorithm immediately shows the high topical similarity even if the lexical similarity between both documents is low, e.g. due to cross-language plagiarism as in this case, or in the case of paraphrases. Although Bibliographic Coupling is a very basic visualization method, the many parallel lines show that structural similarity is given throughout the entire excerpt, which is untypical for original work.

However, structurally similar documents with a high number of matching citations must not necessarily be plagiarisms. Dense BC overlaps can also be the case for highly related publications, such as literature reviews on the same topic. In the case of a chronological literature review, structural similarity may be acceptable. Therefore, in all cases a careful human examination is necessary.

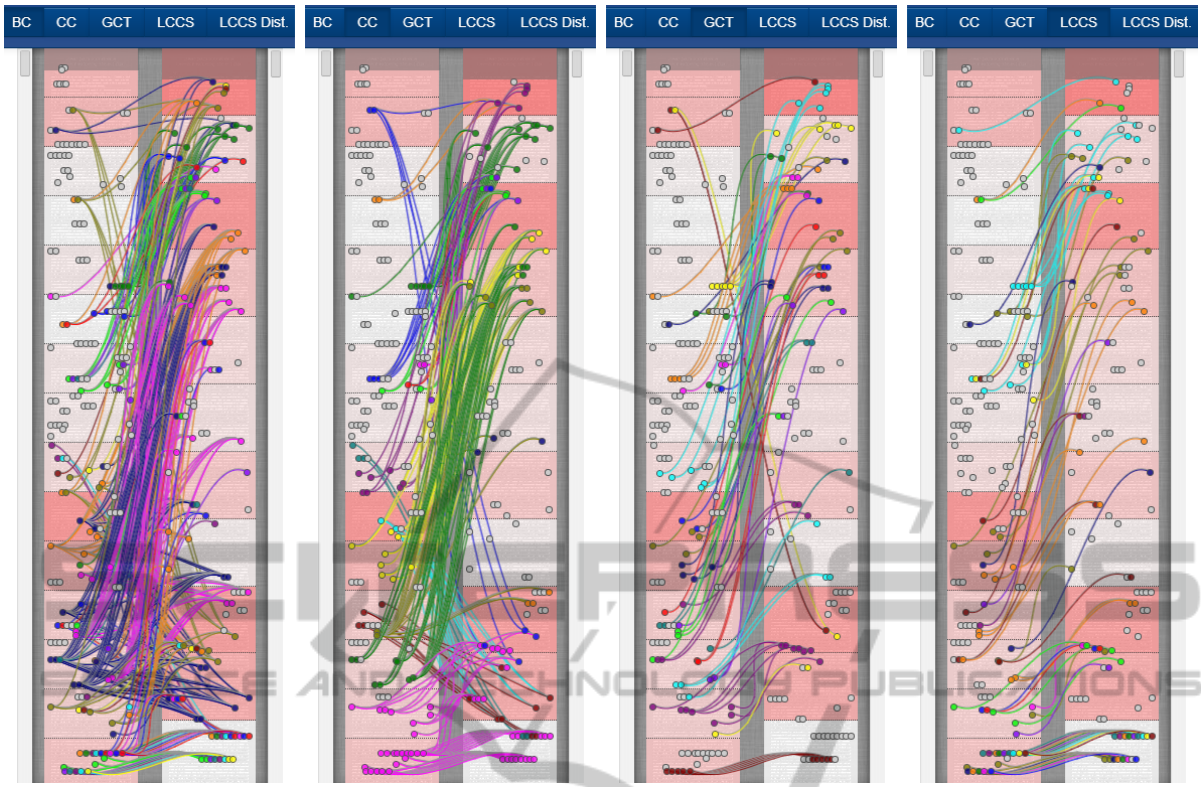

Figure 4: Overview of citation pattern visualization algorithms using the plagiarism detection prototype CitePlag.

4.2 Citation Chunking

Citation Chunking (CC) describes a set of heuristic algorithms that aim to identify matching citation patterns regardless of whether the order of matching citations differs in both documents.

We derived strategies to form citation chunks by observing behaviors of plagiarists and modelling the resulting citation patterns. Chunking means that matching citations are grouped and considered as a single unit (a chunk). The chunking strategy implemented in the CitePlag prototype chunks both documents. A matching citation is included in a chunk if the number of non-matching citations that separate the matching citation from the preceding matching citation is not larger than the number of matching citations already included in the chunk that is currently under construction (Gipp 2013).

Once chunks have been formed for both documents, the order of citations within a chunk is disregarded and each chunk of the first document is compared with each chunk of the second document. The chunk pairs with the highest number of matching citations are permanently related to each other and considered a match. If multiple chunks in the documents share the same number of matching citations, all combinations of chunks with equally many matching citations are stored.

Figure 5 demonstrates the conceptual formation and comparison of chunks for two documents A and B. Numbers in Figure 5 represent matching citations, i.e. citations that occur in both documents. The letter x denotes non matching citations in the two documents.

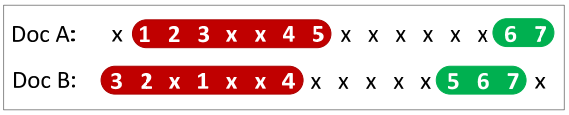

Figure 5: Citation Chunking schematic concept.

The chunking algorithm starts by identifying all citations that occur in both documents regardless of their positions and number of occurrences in the documents. With regard to Figure 5, one could say that the algorithm distinguishes the numbered citations from the remaining citations marked as x.

Next, the algorithm chunks the citation sequence of document A and subsequently the citation sequence of document B according to the rules

explained. Note that only the order of matching and non-matching citations in the document that is chunked determines which chunks are formed for that document. The citation sequence of the other document has no influence on the chunk formation.

For document A in Figure 5, the algorithm starts forming the first chunk shown in red by including the matching citations #1, #2 and #3, because no non-matching citations separate those matching citations. The first chunk contains three matching citations at this point. Therefore, the algorithm will include the next matching citation in the sequence if three or less non-matching citations separate it from the last matching citation in the chunk (#3). Citation #4 fulfills this condition, thus is included in the first chunk, as is citation #5, which directly succeeds #4. At this point, the first chunk contains five matching citations. The next matching citation in the sequence (#6) is separated from the last matching citation in the first chunk (#5) by six non-matching citations. In other words, the number of non-matching citations in between the matching citations #5 and #6 is larger than the number of matching citations in the first chunk. Therefore, the algorithm finalizes the first chunk and includes citation #6 and #7 in a second chunk shown in green thereby completing the processing of document A. By processing the citation sequence of document B in the same manner, the algorithm forms two chunks for document B, although the order of matching citations in document A and B differs.

In the following comparison step, the Citation Chunking algorithm compares all chunks formed for both documents with each other. The chunks with the highest overlap in matching citations are stored as a citation pattern match. For the example shown in Figure 5, the algorithm stores a pattern match of length four between the red chunks and a pattern match of length two for the green chunks.

Citation Chunking aims to uncover potential cases in which text segments or logical structures have been copied or were influenced by another text. The chunking strategy implemented in the CitePlag prototype allows for sporadic non-shared citations that may have been inserted to make the resulting text appear more “genuine”. By allowing an increasing number of non-shared citations within a chunk, given that a certain number of shared citations have already been included, the Citation Chunking algorithm can also detect potential plagiarism cases where text segments and citations from different sources were copied and interwoven (shake&paste plagiarism).

The second image from the left in Figure 4 shows the citation patterns identified in the instance of cross-language plagiarism from the Guttenberg thesis using the Citation Chunking algorithm. The substantial overlap in citations, which was already apparent by visualizing Bibliographic Coupling relations, is also reflected by the numerous and densely linked citation chunks. Visualizing the patterns returned by the Citation Chunking algorithm reveals a number of clusters pointing to similar text segments. Within individual clusters, lines connecting matching citations are mostly parallel with only few overlaps. The pattern suggests that the selection and placement of citations in numerous well defined segments of the Guttenberg thesis is highly similar to the source document.

4.3 Greedy Citation Tiling

The Greedy Citation Tiling (GCT) algorithm identifies all individually longest citation patterns that consist entirely of matching citations in the exact same order. Individually longest patterns refer to sequences of matching citations in the same order that cannot be extended to the left or right without encountering a citation that is not shared by both documents being compared. Such individual longest matches are called citation tiles.

Figure 6 illustrates the formation of Greedy Citation Tiles. Using the notation introduced in Figure 5, Arabic numerals represent matching citations, the letter x denotes non-matching citations. Colored highlights with roman numerals represent citation tiles. Citation tiles are stored as a numeric triplet shown at the bottom of Figure 6. The first element of the triplet indicates the starting position of the citation tile in document A, the second element denotes the starting position in document B and the third element corresponds to the length of the tile.

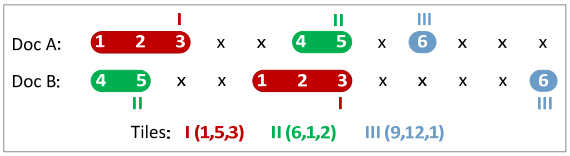

Figure 6: Greedy Citation Tiling schematic concept.

Finding many or long matching citation tiles is rarely a coincidence, and can thus be a strong indicator of plagiarism. In Figure 4, the third image from the left shows the visualization of citation tiles for the longest instance of cross language plagiarism in the Guttenberg thesis. Numerous citation tiles up to a length of five citations were identified in this

excerpt of the thesis. In many sections, even longer sequences of matching citations were only interrupted by a few non-matching citations, resulting in the identification of several shorter citation tiles.

The large number and length of citation tiles in this example is clearly suspicious. Despite the lack of lexical overlap, the long citation tiles found should quickly point an examiner to the text segments that require closer inspection. By examining these text segments, an investigator will quickly identify the instances where content was translated from the source document. Within the translated segments, the citations placed at the end of many sentences were simply copied along with the translated version of that sentence. Citations distributed within a sentence were partially transposed due to different sentence structure in the translation.

4.4 Longest Common Citation Sequence

The Longest Common Citation Sequence (LCCS) algorithm visualizes the longest string of citations that match in both documents in identical order. Transpositions of matching citations, i.e., citations that have been rearranged, are not detected, and any interruptions by non-matching citations are skipped and the string continued upon the next match. Thus, a document pair has either exactly one or no LCCS. Figure 7 illustrates a LCCS pattern using the notation established in Figure 5 and Figure 6.

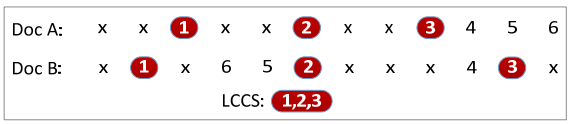

Figure 7: Longest Common Citation Sequence schematic concept.

The LCCS measure is very indicative of originality, because long strings of identical citations, even with interruptions, reflect that both authors presented the content in a similar order. Such parallel structure is often not a coincidence.

In Figure 4, the rightmost image shows the LCCS pattern of the examined cross-language plagiarism excerpt of the Guttenberg thesis. The LCCS pattern reflects the extent of global structural similarity between the original and the plagiarism. The strong parallel structure of matching citations with few interruptions suggests a significant similarity between numerous sections of both documents, thereby amplifying the suspicion raised by the Citation Chunking and GCT patterns. Since almost no literal text overlaps exist between the two documents, the examiner can deduce that the analyzed excerpt is in large parts a directly translated plagiarism.

Although the LCCS algorithm correctly points to the suspicious global document similarity in the Guttenberg case, the algorithm can return misleading results if literature is cited in chronological or alphabetical order. While such similarities are reflective of topical similarity, they are likely genuine.

4.5 Longest Common Sequences of Distinct Citations

The Longest Common Sequence of Distinct Citations (LCCS Dist.) is a more restrictive variant of the LCCS algorithm presented in the previous section. LCCS Dist. includes only the first occurrence of a matching citation and ignores repeated citations to the same source regardless of whether they occur in the same order in both documents. As such, the measure is simply more restrictive than the LCCS measure. Therefore, the outcome for LCCS Dist. is not pictured.

5 CONCLUSION AND FUTURE WORK

As demonstrated using an instance of cross-language plagiarism in the thesis of K. T. zu Guttenberg, citation-pattern visualization is especially beneficial for examining potential plagiarism that features little or no lexical overlap. By visualizing citation patterns in addition to lexical text matches, CitePlag enables a faster inspection, especially for the harder to detect heavily disguised plagiarism forms.

Future systems may benefit from interactive side-by-side visualizations of *both* lexical and non-lexical similarities, as demonstrated by the CitePlag prototype. This paper examined only the visual representation of citation patterns for a certain plagiarism case. A formal evaluation of the citation-based approach is currently awaiting publication in (Gipp et al. 2014) and is scheduled to appear in the course of the next months.

Lastly, it remains worthy to mention that no plagiarism detection system alone can fully automate the identification of plagiarism. The verification of the detected similarities and the final

decision on plagiarism remains with the user. The necessity of humans in the plagiarism detection process demands that the visualizations employed by plagiarism detection systems are interactive, customizable, and intuitive to users if the system is to provide optimal utility.

ACKNOWLEDGEMENTS

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

REFERENCES

FANO, R. M. 1956. *Documentation in Action*. Reinhold Publ. Co., New York, Chapter Information Theory and the Retrieval of Recorded Information, 238–244.

GIPP, B. 2013. Citation-based Plagiarism Detection: Applying Citation Pattern Analysis to Identify Currently Non-Machine-Detectable Disguised Plagiarism in Scientific Publications. Ph.D. thesis, Department of Computer Science, Otto-von-Guericke University Magdeburg, Germany.

GIPP, B. AND MEUSCHKE, N. 2011. 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*. ACM, Mountain View, CA, USA, 249–258.

GIPP, B., MEUSCHKE, N., AND BEEL, J. 2011. 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)*. ACM, Ottawa, Canada, 255–258.

GIPP, B., MEUSCHKE, N., AND BREITINGER, C. 2014. Citation-based Plagiarism Detection: Practicability on a Large-scale Scientific Corpus. *Journal of the American Society for Information Science and Technology (to appear)*.

GIPP, B., MEUSCHKE, N., BREITINGER, C., LIPINSKI, M., AND NÜRNBERGER, A. 2013. Demonstration of the First Citation-based Plagiarism Detection Prototype. In *Proceedings of the 36th International ACM SIGIR conference on research and development in Information Retrieval*. ACM, Dublin, Ireland, 1119–1120.

GUTTENPLAG WIKI. 2011. Eine kritische Auseinandersetzung mit der Dissertation von Karl-Theodor Freiherr zu Guttenberg: Verfassung und Verfassungsvertrag. Konstitutionelle Entwicklungsstufen in den USA und der EU. Online Source. Retrieved Apr. 25, 2012 from: http://de.guttenplag.wikia.com/wiki/GuttenPlag_Wiki.

STEIN, B., LIPKA, N., AND PRETTENHOFER, P. 2011. Intrinsic Plagiarism Analysis. *Language Resources and Evaluation 45,* 1, 63–82.

STEIN, B., MEYER ZU EISSEN, S., AND POTTHAST, M. 2007. Strategies for Retrieving Plagiarized Documents. In *Proceedings of the 30th Annual International ACM SIGIR Conference*. ACM, 825–826.

WEBER-WULFF, D. 2012. Portal Plagiat - Softwaretest Report 2012. Online Source. Retrieved Nov. 27, 2012 from: http://plagiat.htw-berlin.de/collusion-test-2012/.