Comparative Evaluation of Text- and Citation-based Plagiarism Detection Approaches using GuttenPlag

Bela Gipp
OvGU, Germany & UC Berkeley
gipp@berkeley.edu

Norman Meuschke
OvGU, Germany & UC Berkeley
nmeuschk@st.ovgu.de

Joeran Beel
OvGU, Germany & UC Berkeley
j@beel.org

ABSTRACT
Various approaches for plagiarism detection exist. All are based on more or less sophisticated text analysis methods such as string matching, fingerprinting or style comparison. In this paper a new approach called Citation-based Plagiarism Detection is evaluated using a doctoral thesis [8], in which a volunteer crowd-sourcing project called GuttenPlag [1] identified substantial amounts of plagiarism through careful manual inspection. This new approach is able to identify similar and plagiarized documents based on the citations used in the text. It is shown that citation-based plagiarism detection performs significantly better than text-based procedures in identifying strong paraphrasing, translation and some idea plagiarism. Detection rates can be improved by combining citation-based with text-based plagiarism detection.

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

General Terms
Algorithms, Experimentation, Measurement, Languages

Keywords
Plagiarism Detection Systems, Citation-based Plagiarism Detection

1. INTRODUCTION

Plagiarism describes the appropriation of other people's ideas, intellectual or creative work and passing them off as one's own [4]. It is a particularly common problem among college students, but also prevalent among established researchers [19].

Recently, a plagiarism case involving the German minister of defense, Mr. Guttenberg, gained widespread public attention. By chance, a law professor detected plagiarized sections within Mr. Guttenberg's doctoral thesis [8]. After the popular politician repudiated the accusations as "abstruse", volunteers initiated the GuttenPlag project [1] to crowd-source the investigation and determine the true amount of plagiarism present in the work.

As of April 10th 2011, the joint efforts revealed that 371 out of 393 main text pages in the thesis contained plagiarized fragments. In total about 64 % of the text lines were identified as plagiarized. The following barcode illustrates the findings.

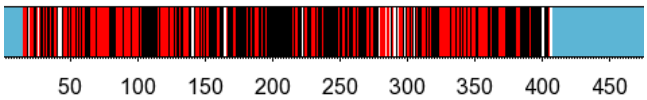

Figure 1: Pages containing Plagiarism (source: [1])

Black lines represent pages containing plagiarism from one source, red lines pages with plagiarism from multiple sources and white lines pages on which no plagiarism was found. The blue sections represent the table of contents and bibliography.

As a result of this analysis, Mr. Guttenberg no longer claimed his thesis to be flawless. He renounced his doctorate and eventually stepped down from his political position.

We applied the citation-based detection approach to Mr. Guttenberg's thesis in order to demonstrate its potential for identifying cleverly disguised plagiarism, especially in cases of otherwise hard to detect translated plagiarisms. After giving an overview of different forms of plagiarism and the currently used detection methods, the concept of Citation-based Plagiarism Detection (CbPD) is elucidated. Afterwards, the methodology of the evaluation is presented. In the final section, the advantages of combining text-based and citation-based methods are discussed.

2. RELATED WORK

2.1 Forms of Plagiarism

Observations of plagiarism behavior in practice reveal a number of commonly found methods for illegitimate text usage, which can briefly be summarized as follows. *Copy&Paste (*c&p*) plagiarism* specifies the act of taking over parts or the entirety of a text verbatim from another author. *Disguised plagiarism* includes practices intended to mask literally copied segments. *Undue paraphrasing* defines the intentional rewriting of foreign thoughts, in the vocabulary and style of the plagiarist without giving due credit in order to conceal the original source [3]. *Translated plagiarism* is the manual or automated conversion of content from one language to another intended to cover its origin. *Idea plagiarism* encompasses the usage of a broader foreign concept without appropriate source acknowledgement. An Example is the appropriation of research approaches, methods, experimental setups, argumentative structures, background sources etc. [5].

2.2 Plagiarism Detection Approaches

Plagiarism Detection (PD) is a hypernym for computer-based procedures supporting the identification of plagiarism incidences. Existing PD systems (PDS) can be categorized into external and intrinsic. *External PDS* compare a suspicious document to a corpus of genuine works. *Intrinsic PDS* statistically examine linguistic features of the suspicious text, a process known as stylometry, without performing comparisons to external documents. While external PDS aim to find literally matching text segments, intrinsic PDS try to recognize changes in writing style [17].

Different comparison strategies have been proposed for external PDS. The most common ones are briefly explained. *Substring matching* procedures aim to identify long pairs of identical strings. Such strings are treated as indicators for potential plagiarism if

their share with regard to the entire text exceeds a chosen threshold. Most commonly suffix document models, such as suffix trees or arrays, have been used for that purpose [13].

Fingerprinting methods, being the most widely used PD approach, aim at forming a representative digest of a document by selecting a set of multiple substrings from it. The set represents the fingerprint; its elements are called minutiae. Mathematical, hash-like functions can be applied on minutiae for transforming them into more space efficient byte strings [9].

More than 1.000 individual style markers have been proposed for usage in stylometry [16]. They range from lexical features, e.g. average word length, to syntactic features, e.g. part-of-speech frequencies, to structural features, e.g. frequency of punctuation. Intrinsic PD systems mostly comprise an individual combination of multiple linguistic features [18].

Citation-based Plagiarism Detection (CbPD) is a fundamentally different approach compared to text-based similarity evaluations. It is especially suitable for scientific publications, since it requires references. In a previous paper [7] we initially proposed employing citation analysis for PD and evaluated its performance using an artificially created dataset.

2.3 Strength and weaknesses of PD Systems

Objective, comparative assessments of the detection performance of PD systems are difficult, since the used collections and evaluation methods differ widely. Two projects address this lack of comparability. Both attempt to benchmark PDS using standardized collections and controlled evaluation environments. The annual PAN International Competition on Plagiarism Detection (PAN-PC) was initiated in 2009, in which competitors present primarily research prototypes [14]. A periodic comparison of productive PDS is performed by a research group at the University of Applied Science Berlin (HTW) since 2004 [10].

The PAN-PC evaluation corpus mainly contains artificially plagiarized sections that were created and partially obfuscated through automated methods such as translation, random shuffles, or semantic substitutions of terms. In addition, 4000 text segments that were manually obfuscated by humans instructed to simulate a plagiarist's behavior are included [15]. In the HTW evaluations a corpus of 42 documents being manually plagiarized or original essays of approx. 1 to 1.5 pages of length is used. The original sources are known and mostly available on the internet [10, 20].

Some results of the two competitions are presented to outline the characteristic strengths and weaknesses of existing PDS.

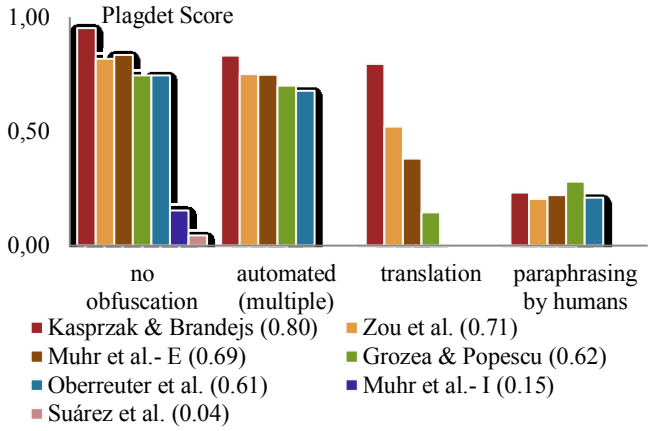

Figure 2: Results of top 5 performing PDS in PAN-PC'10 [14]

Figure 2 displays the plagiarism detection (*plagdet*) scores for the top 5 performing external PDS and the 2 intrinsic PDS of MUHR ET AL. and SUÁREZ ET AL. participating in PAN-PC'10 (see [14] for further results and references regarding individual systems). The *plagdet* score was developed to evaluate systems participating in the PAN-PC (see [15] for details). The scores are plotted according to the obfuscation techniques applied to plagiarized text segments. The overall *plagdet* score for all categories is stated in brackets within each legend entry. In the legend to the figure "-I" is attached to distinguish the system of MUHR ET AL. participating in the intrinsic from the one in the external task.

The results indicate that c&p plagiarism can be detected with high accuracy by state-of-the-art PDS. However, detection rates for disguised plagiarized segments, especially those obfuscated by humans, are substantially lower for all systems. The organizers of the competition judged the results achieved in detecting cross-lingual plagiarism to be misleading. The well-performing systems used automated services for translating foreign-language documents in the reference corpus. Those services were similar or identical to those used for constructing the plagiarized sections. It is hypothesized that the human-made translations obfuscating real-world plagiarism are much more complex and versatile, and hence less detectable by the tested PDS [14].

The findings of the HTW comparisons are in line with those of the PAN-PC. Notably, none of the tested systems was able to identify cases of translated plagiarism [10]. That supports the assumption of unrealistic detection rates for translated segments in the PAN-PC due to the laboratory-like setup of the competition.

Furthermore, it is noteworthy that the performance of any external PDS depends heavily on the reference corpus available to the individual system. Therefore, it is not surprising that tools, which use the extensive indexes of internet search providers, often achieve the best detection results [12]. The same is true for manually performed queries of suspicious keywords and fragments.

2.4 CITATION-BASED PD

To our knowledge it has not been attempted, except for our previous study [7], to identify plagiarism by analyzing citations[1] and references. We propose the following definition:

Citation-based Plagiarism Detection (CbPD) subsumes methods that use citations and references for determining document similarities in order to identify plagiarism.

In the academic environment citations and references of scholarly publications have long been recognized for containing valuable semantic information about the content of a document and its relation to other works [6].

Finding similar patterns in the citations used within two scientific texts is a strong indicator for semantic text similarity. We refer to citation patterns as subsequences in the citation tuples C_A and C_B of two texts A and B that (partially) consist of shared references, and are therefore similar to each other.

The degree of similarity between citation patterns depends, among others factors, mainly on the amount of shared references

[1] Citations are short strings in the body of scientific texts representing sources contained in the bibliography whereas references denote entries in the bibliography.

(bibliographic coupling strength), and the extent to which the order of included citations, as well as their distance towards each other is similar. The idea is to calculate the probability of citation patterns to be the same by chance. For details on our CbPD algorithms, partly published as open-source, please consult [7], the documentation at http://sciplore.org and upcoming publications.

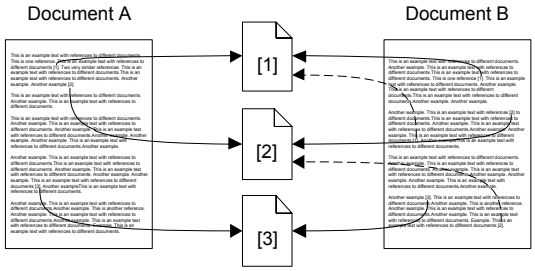

Figure 3: Identifying Citation Patterns for CbPD

3. METHODOLOGY

3.1 Test Corpus

Artificially created evaluation corpora, such as the ones of the PAN-PC do not include citation or reference information. Moreover, they lead to unrealistically high detection rates as machines are not as creative in paraphrasing and disguising plagiarism as humans (see 2.3).

Although the scientific value of Mr. Guttenberg's dissertation is questionable, we consider it as an ideal case study for our evaluation purposes because it:

- has been thoroughly investigated by hundreds of examiners;
- was created by a human author trying to disguise plagiarism;
- provides realistic citation information.

Due to these unique characteristics, the thesis allows for comparative evaluation of commonly applied plagiarism detection and the Citation-based Plagiarism Detection approach.

Previous evaluations (as presented in 2.3) indicate that translated plagiarism is especially hard to detect by conventional text-based PDS. Therefore, the focus of our investigation has been on whether CbPD is better suitable for detecting this form of plagiarism. At the time of our investigation the GuttenPlag project had identified plagiarized passages that represented appropriations of English sources translated to German on 31 pages within the thesis. Those 31 pages were analyzed for matching citations with their identified genuine sources.

3.2 Test PD Systems

To compare citation-based with traditional text-based detection we used three popular PDS. Ephorus, which usually scores among the top 3 PDS in the HTW comparisons [10], the freely available Ferret system [11], both systems use fingerprinting detection, and WCopyFind [2], a PDS that employs substring matching. Since the two latter mentioned systems depend on local availability of possible source documents, all digitally available sources identified by the GuttenPlag project were collected and used.

4. EVALUATION & RESULTS

Our results obtained for text-based PD confirm earlier findings of others presented in 2.3. Manually querying search engines, such as Google, yielded high detection rates with regard to copy&paste plagiarism. Depending on the invested time and selection of keywords, even paraphrased and translated plagiarism can be found.

The text-based PDS, especially Ferret and WCopyfind, which work with local document comparisons, deliver good results for identifying copy&paste plagiarism given that the sources are available, as in our case. The performance of Ephorus in this case study was a little surprising. Only 2 % of the text in the thesis was found to match the sources of plagiarism. Given the large fraction of (almost) verbatim plagiarism in the thesis, and the fact that 77 sources of plagiarized sections, which were identified by the GuttenPlag project, are available on the internet, opposed to 63 that are not [1], this result is disappointing. Not surprisingly, all systems failed to identify almost all stronger paraphrased sections and could not detect any translated plagiarism (see Table 1 for details). However, these figures should be treated with care. Since a real thesis was used, it is uncertain whether all plagiarized fragments are known. Therefore, the stated detection rates might be too high, especially for the very hard to detect idea plagiarism.

Figure 4 shows the citation patterns of all translated plagiarism fragments found by the GuttenPlag project.

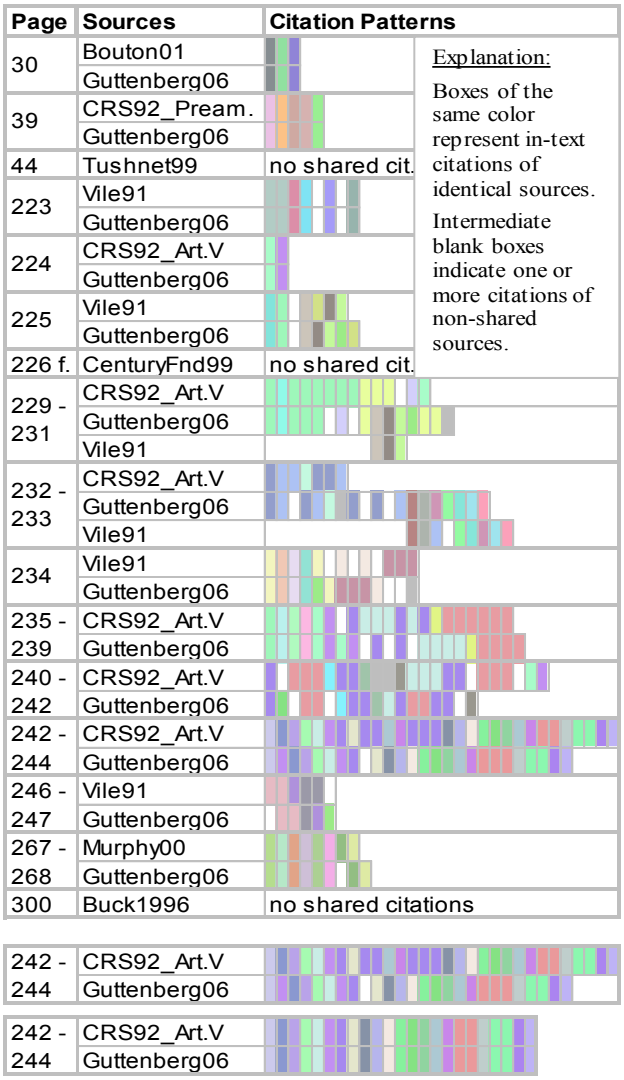

Figure 4: Citation Patterns for translated plagiarism

The figure illustrates that the citation patterns in genuine sources and in Mr. Guttenberg's translated plagiarism are often very similar. With exception of the pages 44, 226 and 300, all other pages share the same references in a similar order in the source document and Guttenberg's translation. This becomes especially obvious after cleaning the citation sequences by removing citations that are not shared by both documents at their corresponding positions. This is exemplified at the bottom of the figure for the pages 242-244.

Plagiarism type	Text-based	Citation-based
Copy&paste	~ 70 % Good results even for short fragments	Unsuitable as short fragments cannot be detected
Disguised plagiarism	< 10 %	Depending on the fragments length ~ 30 %
Idea / structure plagiarism	0 %	Some cases could be identified
Translated plagiarism	< 5 %	~ 80 %. 13 out of 16 fragments could be identified.

Table 1: Comparison of detection results

Whereas the currently used PDS were unable to detect a single translated fragment, the CbPD approach could identify all but three fragments. However, as with every PDS, the findings of CbPD must be carefully verified by humans, especially in cases where only a few citations form the pattern, for example in the fragments on page 30 and 224.

The evaluation indicates to a large extent that the strength of the existing PD systems are the weaknesses of the new citation based PD systems and vice versa. Whereas the strength of existing PDS lies in detecting plagiarism on the sentence level in the form of identifying similar or identical consecutive words, the strength of the citation based approach lies in identifying translation- and idea-plagiarism or disguised paraphrasing. However, since the CbPD relies on citation information, it is unable to identify short paraphrased fragments. By combining the strength of the text- and citation-based approaches the detection rate clearly outperforms currently used techniques.

5. CONCLUSION

It was shown that text-based plagiarism and Citation-based Plagiarism Detection methods have different strengths and weaknesses. Text-based PDS convince in detecting local forms of plagiarism, such as short passages of copied or only slightly paraphrased text. In contrast, they fail, to the here proposed citation-based approach, to detect paraphrased and translated plagiarism. Applying the citation-based approach on Guttenberg's thesis allowed identifying 13 of the 16 plagiarized fragments, whereas the text-based approaches did not identify a single fragment. By combining both approaches the detection rate can be significantly improved.

6. REFERENCES

[1] Guttenplag wiki. Online Resource, 2011. Retrieved Apr. 10, 2011 from http://de.guttenplag.wikia.com.

[2] BLOOMFIELD, L. A. The Plagiarism Resource Site. Online Resource, 2011. Retrieved Mar. 20, 2011 from http://-www.plagiarism.phys.virginia.edu.

[3] CLOUGH, P. Plagiarism in natural and programming languages an overview of current tools and technologies. Tech. rep., Department of Computer Science, University of Sheffield, July 2000.

[4] COCEL. *Concise Oxford Companion to the English Language*. Oxford Reference Online. Oxford University Press, 1998.

[5] FRÖHLICH, G. Plagiate und unethische Autorenschaften. *Information - Wissenschaft & Praxis 57*, 2 (2006), 81–89.

[6] GARFIELD, E. Citation Indexes for Science: A New Dimension in Documentation through Association of Ideas. *Science 122*, 3159 (July 1955), 108–111.

[7] GIPP, B., AND BEEL, J. Citation Based Plagiarism Detection - A New Approach to Identify Plagiarized Work Language Independently. In *Proceedings of the 21st ACM Conference on Hyptertext and Hypermedia (HT'10)* (New York, NY, USA, June 2010), ACM, pp. 273–274.

[8] GUTTENBERG, K.-T. F. *Verfassung und Verfassungsvertrag : Konstitutionelle Entwicklungsstufen in den USA und der EU*. Dissertation, Universität Bayreuth, Berlin, 2009, Retracted as plagiarism.

[9] HOAD, T. C., AND ZOBEL, J. Methods for Identifying Versioned and Plagiarised Documents. *Journal of the American Society for Information Science and Technology 54*, 3 (2003), 203–215.

[10] HOCHSCHULE FÜR TECHNIK UND WIRTSCHAFT BERLIN. Portal Plagiat - Test von Plagiatserkennungssoftware. Online Resource. Retrieved Apr. 08, 2011 from http://plagiat.htw-berlin.de/software/.

[11] LYON, C., MALCOLM, J., AND DICKERSON, B. Detecting Short Passages of Similar Text in Large Document Collections. In *Proceedings of the 2001 Conference on Empirical Methods in Natural Language Processing* (2001), L. Lee and D. Harman, Eds., pp. 118–125.

[12] MAURER, H., KAPPE, F., AND ZAKA, B. Plagiarism - A Survey. *Journal of Universal Computer Science 12*, 8 (Aug. 2006), 1050–1084.

[13] MONOSTORI, K., ZASLAVSKY, A., AND SCHMIDT, H. Document Overlap Detection System for Distributed Digital Libraries. In *DL '00: Proceedings of the fifth ACM conference on Digital libraries* (New York, NY, USA, 2000), ACM, pp. 226–227.

[14] POTTHAST, M., BARRÓN-CEDEÑO, A., EISELT, A., STEIN, B., AND ROSSO, P. Overview of the 2nd international competition on plagiarism detection. In *Notebook Papers of CLEF 2010 LABs and Workshops, 22-23 September, Padua, Italy* (Sept. 2010), M. Braschler, D. Harman, and E. Pianta, Eds.

[15] POTTHAST, M., STEIN, B., BARRÓN-CEDEÑO, A., AND ROSSO, P. An evaluation framework for plagiarism detection. In *Proceedings of the 23rd International Conference on Computational Linguistics (COLING 2010)* (Block A, Xue Yan Building, Tsinghua University, Beijing 100084, China, Aug. 2010), C.-R. Huang and D. Jurafsky, Eds., Tsinghua University Press.

[16] RUDMAN, J. The state of authorship attribution studies: Some problems and solutions. *Computers and the Humanities 31* (1997), 351–365.

[17] STEIN, B., KOPPEL, M., AND STAMATATOS, E., Eds. *Proceedings of the SIGIR 2007 International Workshop on Plagiarism Analysis, Authorship Identification, and Near Duplicate Detection, PAN 2007, Amsterdam, Netherlands, July 27, 2007* (2007), vol. 276 of *CEUR Workshop Proceedings*, CEUR-WS.org.

[18] STEIN, B., LIPKA, N., AND PRETTENHOFER, P. Intrinsic Plagiarism Analysis. *Language Resources and Evaluation* (2010), 1–20.

[19] SUN, Z., ERRAMI, M., LONG, T., RENARD, C., CHORADIA, N., AND GARNER, H. Systematic characterizations of text similarity in full text biomedical publications. *PLoS ONE 5*, 9 (Sept. 2010), e12704.

[20] WEBER-WULFF, D. Test cases for plagiarism detection software. In *Proceedings of the 4th International Plagiarism Conference* (Newcastle Upon Tyne, 2010).