Erschienen in: Journal of the Association for Information Science and Technology ; 65 (2014), 8. - S. 1527-1540

Citation-Based Plagiarism Detection: Practicability on a Large-Scale Scientific Corpus[1]

Bela Gipp
Department of Statistics, University of California, Berkeley, 493 Evans Hall, Berkeley, CA 94720.
E-mail: gipp@berkeley.edu

Norman Meuschke
Department of Statistics, University of California, Berkeley, 493 Evans Hall, Berkeley, CA 94720.
E-mail: meuschke@berkeley.edu

Corinna Breitinger
SciPlore Research Group, University of California, Berkeley, 493 Evans Hall, Berkeley, CA 94720.
E-mail: breitinger@berkeley.edu

The automated detection of plagiarism is an information retrieval task of increasing importance as the volume of readily accessible information on the web expands. A major shortcoming of current automated plagiarism detection approaches is their dependence on high character-based similarity. As a result, heavily disguised plagiarism forms, such as paraphrases, translated plagiarism, or structural and idea plagiarism, remain undetected. A recently proposed language-independent approach to plagiarism detection, Citation-based Plagiarism Detection (CbPD), allows the detection of semantic similarity even in the absence of text overlap by analyzing the citation placement in a document's full text to determine similarity. This article evaluates the performance of CbPD in detecting plagiarism with various degrees of disguise in a collection of 185,000 biomedical articles. We benchmark CbPD against two character-based detection approaches using a ground truth approximated in a user study. Our evaluation shows that the citation-based approach achieves superior ranking performance for heavily disguised plagiarism forms. Additionally, we demonstrate CbPD to be computationally more efficient than character-based approaches. Finally, upon combining the citation-based with the traditional character-based document similarity visualization methods in a hybrid detection prototype, we observe a reduction in the required user effort for document verification.

[1]This paper presents research results from the currently unpublished doctoral thesis of the primary author (Gipp, 2013).

Introduction

Automated plagiarism detection (PD) is a task supported by specialized information retrieval systems termed *plagiarism detection systems* (PDS). PDS employ one of two detection approaches, intrinsic or extrinsic. Today's commercially available PDS rely exclusively on the extrinsic approach, meaning that they consult an external collection, typically a subset of the web, against which to compare suspicious text. The retrieval task is then to return from this collection all documents that contain text passages similar above a chosen threshold to segments in the suspicious document (Stein, Lipka, & Prettenhofer, 2011).

Intrinsic detection approaches statistically examine the linguistic characteristics of a text without comparisons with an external collection and have been explored less frequently (Meyer zu Eissen, Stein, & Kulig, 2007; Stein et al., 2011). Intrinsic approaches have not been commercially adopted, mainly because of the obstacles posed by the minimum required document length and the possibility of legitimate style differences through author collaboration, which can lead to false positives. In an evaluation of intrinsic approaches by Stein et al., documents under 35,000 words were excluded for not being reliably analyzable (Stein et al., 2011).

Extrinsic PDS typically follow a retrieval process that comprises several phases during which the systems successively narrow down the retrieval space to allow for increasingly fine-grained and computationally more expensive text comparisons. The initial phase typically involves some form of computationally moderate heuristic retrieval step, for example, using fingerprinting indices or vector space models

at the sentence or document level. In subsequent phases, PDS commonly perform more detailed comparisons based on exact (Goan, Fujioka, Kaneshiro, & Gasch, 2006) or approximate (Zhan et al., 2008) string matching. Cross-language plagiarism detection (CLPD) has received increasing attention. However, Potthast et al. (2010) view the cross-language field, in comparison with monolingual PD, as being ". . . still in its infancy" (Potthast, Barrón-Cedeño, Stein, & Rosso, 2010).

A shared weakness of existing extrinsic detection approaches is their exclusive reliance on textual overlap, or "character-based similarity," to identify plagiarism. This characteristic leaves current detection approaches unable to identify reliably heavily disguised plagiarism, such as paraphrases, translated plagiarism, or plagiarism of ideas or document structure, which feature little or no shared text. Evaluations of state-of-the-art PDS in the PAN Competition on Plagiarism Detection (Potthast, Eiselt, Barrón-Cedeño, Stein, & Rosso, 2011), the HTW PDS Tests (Weber-Wulff, 2010a), and other studies (Kakkonen & Mozgovoy, 2010) showed that, when text has low character-based similarity, available detection approaches fail. In the words of Weber-Wulff, an organizer of PDS performance evaluations, available systems "find copies, not plagiarism" (Weber-Wulff, 2010b).

Ideally, a PDS should detect both lexical and semantic similarities between documents. The need to incorporate semantic information into similarity checks to allow for the detection of disguised plagiarism has been acknowledged. In experiments by Bao et al. (2007), taking into account synonyms increased detection performance by a factor of two to three but also increased processing time by a factor of 27 (Bao, Lyon, Lane, Wei, & Malcolm, 2007). Thus, the already computationally expensive character-based text analysis quickly becomes unfeasible for most practical PD tasks.

Citation analysis has long been used as an indicator of semantic similarity among documents (Garfield, 1955). However, using citations for automated plagiarism detection was not considered until the introduction of Citation-based Plagiarism Detection (CbPD; Gipp & Meuschke, 2011). CbPD analyzes the selection and placement of citations in the full text of documents to form a unique language-independent fingerprint of document semantic similarity. CbPD exploits the tendency of plagiarists to copy in-text citations rather than researching their own. Observations confirmed that citation patterns often remain noticeably similar, even when plagiarists translate or strongly paraphrase the source text using synonyms and word rearrangement (Gipp, 2013; Gipp, Meuschke, & Beel, 2011). In such cases, the citation-based approach allows detecting both local and global instances of semantic similarity among publications even when character-based similarity is unsuspicious or lacking.

This article evaluates the CbPD approach for its ability to detect various degrees of plagiarism disguise in a large collection of 185,000 scientific documents. Thus far, we have tested CbPD only on a small scale to detect translated plagiarism in the GuttenPlag Wiki (Gipp et al., 2011), a collection of plagiarism instances from a single author. This article makes the following contributions.

- We demonstrate the superior ranking performance of the CbPD algorithms for the top n most suspicious user-ranked disguised plagiarism forms compared with two representative character-based detection approaches.
- We show a significant reduction in computational effort in the average case for CbPD compared with the two character-based PD approaches.
- We observe a reduction in user effort for manual document verification when the citation-based document similarity is visualized with a PDS prototype implementing CbPD.

The remainder of this article is structured as follows. The Related Work section discusses prior large-scale evaluations of PD approaches. The Methods section explains test collection requirements, introduces the detection algorithms to be evaluated, and explains our evaluation procedure. The Results section reports for the detection algorithms' retrieval and ranking performance, computational efficiency, and user utility. Limitations and future work are discussed prior to the Conclusions.

Related Work

Previous large-scale evaluations of plagiarism detection performance have centered on character-based detection approaches. The CbPD approach demands characteristics of the test collection that are not fulfilled by available artificially fabricated document collections. Primarily, existing test collections lack sufficient academic citations and realistic plagiarism disguise, as we will discuss in more detail. Nonetheless, we point out two projects that stand out for their contributions to standardizing PD evaluations.

The first project is the annual International PAN Competition on Plagiarism Detection (PAN-PC) initiated in 2009 (Potthast et al., 2012). The PAN-PC offers tracks for extrinsic and intrinsic plagiarism detection. We refer to the PAN-PC competition in 2011 (Potthast et al., 2011), because the competitions in more recent years evaluated the phases of the initial heuristic retrieval and the subsequent detailed comparison separately using different collections. This division of tasks makes it difficult to estimate the overall performance achievable by a PDS in a real-world setting.

The corpus of the PAN-PC in 2011 contained 26,939 documents, of which 50% were suspicious texts and the remainder formed the reference collection. Suspicious documents contained 61,064 artificially plagiarized sections, with 82% of sections being disguised using automated or manual English translations of German and Spanish text sections, random shuffles, insertions, deletions, or synonym replacements of terms, as well as paraphrases created by paid writers. The remaining plagiarized sections were literal copies (Potthast et al., 2011). In-text citations or reference lists are virtually nonexistent, because the test documents

were created by randomly copying text passages from books. Thus, when the fabricated texts do by chance contain citations or bibliographies, they tend to be incomplete.

The second project is a regular test of PDS by a research group at the HTW University of Applied Sciences in Berlin (Weber-Wulff, 2013). We will henceforth refer to this project as the HTW PDS tests. Whereas the competitors in the PAN-PC primarily present research prototypes, the HTW PDS tests compare commercial and otherwise publicly available PDS. The 2010 HTW PDS test evaluated 26 publicly available systems using a test collection of 40 manually fabricated essays, 30 in German and 10 in English. Five plagiarism cases were manually or machine translated from English to German and one from French to English (Weber-Wulff, 2010a). Most documents contained copy and paste or shake and paste types of plagiarism and only moderate text alterations. The obfuscation of plagiarism test cases resembles the expected plagiarism behavior among students.

Both the PAN-PC and the HTW PDS tests found that PDS retrieved verbatim copies and slightly disguised plagiarism with high accuracy, but the detection rates for manually paraphrased and manually translated plagiarism were significantly lower. In the 2010 HTW PDS test, only a single system, Turnitin, could partially identify one of the five translations in the test collection and only because the translation contained several unique place names. Similarly, the system by Grman and Ravas (Grman & Ravas, 2011), which performed best overall in the 2011 PAN-PC, achieved a recall of 0.33 for manually paraphrased segments and a recall of 0.26 for manually translated segments (Potthast et al., 2011). In other words, the best system failed to identify 67% to 74% of manually disguised plagiarism cases.

Comparably high detection rates for machine-fabricated translated plagiarism in the 2011 PAN-PC are misleading, because the systems employed translation services, such as Google Translate, which were similar or identical to the ones used to construct the automated translations in the first place. Subsequent PAN-PC competitions improved on this weakness by using only manual translations (Potthast et al., 2012, 2013). However, the approach of the best-performing systems in the PAN-PC competitions to machine translate all documents not in the target language is currently not scalable to a realistic PD scenario, in which the reference collection is a subset of the web.

Methods

Test Collection

The novelty of the citation-based approach demands a set of characteristics of the test collection that corpora of prior PDS evaluations do not offer. This prevents the comparison of citation-based algorithms with the performance metrics established, for example, in the PAN-PC competitions or the HTW PDS tests. An evaluation corpus suitable for CbPD ideally features the following.

1. Real plagiarism. Test cases should not be fabricated, either manually or automatically, when the goal is to evaluate performance on realistically disguised plagiarism containing potential citation copying.
2. Citations. The full-text of documents must contain readily available citations to allow parsing of citation position.
3. Large-scale and diversity. Documents should originate from a variety of authors to reflect different writing and citing styles.

Finally, a ground truth, in our case in the form of verified plagiarism instances, is required to gauge the retrieval performance of PD algorithms.

Given that the test collections of prior evaluations were created for character-based PDS, academic citations were not purposefully included. Furthermore, with plagiarized sections artificially fabricated, available corpora miss the full range of realistically disguised plagiarism we expect to find in real-world collections. For the corpora of the 2010–2013 PAN-PC competitions, Potthast et al. made a significant effort to create "real" plagiarism by contracting writers to produce plagiarized articles using the crowd-sourcing platforms Amazon Mechanical Turk (PAN-PC 2010 and 2011), and oDesk (PAN-PC 2012 and 2013). This approach produced the most realistic test cases available thus far, especially for the 2012 and 2013 collections, which contain about 300 articles featuring disguised plagiarism (Potthast et al., 2012, 2013); however, none of the test cases contained citations.

We argue that it remains doubtful whether articles written by contractors lacking expert knowledge are comparable in their degree of plagiarism disguise to the types of plagiarism potentially found in scientific publications. The motivation for disguise likely differs in a setting in which authors work for months or even years on a publication. Since the strength of the CbPD approach lies in the detection of heavily disguised plagiarism, we made use of a real-world scientific document collection in our evaluation to reflect the full range of disguised plagiarism forms and potential citation copying.

We chose the PubMed Central Open Access Subset (PMC OAS), a collection of biomedical full texts, for a large-scale evaluation of CbPD. The PMC OAS contains peer-reviewed publications, which leads us to assume a low level of plagiarism containment. If present, however, we assume that several plagiarism cases have been disguised, which allowed them to remain undetected, thus fulfilling the real plagiarism requirement number one. Given that the PMC OAS contains scientific publications, citations are readily available, fulfilling requirement number two. The PMC OAS contains 234,591 articles by approximately 975,000 authors from 1,972 peer-reviewed journals (as of April, 2011). This fulfills the third suitability requirement of a large-scale and diverse collection. A desirable bonus of the PMC OAS is its XML document format, which offers machine-readable markup for metadata and citations.

We established a user study-derived ground truth approximation of plagiarism and its severity for a finite pool of documents. In summary, when combined with a user-study, the PMC OAS collection is ideally suited for evaluating CbPD performance.

Detection Algorithms

On an abstract level, CbPD employs a concept we term *sequential pattern analysis* (Gipp, 2013). *Sequential pattern analysis* describes the identification of patterns for any range of markers within a document's full text by analyzing the marker's characteristics, including proximity, overlap, order, frequency, and distinctiveness. Suitable markers can be either language dependent, for example, character sequences, or language independent, for example, citations.

CbPD implements three citation pattern analysis algorithms at its core, namely, greedy citation tiling (GCT), longest common citation sequence (LCCS), and citation chunking (CC). Additionally, CbPD considers an adaption of bibliographic coupling (BC) applied for the first time to plagiarism detection. For details on algorithm implementation, refer to Gipp and Meuschke (2011).

In short, the CbPD algorithms consider citation proximity, overlap, order, frequency, and distinctiveness to varying degrees to cover the possible citation pattern rearrangements that can occur for different plagiarism forms. GCT, for instance, focuses on the order characteristic by identifying all sequences of matching citations in the same order without interruptions by nonmatching citations. Thus, GCT patterns may indicate copy and paste or directly translated plagiarism. The LCCS algorithm identifies the longest sequence of matching citations in a document while skipping over pattern interruptions caused by nonmatching or scaled citations. Citations are said to be *scaled* when the same source is cited multiple times in the full text. The CC algorithm implements heuristics that focus on the proximity characteristic of citations to allow identification of citation patterns in which matching citations are transposed, scaled, or separated by nonmatching citations. Citation chunking patterns can point to heavily disguised paraphrases, freely translated plagiarism, or structural and idea similarity, which can point to potential plagiarism. Note that we group suspicious similarities of document structure and ideas in a single category for evaluation purposes; it is extremely difficult to judge whether structural plagiarism also copies ideas.

To better gauge performance of the four core algorithms, we developed seven variations for evaluation, as listed in Table 1. We used the two, state-of-the-art character-based approaches Encoplot (ENCO; Grozea, Gehl, & Popescu, 2009), which received the highest overall score in the PAN 2009 competition, and Sherlock (Lachlan, 2012) as baselines.

Detection results are stored in a database and visualized by using CitePlag, the first citation-based PDS prototype presented by Gipp, Meuschke, Breitinger, Lipinski, and Nürnberger (2013). Figure 1 shows the ability of CitePlag to visualize both citation-based and character-based similarities to aid the user in the document verification task. Matching text strings and matching in-text citation patterns are highlighted in the bodies of the two documents (left and right columns) in identical colors. The central column displays a schematic overview of both documents, in which matching citation patterns are colored identically and connected by lines. Nonmatching in-text citations are shown in gray. Additionally, the extent of matching text in a document section is indicated by shading the background of the central column for that section in different shades of red (the higher the literal text overlap, the darker the shade of red).

TABLE 1. Detection approaches evaluated.

Citation-based approaches	
BC abs.	Absolute bibliographic coupling strength
BC rel.	Relative bibliographic coupling strength
LCCS	Longest common citation sequence
LCCS dist.	Longest common sequence of distinct citations
Max. GCT	Longest greedy citation tile
CC40	Longest citation chunk, both documents chunked, considering consecutive shared citations only without merging of chunks
CC42	Longest citation chunk, both documents chunked, considering shared citations depending on predecessor with merging of chunks
Character-based approaches	
ENCO	Encoplot, exact 16-character gram string matching
Sherlock	Sherlock, probabilistic word-based fingerprinting

Evaluation Procedure

Our evaluation procedure consists of four main steps: (a) corpus preprocessing, (b) applying the detection algorithms to the data set and pooling results, (c) addressing collection-specific false positives, and (d) performing a user study to collect relevance judgments regarding document suspiciousness.

Corpus preprocessing. The PMC OAS collection comprised 234,591 documents before preprocessing. We excluded 13,371 documents for being either unprocessable, for example, no processable text body as in the case of scanned articles in image file formats, or for being duplicates, or for containing multiple text bodies, for example, summaries of all articles in conference proceedings. Table 2 gives an overview of the excluded documents.

From the set of 221,220 processable documents, we removed an additional 36,118 documents with no references and/or citations and 68 with inconsistent citations after parsing. Documents with no references and/or citations were typically short comments, letters, reviews, or editorial notes that cited no other documents. The resulting test collection comprised 185,170 documents. Table 3 gives an overview of preprocessing results.

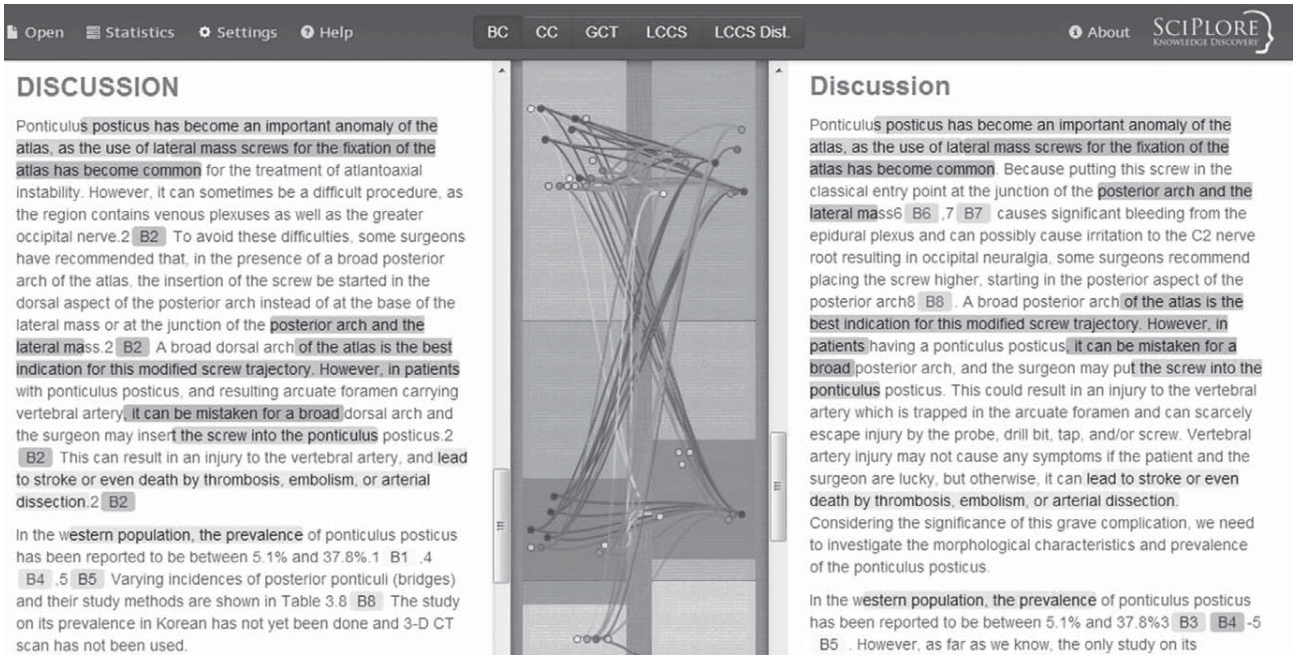


FIG. 1. CitePlag's front-end visualization. [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

TABLE 2. Number of PMC OAS documents excluded from evaluation.

Criterion	No.
PMC OAS	234,591
Excluded documents	13,371
No text body	12,783
Duplicate files	471
Multiple text bodies	117
Processable documents	221,220

TABLE 3. Preprocessing results for the PMC OAS collection.

Criterion	No. of documents	Citations	References
Processable documents	221,220	10,976,338	6,921,249
No references and/or citations	36,118	0	6,447
Inconsistent citations	68	11,405	4,722
Test collection	185,170	10,964,933	6,910,080
References w/o citations	16,866	–	65,588
Citations w/o references	59	474	–
Non-unique references	10,746	–	32,122

We were unable to acquire citation placement information fully for 16,866 documents, because citations were not marked up in the XML source file or because the original text stated citations within figures or captions. An additional 10,746 documents listed the same reference multiple times in their bibliography, and 59 documents listed references that were not cited in the main text. We did not exclude these documents, because the likelihood of false negatives, that is, unidentified cases of true plagiarism, is higher when removing the documents entirely than if the incomplete citation information is retained.

Applying algorithms and pooling. The typical PD retrieval task requires a *1:n* analysis, in which a single suspicious document is compared against a reference collection. Because the set of potentially suspicious documents is unknown in our evaluation, the PMC OAS collection calls for an *n:n* analysis. Analyzing the PMC OAS in an *n:n* fashion would require

$$\binom{n}{2} = \binom{185{,}170}{2} = 17{,}143{,}871{,}865$$

comparisons. This amount is practically infeasible to perform by any PDS in a sensible time frame and thus requires an initial limitation of the test collection.

Character-based approaches typically reduce the retrieval space by comparing heuristically selected text fragments and imposing a minimum threshold of shared text. Such heuristics, however, have the inherent disadvantage of decreasing detection accuracy. The citation-based detection approach, on the other hand, allows limiting the document collection without compromising detection accuracy. Because documents must be bibliographically coupled, that is, share at least one reference, to qualify for a citation-based analysis, we reduced collection size by filtering for bibliographic coupling (BC) strength, $s_{BC} \geq 1$.

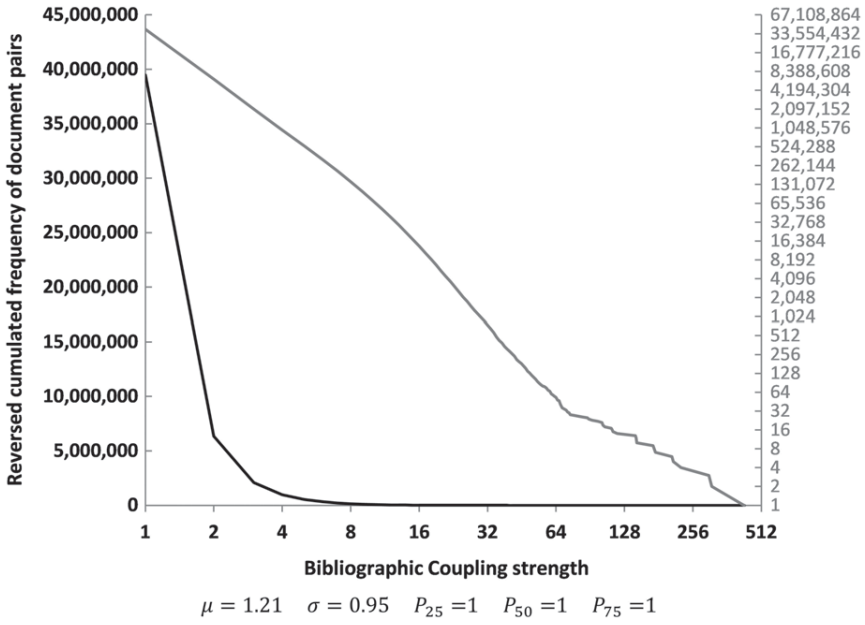


FIG. 2. BC strength for documents in the PMC OAS collection.

Figure 2 shows the distribution of BC strength s_{BC} for pairs of documents (d_1,d_2) in terms of the reverse cumulative frequency,

$$\overline{f}_c = \sum_{i=n}^{1} (d_1, d_2) \mid (s_{BC}) \geq i$$

plotted on an absolute (black line) and log 10 scale (gray line).

This restriction reduced the PMC OAS size to 39,463,660 document pairs requiring analysis. Because of the practical infeasibility of a collection-wide character-based *n:n* analysis, we applied ENCO and Sherlock only to the 6,219,504 document pairs with an $s_{BC} \geq 1$. To our knowledge, bibliographic coupling has thus far not been used as a criterion to limit collection size for PD purposes. This limitation may lead to the exclusion of some true positives.

We argue that limiting collection size using BC strength is unlikely to have a significant negative effect on character-based detection performance. To substantiate this hypothesis, we performed an ex post *n:n* analysis of the top 20 most suspicious documents as identified in the user study. Because we did not filter for BC strength, it took several weeks on a quad-core system to compute the character-based Encoplot scores for these 20 documents with all other documents in the PMC OAS collection. Figure 3 plots BC strength and ENCO score (a measure of character-based document similarity). The smallest dots represent single occurrences; the largest dots represent up to 20 occurrences.

The sample contained no publication pair with an Encoplot score above 3 that was not bibliographically coupled. Given the correlation between bibliographic similarity and character-based similarity, we hypothesize that the loss of detection performance is minimal and that using a minimum BC strength is an acceptable compromise to limit collection size and allow an *n:n* analysis of documents.

Because judging all retrieval results is infeasible, the top 30 ranked document pairs for the nine detection algorithms are pooled, as is common practice in information retrieval (IR) evaluations, such as TREC, NTCIR, or CLEF (Buckley, Dimmick, Soboroff, & Voorhees, 2007). We removed duplicates and PMC OAS-specific false positives before collecting relevance judgments.

Addressing false positives. The retrieval of false positives is a universal problem for PDS. In the case of the PMC OAS corpus, false positives presented a larger challenge to character-based approaches than to citation-based approaches, because specific document types reused standardized expressions or boilerplate text. Several instances of high textual similarity were thus justified in the PMC OAS. For this reason, we applied a false-positive-reduction strategy to the pooled documents prior to collecting relevance judgments.

We excluded the collection-specific document types editorials and updates. Editorials were typically nonscientific texts written by journal editors or publishers, which provide publishing guidelines or descriptions of the journal's purpose and policies. Such text is often "recycled" as boilerplate text among journals without citing the source. Updates included revisions to published material, as well as slight changes to annually published medical standards, best practices, or procedures, commonly published by medical associations, such as the American Diabetes Association.

This exclusion of documents was necessary for a meaningful performance evaluation. Without it, the character-based approaches, in particular Encoplot, would have retrieved among its top ranks almost exclusively legitimately similar documents, which would have resulted

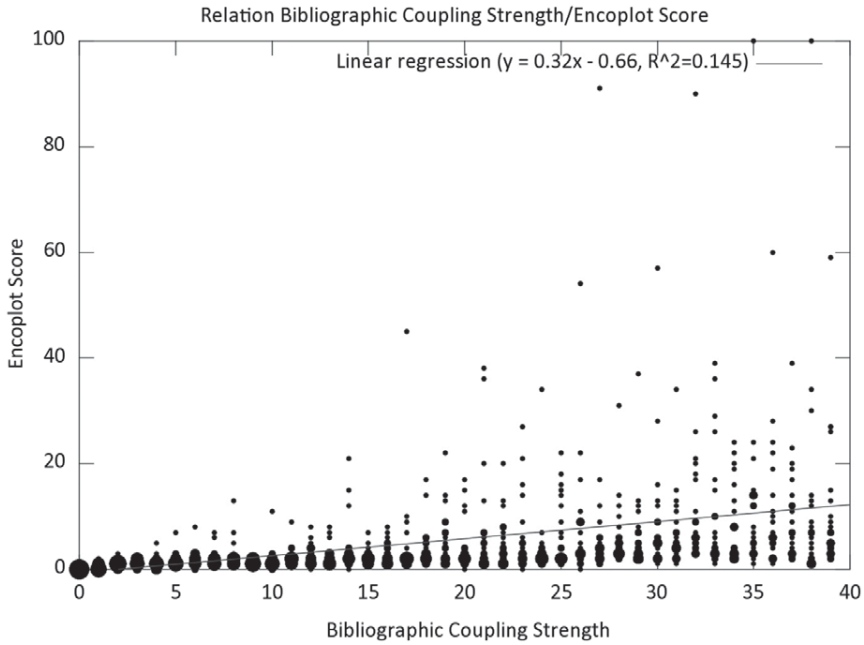


FIG. 3. Correlation between BC strength and Enco score. [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

in an unwarranted high rate of false positives for the character-based approaches unjustly resulting from the properties of the test collection. We also excluded PMC OAS publications that cited each other or had shared author sets, to reduce false positives that referenced the source or were likely examples of legitimate collaboration.

Two additional factors contributed to a higher false-positive rate in the case of the PMC OAS. First, we carried out the pooling process as an *n:n* document comparison, whereas in the typical PD use case a *1:n* comparison is performed. An *n:n* comparison of a very large collection naturally results in the retrieval of high numbers of legitimate similarities. Second, the relatively sparse amount of plagiarism in the PMC OAS makes the retrieval of legitimate document similarities more likely.

Despite our strategy to reduce false positives, some cases of legitimate text similarity remained, which we identified during the pooling step and removed before the user study. Collecting the top 30 similar documents for the character-based approach, Encoplot required examining 235 documents, because Encoplot retrieved 205 collection-specific false positives, such as editorials and updates. Collecting the top 30 similar documents for the LCCS approach required the examination of 31 documents, because only one collection-specific false positive was retrieved. The citation-based approaches retrieved fewer false positives in the case of the PMC OAS collection, because these documents featured unique citation patterns despite high textual overlap, for example, in medical case studies and editorials, or they had insufficient citations because of their nonscientific nature. In conclusion, false-positive rates are highly corpus dependent. Every corpus contains different document formats and text from different disciplines, meaning that the reuse of text or citations may be seen as legitimate in some cases but not in others.

Participants. We performed a user study to collect judgments on the dominant plagiarism form and the level of document suspiciousness, that is, document relevance to a PD scenario. The top 30 pooling method yielded 270 document pairs, of which 181 were unique. To obtain relevance judgments, we presented the unique pairs to 26 participants using the web-based prototype CitePlag. We divided participants according to their level of biomedical expertise into the following three groups: five medical experts, 10 graduate students from the medical and life sciences, and 11 undergraduate students from a variety of majors. Because no standard guidelines or thresholds exist for classifying a document as "plagiarism," we asked participants to assess documents keeping in mind the information need in a real plagiarism detection scenario: "Consider viewing a retrieved document pair as relevant if similarities exist that an examiner in a real check for plagiarism would likely find valuable to be made aware of."

We instructed participants to rate presented document pairs on a scale from 0 to 5, where a score of 0 indicated a false positive and scores 1 through 5 described various levels of document suspiciousness. An online submission form provided uniform guidelines, including definitions of the four examined plagiarism forms and verbal descriptions of the suspiciousness scores, that is, relevance to a plagiarism-detection scenario. For example, a score of 5

indicated extremely suspicious similarities with obvious plagiarism intent, whereas a score of 1 described noticeable similarities in some sections when an author might have found inspiration from the source but most likely did not plagiarize.

A participant from each of the three knowledge groups examined each document pair. If examiners found the document pair presented to fulfill the given information need, that is, suspiciousness score, s, >0, we asked them to (a) indicate the most prevalent form of plagiarism, (b) judge relevance for a plagiarism detection scenario (suspiciousness scores 1–5), and (c) indicate whether a character-based, citation-based, or hybrid document similarity visualization was most suitable to assess document suspiciousness.

Our evaluation procedure was as follows. We retained and grouped by plagiarism form all document pairs assigned $s > 0$ by at least one examiner. If examiners disagreed on the prevailing plagiarism form, we used the expert response. For each document pair to arrive at a single score, we calculated a weighted average of the suspiciousness scores assigned by the examiner groups as

$$\bar{s} = (s_u + 1.25 s_g + 1.5 s_e)/3.75,$$

where s_u denotes the score assigned by undergraduates, s_g the score assigned by graduate students, and s_e the score assigned by medical experts. Finally, to derive a ground truth for the four examined plagiarism forms, we ordered the document pairs in each of the plagiarism categories by decreasing $\bar{s}$ and selected the top 10 documents with the highest user-assigned suspiciousness scores.

To confirm agreement on document suspiciousness among participants above the agreement rate to be expected by chance, we calculated interrater agreement using Fleiss's kappa, κ as follows.

$$\kappa = \frac{\bar{P} - \bar{P}_e}{1 - \bar{P}_e}$$

In this notation, $\bar{P}_e$ represents observed agreement and $\bar{P}_e$ represents the probability of chance agreement. Thus, $\bar{P} - \bar{P}_e$ is the degree of agreement achieved above chance and $1 - \bar{P}_e$ the degree of agreement that is attainable above chance. Fleiss's kappa, κ for all assigned document scores was 0.65, indicating substantial interrater agreement on suspiciousness. Agreement was highest for copy and paste, $\kappa = 0.73$, and lowest for structural and idea similarity, $\kappa = 0.59$. This observation was in line with our expectation of higher discrepancies in judgment for disguised plagiarism forms, which are often more controversial.

Results

Retrieval and Ranking Performance

In the typical use case, manual verification of suspicious documents can reasonably be performed only for documents retrieved at the highest ranks. We therefore consider the rank at which a detection algorithm retrieves the top n relevant results as a crucial measure of the retrieval effectiveness of a PDS. We evaluated detection performance of the nine approaches by comparing their ranking performance with the ground truth approximation on document suspiciousness derived in the user study for the four plagiarism forms examined.

For each of the top 10 user-rated document pairs, we selected the more recent publication and checked at which rank, if at all, a detection algorithm identified the recent publication as similar to the earlier publication. If detection approaches assigned the same score, and thus the same rank, i, to multiple documents, the midrank, calculated as

$$\bar{r}_i = r_{i-1} + (|d_i| - 1)/2$$

was assigned to all documents, d_i, with initial rank i. We found that the best-performing approach is highly dependent on plagiarism form. The following subsections describe algorithm retrieval performance for the four examined plagiarism forms in detail. Figure 4 plots the distribution of ranks for all levels of plagiarism disguise examined.

Copy and paste. The raking distribution of detection approaches for the minimally disguised copy and paste type of plagiarism shows that the character-based detection approach ENCO performed best at highly ranking this form of plagiarism. The citation-based LCCS algorithm performed second best, and the character-based PDS Sherlock ranked third. The upper quartile of the three best-performing approaches equals 1. This means that, for at least 75% of the examined top 10 document pairs, the approaches retrieved the source document at rank 1.

Among the top 10 copy and paste document pairs, ENCO identified all at rank 1. LCCS and Sherlock retrieved nine at rank 1. The results confirm that current detection approaches have no difficulty in retrieving at high ranks documents that contain verbatim text overlap. The citation-based approaches, especially LCCS, performed better than expected for nondisguised plagiarism. The reason for this might have been collection specific, in that many document pairs with extensive text overlaps also featured long instances of shared citation patterns.

Shake and paste. The distribution of ranks for shake and paste plagiarism shows that ENCO identified the more recent document that users rated as suspiciously similar to the earlier published document at rank 1 for all 10 document pairs. Sherlock and the two LCCS measures each identified nine pairs at rank 1. The remaining citation-based approaches showed slightly lower retrieval performance but could identify the source document for each of the user-classified top 10 document pairs. No third quartile of any citation-based approach exceeded rank 2.

The good performance of ENCO's exact 16-character-gram string matching and Sherlock's probabilistic word-based fingerprinting approach in identifying shake and

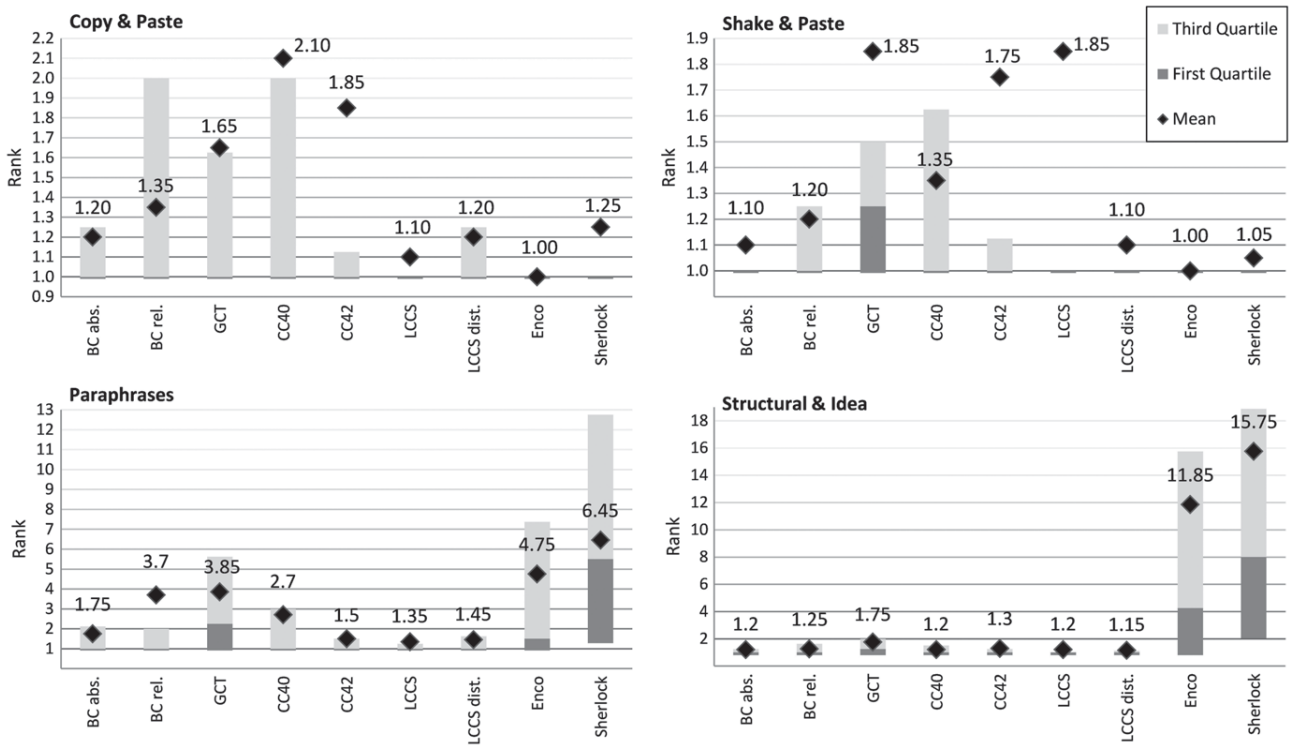


FIG. 4. Overview of ranking performance for plagiarism forms.

paste similarities was no surprise, given that many of the identified instances have high verbatim text overlap. The citation-based measures performed better than expected, which was due mainly to most shake and paste similarities being concentrated in the introductory and background sections of publications, which also included a high number of shared citations.

Paraphrases. The box plots for both paraphrases and structural and idea similarity show that the CbPD approach outperformed character-based approaches in identifying these forms of plagiarism. The two best-performing algorithms for paraphrases, LCCS and LCCS dist., identified eight and seven of the top 10 document pairs at rank 1 and ranked no document pair below rank 4. ENCO identified six, and Sherlock eight of the document pairs below the top rank of 1. The lowest ranks at which the two character-based approaches retrieved one of the top 10 document pairs was at rank 18 for Encoplot and at rank 14.5 for Sherlock.

Structural and idea similarity. For structural and idea similarity, the advantage of CbPD was even more apparent than for paraphrases. The citation-based approach, especially the variations of LCCS (LCCS and LCCS dist.) significantly outperformed the character-based approach in prominently ranking structural and idea similarity. LCCS identified nine and LCCS dist. eight document pairs at rank 1 and the remaining document pairs no lower than rank 3. ENCO, on the other hand, ranked six and Sherlock ranked nine document pairs at rank 4 or below. Figure 5 shows the ranking distribution for structural and idea similarity in detail. One can see that the lowest ranks at which ENCO and Sherlock retrieved the document pairs were at rank 57.5 for ENCO and rank 79.5 for Sherlock. Note how the retrieved documents cluster around rank 1 for the citation-based algorithms but are widely distributed for ENCO and Sherlock.

Our results indicate that character-based and citation-based approaches have complementary strengths. The ranking distribution for the top 10 suspicious documents for each of the four examined plagiarism forms confirmed that CbPD more effectively detects disguised plagiarism forms with low textual similarity, that is, paraphrases and similarities in document structure or ideas, which can indicate unoriginality and potential plagiarism. Character-based approaches, on the other hand, perform effectively in retrieving plagiarism in the copy and paste and shake and paste categories.

Computational Efficiency

Computational efficiency is crucial to PDS performance, because exhaustive *n:n* comparisons quickly become unfeasible for large collections. Although character-based approaches require a pairwise comparison for the entire collection to prevent a loss in detection performance, the CbPD approach retains only those documents of the reference collection that share at least one citation in

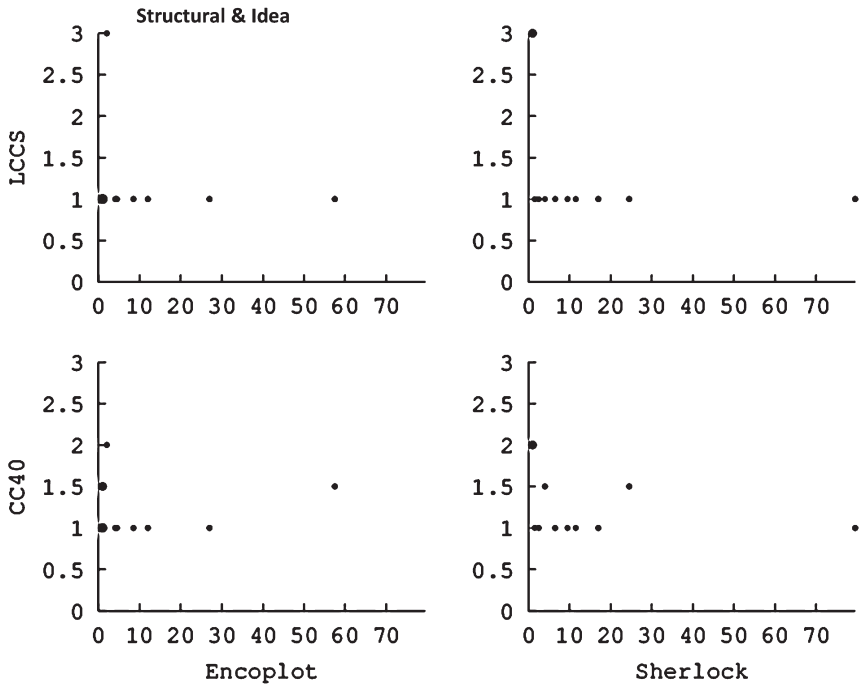


FIG. 5. Ranking distribution for structural and idea similarity.

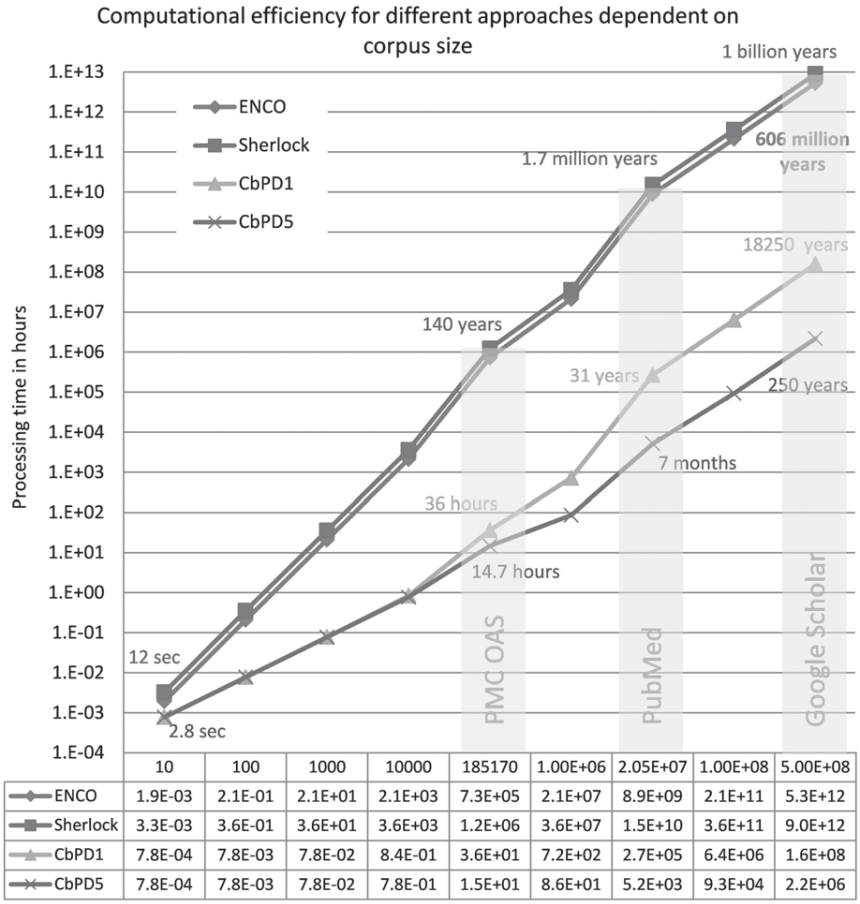
common with the examined document. Thus, we retained only the documents with $s_{BC} \geq 1$ in the PMC OAS collection for further analysis.

In general, the processing time for automated plagiarism detection is composed of two elements, first, the time required for preprocessing and, second, the time required for document comparison. Preprocessing encompasses file system and/or database operations and document type conversions. For example, ENCO and Sherlock required converting PMC OAS's NXML format to plain text. Preprocessing for citation-based approaches includes text parsing to acquire references, determining citation distribution in the full text, and extracting document metadata. Extracted data must be cleaned and disambiguated before being stored in the database. Because the restriction $s_{BC} \geq 1$ limits collection size, we included the time required for computing BC strength to the citation-based algorithms' preprocessing time.

Character-based detection approaches require $O(n)$ time for preprocessing, because n documents must be converted from NXML to plain text. The citation-based approach also requires $O(n)$ time for converting and parsing documents and for cleaning and disambiguating the parsed data. The additional calculation of s_{BC} requires $O(n \cdot \log(n))$ time when using an index that allows comparing the references in documents in $O(\log(n))$ time. All citation-based algorithms had similar overall run times. We therefore summarized all seven citation-based methods under the label *CbPD* and examined their mean processing time.

Figure 6 plots both measured and extrapolated average case processing times for ENCO, Sherlock, CbPD1, and CbPD5, where CbPDn stands for any citation-based approach using an s_{BC} threshold $\geq n$. Processing time in hours is plotted on a log 10 scale and assumes a 3.40 GHz quad core processor with 16 GB of RAM. Shaded columns depict the size ranges of well-known, large-scale corpora, namely, PMC OAS, PubMed, and Google Scholar. Figure 6 shows that, to process the PMC OAS, the CbPD5 algorithm required 14.7 hours, whereas Sherlock would require an estimated 140 years. For ENCO and Sherlock, we measured processing times for sample sizes 10, 100, and 1,000 and extrapolated the processing times for the larger collections with unfeasible runtime requirements. For the CbPD algorithms, we calculated processing times up to the size of the PMC OAS and extrapolated the times for larger collections.

The efficiency of approaches depends heavily on corpus size. If a single document pair (*1:1*) is analyzed, the character-based approach is comparatively less expensive than the citation-based approach. This is true because for smaller corpora citation parsing is computationally more intensive than a character-based *n:n* comparison. However, the break-even point, depending on document length and number of citations, is commonly reached at about five documents. For larger collections, character-based approaches are typically more expensive, given that they require $\binom{n}{2}$ comparisons. In summary, the superior computational efficiency of the citation-based approach is



	10	100	1000	10000	185170	1.00E+06	2.05E+07	1.00E+08	5.00E+08
ENCO	1.9E-03	2.1E-01	2.1E+01	2.1E+03	7.3E+05	2.1E+07	8.9E+09	2.1E+11	5.3E+12
Sherlock	3.3E-03	3.6E-01	3.6E+01	3.6E+03	1.2E+06	3.6E+07	1.5E+10	3.6E+11	9.0E+12
CbPD1	7.8E-04	7.8E-03	7.8E-02	8.4E-01	3.6E+01	7.2E+02	2.7E+05	6.4E+06	1.6E+08
CbPD5	7.8E-04	7.8E-03	7.8E-02	7.8E-01	1.5E+01	8.6E+01	5.2E+03	9.3E+04	2.2E+06

The first row lists the collection size. The rows underneath show processing times in hours (partially extrapolated).

FIG. 6. Computational efficiency. [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

TABLE 4. User-perceived suitability of methods.

	Copy and paste	Shake and paste	Paraphrased	Structural and idea	Translated[a]
Character-based	51%	27%	6%	1%	0%
Citation-based	1%	5%	32%	86%	54%
Hybrid	47%	68%	62%	13%	46%

[a]Examination of zu Guttenberg thesis only.

advantageous especially for large document collections, which remain analyzable in an *n:n* fashion without a loss in detection accuracy.

User Utility

An effective, automated detection approach maximizes user utility by addressing user information need and minimizing effort. We assessed utility by questioning the 26 participants on the similarity visualization method, character-based, citation-based, or hybrid, that they perceived most suitable for the various plagiarism forms. We additionally examined whether a reduction in user effort is attainable if the citation-based approach is combined with the strictly character-based similarity visualization of current PDS.

Table 4 shows the document similarity visualization approaches participants indicated as most suitable depending on the dominating plagiarism form. We collected these responses for the 461 document pair judgments from all three examiner groups, where $s > 0$. The majority of participants indicated traditional text highlights as the single most suitable similarity visualization method to assist in document verification for the copy and paste type of plagiarism. For the heavily

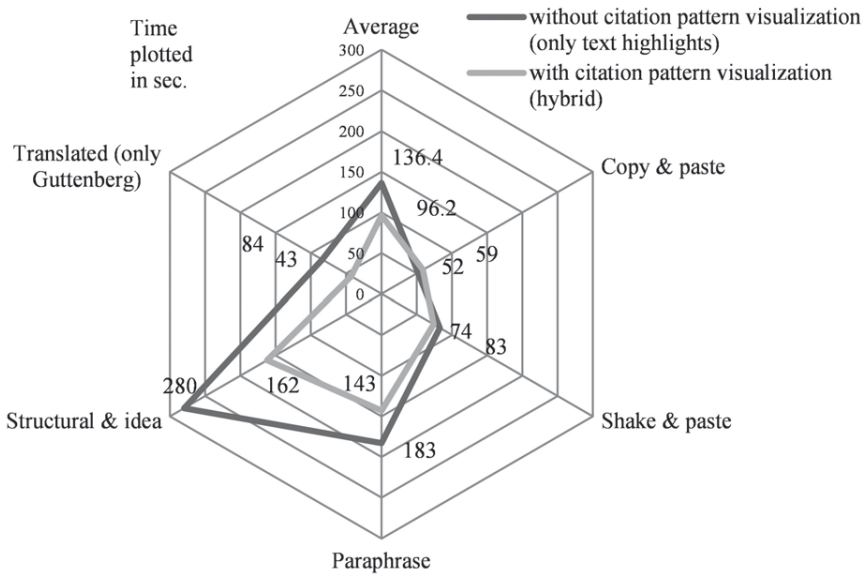


FIG. 7. Mean times in seconds required for document verification.

disguised structural and idea similarity, the majority of participants rated the citation-based approach as the most effective visualization method. A hybrid approach combining both text and citation pattern visualization was perceived as most suitable to detect paraphrases and shake and paste plagiarism. Because the PMC OAS contains English-language publications only, we additionally asked 13 volunteers (of 26 participants) to indicate the suitability of visualization methods for an excerpt of the zu Guttenberg thesis (zu Guttenberg, 2009), a prominent plagiarism case of former German defense Minister K.-T. zu Guttenberg containing several instances of translated plagiarism as identified by the GuttenPlag Wiki (GuttenPlag Wiki, 2011). With opinions on translated plagiarism collected only for a single plagiarism case, these results, however, cannot be generalized.

In a subsequent evaluation, we examined whether a reduction in user effort, measured as a time saving, is observable upon citation pattern visualization. We recruited a subset of eight participants and divided them into two groups of four to judge document suspiciousness once with text similarity visualized, and once with both text and citation pattern similarity visualized. For each of the visualization methods, we recorded the time examiners required to verify the first two instances that they deemed likely plagiarism.

Each participant rated 25 document pairs, six pairs in each of the four assigned plagiarism categories as well as a single document to represent translated plagiarism, an excerpt of the zu Guttenberg thesis. The six documents for each of the four plagiarism forms were a random sample of the top 30 documents yielded by the pooling approach.

Figure 7 shows the mean times in seconds recorded for document verification with and without citation pattern visualization. We observed a notable reduction in the mean times required to identify suspicious similarity upon citation pattern visualization for the heavily disguised structural and idea similarity, with a 42.1% time reduction; followed by paraphrases, with a 21.8% reduction; and shake and paste-type plagiarism, with a 10.8% reduction. We also observed a lower user effort to verify translated plagiarism; however, these data should not be generalized, given that they represent only a single examined case.

The recorded time savings were in line with the user-classified suitability of the approaches. The citation pattern visualization of CbPD was most helpful for verifying structural and idea similarities. For plagiarism forms with very high textual similarity, such as copy and paste, citation pattern visualization interestingly had a negative effect on time reduction over text-only visualization. We suspect that some examiners clicked through sections with high citation pattern similarity more thoroughly and thus took longer to submit the first two instances of suspected plagiarism.

Identified Cases

For user-perceived cases of plagiarism, we contacted the authors of the earlier published article. Thus far, three plagiarized medical studies have been retracted by the issuing journal, and six further publications were confirmed to contain plagiarism by the earlier authors. Additional cases are still under examination. Due to the sensitivity of the issue, we do not disclose nonconfirmed cases of potential plagiarism. Please refer to http://citeplag.org for the most recent user-identified cases using the prototype as well as updates on retraction notices.

Limitations

General challenges faced when evaluating the performance of PD algorithms using nonartificially created test

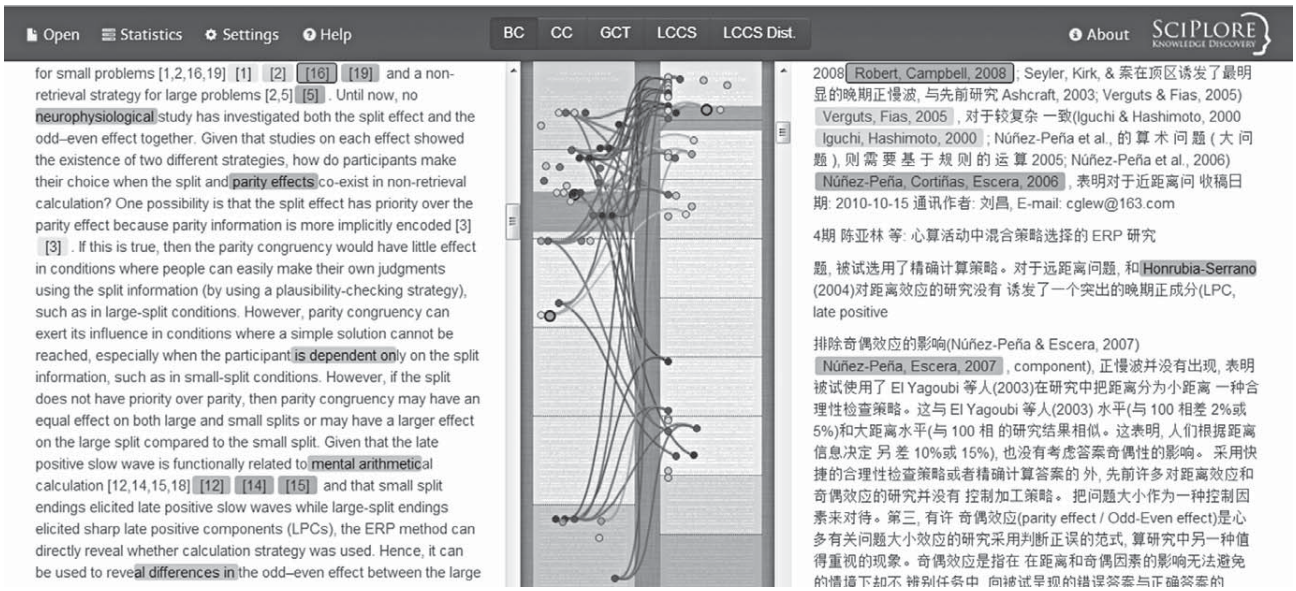


FIG. 8. Example of plagiarism in an alphabet different from that of the source. [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

collections are the lack of ground truth and the subjectivity of human judgment. We addressed the first challenge by establishing a ground truth approximation for a pooled set of documents. We addressed the second challenge as well as this is possible, by providing uniform definitions and guidelines to participants.

Although the CbPD algorithms took into account citations and references to sources cited outside of the PMC OAS, the restricted access to full texts allowed searching for plagiarism only if the similar document was included in the PMC OAS. We assumed the PMC OAS corpus to exhibit relatively low plagiarism content, because plagiarism was likely detected in the journals' peer-review process, by character-based PDS, or as a result of prior PMC OAS corpus testing, such as the character-based examinations conducted by a research group at the Harold Garner laboratory (Sun et al., 2010). Because nondisguised plagiarism, in particular, is more likely to be detected and removed, the results obtained from the PMC OAS may not be representative of other collections.

Future Work

Several interesting areas of research remain. Integrating CbPD with current technologies for cross-language information retrieval and information visualization methods is one such application. Applying the citation-based approach to large-scale heterogeneous collections even when they contain different languages and alphabets is another promising use. Figure 8 shows CbPD applied to a retracted publication (Chen, Liu, Xu, Zhang, & Shen, 2012), which translated a Chinese publication without attributing the source.

The citation-based approach also raises the question of how to define and address newly detectable plagiarism forms. No consensus exists on the levels of structural similarity in documents that may adequately represent critical thresholds. One new form, for example, may be termed citation composition plagiarism.

High levels of citation-based document similarity need not necessarily point to plagiarism. For this reason, we propose citation-based similarity as a supplemental indicator to determine whether a work exhibits a high degree of novelty. For example, in evaluating the merits of a grant proposal, a reviewer likely is interested in similarity to other proposals, patents, or published ideas, to cross check the level of originality. Citation-based document similarity may also be used to identify related work or make visible author inspiration trails, defined as the texts consulted but not cited by an author (Gipp, 2013).

Conclusions

Our evaluation demonstrates the effectiveness and practicability of CbPD with a large-scale scientific document collection containing various degrees of plagiarism disguise. We evaluated the effectiveness of seven CbPD algorithms and two popular character-based approaches using human judgment and a top n results pooling approach. Our test collection derived from the PMC OAS contained 185,170 publications. In comparing the ranks at which each detection approach identified the top 30 suspicious document pairs for each plagiarism form against human judgment, we found that the citation-based detection approaches significantly outperformed the character-based approaches in retrieving among their top ranks those documents, which contained

paraphrases and structural and idea similarity. The character-based approaches ranked the top results as judged by humans highest for copy and paste and shake and paste plagiarism.

For the PMC OAS collection, we approximated the advantage in computational efficiency to be on the order of 3.6×10^4, that is, 14.7 hours for the CbPD approach compared with ~140 years for character-based approaches. The citation-based visualization method reduced user effort as measured in a time savings for examiners, which was especially noticeable for heavily disguised plagiarism forms. Moreover, CbPD discovered several cases of previously unidentified plagiarism in the PMC OAS collection some of which would have remained undetected by today's character-based detection approaches.

In summary, citation-based and character-based approaches to automated plagiarism detection have complementary strengths and weaknesses. We conclude that a hybrid detection approach as explored in our prototype would represent a significant improvement upon current detection approaches.

Acknowledgments

We acknowledge Mario Lipinski, André Gernandt, Leif Timm, Markus Bruns, Markus Föllmer, and Rebecca Böttche for their contributions to improving the CitePlag prototype.

References

Bao, J., Lyon, C., Lane, P.C.R., Wei, J., & Malcolm, J.A. (2007). Comparing Different Text Similarity Methods. Technical Report 461, Science and Technology Research Institute, University of Hertfordshire.

Buckley, C., Dimmick, D., Soboroff, I., & Voorhees, E. (2007). Bias and the limits of pooling for large collections. Information Retrieval, 10, 491–508.

Chen, Y., Liu, C., Xu, X., Zhang, X., & Shen, W. (2012). Simple mental arithmetic is not so simple: An ERP study of the split and odd-even effects in mental arithmetic. Neuroscience Letters, 510, 62–66.

Garfield, E. (1955). Citation indexes for science: A new dimension in documentation through association of ideas. Science, 122(3159), 108–111.

Gipp, B. (2013). Citation-based Plagiarism Detection: Applying Citation Pattern Analysis to Identify Currently Non-Machine-Detectable Disguised Plagiarism in Scientific Publications. Unpublished doctoral thesis, Otto-von-Guericke University, Magdeburg, Germany.

Gipp, B., & 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 (DocEng '11), Mountain View, CA: ACM.

Gipp, B., Meuschke, N., & Beel, J. (2011). Comparative Evaluation of Text- and Citation-based Plagiarism Detection Approaches Using GuttenPlag. In Proceedings of the 11th ACM/IEEE-CS Joint Conference on Digital Libraries (JCDL'11). Ottawa, Canada: ACM.

Gipp, B., Meuschke, N., Breitinger, C., Lipinski, M., & Nürnberger, A. (2013). Demonstration of Citation Pattern Analysis for Plagiarism Detection. In Proceedings of the 36th International ACM SIGIR Conference on Research & Development on Information Retrieval. Dublin, Ireland: ACM.

Goan, T., Fujioka, E., Kaneshiro, R., & Gasch, L. (2006). Identifying information provenance in support of intelligence analysis, sharing, and protection. In S. Mehrotra, D. Zeng, H. Chen, B. Thuraisingham, & F.-Y. Wang (Eds.), Intelligence and Security Informatics (Vol. 3975, pp. 692–693). Berlin: Springer.

Grman, J., & Ravas, R. (2011). Improved Implementation for Finding Text Similarities in Large Collections of Data. In Proceedings of the Notebook Papers of CLEF 2011 LABs and Workshops. Amsterdam. Retrieved from http://dblp.uni-trier.de/rec/bibtex/conf/clef/GrmanR11

Grozea, C., Gehl, C., & Popescu, M. (2009). ENCOPLOT: Pairwise Sequence Matching in Linear Time Applied to Plagiarism Detection. In Proceedings of the the 3rd PAN Workshop. Uncovering Plagiarism, Authorship and Social Software Misuse. Aachen, Germany: CEUR-WS.org.

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. Retrieved from http://de.guttenplag.wikia.com/wiki/GuttenPlag_Wiki

Kakkonen, T., & Mozgovoy, M. (2010). Hermetic and web plagiarism detection systems for student essays—An evaluation of the state-of-the-art. Journal of Educational Computing Research, 42, 135–159.

Lachlan, P. (2012). *The Sherlock plagiarism detector*. Retrieved from http://sydney.edu.au/engineering/it/ scilect/sherlock/

Meyer zu Eissen, S., Stein, B., & Kulig, M. (2007). Plagiarism Detection Without Reference Collections. In Proceedings of the the 30th Annual Conference of the German Classification Society (GfKl). Berlin: Springer.

Potthast, M., Barrón-Cedeño, A., Stein, B., & Rosso, P. (2010). Cross-language plagiarism detection. Language Resources and Evaluation, 45, 45–62.

Potthast, M., Eiselt, A., Barrón-Cedeño, A., Stein, B., & Rosso, P. (2011). Overview of the 3rd International Competition on Plagiarism Detection. In V. Petras & P. Clough (Eds.), CLEF (Notebook Papers/Labs/Workshop). Amsterdam. Retrieved from http://www.clef-initiative.eu/publication/working-notes

Potthast, M., Gollub, T., Hagen, M., Kiesel, J., Michel, M., Oberländer, A., et al. (2012). Overview of the 4th International Competition on Plagiarism Detection. In P. Froner, J. Karlgren, & C. Womser-Hacker (Eds.), CLEF (Online Working Notes/Labs/Workshop). Rome. Retrieved from http://www.clef-initiative.eu/publication/working-notes

Potthast, M., Hagen, M., Gullub, T., Tippmann, M., Kiesel, J., Rosso, P., et al. (2013). Overview of the 5th International Competition on Plagiarism Detection. In P. Forner, R. Navigli, & D. Tufis (Eds.), CLEF 2013 Evaluation Labs and Workshop. Valencia, Spain. Retrieved from http://www.clef-initiative.eu/publication/working-notes

Stein, B., Lipka, N., & Prettenhofer, P. (2011). Intrinsic plagiarism analysis. Language Resources and Evaluation, 45, 63–82.

Sun, Z., Errami, M., Long, T., Renard, C., Choradia, N., & Garner, H. (2010). Systematic characterizations of text similarity in full text biomedical publications. PLoS ONE, 5(9). Retrieved from http://www.plosone.org/article/info%3Adoi%2F10.1371%2Fjournal.pone.0012704

Weber-Wulff, D. (2010a). Portal Plagiat—Softwaretest 2010. Retrieved from http://plagiat.htw-berlin.de/software-en/2010-2/

Weber-Wulff, D. (2010b). Test Cases for Plagiarism Detection Software. In Proceedings of the 4th International Plagiarism Conference. Newcastle upon Tyne, UK. Retrieved from http://www.plagiarismadvice.org/research-papers/item/test-case-for-plagiarism-detection-software

Weber-Wulff, D. (2013). *Test of plagiarism software*. Retrieved from http://plagiat.htw-berlin.de/software-en/

zu Guttenberg, K.-T. (2009). Verfassung und Verfassungsvertrag: Konstitutionelle Entwicklungsstufen in den USA und der EU. Berlin: Duncker & Humblot.

Zhan, S., Byung-Ryul, A., Ki-Yol, E., Min-Koo, K., Jin-Pyung, K., & Moon-Kyun, K. (2008). Plagiarism Detection Using the Levenshtein Distance and Smith-Waterman Algorithm. In Proceedings of the 3rd International Conference on Innovative Computing Information and Control. New York, U.S.A.: IEEE.