ABSTRACT
This paper proposes a hybrid approach to plagiarism detection in academic documents that integrates detection methods using citations, semantic argument structure, and semantic word similarity with character-based methods to achieve a higher detection performance for disguised plagiarism forms. Currently available software for plagiarism detection exclusively performs text string comparisons. These systems find copies, but fail to identify disguised plagiarism, such as paraphrases, translations, or idea plagiarism. Detection approaches that consider semantic similarity on word and sentence level exist and have consistently achieved higher detection accuracy for disguised plagiarism forms compared to character-based approaches. However, the high computational effort of these semantic approaches makes them infeasible for use in real-world plagiarism detection scenarios. The proposed hybrid approach uses citation-based methods as a preliminary heuristic to reduce the retrieval space with a relatively low loss in detection accuracy. This preliminary step can then be followed by a computationally more expensive semantic and character-based analysis. We show that such a hybrid approach allows semantic plagiarism detection to become feasible even on large collections for the first time.

Categories and Subject Descriptors

General Terms
Algorithms, Performance.

Keywords
Information Retrieval, Plagiarism Detection, Semantic Analysis, Citation Analysis, Disguised Plagiarism, Large Scale Collections

1. INTRODUCTION
Today’s plagiarism detection (PD) systems exclusively compare text strings to identify suspicious similarity between documents. These systems successfully retrieve copied text, but fail to identify disguised plagiarism, such as paraphrases, translations, or idea plagiarism [21]. Prof. Weber-Wulff, who organizes a regular benchmark test for plagiarism detection systems (PDS), summarizes the capabilities of available systems as follows: “[…] Plagiarism Detection Systems find copies, not plagiarism.” [20] and “[…] for translations or heavily edited material, the systems are powerless […]” [21]. Due to the limitations of available PDS, a large fraction of disguised plagiarism currently goes undetected.

Most PDS follow a three-stage retrieval process shown in Figure 1 [18]. In the first stage, PDS apply computationally inexpensive heuristics to identify a small fraction of the reference collection as candidate documents from which the input text could originate.

In the second stage, candidate documents undergo a computationally more expensive detailed comparison. In the third stage, PDS apply knowledge-based post-processing to text segments retrieved in the second stage. The goal is to eliminate typical false positives, which the detection procedures in the previous stages are prone to produce. Finally, a human examiner must review the text segments retrieved and make a judgment.

To identify plagiarism forms beyond copy and paste, non-character-based detection approaches have been explored; however, thus far, only in prototypical settings. The approaches focus on the detailed comparison stage of the plagiarism detection process (see Figure 1). For this stage, non-character-based approaches typically outperform character-based methods in identifying disguised plagiarism, as we present in Section 2.

However, prior research did not answer the question how to effectively limit a large-scale, real-world reference collection in the first stage of the detection process without relying on character-based heuristics. These character-based approaches fail to detect plagiarism with little to no identical text. In practice, PDS must compare an input text to a large reference collection, which typically includes a subset of the Internet. Detecting
disguised plagiarism in this type of real-world use case requires computationally modest detection approaches that reduce the retrieval space to the semantically most similar documents regardless of literal text overlap. We propose that citation-based PD methods fulfill these requirements and can serve as an initial heuristic to identify semantically similar candidate documents during a search for plagiarism. We suggest that combining citation-based detection methods with existing, computationally expensive semantic detection approaches can enable the detection of disguised plagiarism in real-world, large-scale collections.

2. NON-CHARACTER-BASED PD

Researchers have recognized that identifying disguised plagiarism requires detection approaches that go beyond comparing literal text overlap. Non-character-based detection approaches proposed for this purpose employ cross-language analysis, and semantic analysis. Cross-language detection approaches typically use machine translation [15]. Since machine translating entire texts is computationally expensive, cross-language PD methods typically extract key words from the input text and query these key words against an index of key words extracted from documents in the reference collection. Either the keywords of the input text, the index of key words representing the reference collection or both are machine-translated prior to being matched. Despite advances in machine translation, cross-language PD is currently not reliable enough for practical use cases [15].

Several researchers addressed monolingual paraphrase detection by analyzing semantic word similarity [1, 9, 14, 16, 19]. Commonly, these detection approaches employ pairwise sentence comparisons and use the WordNet\(^1\) thesaurus to retrieve semantically related terms for words in the sentences compared. Using the set of exactly matching and related words, the detection approaches derive similarity measures and flag documents as suspicious if the texts exceed a certain similarity threshold.

Other works go beyond comparing word-based semantic similarity by also considering similarity in the argument structure of the sentences [10, 13]. These approaches apply semantic role labeling based on lexical resources such as PropBank\(^2\). Semantic role labeling is an automated process to identify the arguments of a sentence, i.e. the subject, object, events, and relations between these entities using a pre-defined set of roles. The detection approaches then combine the information on semantic arguments with the word-based semantic similarity derived from thesauri such as WordNet or corpora such as Wikipedia. For instance, Osman et al. only consider exactly matching words and WordNet-derived synonyms for the similarity assessment if they belong to the same argument in both sentences [12].

Semantic plagiarism detection approaches have consistently outperformed character-based approaches in terms of detection accuracy. Osman et al. recently showed that their approach achieved a better detection performance in terms of precision, recall, and F-measure than state-of-the-art algorithms from all other classes of detection approaches [12]. However, the computational effort of semantic approaches is significantly higher than that of character-based approaches. For example, Bao et al. showed that considering WordNet synonyms, which exemplify a relatively straightforward semantic analysis, increased processing times by factor 27 compared to character-based approaches [1]. Their computational effort limits the applicability of semantic detection approaches to the detailed comparison phase of the detection process (see Figure 1). Without a suitable reduction in the retrieval space, semantic detection approaches are infeasible to be applied in large collections, and hence unsuitable for most real-world use cases in PD.

To identify disguised plagiarism in real-world, large-scale collections, we developed the Citation-based Plagiarism Detection (CbPD) approach [3, 5]. We use the general term citation-based for approaches that use citations, citation markers, references or combinations thereof for similarity assessment. We define these terms as follows: citation expresses that a document is cited, reference denotes an entry in the bibliography, and citation marker describes a token in the text linking to references in the bibliography.

CbPD uses patterns of citation markers within academic documents as language independent characteristics to identify semantic similarity. Figure 2 depicts the concept of citation pattern analysis for PD. Documents A and B are shown as citing the documents C, D and E. Given their shared references, documents A and B likely discuss semantically similar content. More interestingly, however, they cite the three sources in a similar order. When comparing the citation patterns of documents A and B a pattern agreement of length three results, see gray highlights in Figure 2. Document B simply repeats the citations to document C and D, see dashed lines in Figure 2.

![Figure 2: Citation-based PD concept. Source: [5]](image)

The concept of CbPD thus allows for a document similarity computation even in the absence of a character-based similarity among the documents. Our evaluations of real-world plagiarism cases have shown that plagiarists commonly disguise academic misconduct by paraphrasing copied text, but typically do not change the citations copied from the source document [6]. Subsequently, we demonstrated the practicability of the CbPD approach on a large-scale corpus using the biomedical full-text collection PubMed Central Open Access Subset (PMC OAS), which at the time of analysis contained 185,170 publications. Our analysis identified several instances of previously undiscovered plagiarism. Furthermore, the CbPD algorithms outperformed character-based approaches in ranking more highly the heavily disguised plagiarism forms [7].

3. CITATION-BASED RETRIEVAL FOR PD PURPOSES

Prior research in other fields than PD has shown that citations are valuable indicators for semantic document similarity [2, 4-7, 11, 12].
Given these findings, we propose that analyzing citations for PD purposes is an effective and computationally efficient heuristic to reduce a large reference collection to likely sources of plagiarism. The idea is to consider citation patterns as similarity characteristics and exclude documents below a certain similarity threshold. In a basic case, documents that do not share references can be excluded from the analysis. If documents share references, they are said to be bibliographically coupled [11].

We tested the basic approach of excluding documents that are not bibliographically coupled in our analysis of the PMC OAS described in the last paragraph of Section 2. The PMC OAS contained 185,170 documents. Without reducing the collection, a pairwise comparison of all documents \( n \choose 2 \) would require \( \frac{n(n-1)}{2} = \frac{185170 
choose 2} = 17,143,871,865 \) comparisons, a number that is infeasible to perform in real-world PD use cases. Excluding documents that did not share at least one reference reduced the number of document pairs to 39,463,660, i.e. to 0.23% of the comparisons required for a n:n-analysis. Thus, if one was to compare all documents that share references in the PMC OAS about 39.5 million comparisons were necessary. However, this analysis does not reflect a typical PD use case.

In a typical use case, a single document, or a small number of input documents, for example, papers submitted to a conference, are compared to a much larger reference collection. In this use case, excluding the documents that share no references with the input document(s) reduces the number of necessary comparisons much more than if all bibliographically coupled document pairs in the collection must be compared. Figure 3 illustrates the reduction in document comparisons that a basic citation-based analysis can achieve in a typical plagiarism detection scenario.

The results shown in Figure 4 support our hypothesis and demonstrate the benefit of a citation-based reduction over a character-based reduction of the retrieval space. The sample did not contain a single article pair with an Encoplot score >3 that was not bibliographically coupled. A common threshold for plagiarism suspicion is an Encoplot score of 15 or higher. Thus, using citation-based indicators retains most of the suspicious documents that character-based methods can identify, while also retrieving the suspicious documents with little or no textual similarity that character-based methods often miss.

4. HYBRID PD PROCESS

Given the findings presented in Section 3, we propose a hybrid plagiarism detection approach that integrates citation-based, character-based and semantic detection methods as shown in Figure 5. The hybrid PD approach follows the concept of a three-stage retrieval process as presented in the introduction.

We suggest improving the first heuristic retrieval stage and the second detailed comparison stage of the process as follows. In the first stage, the hybrid approach employs CbPD and character-based heuristics to retrieve candidate documents from the reference collection. Combining the two retrieval heuristics...
allows identifying documents with both literal text overlaps, which may point to copy-and-paste plagiarism and non-lexical semantic similarity, which may point to disguised plagiarism.

Figure 5: Retrieval process for hybrid PD approach.

In the second stage, semantic analysis, string matching, cross-language (CL) PD and approximate citation-based methods can be applied. Semantic detection methods use lexica or Latent Semantic Analysis. Semantic detection methods have been shown to perform well in identifying disguised mono-lingual plagiarism (see Section 2). Elaborate string matching algorithms can identify all textual overlap to identify copy & paste-type plagiarism. Performing extensive machine translation and cross-language PD can improve the identification of translated plagiarism. An approximate citation-based analysis uses similarity measures, such as the Co-Citation-Probability Analysis [4], to identify semantically similar sources and include these documents in the assessment. The goal is to detect citation substitutions, which plagiarists may use to counter detection if citation-based methods find wide-spread adoption. All of the aforementioned detection methods are computationally expensive and only become feasible after the combined heuristics of the first stage reduced the reference collection to a small set of candidate documents.

The hybrid detection process combines the individual strengths of the different detection approaches to enable the identification of disguised plagiarism in real-world settings. Our future work will focus on implementing and evaluating the detection process introduced in this paper as part of our prototype of a hybrid plagiarism detection system citePlag (http://www.citeplag.org).

5. ACKNOWLEDGMENTS

The authors thank the German Academic Exchange Service (DAAD) for its support.

6. REFERENCES

[5] B. Gipp and N. Meuschke, “Citation Pattern Matching Algorithms for Citation-based Plagiarism Detection: Greedy Citation Tiling, Citation Chunking and Longest Common Title Sequence,” in Proc. 11th ACM Symp. on Document Eng. Mountain View, CA, USA: ACM, 2011, pp. 249–258.
Citation for this Paper

Citation Example:

Bibliographic Data:

<table>
<thead>
<tr>
<th>RIS Format</th>
<th>BibTeX Format</th>
</tr>
</thead>
<tbody>
<tr>
<td>TY - CPAPER</td>
<td>@INPROCEEDINGS{Meuschke14,</td>
</tr>
<tr>
<td>AU - Meuschke, Norman</td>
<td>author = {Meuschke, Norman and Gipp, Bela},</td>
</tr>
<tr>
<td>AU - Gipp, Bela</td>
<td>title = {Reducing Computational Effort for Plagiarism</td>
</tr>
<tr>
<td>T1 - Reducing Computational Effort for Plagiarism Detection by using Citation Characteristics to Limit Retrieval Space</td>
<td>Detection by using Citation Characteristics to Limit Retrieval Space},</td>
</tr>
<tr>
<td>Y1 - 2014/september 8-12</td>
<td>year = {2014},</td>
</tr>
<tr>
<td>CY - London, UK</td>
<td>month = sep # { 8-12},</td>
</tr>
<tr>
<td>SP - 197</td>
<td>address = {London, UK},</td>
</tr>
<tr>
<td>EP - 200</td>
<td>pages = {197 - 200},</td>
</tr>
<tr>
<td>PB - IEEE</td>
<td>publisher = {IEEE}},</td>
</tr>
<tr>
<td>DO - 10.1109/JCDL.2014.6970168</td>
<td>doi = {10.1109/JCDL.2014.6970168}</td>
</tr>
</tbody>
</table>