Plagiarism Detection Using Stopword n-grams

Efstathios Stamatatos

Dept. of Information and Communication Systems Eng.
University of the Aegean
83200 – Karlovassi, Greece
stamatatos@aegean.gr

Abstract

In this paper, a novel method for detecting plagiarized passages in document collections is presented. In contrast to previous work in this field that uses content terms to represent documents, the proposed method is based on a small list of stopwords (i.e., very frequent words). We show that stopword *n*-grams reveal important information for plagiarism detection since they are able to capture syntactic similarities between suspicious and original documents and they can be used to detect the exact plagiarized passage boundaries. Experimental results on a publicly-available corpus demonstrate that the performance of the proposed approach is competitive when compared with the best reported results. More importantly, it achieves significantly better results when dealing with difficult plagiarism cases where the plagiarized passages are highly modified and most of the words or phrases have been replaced with synonyms.

This is a preprint of an article published in the *Journal of the American Society for Information Science and Technology*, 62(12), pp. 2512-2527, Wiley, 2011.

1. Introduction

Vấn đề đạo văn ngày càng dễ dàng

According to Hannabuss (2001), plagiarism is the "unauthorized use or close imitation of the ideas and language/expression of someone else and involves representing their work as your own". Given the rapid growth of online publishing of text, the act of plagiarism becomes easier than ever. The problem of plagiarism is particularly evident in journalism (i.e., newspapers, blogs) and academia (i.e., student reports, theses) (Clough, 2003). In such cases significant parts or even entire documents are plagiarized from a single or multiple sources (i.e., patchwork plagiarism). While many plagiarism cases are easy to be found by human readers, the great volumes of suspicious and source texts demand automatic plagiarism detection tools to facilitate this process.

There are several plagiarism types according to the similarity of the plagiarized passage with the source document. The verbatim (aka copy-paste) case regards the direct copying of a passage from a source document. However, in most of the cases, plagiarists attempt to hide the similarity with the original document by modifying the plagiarized passage. This can be done by removing, adding, or replacing words/phrases and rewriting short parts of the passage affecting its syntax. A more difficult case is when the plagiarized and the source document may share the same ideas but the expressions and the language are different. Finally, the plagiarized and source documents may be written in different natural languages. Provided the availability of machine translation tools, this process is facilitated (Potthast, Barrón-Cedeño, Stein, & Rosso, 2011).

đạo văn ngày càng tinh vi

Automatic plagiarism detection comprises several tasks. The default scenario (aka *external plagiarism detection*) regards the identification of passages in suspicious documents as likely plagiarized and associate these passages with certain passages of source documents in a given reference collection. *Intrinsic plagiarism detection* considers the case where no reference collection is available and the likely plagiarized passages in a suspicious document have to be extracted based on stylistic inconsistencies (Stamatatos, 2009). This task has many similarities with the authorship verification problem (Stein, Lipka, & Prettenhofer, 2011). *Cross-lingual plagiarism detection* deals with the case where the suspicious and source documents are written in different natural languages (Potthast, Barrón-Cedeño, Stein, & Rosso, 2011). *Text reuse* or *near-duplicate detection* is associated with plagiarism detection since it attempts to find documents that share most of their content and are derivatives of an original source (Hoad & Zobel, 2003; Bendersky & Croft, 2009). However, it examines similarity on the document level. *Local text reuse* or *partial-duplicate detection* is closer to plagiarism detection where a very short passage may be copied in a long document (Seo & Croft, 2008; Zhang, Zhang, Yu, & Huang, 2010). In this task, the similarity is considered legitimate, so usually there is no attempt to hide it. As a result, it resembles the verbatim case of plagiarism detection.

One major issue in plagiarism detection is efficiency (Schleimer, Wilkerson, & Aiken, 2003; Stein & Meyer zu Eissen, 2006). The suspicious documents should be compared with any document in the reference collection which may be very large (i.e., the whole indexed Web). Therefore, similarity estimation between a pair of documents should be based on simple measures. Additionally, they should be able to capture local similarities where only a likely short passage is common in both documents. Given that the plagiarized and the original passages may not be exactly the same in case the plagiarist performed some kind of paraphrasing, the information used to represent texts should capture the similarity even when most of the words and word ordering are different (Gustafson, Pera, & Ng, 2008). Existing approaches in plagiarism detection are based on sequences of words or characters to represent texts (Schleimer, et al., 2003; Lyon, Malcolm, & Dickerson, 2001; Barrón-Cedeño & Rosso, 2009; Hoad & Zobel, 2003). Since the content information is considered more important, very frequent words conveying no meaning (i.e., stopwords) are usually excluded (Gustafson, et al., 2008; Hoad & Zobel, 2003; Chowdhury, Frieder, Grossman, & McCabe, 2002; Potthast,

Tính hiệu quả
Vì đạo văn thường viết lại ý, nên khi phát hiện đạo văn, các từ không truyền đạt nội dùng cần được loại bỏ như the, is, an... để tăng performancce.

Barrón-Cedeño, Eiselt, Stein, & Rosso, 2010) or used to identify the position of important content terms (Theobald, Siddharth, & Paepcke, 2008).

It is a common practice in Information Retrieval (IR) to discard stopwords since they increase the size of index with many postings, corresponding to their appearances in documents. According to the *rule of 30*, the 30 most common words account for (roughly) 30% of the word tokens in a corpus (Manning, Raghavan, & Schütze, 2008). However, efficient index compression methods can considerably decrease the size required by these postings. Moreover, the elimination of stopwords makes phrase queries more difficult or even impossible to be processed. As a result, modern IR systems, including many Web search engines, adopt full-text indexing (Manning, et al., 2008). Stopwords have been proved to be extremely useful in text mining tasks including authorship attribution (Arun, Suresh, & Madhavan, 2009) and text-genre detection (Stamatatos, Fakotakis, & Kokkinakis, 2000) where the aim is to represent style rather than content. In plagiarism detection, it has been demonstrated that stopword removal considerably hurts the performance (Ceska, & Fox, 2009).

Việc loại bỏ các stopword làm giảm đáng kể performance. phản biện cái trên

In this paper, we propose a novel plagiarism detection method that takes full advantage of stopword occurrences in texts. Instead of following the common practice of eliminating stopwords, the proposed method eliminates all the other tokens and is entirely based on the remaining stopword sequences. Therefore, it is a method based exclusively on structural information rather than content information. We show that stopword *n*-grams are able to capture syntactic similarities between suspicious and original documents and they can be used to detect the plagiarized passage boundaries. Results on a publicly-available corpus demonstrate that the performance of the proposed approach is competitive when compared with the best reported results on the same corpus. More importantly, our method achieves significantly better results when dealing with difficult plagiarism cases where the plagiarized passages are highly modified and most of the words or phrases have been replaced with synonyms.

đề xuất phương pháp dùng stopword, phát hiện tương tự cấu trúc, hiệu quả khi ng đạo dùng từ đồng nghĩa để thay thế.

The rest of this paper is organized as follows. The next section describes previous related work. Section 3 presents the proposed method in detail. The experimental settings and results are included in Section 4 while the conclusions drawn from this study and suggested future work directions are given in Section 5.

2. Related Work

The majority of approaches to plagiarism detection adopt the same architecture (Potthast, Barrón-Cedeño, Eiselt, Stein, & Rosso, 2010). First, to improve efficiency in large document collections, for each suspicious document a small set of candidate source documents is retrieved. This set is either of predefined or variable size according to the similarity between the documents. Then, a more detailed analysis between the suspicious document and each of the retrieved documents provides the requested passage boundaries. Finally, a post-processing step checks these detections and removes or merges some of them.

In order to detect the degree of similarity between documents, two basic approaches have been proposed. The first follows the typical IR methodology that considers the suspicious document (or parts of the document) as a query and attempts to rank documents in the reference collection according to their similarity with the query (Shivakumar & Garcia-Molina, 1995; Hoad & Zobel, 2003; Gustafson, et al., 2008; Muhr, Kern, Zechner, & Granitzer, 2010). The similarity measures take into account relative word frequencies, document frequencies, and document lengths (Metzler, Bernstein, Croft, Moffat, & Zobel, 2005) while stopwords are usually discarded (Hoad & Zobel, 2003; Gustafson, et al., 2008). To take into account word substitutions by synonyms Gustafson et al. (2008) proposes the use of word-correlation factors that measure frequency of co-occurrence and relative distance between pairs of terms in Wikipedia documents. The syntactic structure of sentences is more robust in cases of paraphrasing the plagiarized passages (Uzuner, Katz, & Nahnsen, 2005) but the required syntactic analysis considerably harms the efficiency.

The second basic family of approaches relies on document fingerprints comprising hashes of fixed-length chunks (aka *shingles*) in documents (Brin, Davis, & Garcia-Molina, 1995; Lyon, et al., 2001; Seo & Croft, 2008; Schleimer, et al., 2003; Stein & Meyer zu Eissen, 2006). Either the complete set of chunks can be included in the document fingerprint (full fingerprinting) to optimize effectiveness or a chunk selection method can be applied to decrease storage requirements and optimize efficiency (Schleimer, et al., 2003). Some approaches define chunks so that to capture information about the content and the structure of a short piece of text. Usually they are character *n*-grams (Schleimer, et al., 2003), word *n*-grams (Lyon, et al., 2001; Barrón-Cedeño & Rosso, 2009) or sentences (Gustafson, et al., 2008; Zhang, et al., 2010). Word *n*-grams can be sorted to be more flexible in small changes between the plagiarized and the source passages, e.g., the phrases ‘plagiarism detection in documents’ and ‘detection of plagiarism in documents’ share the same sorted word 3-gram after the removal of short words (Kasprzak, & Brandejs, 2010). Theobald, et al., (2008) use stopword positions to identify useful chains of content words in web pages. Chowdhury, et al. (2002) eliminate stopwords and infrequently occurring terms and considers a single chunk comprising the remaining content words. In contrast, Basile, Benedetto, Caglioti, Cristadoro, & Esposti (2009) consider chunks that are based exclusively on structural information (i.e., word-length sequences).

Provided a suspicious document is found to be similar with a source document, a scatter plot of the positions of all the matches found between the two documents can reveal the approximate passage boundaries (Zhang, et al., 2010; Zou, Long, & Ling, 2010). This resembles the detection of similarity in DNA sequences (Church & Helfman, 1993) and the procedure of mapping *bitexts*, i.e., texts available in two languages (Melamed, 1999). In case of verbatim plagiarism or partial-duplicate detection, these passages will be straight diagonal lines in the scatter plot. To detect such passage boundaries, algorithms for finding diagonals of maximal length are appropriate (Zhang, et al., 2010). However, in cases when the plagiarized passage is modified there is noise in the diagonal lines. Essentially, a cluster of matches is produced and it is usual to have small gaps between adjacent areas that correspond to the same passage. To solve this problem, several methods have been proposed including sets of heuristic rules to identify and merge adjacent passages (Kasprzak, & Brandejs, 2010; Basile, et al., 2009; Kolak & Schilit, 2008), Monte Carlo optimization to join adjacent matches (Grozea, Gehl, & Popescu, 2009), and application of clustering methods (Zou, et al., 2010). Although this kind of analysis has to be performed for relatively few source documents per suspicious document, it can harm the efficiency of the approach when its computational cost is high.

After the detection of passage boundaries, the post-processing step is used to filter the passage detections and eliminate or merge cases of short passages and overlapping or ambiguous (e.g., indicating the same plagiarized passage and different source passages) detections (Kasprzak, & Brandejs, 2010; Mhur, et al., 2010; Zou, et al., 2010; Kolak & Schilit, 2008). A final verification of similarity between the passages in the suspicious and the source documents has also been proposed (Muhr, et al., 2010). The post-processing step is especially important for improving the precision of the plagiarism detection methods.

Recently, two competitions on plagiarism detection were organized addressing several plagiarism types, including external plagiarism, intrinsic plagiarism, and cross-lingual plagiarism (Potthast, Stein, Eiselt, Barrón-Cedeño, & Rosso, 2009; Potthast, Barrón-Cedeño, Eiselt, Stein, & Rosso, 2010). Evaluation corpora and methodologies have been released (Potthast, Stein, Barrón-Cedeño, & Rosso, 2010) providing the possibility to compare different approaches on the same testing ground. The focus of the evaluation in these competitions is on the exact detection of passage boundaries in plagiarized and source documents. Although the majority of the participants eliminated stopwords to increase the

TABLE 1. A summary of the properties of the four top-performing PAN-10 methods.

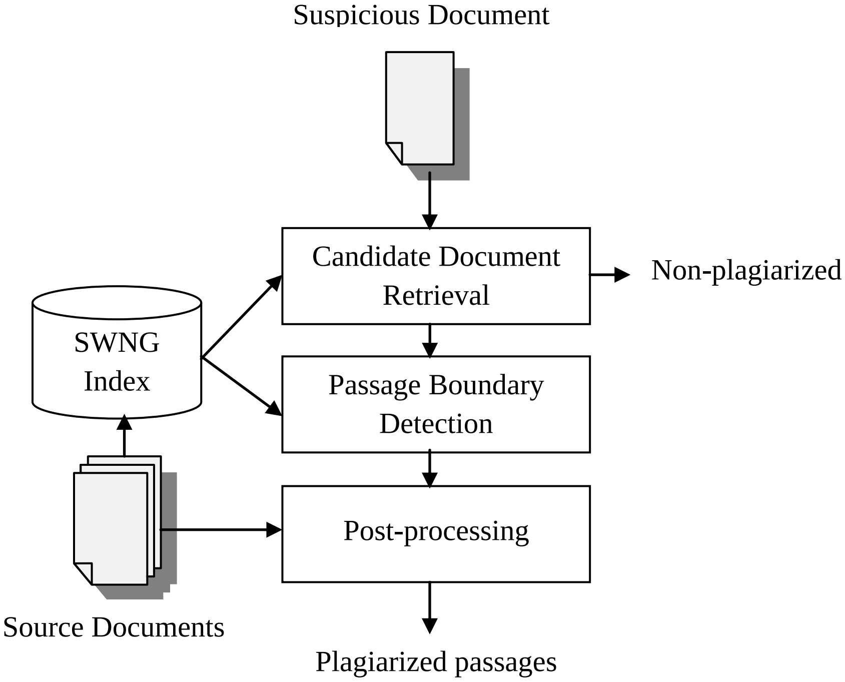
	Representation	Candidate Document Retrieval	Passage Boundary Detection	Post-Processing
PAN-10-1 (Kasprzak & Brandejs, 2010)	Chunks of sorted word 5-grams (short words are excluded)	Similarity threshold (20 common chunks)	Heuristic rules	Heuristics for removing short or overlapping detections
PAN-10-2 (Zou, et al., 2010)	Word 5-grams	Winnowing	Clustering	Heuristics for merging detections
PAN-10-3 (Muhr, et al., 2010)	Overlapping blocks of 40 tokens	Boolean queries and heuristic rules	Word sequence analysis and heuristic rules	Heuristics for merging detections, final check (Jaccard similarity)
PAN-10-4 (Grozea, et al., 2009)	Character 16-grams	Similarity estimation using a kernel function	Monte Carlo optimization	Heuristics for removing short or imbalanced detections

efficiency of document representation, the winning methods avoided explicit removal of stopwords. The winner of the 2009 competition used character 16-grams (Grozea, et al., 2009) while the winner of the 2010 competition used (sorted) word 5-grams including all words with at least three characters (Kasprzak, & Brandejs, 2010).

Table 1 presents a summary of the properties of the four top-performing methods of the 2010 competition. The four participants are denoted as: PAN-10-1 (Kasprzak & Brandejs, 2010), PAN-10-2 (Zou, et al., 2010), PAN-10-3 (Muhr, et al., 2010), and PAN-10-4 (Grozea, et al., 2009). The latter was the winner of the 2009 competition using the same method in both competitions. PAN-10-1 is based on chunks of sorted word 5-grams after the removal of short words (less than 3 characters). MD5 hashes are produced to index these chunks. The candidate documents are retrieved according to the number of chunks they have in common with the suspicious document. At least 20 common chunks are required without caring about their position in the document. Therefore, long source documents are likely to join the candidate set for many suspicious documents. Then, for each candidate document an evaluation of similar passages with the suspicious document is performed based on heuristic rules (allowing some gaps between the matched chunks). In the post-processing step, short (less than 600 characters) overlapping detections are removed. PAN-10-2 is based on word 5-grams and the winnowing method (Schleimer, et al., 2003) to select the fingerprints of each document. Then, the candidate documents are retrieved according to the number of their successive same fingerprints with the suspicious document. In each candidate document of a suspicious document, the longest common substring algorithm is used to merge common substrings and then a clustering algorithm is used to detect the passage boundaries. Finally, a set of heuristic rules is applied to the detected passages in order to handle merging errors.

PAN-10-3 follows the traditional IR model. It first segments the source documents into overlapping blocks of 40 tokens and indexes them. Then, each suspicious document is also transformed into a set of blocks and Boolean queries in combination with some heuristic rules are used to retrieve the candidate documents with high similarity to the suspicious documents.

FIG 1. Overview of the presented method.



This approach is better able to capture the similarity in modified plagiarized passages. For each candidate document, the matched blocks are enlarged with neighboring word sequences according to heuristic rules. In the post-processing step, neighboring detections are merged using heuristics and a final check is performed based on the Jaccard similarity where detections shorter than 5,000 characters are removed given that their similarity is less than 0.55 while for longer detections a similarity score of at least 0.7 is required. PAN-10-4 is based on character *n*-grams (16-grams) which produce a detailed representation of texts. Then, a linear kernel function is used to calculate the similarity between a suspicious document and each of the source documents. To select the candidate documents, it is possible to sort source documents under each suspicious document and get the most similar ones. However, Grozea, et al., (2009) propose the opposite: sort the suspicious document under each source document according to their similarity and get a fixed number (51) of the most similar ones for each source document. This method produces many candidate documents and heavily depends on the size of the source document set. For each pair of suspicious-source document a Monte-Carlo optimization procedure is called to find the largest group of matches that correspond to the detected passages. In the post-processing step, short detections (less than 256 characters) or imbalanced detections (the absolute size difference is less than half of the mean) are removed.

3. The Proposed Method

In this study, we deal with monolingual plagiarism detection. Let D_x be a set of suspicious documents we want to examine and D_s be the set of source documents. The first task is to decide whether or not a suspicious document is plagiarized or non-plagiarized. In the former case, all the sources of plagiarism should be identified including a subset of D_s and the exact boundaries of the plagiarized passages in both the suspicious and source documents. Furthermore, it is desirable to assign a score to each detected plagiarized passage to indicate the degree of plagiarism. This score can be used to sort the detected passages from exact copies to somehow related passages. The architecture of the presented method is depicted in Figure 1 and follows the state-of-the-art in this field (Potthast, Stein, Barrón-Cedeño, & Rosso, 2010).

TABLE 2. The list of 50 most frequent words of BNC corpus used in this study.

1. the	11. with	21. are	31. or	41. her
2. of	12. he	22. not	32. an	42. n’t
3. and	13. be	23. his	33. were	43. there
4. a	14. on	24. this	34. we	44. can
5. in	15. i	25. from	35. their	45. all
6. to	16. that	26. but	36. been	46. as
7. is	17. by	27. had	37. has	47. if
8. was	18. at	28. which	38. have	48. who
9. it	19. you	29. she	39. will	49. what
10. for	20. ‘s	30. they	40. would	50. said

These savage birds are very common in Maine, where they make great havoc among the flocks of wild-ducks and Canada grouse, and will even, when driven by hunger, venture an attack on the fowls of the farm-yard.

(a) A text passage

are in they the of and and will by an on the of the

(b) The text after removing all tokens not found in the stopword list

[are, in, they, the, of, and, and, will]
[in, they, the, of, and, and, will, by]
[they, the, of, and, and, will, by, an]
[the, of, and, and, will, by, an, on]
[of, and, and, will, by, an, on, the]
[and, and, will, by, an, on, the, of]
[and, will, by, an, on, the, of, the]

(c) The stopword 8-grams of the text

FIG 2. An example of transforming a text to stopword *n*-grams.

3.1 Text Representation

The representation of texts according to the proposed method is based on stopword *n*-grams (SWNG). Given a document and a list of stopwords, the text is reduced to the appearances of the stopwords in the document. All the other tokens are discarded. As stopwords, in this study, we use a list of the 50 most frequent words of the English language provided by the British National Corpus which includes about 90 millions tokens. This list is shown in Table 2 and has also been used previously for text genre detection (Stamatatos, et al., 2000). Therefore, a text is first transformed to lowercase, then it is tokenized and all the tokens not belonging to the list of stopwords are removed. Finally, the *n*-grams of the remaining stopwords are produced. We call this set of SWNGs the *profile* of the document. Given a document *d*, the profile *P*(*n*,*d*) comprises all the stopword *n*-grams, i.e., analogous to the full-fingerprinting method (Hoad & Zobel, 2003). The SWNGs in *P*(*n*,*d*) are ordered according to their first appearance in the document. The procedure of transforming a text passage to a set of stopword *n*-grams is demonstrated in Figure 2.

Với 1 tài liệu được cung cấp, các từ khóa sẽ bị loại bỏ, chỉ giữ lại các stopword, sau đó sinh ra n-gram (xếp các stopword đấy theo thứ tự đầu tiên xuất hiện)

The intuition behind this representation is that stopword occurrences are usually associated with syntactic patterns. Therefore, sequences of stopwords reveal hints of the syntactic structure of the document that is likely to remain stable during the procedure of plagiarizing a

This came into existence likely from the deviance in the time-period of the particular billet. As the premier is to be nominated for not more than a period of four years, it can infrequently happen that an ample wage, fixed at the embarkation of that period, will not endure to be such to its end.

(a) The plagiarized passage.

This probably arose from the difference in the duration of the respective offices. As the President is to be elected for no more than four years, it can rarely happen that an adequate salary, fixed at the commencement of that period, will not continue to be such to its end.

(b) The original passage.

FIG 3. An example of a difficult plagiarism case where stopword *n*-grams capture the similarity between the plagiarized and the original texts.

passage. That is, when one attempts to plagiarize a particular passage of text and wants to cover their traits, the most usual act is to replace words and phrases with synonyms. It is much more difficult to change the basic syntactic structure or rewrite large parts of the text. Stopwords are function words, that is they are content-independent and they do not convey any semantic information. They can usually be removed/replaced when the syntactic structure changes. According to the terminology introduced in the work of Koppel, Akiva, and Dagan (2006), a language element (i.e., a word or a syntactic structure) is *unstable* when it can be replaced by other semantically equivalent elements. *Stability* of words can be regarded as the availability of synonyms. Given that definition, stopwords are words with high stability and, therefore, are likely to remain intact when someone attempts to slightly modify a text passage. In case the modification does not involve significant reordering of contents, long sequences of stopwords of the original passage are likely to also be included in the modified passage. Moreover, language diversity and language errors especially when the authors are non-native speakers can affect the stability of words. For example, the tokens ‘plagiarize’, ‘plagiarise’, ‘pladgiarize’, and ‘plagarize’ are some different (correct or erroneous) versions of the same content word. On the other hand, most speakers of the language are familiar with stopwords and since they are relatively short, they are less likely to contain errors.

người đạo văn thường sử dụng các từ đồng nghĩa để thay thế, con stopword n-gram chứa cấu trúc của đoạn văn đó. Mà cấu trúc thì rất khó bị thay đổi, các stopword có tính ổn định rất cao (khó bị thay thế bởi từ đồng nghĩa)

The stability of stopwords is demonstrated in the example of Figure 3 where an original piece of text and a plagiarized version of it are given. Despite the fact that the plagiarized version is highly modified, most of the sequences of our list of 50 stopwords remain the same with those of the original document (the original and the plagiarized passage have 18 common 5-grams, 12 common 8-grams, and 6 common 11-grams of stopwords). This similarity is affected only in the case the plagiarist rewrites significant part of the passage. On the other hand, texts that are not associated are unusual to share long sequences of stopwords since that would mean they share the same syntactic structure in consecutive sentences or entire paragraphs.

Nói về ví dụ hình 3

To verify that such coincidental similarity of SWNGs is rare, the Reuters Corpus Volume 1 (RCV1)[1] was used. This corpus contains over 800,000 newswire stories produced between August 20, 1996 and August 19, 1997. According to Khmelev & Teahan (2003) a significant proportion of the RCV1 articles are either exact duplicates (3.4%) or extensively plagiarized (7.9%). There are also multiple cases where two unrelated documents share some standardized sentences, such as *‘The following are top headlines from selected Canadian newspapers. Reuters has not verified these stories and does not vouch for their accuracy.’* Unfortunately, there is no available annotation of plagiarism cases in this corpus.

Minh chứng về việc trùng ngẫu nhiên SWNGs là rất hiếm. 800000 câu chuyện thì trùng ngẫu nhiên là 3.4%, đạo là 7.9%

[1] http://trec.nist.gov/data/reuters/reuters.html

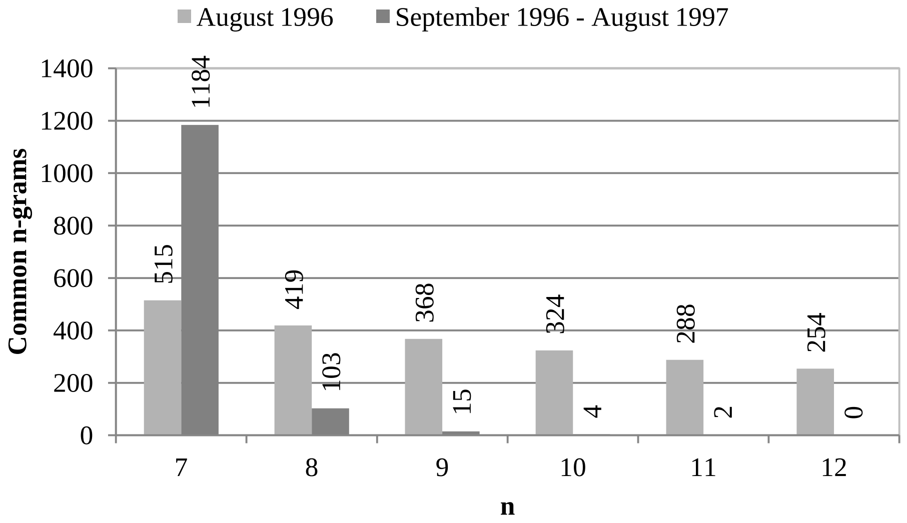


FIG 4. Common stopword *n*-grams between 10 RCV1 stories published in August 20, 1996 and other stories published in either the same month or from September 1996 till August 1997.

Nevertheless, it is expected for the plagiarism cases to appear in newswire stories produced in the same day or within a short period (i.e., a week) from the publication of the first version. To take advantage of this, a set of 10 RCV1 stories all published on August 20, 1996 were selected. None of these texts include standardized sentences like the ones mentioned above. Then, the stopword *n*-grams of these texts were extracted and compared with the stopword *n*-grams of all the other texts published in August 1996 (23,297 stories). Figure 4 shows the number of common *n*-grams found for varying values of *n*. It is evident that when *n* increases the number of common *n*-grams slightly decreases indicating that there are significant similarities in some documents (i.e., likely plagiarized cases since they are produced in the same month). This experiment was repeated this time comparing the 10 selected texts from August 20, 1996 with all the RCV1 texts published from September 1, 1996 till August 19, 1997 (783,484 stories). The number of common *n*-grams is also given in Figure 4. It is obvious that when *n* increases, the number of common *n*-grams is drastically reduced indicating that there is no plagiarism case. Note that there is not a single match for *n*-grams longer than 11 despite the very large volume of texts.

3.2 Candidate Document Retrieval

As shown in Figure 1, the first important step is to retrieve a subset of D_s that comprises the sources of likely plagiarism in a suspicious document. This procedure includes the comparison of the suspicious document with any member of D_s to identify any local similarities. It is not known a priori what the number of source documents is for each suspicious document. It could be none, a single, or multiple source documents. The most important issue here is to achieve a high recall since it is just the first step in the detection process and any source document missed will no further examined. A low precision will affect the efficiency of the subsequent steps.

Trích xuất tập tài liệu nguồn

Given the SWNG representation, our aim is to find common *n*-grams of stopwords between the suspicious and the source documents. The main question here regards the definition of an appropriate value of *n*. That is how long the sequences of stopwords should be so that to detect a similarity between a suspicious and a source document. Let n_1 be this value. Any common *n*-gram between a pair of documents with $n<n_1$ is considered not significant. A common *n*-gram with $n>=n_1$ suggests a match that is not coincidental. In that sense, the value of n_1 should be relatively high (see Figure 4). On the other hand, beyond the case of verbatim

The minutes of the committee record the motion of appreciation to the owners. Mr. Robert Bell of the old printing firm of that name made…

…the Fathers of the Church; the aesthetic mysticism of Plotinus, reborn to its greatest triumphs, during the classic period of German thought. Through the midst of these variously erroneous theories, that traverse…

FIG 5. Two unrelated text passages with the same sequence of stopwords.

Minh họa trường hợp đặc biệt có cùng n-gram nhưng méo liên quan nhau

plagiarism, when the plagiarized passages have been highly modified, we should not expect to find too long common sequences of stopwords. In those cases, a high value of n_1 would miss source documents including the originals of either short or highly modified plagiarized passages. Therefore, there is a trade-off between low and high values of n_1 for the candidate document retrieval task.

Mục tiêu tìm ra n1-gram tương đồng giữa input và nguồn, kì cái n-gram có n<n1 thì có thể coi là không đạo, vì cấu trúc câu ngắn nên việc giống nhau là bình thường, chỉ có n>=n1 thì thì bị xem là sao chép. Nhưng có trường hợp người đạo tinh vi có thể sao chép những cấu trúc câu ngắn. Vì vậy, việc chọn n1 hợp lí, ko nhỏ để nhầm và cũng không lớn để bỏ xót là rất quan trọng

One common case of coincidental similarity between the sequences of stopwords of unrelated documents is when the sequence contains only specific, very frequent stopwords. These words are the first 6 most frequent stopwords (*the*, *of*, *and*, *a*, *in*, *to*) plus *'s*. An example is shown in Figure 5, where two unrelated text passages have exactly the same sequence of stopwords (11-gram). Such cases considerably increase the false positives of our approach. To avoid them, we need an additional constraint on the contents of the common *n*-grams found in the profiles of two documents. This constraint should not be too rigid so that similarities of short plagiarized passages are not filtered out. Let C={*the*, *of*, *and*, *a*, *in*, *to*, *'s*} be the set of the stopwords usually appear in coincidental matches. Let $d_x \in D_x$ and $d_s \in D_s$ while $P(n_1, d_x)$ and $P(n_1, d_s)$ are the corresponding profiles of these documents comprising SWNGs of length n_1. A match between these documents is detected when the following criterion is satisfied:

$$\exists g \in P(n_1,d_x) \cap P(n_1,d_s): member(g,C) < n_1\text{-}1 \wedge maxseq(g,C) < n_1\text{-}2 \qquad (1)$$

where the functions $member(g,C)$ and $maxseq(g,C)$ return the number of stopwords of the *n*-gram g that belong to C and the maximal sequence of words of g that belong to C, respectively. In other words, when n_1=11, if a match of a common 11-gram is detected in the profiles of a suspicious and a source document, it would indicate a possible plagiarism case given that g contains at least 2 stopwords not belonging to C (i.e., $member(g,C)<10$) and the maximal sequence in g of stopwords belonging to C is less than 9. Note that the example of Figure 5 fails to satisfy both of these constraints since $member(g,C)=10$ and $maxseq(g,C)=10$.

có một số n-gram phổ biến vị trùng ngẫu nhiên giồm (the, of, and, a, in, to) + 's. Để tránh bị trùng thì ta có thêm ràng buộc cho các n-gram. Như ban đầu, 2 đoạn văn bị coi là đạo nếu có cũng n-gram. Siết chặt hơn, trong n-gram số lượng các stopword phổ biến nhất phải xuất hiện vs tần số <n-1 và 1 dãy liên tiếp các stopword phổ biến dài nhất có thể phải có độ dài <n-2 thì mới coi là đạo.

Figure 6 depicts the amount of common *n*-grams in a collection of 1,000 documents without any known case of plagiarism before and after the application of the criterion (1). The document length in this collection varies from 3,000 to 2.5 million characters. Apparently, this criterion significantly reduces the amount of common *n*-grams. Figure 7 shows the percentage of document pairs in this collection retrieved based on the criterion (1) for varying *n*-gram length. In the case of 11-grams less than 0.1% of the possible document pairs are retrieved. Note that there are many cases where two documents may share the same passage (e.g. famous quotations) (Kolak & Schilit, 2008). So, some document pairs are likely to be detected in any collection. We discuss this further in Section 3.4.

3.3 Passage Boundary Detection

In case we find a set of source documents that match a suspicious document, the next step is to perform a more detailed analysis to estimate the exact boundaries of plagiarized passages in both the suspicious and the source documents. Let $D_{rx} \subseteq D_s$ denote the set of source documents that have been retrieved for the suspicious document d_x. Our aim is to find the common SWNGs between the profiles of d_x and each $d_s \in D_{rx}$ and build maximal sequences of them that correspond to text passages.

Sau khi có tập nguồn rồi, chúng ta phải tìm đường biên,...

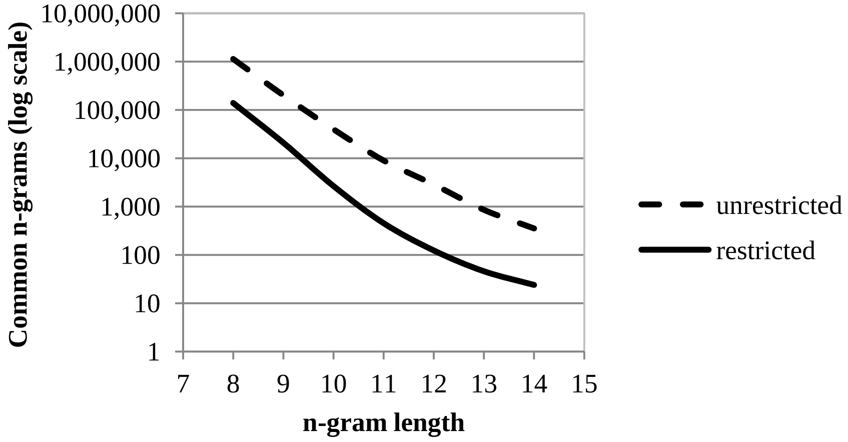


FIG 6. Amount of common *n*-grams in a collection of 1,000 documents without any known case of plagiarism before and after applying criterion (1).

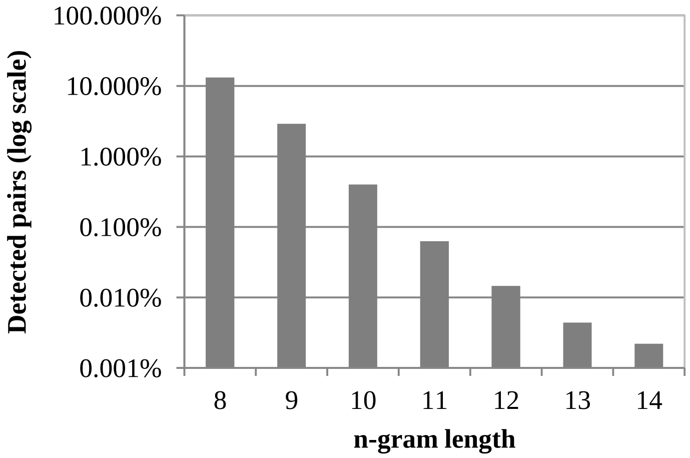


FIG 7. Percentage of detected document pairs for varying *n*-gram length in a collection of 1,000 documents without any known case of plagiarism.

In case the plagiarized passage is an exact copy of the source document, the task is quite easy since the same sequence of SWNGs will be included in both profiles in the same order. Then, the scatter plot showing the matches between a suspicious and source document will be composed of diagonal lines. An example of verbatim plagiarism is given in Figure 8. However, when the plagiarized passage is highly modified there will be considerable noise and gaps between common SWNGs of the two profiles. An example is given in Figure 9. The amount of noise and gaps depends on the value of *n* (order of *n*-grams) used in producing the profiles of the documents. The higher *n* is, the more gaps and noise will appear. Therefore, the long n_1-grams used to identify similarity between documents in the previous step are not appropriate in the current step. We need shorter *n*-grams (of order $n_2<n_1$) so that more detailed matches between the documents to be captured. In order to avoid noise of coincidental matches of SWNGs due to *n*-grams containing only stopwords of *C*, we also need a criterion similar to (1) to exclude some uninformative SWNGs. However, to keep the gaps between common SWNGs low, this criterion should be more relaxed in comparison to (1). Let $P(n_2, d_x)$ and $P(n_2, d_s)$ be the profiles of the suspicious and source documents comprising stopword n_2-grams. A n_2-gram *g* is a match between these documents when the following criterion is satisfied:

Trong trường hợp copy toàn văn thì một chuỗi SWNGs sẽ có cùng thứ tự, đem phân tích sẽ ra những đường chéo. Nếu đạo cao hơn thì sẽ có nhiễu, việc chọn n cao sẽ tạo ra nhiễu và khoảng trống trong đồ thị, nên phải chọn n ngắn hơn. Để tránh nhiễu do trùng vs các SWNGs ngẫu nhiên thì cần (1) nhưng lỏng hơn

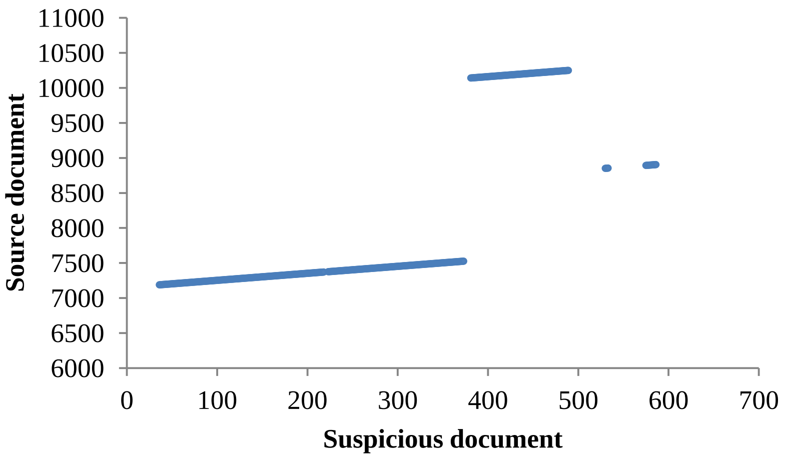


FIG 8. Scatter plot of the matched *n*-grams in verbatim plagiarism cases where the plagiarized passages are next to each other.

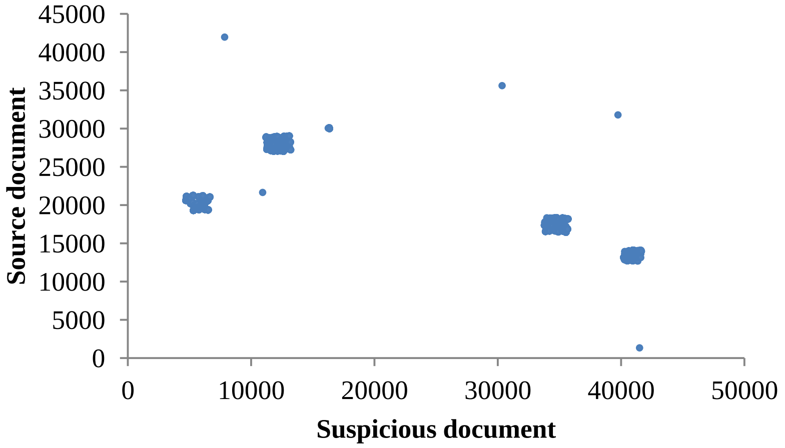


FIG 9. Scatter plot of the matched *n*-grams in cases where the plagiarized passage is significantly modified.

g là 1 n2-gram

$$g \in P(n_2,d_x) \cap P(n_2,d_s) \wedge member(g,C) < n_2 \qquad (2)$$

where the function *member*(*g*,*C*) returns the number of stopwords of the *n*-gram *g* that belong to *C*. Let $M(d_x,d_s)$ be the set of the matched *n*-grams between the profiles $P(n_2,d_x)$ and $P(n_2,d_s)$ of the suspicious and the source documents. Members of $M(d_x,d_s)$ are ordered according to the first appearance of a match in the suspicious document. For example, in the case of the text passages of Figure 3, the ordered set of matches between the 8-grams of the plagiarized and the original passages is the following:

$$M(d_x,d_s)=\{(1,1), (2,2), (3,3), (4,4), (5,5), (6,6), (17,14), (18,15), (19,16), (20,17), (21,18), (22,19)\}$$

that is, the first 8-gram of the plagiarized passage is identical with the first 8-gram of the original document, the 17th 8-gram of the plagiarized passage is identical with the 14th 8-gram of the original document, etc. Moreover, let M_1 and M_2 be the parts of M that correspond to the suspicious document and the source document, respectively. Therefore, consecutive M_1 values always increase while consecutive M_2 values may decrease as well. As shown in Figures 8 and 9 (scatter plots of M_1 vs. M_2) the boundaries of plagiarized passages are associated with big changes in consecutive values of M_1 and M_2. However, if these changes are not big enough they may correspond to gaps in noisy cases where the plagiarized passage is heavily modified.

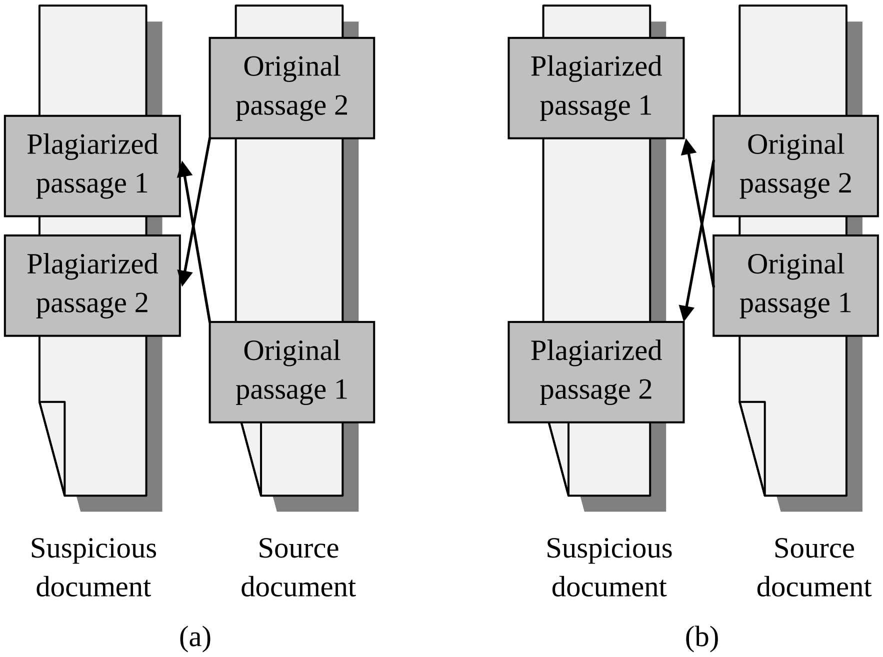


FIG 10. Examples of plagiarism cases with multiple passages in the same document. (a) Neighboring passages in the suspicious document. (b) Neighboring passages in the source document.

Another important problem in this task is when there are multiple plagiarized passages in a suspicious document and the distance between them is relatively low. This case is depicted in Figure 10a, where the distance (in characters) between the plagiarized passages in the suspicious document is too low. Note that this is not necessarily related with the distance of the original passages in the source document. This is also the case in the example of Figure 8 where the plagiarized passages are next to each other in the suspicious document (*x*-dimension). Similarly, two original passages in the same source document can be close enough (as depicted in Figure 10b) while the distance between the corresponding plagiarized passages in the suspicious document may be high. To handle this problem in the detection of passage boundaries, we propose the following procedure. First, an initial set of passage boundaries of maximal length is detected in the suspicious document allowing small gaps to be included. Then, the corresponding passages in the source document are examined. In case a passage in the source document is not homogeneous (i.e., comprises parts of the document with significant gaps between them) it splits into smaller passages. Finally, the passage boundaries in the suspicious document are determined based on these smaller passages of the source document.

Các đoạn văn bị ăn cắp nằm cạnh nhau trong tài liệu trong tài liệu copy H10a, H10b nói các đoạn văn ăn cấp nằm cách xa nhau. Phương pháp đề xuất:

In more detail, the initial set of passage boundaries in the suspicious document is detected according to the following criterion:

$$m_i \in M_1(d_x,d_s): abs(diff(m_i)) > \theta_g \qquad (3)$$

where the functions *abs* and *diff* return the absolute value and the difference (derivative) and θ_g is a threshold that permits relatively small gaps to be included in the detected passage. If there are adjacent boundaries, they are joined to a single boundary. Each detected passage in the suspicious document (a subset of M_1 values) corresponds to a subset of M_2 values. However, a subset of M_2 values may correspond to different passages of the original document (i.e., the case depicted in Figure 10a). Then, each $M_{2i} \subseteq M_2$, corresponding to a maximal subset of a detected passage in M_1 values, is examined to detect maximal passages of the original document. The boundaries of the source document passages are detected according to the following criterion:

$$m_i \in M_{2i}(d_x,d_s): abs(diff(m_i)) > \theta_g \quad (4)$$

where $M_{2i}(d_x,d_s)$ is a subset of M_2 that corresponds to an already detected plagiarized passage in the suspicious document. Gaps lower than θ_g are allowed in a passage. Again, if there are adjacent boundaries, they are joined to a single boundary. Finally, in case multiple passages are detected in the original document, the corresponding passage in the suspicious document is split accordingly to produce the final boundaries of the plagiarized passages. Note that this procedure detects boundaries in the sequence of *n*-grams. Let $<S_i, E_i>$ be the start and ending *n*-gram boundaries of a detected passage. These can be transformed into character boundaries by taking the position of the first character of the first word of S_i and the position of the last character of the last word of E_i.

In the example of Figure 8, a single passage <36,586> is initially detected in M_1 (i.e., the *x*-dimension) according to criterion (3). Then, the initial M_2 subset <7189,10250> (*y*-dimension) corresponding to the single passage detected in M_1 is divided into three passages <7189,7525>, <8852,8905>, and <10142,10250> according to criterion (4). Finally, using these passages of the source document, three plagiarized passages are formed in the suspicious document, namely <36,373>, <530,586>, and <381,489>. Note that the second detected passage incorporates the small gap depicted in Figure 8.

3.4 Post-processing

The procedure described so far, is based on SWNG representation and disregards all the words of the text not belonging to the set of the 50 stopwords. The detections obtained, especially in case they are short, should be checked to verify that the similarity of the detected plagiarized passage with the detected original passage is high, when the full text of the passages is taken into account. Moreover, we need a mechanism to assign scores to the detected plagiarism cases according to the degree of similarity with the original passages. This procedure should not be computationally expensive since it will be applied to full text of multiple passages. In addition, it should be flexible so that to capture the similarity even in cases where the plagiarized passage is highly modified and contains many different words with respect to the original passage (i.e., the case of Figure 3).

Each detection is a 4-tuple $<t_x, d_x, t_s, d_s>$ that associates a plagiarized passage t_x in a suspicious document d_x with a passage t_s in an original document d_s. The presented approach examines the similarity between these passages by extracting the profile of character *n*-grams of each passage and calculating the amount of common *n*-grams in the two profiles. To normalize the form of the passages, all characters are transformed into lowercase and punctuation marks are removed. Let $P_c(n,t_x)$ and $P_c(n,t_s)$ be the character *n*-gram profiles (where multiple occurrences of the same *n*-gram are replaced by one single occurrence) of the detected passages in the suspicious and the original document, respectively. Then, the similarity between t_x and t_s is calculated as follows:

$$Sim(t_x,t_s) = \frac{|P_c(n_c,t_x) \cap P_c(n_c,t_s)|}{\max(|P_c(n_c,t_x)|,|P_c(n_c,t_s)|)} \quad (5)$$

where $|a|$ is the size of *a*. Note that in case the $P_c(n_c,t_x)$ and $P_c(n_c,t_s)$ are identical, the similarity measure is 1. This similarity measure resembles the *containment* measure (Broder, 1997). However, the denominator ensures that if one of the profiles is much longer than the other, the similarity score is considerably reduced. This is especially useful to filter out cases where adjacent passages were erroneously merged. The choice of n_c is associated with the flexibility of the similarity measure. The longer the character *n*-grams are, the more they will be affected by changes in the plagiarized passage with respect to the original passage. Then, in case the similarity score is above a threshold θ_c the detected plagiarism case is considered true.

Otherwise, it is removed from the set of detections. For n_c=3, the similarity of the text passages of the highly modified plagiarism case of Figure 3 is 0.59 while the similarity score of the two unrelated passages of Figure 5 is just 0.18.

Another problem that should be faced in the post-processing stage is the existence of many short passages in both the suspicious and source documents that are not plagiarized. Such passages are short and refer to famous quotations, sayings, poems, parts of the Bible, etc. (Kolak & Schilit, 2008). A couple of examples are given below.

…for we have heard Him ourselves, and know that this is indeed the Christ, the Saviour of the world.

…He who of old would rend the oak, Deemed not of the rebound; Chained by the trunk he vainly broke, Alone, how looked he round!"

Ideally, such cases should not be reported as plagiarism acts. However, their identification among the set of detections is very difficult. Since they are usually almost identical in both the suspicious and the source documents, their similarity score would be very high. The same is true for verbatim plagiarism cases. As already mentioned, such passages are usually very short. Therefore, it is possible to apply a threshold θ_L to the length of the detected passages and filter out the vast majority of these. The length threshold is expected to also hurt the recall of the proposed approach since detected plagiarism cases of very short length will also be eliminated. If the aim is to find any similarities between a suspicious document and a set of source documents, no matter if they are plagiarism cases or not, this length threshold should not be applied.

4. Evaluation

4.1 Corpora

Recently, in the framework of the PAN Workshop series, evaluation campaigns for plagiarism detectors were initiated (Potthast, Stein, Eiselt, Barrón-Cedeño, & Rosso, 2009; Potthast, Barrón-Cedeño, Eiselt, Stein, & Rosso, 2010). A corpus including multiple suspicious and source documents as well as many types of plagiarism cases was released in 2010 (Potthast, Stein, Barrón-Cedeño, & Rosso, 2010). More specifically, the PAN 2010 Plagiarism Competition corpus[2] (PAN-PC-10) comprises 27,073 documents divided into a set of 15,925 suspicious documents and a set of 11,148 source documents. The length of the documents varies from one page to an entire book of several hundred pages. Half (7,972) of the suspicious documents are non-plagiarized. The other half of the suspicious documents contains 68,558 plagiarism cases that were inserted into randomly selected parts of the suspicious documents. Therefore, there are suspicious documents with only one plagiarized passage and other suspicious documents with dozens of plagiarized passages. 70% of the plagiarism cases refer to the external plagiarism detection task and the rest 30% refer to the intrinsic plagiarism detection task (the originals of the plagiarized passages were not taken from the source documents).

The external plagiarism detection cases have been produced either by humans (simulated) or computational tools (artificial) able to obfuscate a passage by replacing words and phrases with synonyms. In the latter case, it is possible to estimate the degree of obfuscation (high, low, or none). Additionally, 14% of the external plagiarism cases were produced by automatic translation tools that used source documents in Spanish and German. Since the proposed approach aims at the monolingual external plagiarism detection task we used the part of the PAN-PC-10 corpus that refers to this, that is, we exclude the suspicious documents with intrinsic or cross-lingual plagiarism cases. Note that each plagiarized document of PAN-PC-

[2] http://www.uni-weimar.de/cms/medien/webis/research/corpora/pan-pc-10.html

TABLE 3. Details about the PAN-PC-10 corpus.

Plagiarism type	Documents	Plagiarism Cases
Simulated	598	2,347
Artificial: High obfuscation	1,337	14,756
Artificial: Low obfuscation	1,354	14,883
Verbatim	1,728	17,423
Non-plagiarized	7,972	0
Total	12,989	49,409

TABLE 4. Details about the CS11 corpus.

Category	Documents
Original	5
Heavy revision	19
Light revision	19
Verbatim	19
Non-plagiarized	38
Total	100

10 contains only one type of plagiarism to facilitate the extraction of a sub-corpus with a certain type of plagiarism cases. Some statistics of the corpus we used in this study are shown in Table 3.

PAN-PC-10 is the largest available corpus for evaluating plagiarism detection approaches. Moreover, it covers a wide variety of topics and a wide range of document lengths. On the other hand, an obvious weakness of PAN-PC-10 is that most of the plagiarism cases are artificially generated. Another more focused corpus is presented by Clough & Stevenson (2011). This corpus[3] (henceforth, it will be called CS11) comprises answers to short questions on Computer Science topics. Here, plagiarism is simulated by asking authors to intentionally reuse an original document (Wikipedia article). Moreover, plagiarism is only considered on the document level (i.e., the whole document is either plagiarized or non-plagiarized). CS11 contains 100 documents as shown in Table 4. All texts are relatively short (average text-length is 208 words). Despite the fact that the source document set is extremely small this corpus comprises some difficult plagiarism cases simulating the strategies used by students.

4.2 Measures

For evaluating the produced detections, we use the recently proposed measures of macro-average *precision*, *recall* and *granularity* (Potthast, Stein, Barrón-Cedeño, & Rosso, 2010). In more detail, let *S* denote the set of plagiarism cases and *R* denote the set of detections. Then, macro-average precision and recall are defined as follows:

$$Prec(S,R) = \frac{1}{|R|}\sum_{r\in R}\frac{|\bigcup_{s\in S} s\cap r|}{|r|} \quad (6)$$

$$\mathrm{Rec}(\mathrm{S},\mathrm{R}) = \frac{1}{|\mathrm{S}|}\sum_{\mathrm{s}\in \mathrm{S}}\frac{|\bigcup_{\mathrm{r}\in \mathrm{R}} \mathrm{s}\cap \mathrm{r}|}{|\mathrm{s}|} \quad (7)$$

[3] http://ir.shef.ac.uk/cloughie/resources/plagiarism_corpus.html

TABLE 5. The parameter values used in the PAN-PC-10 experiments.

Parameter	Value	Function
n_1	11	Stopword *n*-gram length to retrieve candidate documents
n_2	8	Stopword *n*-gram length to detect passage boundaries
n_c	3	Character *n*-gram length to measure similarity between passages
θ_g	100	Upper threshold (in SWNGs) of gap-length allowed in a passage
θ_c	0.5	Lower threshold of the similarity measure to keep a detection
θ_L	200	Lower limit (in characters) of the detected passage length

where $s \cap r$ is the amount of overlapping characters between *s* and *r* when they share at least one character in both the suspicious and the source passage. Otherwise it is 0. These measures give equal weight to each plagiarism case regardless of its length. Additionally, they do not take into account the similarity score assigned by detectors to each plagiarism case.

In plagiarism detection, recall and precision do not give a complete picture of the effectiveness. In case a detector reports overlapping passages for the same plagiarism case or divides a long passage into shorter segments, recall and precision may be affected (increase). Therefore, we need an additional measure that takes these cases into account. Let $S_R \subseteq S$ be the cases detected in *R* and $R_s \subseteq R$ be the detections regarding the passage *s*. Then, the granularity measure is defined as follows:

$$\mathrm{Gran}(\mathrm{S},\mathrm{R}) = \frac{1}{|\mathrm{S_R}|} \sum_{\mathrm{s} \in \mathrm{S_R}} |\mathrm{R_s}| \qquad (8)$$

The minimum and ideal granularity value is 1. The larger the granularity is, the more (possibly overlapping) segments are detected for the same plagiarized passage. Precision, recall, and granularity can be combined to a single measure, *plagdet*, defined as follows:

$$\mathrm{plagdet}(\mathrm{S},\mathrm{R}) = \frac{\mathrm{F_1}}{\log_2(1+\mathrm{Gran}(\mathrm{S},\mathrm{R}))} \qquad (10)$$

where F_1 is the harmonic mean of precision and recall. Note that the *plagdet* measure was used to rank the candidates in the PAN competitions on plagiarism detection (Potthast, Stein, Eiselt, Barrón-Cedeño, Rosso, 2009; Potthast, Barrón-Cedeño, Eiselt, Stein, & Rosso, 2010).

4.3 Experimental Results

To apply the presented approach to PAN-PC-10 corpus, a small part of it was first used to estimate the appropriate parameter settings. In more detail, the first 100 suspicious documents and their corresponding source documents were used and various values of *n*-gram length and thresholds were tested. Our aim in these preliminary experiments was not to optimize the results for this specific sub-corpus but to estimate general parameter values that increase recall of the first steps and precision of the last steps. The parameter settings shown in Table 5 were selected and used in the experiments described below.

First, we examine the performance in each processing step. Figure 11 shows the results after applying the candidate document retrieval, the passage boundary detection and the post-processing steps. Note that, for the candidate retrieval task, recall and precision are calculated on the document level while granularity and *plagdet* are not defined. The final precision is very high while recall is lower indicating that many plagiarism cases are not detected but the provided detections are usually correct. Granularity remains low indicating that in the vast majority of the cases one passage is detected per plagiarism case. The first two steps achieve poor precision scores. However, the post-processing step significantly improves precision. A more detailed look in the usefulness of the post-processing step is depicted in Figure 12. The performance attained by applying the similarity threshold and the length threshold separately

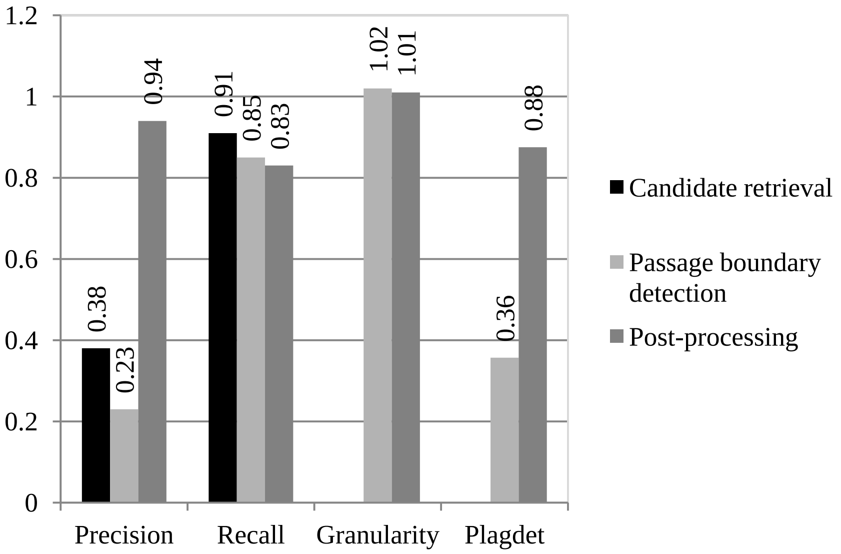


FIG 11. Evaluation results for the processing steps of the presented method.

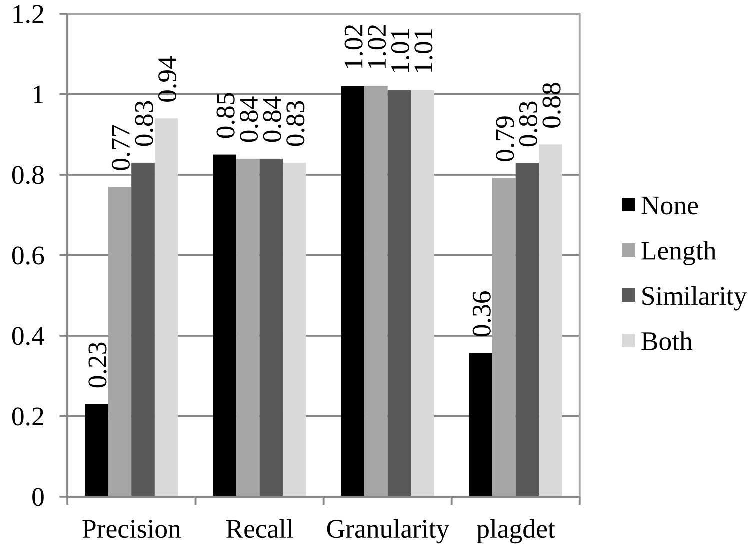


FIG 12. The contribution of the post-processing criteria (length threshold and similarity threshold) to the performance of the presented method.

or in combination is given. Apparently, each of these criteria is very important to significantly improve precision. This means that the vast majority of the wrong predictions of the passage boundary detection step correspond to very short passages with similar sequence of stopwords but essentially different content. The combination of these criteria further improves precision due to the elimination of short near-identical passages in suspicious and source documents that are not plagiarism cases (quotations, sayings, etc). Granularity is also improved. On the other hand, recall is slightly reduced.

Next, we examine the performance of the proposed approach in detecting certain plagiarism types. Table 6 shows the results when only simulated plagiarism, artificial plagiarism with high obfuscation, artificial plagiarism with low obfuscation, and verbatim cases are considered. For each type, we use the documents containing this kind of plagiarism (see Table 3) plus an equal number of non-plagiarized documents. This procedure was also followed in

TABLE 6. Comparative performance results on PAN-PC-10 for several plagiarism types.

Plagiarism Type		SWNG	PAN-10-1	PAN-10-2	PAN-10-3	PAN-10-4
Simulated	Prec.	**0.89**	0.33	0.19	0.19	0.33
	Rec.	**0.27**	0.18	0.22	0.26	0.25
	Gran.	**1.00**	**1.00**	**1.00**	**1.00**	1.03
	plagdet	**0.41**	0.23	0.20	0.22	0.28
Artificial: High	Prec.	**0.97**	0.93	0.76	0.77	0.85
	Rec.	0.79	0.75	0.76	**0.81**	0.61
	Gran.	1.03	**1.00**	1.02	1.08	1.02
	plagdet	**0.85**	0.83	0.75	0.75	0.70
Artificial: Low	Prec.	**0.95**	0.93	0.81	0.78	0.82
	Rec.	0.84	**0.92**	0.85	**0.92**	0.66
	Gran.	**1.00**	**1.00**	1.22	1.10	1.01
	plagdet	0.89	**0.92**	0.72	0.79	0.73
Verbatim	Prec.	**0.96**	0.94	0.78	0.76	0.82
	Rec.	0.93	**0.96**	0.86	0.92	0.68
	Gran.	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**
	plagdet	0.94	**0.95**	0.82	0.83	0.74

TABLE 7. Comparative performance results for intra-topic and inter-topic plagiarism cases.

Topic agreement		SWNG	PAN-10-1	PAN-10-2	PAN-10-3	PAN-10-4
Intra-topic	Prec.	**0.95**	0.92	0.76	0.74	0.79
	Rec.	0.86	**0.87**	0.81	0.86	0.66
	Gran.	1.01	**1.00**	1.08	1.05	1.01
	plagdet	**0.90**	0.89	0.74	0.77	0.71
Inter-topic	Prec.	**0.96**	0.94	0.84	0.83	0.88
	Rec.	0.82	**0.84**	0.76	0.82	0.57
	Gran.	1.01	**1.00**	1.06	1.19	1.02
	plagdet	0.88	**0.89**	0.77	0.73	0.68

(Potthast, Barrón-Cedeño, Eiselt, Stein, & Rosso, 2010), so the presented results are directly compared with the performance of the four top-performing participants in the PAN-10 plagiarism detection competition (see Table 1). It should be underlined that the PAN-10 results were produced in a blind experiment where the ground truth was not available to researchers so they were unable to make any training or optimization in this specific corpus. As can be seen, the proposed approach is very competitive in all plagiarism types. It achieves better precision results in any case in comparison to the PAN-10 participants. On the other hand, recall is usually lower in comparison to top-performing approaches. Interestingly, in the most difficult cases of simulated plagiarism and artificial plagiarism with high obfuscation the attained performance is considerably better than the other approaches. This shows that the SWNG representation is better able to capture the structure of a text that remains roughly the same despite significant and deep changes to hide the origin of the plagiarized passages.

The PAN-PC-10 corpus also provides interesting information concerning agreement in topic between the suspicious and source documents. In more detail, the artificial plagiarism cases are divided into two categories: intra-topic where the passages inserted in a suspicious document were taken from source documents that belong to the same thematic cluster with the suspicious document, and inter-topic where the suspicious and the source document belong to different thematic clusters. Table 7 presents the performance of our approach when

TABLE 8. Performance results of the presented approach for different text-length ranges.

Passage length	Prec.	Rec.	Gran.	plagdet
Long (>10K chars)	0.89	1.00	1.02	0.93
Medium (1K-10K chars)	0.87	0.92	1.00	0.89
Short (<1K chars)	0.72	0.48	1.00	0.58

TABLE 9. Comparative performance results for different text-length ranges.

Passage length		SWNG	PAN-10-1	PAN-10-2	PAN-10-3	PAN-10-4
Long (>10K chars)	Prec.	**0.88**	0.84	0.49	0.50	0.61
	Rec.	0.89	0.90	0.84	**0.91**	0.61
	Gran.	1.02	**1.00**	1.15	1.31	1.03
	plagdet	**0.87**	**0.87**	0.56	0.53	0.60
Medium (1K-10K chars)	Prec.	**0.86**	0.82	0.38	0.35	0.55
	Rec.	0.71	**0.73**	0.68	0.72	0.58
	Gran.	**1.00**	**1.00**	**1.00**	1.02	1.01
	plagdet	**0.78**	0.77	0.49	0.46	0.56
Short (<1K chars)	Prec.	**0.67**	0.57	0.12	0.14	0.14
	Rec.	0.33	0.35	0.28	**0.40**	0.15
	Gran.	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**
	plagdet	**0.44**	0.43	0.17	0.21	0.14

considering intra-topic and inter-topic artificial plagiarism. While recall is reduced in the inter-topic type with respect to the intra-topic cases, the precision is slightly improved. The same pattern is noticed for the other methods.

A crucial parameter is the length of the plagiarized passage. Table 8 shows the performance of the presented approach when considering three passage length types: long (i.e., more than 10,000 characters), medium (i.e., between 1,000 and 10,000 characters) and short (i.e., less than 1,000 characters). As expected, the performance worsens when moving from long to short passages. Notably, precision remains relatively high even for short plagiarism cases. In the case of long passages, the recall is perfect but the increased granularity indicates broken detections for the same plagiarism case. To be able to compare the performance of the proposed method with the results reported by Potthast, Barrón-Cedeño, Eiselt, Stein, & Rosso (2010) in another experiment we included in the suspicious document corpus additional documents comprising intrinsic plagiarism and cross-lingual plagiarism cases. These are 2,936 documents and 10,245 cases. Note that our method makes no attempt to detect such cases of plagiarism (the same is true for some of the PAN-10 participants). Table 9 presents the comparative results. As expected, the recall of our approach is considerably lower in comparison with the results of Table 8 since additional unknown plagiarism cases were added. However, the precision is not considerably hurt with the exception of the short passages. In any case, the SWNG approach achieves better precision scores from the best PAN-10 participants and a better overall plagdet score. The recall results are slightly worse in comparison with the best performing approaches since they are also able to detect some intrinsic or multilingual plagiarism cases.

The presented approach was also applied to CS11 corpus. Since this corpus regards plagiarism on the document level, only the candidate document retrieval task can be tested. Figure 13 shows the recall of the detections for the categories of plagiarism and various values of n_1 (SWNG length used to detect similarity in documents). In all cases, the detection of non-plagiarized documents and near-copies was very successful. On the other hand, the

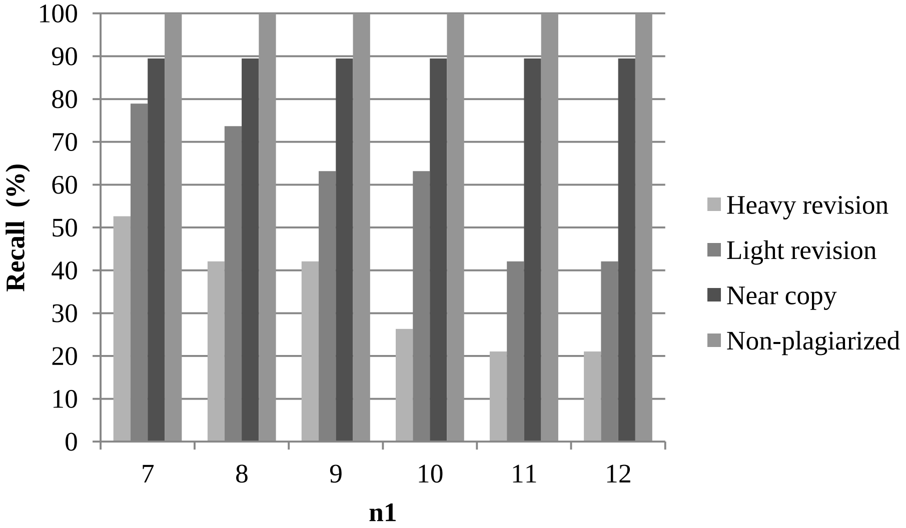


FIG 13. The performance of the candidate retrieval task on CS11.

detection of plagiarized documents with light revision or high revision of the original document decreases as n_1 increases. It seems that SWNG length should be lower than 11 (i.e., value used in PAN-PC-10 experiments) for increasing the potential of detecting simulated plagiarism cases. However, such a choice may harm the precision. In experiments on CS11, precision was 100% in all cases since it is a small corpus with only a few source documents. Note that the presented performance results cannot be compared with the results reported by Clough & Stevenson (2011) since their method is based on supervised classification trained using parts of the corpus and evaluated based on a cross-validation procedure.

5. Conclusions

Plagiarism detection in large document collections should be both efficient and effective. The former requires that the measures used to represent documents are easily available and capture local similarities so that to enable the identification of a short plagiarized passage within a long document. Moreover, the document representation measures should be flexible in modifications intentionally made by plagiarists to hide the similarity with the original passages. In contrast to the vast majority of the existing approaches that are (entirely or in part) based on content terms, in this paper we presented a method that uses only a small list of stopwords to represent documents. It has been demonstrated that the stopword *n*-gram method is reliable when it is used to identify similarity in the document level as well the exact passage boundaries in the plagiarized and the source documents.

Experiments using publicly-available corpora for plagiarism detection show that the performance of the presented method is very competitive when compared with methods based on content information. Interestingly, the proposed method achieves significantly better performance when it deals with plagiarism cases where the plagiarized passage has been extensively modified. In such cases, usually most of the content words/phrases are replaced by synonyms. This type of modification is relatively easy for plagiarists while rephrasing is much harder. However, usually this act does not change the main syntactic structure of the sentences and consequently the stopword sequences are not heavily affected. Note that in these difficult plagiarism cases, content-based methods either cannot capture the similarity (since most of the words are different) or require a more elaborate (and inefficient) analysis of texts involving thesauri or other specialized and language-dependent resources to detect terms with the same meaning.

Our method supposes that the suspicious and source documents share the same syntactic form. It has been demonstrated that sharing the same long sequence of stopwords is extremely unlikely (especially when the appearances of the most frequent stopwords are limited). On the other hand, when the plagiarist just borrows the ideas of some source documents and rephrases large parts of the passages, the stopword sequences are heavily affected. In this case, the existence of proper names or other content-based information is likely to be included in both plagiarized and source documents though not necessarily in the same order. In that case, methods that are based on text similarity disregarding word ordering seem to be more appropriate.

The SWNG representation reduces text size since only the stopword appearances are kept. It is therefore an efficient representation for large document collections. In this paper, we followed the full-fingerprinting approach where all the stopword *n*-grams are included in the fingerprint of a document. However, techniques that select a subset of stopword *n*-grams can also be applied to reduce the storage requirements and increase efficiency in very large document collections (Schleimer, et al., 2003). Moreover, provided that modern IR systems adopt full-text indexing, the presented method indicates an additional exploitation of the available information about stopword postings. Beyond the improvement in phrase queries, stopword occurrences can also be used to detect plagiarism.

The proposed method is very easy to follow and requires minimal text pre-processing cost. In order to apply it to PAN-PC-10 corpus that comprises a wide variety of text lengths (from one page to an entire book), a set of appropriate parameter settings is proposed (see Table 5). However, in case this method is going to be applied to a more homogeneous and perhaps easier corpus (e.g., CS11) more relaxed parameter values would give better results. Machine learning technology can also be used to extract the most effective parameter setting for a specific corpus.

The plagiarism detection method presented in this paper can also be applied to detect near-duplicates. The SWNG document representation method can be combined with traditional content-based methods to improve the detection results. An open question regards the minimum number of stopwords required to provide accurate results. This should be examined for several natural languages since the use and definition of stopwords may differ. Another interesting future work dimension is the use of stopword *n*-gram information in the framework of intrinsic plagiarism detection where there is no reference collection. In this case the question is whether stopword *n*-grams are able to capture stylistic inconsistencies within a document.

References

Arun, R., Suresh, V., & Madhavan, C.E.V. (2009). Stopword graphs and authorship attribution in text corpora. In Proceedings of the IEEE International Conference on Semantic Computing (pp. 192-196).

Barrón-Cedeño, A., & Rosso, P. (2009). On automatic plagiarism detection based on n-grams comparison. In Proceedings of the 31th European Conference on IR Research on Advances in Information Retrieval (pp. 696-700).

Basile, C., Benedetto, D., Caglioti, E., Cristadoro, G., & Esposti, M.D. (2009). A plagiarism detection procedure in three steps: Selection, matches and “squares”. In Proceedings of the 3rd Workshop on Uncovering Plagiarism, Authorship and Social Software Misuse, (pp. 19–23).

Bendersky, M. & Croft, W.B. (2009). Finding text reuse on the web. In Proceedings of the 2nd International Conference on Web Search and Web Data Mining, (pp. 262-271).

Brin, S., Davis, J., & Garcia-Molina, H. (1995). Copy detection mechanisms for digital documents. In Proceedings of the ACM SIGMOD International Conference on Management of Data (pp. 398-409).

Broder, A.Z. (1997). On the resemblance and containment of documents. In Proceedings of the Compression and Complexity of Sequences (pp. 21-29).

Ceska, Z. & Fox, C. (2009). The influence of text pre-processing on plagiarism detection. In Proceedings of the Int. Conf. on Recent Advances in Natural Language Processing (pp. 55-59).

Chowdhury, A., Frieder, O., Grossman, D., & McCabe, M.C. (2002). Collection statistics for fast duplicate document detection. ACM Transactions on Information Systems, 20(2), 171–191.

Church, K.W. & Helfman, J.I. (1993). Dotplot: A Program for exploring self-similarity in millions of lines of text and code. Journal of Computational and Graphical Statistics, 2(2), 153-174.

Clough, P. (2003). Old and new challenges in automatic plagiarism detection. National UK Plagiarism Advisory Service.

Clough, P. & Stevenson, M. (2011). Developing a corpus of plagiarised short answers. Language Resources and Evaluation, 45(1), 5-24.

Grozea, C., Gehl, C., & Popescu, M. (2009). ENCOPLOT: Pairwise sequence matching in linear time applied to plagiarism detection. In Proceedings of the 3rd Workshop on Uncovering Plagiarism, Authorship and Social Software Misuse (pp. 10-18).

Gustafson, N., Pera, M.S., & Ng, Y.K. (2008). Nowhere to hide: Finding plagiarized documents based on sentence similarity. In Proceedings of the IEEE/WIC/ACM Int. Conference on Web Intelligence and Intelligent Agent Technology, (pp. 690-696).

Hannabuss, S. (2001). Contested texts: Issues of plagiarism. Library Management, 22(6-7), 311-318.

Hoad, T.C. & Zobel, J. (2003). Methods for identifying versioned and plagiarized documents. Journal of the American Society for Information Science and Technology, 54(3), 203-215.

Kasprzak, J. & Brandejs, M. (2010). Improving the reliability of the plagiarism detection system - Lab report for PAN at CLEF 2010. In Proceedings of the 4th Workshop on Uncovering Plagiarism, Authorship, and Social Software Misuse.

Khmelev, D.V., & Teahan, W.J. (2003a). A repetition based measure for verification of text collections and for text categorization. In Proceedings of the 26th ACM SIGIR, (pp. 104–110).

Kolak, O. & Schilit, B.N. (2008). Generating links by mining quotations. In Proceedings of HT 2008, (pp. 117–126).

Koppel, M., Akiva, N., & Dagan, I. (2006). Feature instability as a criterion for selecting potential style markers. Journal of the American Society for Information Science and Technology, 57(11), 1519–1525.

Lyon, C., Malcolm, J., & Dickerson, B. (2001). Detecting short passages of similar text in large document collections. In Proceedings of the Conference on Empirical Methods in Natural Language Processing (pp. 118-125).

Manning, C.D., Raghavan, P., & Schütze, H. (2008). Introduction to information retrieval. Cambridge University Press.

Melamed, I.D. (1999). Bitext maps and alignment via pattern recognition, Computational Linguistics, 25(1), 107-130.

Metzler, D., Bernstein, Y., Croft, W.B., Moffat, A., & Zobel, J. (2005). Similarity measures for tracking information flow. In Proceedings of the ACM Conference on Information and Knowledge Management (pp. 517-524).

Muhr, M., Kern, R., Zechner, M., & Granitzer, M. (2010). External and intrinsic plagiarism detection using a cross-lingual retrieval and segmentation system - Lab report for PAN at CLEF 2010. In Proceedings of the 4th Workshop on Uncovering Plagiarism, Authorship, and Social Software Misuse.

Potthast, M., Barrón-Cedeño, A., Stein, B., & Rosso, P (2011). Cross-language plagiarism detection. Language Resources & Evaluation, 45(1), 45-62.

Potthast, M., Barrón-Cedeño, A., Eiselt, A., Stein, B., & Rosso, P. (2010). Overview of the 2nd international competition on plagiarism detection. In Proceedings of the 4th Workshop on Uncovering Plagiarism, Authorship, and Social Software Misuse.

Potthast, M., Stein, B., Barrón-Cedeño, A., & Rosso, P. (2010). An evaluation framework for plagiarism detection. In Proceedings of the 23rd International Conference on Computational Linguistics.

Potthast, M., Stein, B., Eiselt, A., Barrón-Cedeño, A., & Rosso, P. (2009). Overview of the 1st international competition on plagiarism detection. In Proceedings of the 3rd Workshop on Uncovering Plagiarism, Authorship, and Social Software Misuse (pp. 1-9).

Schleimer, S., Wilkerson, D.S., & Aiken, A. (2003). Winnowing: Local algorithms for document fingerprinting. In Proceedings of the ACM SIGMOD International Conference on Management of Data (pp. 76-85).

Seo, J., & Croft, W.B. (2008). Local text reuse detection. In Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (pp. 571-578).

Shivakumar, N. & Garcia-Molina, H. (1995). SCAM: A copy detection mechanism for digital documents. In Proceedings of the Int. Conference on Theory and Practice of Digital Documents.

Stamatatos, E. (2009). Intrinsic plagiarism detection using character n-gram profiles. In Proceedings of the 3rd Int. Workshop on Uncovering Plagiarism, Authorship, and Social Software Misuse.

Stamatatos, E., Fakotakis, N., & Kokkinakis, G. (2000). Text genre detection using common word frequencies. In Proceedings of the 18th Int. Conf. on Computational Linguistics (pp. 808-814).

Stein, B., Lipka, N. & Prettenhofer, P. (2011). Intrinsic plagiarism analysis. Language Resources & Evaluation, 45(1), 63-82.

Stein, B., & Meyer zu Eissen, S. (2006). Near similarity search and plagiarism analysis. In M. Spiliopoulou, et al. (eds), From Data and Information Analysis to Knowledge Engineering (pp. 430-437).

Theobald, M., Siddharth, J., & Paepcke, A. (2008). Spotsigs: Robust and efficient near duplicate detection in large web collections. In Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (pp. 563–570).

Uzuner, O., Katz, B., & Nahnsen, T. (2005). Using syntactic information to identify plagiarism. In Proceedings of the ACL Workshop on Educational Applications (pp. 37–44).

Zhang, Q., Zhang, Y., Yu, H., & Huang, X. (2010). Efficient partial-duplicate detection based on sequence matching. In Proceedings of the 33rd Int. ACM SIGIR Conference on Research and Development (pp. 675-682).

Zou, D. Long, W., & Ling, Z. (2010). A cluster-based plagiarism detection method - Lab report for PAN at CLEF 2010. In Proceedings of the 4th Workshop on Uncovering Plagiarism, Authorship, and Social Software Misuse.