Decker and Lenz (Eds.): Advances in Data Analysis
Selected Papers from the 30th Annual Conference of the German Classification Society (GfKl)
Berlin, ISBN 978-3-540-70980-0, pp. 359-366, ©Springer 2007.

Plagiarism Detection without Reference Collections

Sven Meyer zu Eissen, Benno Stein, and Marion Kulig

Faculty of Media, Media Systems
Bauhaus University Weimar, 99421 Weimar, Germany
sven.meyer-zu-eissen@medien.uni-weimar.de
benno.stein@medien.uni-weimar.de

Abstract. Current research in the field of automatic plagiarism detection for text documents focuses on the development of algorithms that compare suspicious documents against potential original documents. Although recent approaches perform well in identifying copied or even modified passages [Brin 1995, Stein 2005], they assume a closed world where a reference collection must be given [Finkel 2002]. Recall that a human reader can identify suspicious passages within a document without having a library of potential original documents as reference in mind.

This raises the question whether plagiarized passages within a document can be detected automatically if no reference is given, e. g. if the plagiarized passages stem from a book that is not available in digital form. This paper contributes right here; it proposes a method to identify potentially plagiarized passages by analyzing a single document with respect to changes in writing style. Such passages then can be used as a starting point for an Internet search for potential sources. As well as that, such passages can be preselected for inspection by a human referee. Among others, we will present new style features that can be computed efficiently and which provide highly discriminative information: Our experiments, which base on a test corpus that will be published, show encouraging results.

1 Introduction

A recent large-scale study on 18,000 students by McCabe reveals that about 50% of the students admit to plagiarize from extraneous documents [10]. Plagiarism in text documents happens in several forms: one-to-one copies, passages that are modified to a greater or lesser extent, or even translated passages. Figure 1, which is taken from [15], shows a taxonomy of plagiarism delicts along with possible detection methods.

1.1 Some Background on Plagiarism Detection

The success of current approaches in plagiarism detection varies according to the underlying plagiarism delict. The approaches stated in [1,6] employ

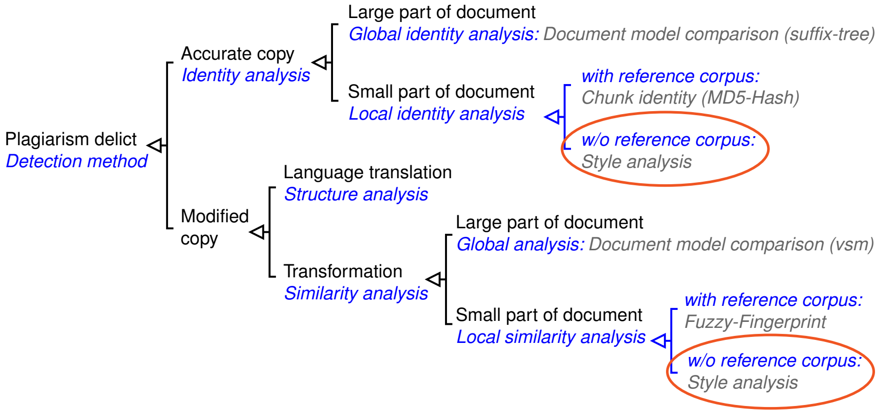

Fig. 1. A taxonomy of plagiarism delicts and analysis methods [15]. The encircled parts indicate our contributions: the detection of a plagiarism delict without having a reference corpus at hand.

cryptographic hash functions to generate digital fingerprints of so-called text chunks, which are compared against a database of original text passage fingerprints. Since cryptographic fingerprints identify a text chunk exactly, the quality of these approaches depends on offsets and sizes of chunks within both plagiarized and original texts. An approach introduced in [14] overcomes these limitations: unlike cryptographic fingerprints, the proposed method generates fingerprints that are robust against modifications to some extent.

However, the mentioned approaches have one constraint in common: they require a reference collection with original documents. Observe that human readers may identify suspicious passages within a document without having a library of reference documents in mind: changes between brilliant and baffling passages, or the change of person narrative give hints to plagiarism. Situations where such an *intrinsic plagiarism detection* can be applied are shown encircled in Figure 1.

1.2 Contributions of the Paper

Basically, the power of a plagiarism approach depends on the quality of the quantified linguistic features. We introduce features which measure—simply put—the customariness of word usage, and which are able to capture a significant part of style information. To analyze the phenomenon of intrinsic plagiarism detection we have constructed a base corpus from which various application corpora can be compiled, each of which modeling plagiarism delicts of different severity. Section 3 reports on experiments that we have conducted with this corpus.

2 Quantification of Writing Style

Intrinsic plagiarism detection can be operationalized by dividing a document into "natural" parts, such as sentences, paragraphs, or sections, and analyzing the variance of certain style features. Note in this connection that within the experiments presented in Section 3 the size of a part is chosen rather small (40-200 words), which is ambitious from the analysis standpoint, but which corresponds to realistic situations.

2.1 Stylometric Features

Each author develops an individual writing style; i. e. he or she employs consciously or subconsciously patterns to construct sentences and uses an individual vocabulary. Stylometric features quantify style aspects, and some of them have been used successfully in the past to discriminate between texts with respect to authorship [9,12]. Most stylometric features are based on the following semiotic features:

1. Text statistics, which operate at the character level.
   Examples: number of commas, question marks, word lengths.
2. Syntactic features, which measure writing style at the sentence level.
   Examples: sentence lengths, use of function words
3. Part-of-speech features to quantify the use of word classes.
   Examples: number of adjectives or pronouns
4. Closed-class word sets to count special words.
   Examples: number of stopwords, foreign words, "difficult" words
5. Structural features, which reflect text organization.
   Examples: paragraph lengths, chapter lengths

Based on these features, formulas can be constructed that quantify the characteristic trait of an author's writing style. Almost all of the developed formulas aim at a quantification of the educational background, i. e., they quantify an author's vocabulary richness or style complexity, or a reader's grading level that is required to understand a text. Figure 2 classifies style-quantifying formulas according to their intention.

Widely employed grading measures include the Flesch Kincaid Grade Level [4,8] and the Dale-Chall formula [3,2]. The former, which is used among others by the US Government Department of Defense, combines the average number of syllables per word (here denoted as ASW) with average sentence length (denoted as ASL) as follows: $FK = 0.39 \cdot ASL + 11.8 \cdot ASW - 15.59$. The resulting number shall be an estimate for the number of years a reader has to spend in school before being able to understand the text.

The Dale-Chall formula employs a closed-class word list containing 3000 familiar words usually known by 4th grade children. The formula combines the percentage of difficult words that do not appear in the list with the average sentence length and defines a monotonic function that maps onto a grading level.

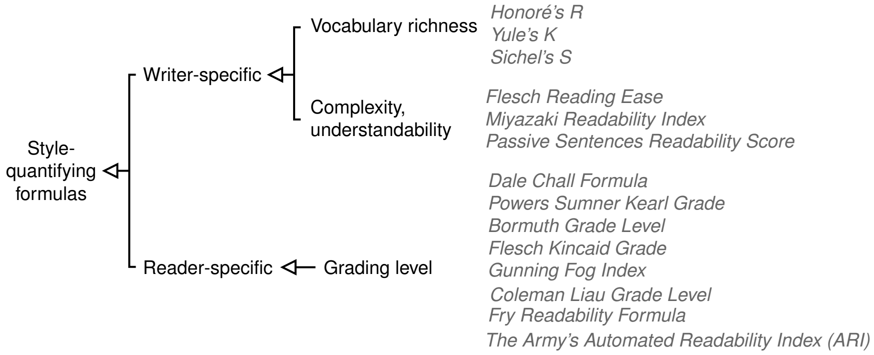

Fig. 2. A classification of the most well-known style-quantifying formulas with respect to their application range and underlying concept.

Methods to measure an author's vocabulary richness are often based on the ratio between the number of different words and the total number of words within a document; well-known examples include Yule's K [18] and Honore's R [7]. However, it was reported that these measures depend significantly on document length or passage length [16,13]. As a consequence, they are not suited to compare passages of varying lengths and deliver unreliable results for short passages, which is a disqualifying criterion for plagiarism analysis.

We now introduce a new vocabulary richness statistic, the averaged word frequency class, which turned out to be the most powerful and stable concept with respect to intrinsic plagiarism detection that we have encountered so far.

2.2 Averaged Word Frequency Class

The frequency class of a word is directly connected to Zipf's law and can be used as an indicator of a word's customariness. Let $\mathcal{C}$ be a text corpus, and let $|\mathcal{C}|$ be the number of words in $\mathcal{C}$. Moreover, let $f(w)$ denote the frequency of a word $w \in \mathcal{C}$, and let $r(w)$ denote the rank of w in a word list of $\mathcal{C}$, which is sorted by decreasing frequency.

In accordance with [17] we define the word frequency class $c(w)$ of a word $w \in \mathcal{C}$ as $\lfloor \log_2(f(w^*)/f(w)) \rfloor$, where w^* denotes the most frequently used word in $\mathcal{C}$. In the Sydney Morning Herald Corpus, w^* denotes the word "the", which corresponds to the word frequency class 0; the most uncommonly used words within this corpus have a word frequency class of 19. A document's averaged word frequency class tells us something about style complexity and the size of an author's vocabulary—both of which are highly individual characteristics [11].

Note that, based on a lookup-table, the averaged word frequency class of a text passage can be computed in linear time in the number of words. Another salient property is its small variance with respect to text length, which renders it ideal for our purposes.

3 Experimental Analysis

This section reports on experiments related to a plagiarism analysis without reference collections; it addresses the following questions:

1. Which vocabulary richness measure is suited best?—which leads us to the question: How stable is a measure with respect to text length?
2. To which extent is the detection of plagiarized text portions possible?

The first question can be answered by analyzing the characteristic of the vocabulary richness measures concerning single author (= non plagiarized) documents. The second question can be reformulated as a document classification task, given a reference corpus with plagiarized and non plagiarized documents.

3.1 Evaluation of Vocabulary Richness Measures

As pointed out above, changes in vocabulary richness across paragraphs are a good indicator for plagiarism. Confer in this connection the left plot in Figure 3, which contrasts the averaged word frequency class of four different authors.

Plagiarism analysis requires a measure that works reliably at the *paragraph level.* Put another way, when analyzing a portion of text from a single-author document the ideal vocabulary richness measure should behave fairly constant—regardless of the portion's position and size. An according comparison of Honore's R, Yule's K, and the average word frequency class is shown in the right plot of Figure 3; here, the analyzed text portion varies between 10% and 100% of the entire document. Observe that the average word frequency class is stable even for small paragraphs, which qualifies the measure as a powerful instrument for intrinsic plagiarism analysis.

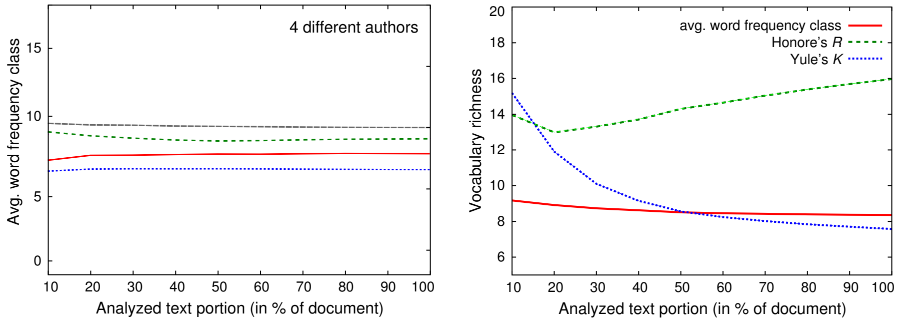

Fig. 3. Average word frequency class of four different authors (left plot). The right plot shows the development of Honore's R, Yule's K, and the average word frequency class of a single-author document for different text portions. For a better readability the values of Honore's R and Yule's K are divided by 100 and 10 respectively.

3.2 Corpus Construction

Since no reference collection is available for classification experiments, we have compiled a new corpus, which will be made available for all interested researchers. Its construction is oriented at the following corpus-linguistic criteria [5]:

1. authenticity and homogeneity
2. possibility to include many types of plagiarism
3. easy processable for both human and machine
4. clear separation of text and annotations

We chose genuine computer science articles from the ACM digital library, which were "plagiarized" by hand with both copied as well as reformulated passages from other ACM computer science articles, contributing to criterion 1. To separate annotations from text and to allow both maintenance for human editors and standardized processing for machines, all documents in the corpus are represented in XML-syntax (cf. criteria 2-4). They validate against the following DTD, which declares a mixed content model and provides element types for plagiarism delict and plagiarism source among others.

```
<!ELEMENT document (#PCDATA|plagiarized)*>
<!ATTLIST document source CDATA #REQUIRED>
<!ELEMENT plagiarized (#PCDATA)>
<!ATTLIST plagiarized type (copied|mod|trans) source CDATA #REQUIRED>
```

An XML document with k plagiarized passages defines a template from which 2^k instance documents can be generated, depending on which of the k plagiarized parts are actually included. Instance documents contain no XML tags in order to ensure that they can be processed by standard algorithms. Instead, a meta information file is generated for each, specifying the exact position of plagiarized passages.

3.3 Classification Experiments

For the results presented here more than 450 instance documents were generated each of which containing between 3 and 6 plagiarized passages of different lengths. During the plagiarism analysis each instance document was decomposed into 50 - 100 passages, and for each passage a paragraph-specific feature vector $\mathbf{f}_p$ was computed. The feature set includes average sentence length, 18 part-of-speech features, average stopword number, the Gunning Fog index, Flesch-Kincaid Grade Level, the Dale-Chall formula, Honore's R, Yule's K, and the averaged word frequency class.

Since we are interested in the detection of writing style variations, a document-specific feature vector, $\mathbf{f}_d$, was computed and compared to each of the $\mathbf{f}_p$. The rationale is that the relative differences between $\mathbf{f}_d$ and the feature vectors of the plagiarized passages reflect possible writing style changes.

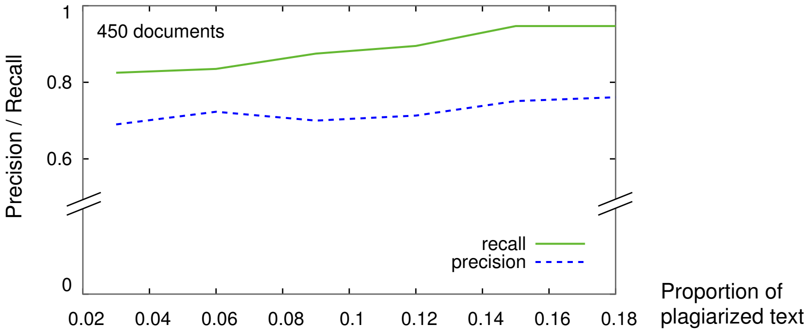

Fig. 4. Detection performance versus severity of plagiarism delict: The plot shows the averaged values for precision and recall of a series of experiments, where the sizes of the plagiarized passages are successively increased. The values are averaged using a ten-fold cross-validation.

The vector of these relative differences along with the class information (plagiarized or not) formed the input for different machine learning approaches. Figure 4 summarizes the results: We obtained good detection rates for plagiarism delicts in terms of precision and recall, irrespective of the plagiarism severity. These results were achieved using a classical discriminant analysis; however, an SVM classification showed similar results. Table 1 quantifies the discrimination power of the best features.

	Wilks Lambda	F-Ratio	significant
av. word frequency class	0.723	152.6	yes
av. preposition number	0.866	61.4	yes
av. sentence length	0.880	54.0	yes

Table 1. Significance scores for the three best-discriminating features. Lower Lambda-values and higher F-ratios indicate a better performance.

4 Summary

This paper presented an approach to detect plagiarized passages within a document if no reference collection is given against which the suspicious document can be matched. This problem, which we call “intrinsic plagiarism detection”, is related to the identification of an author’s writing style, for which various measures have been developed in the past. We presented new style features and showed their usability with respect to plagiarism detection: Classification experiments on a manually constructed corpus delivered promising precision and recall values, even for small plagiarized paragraphs.

Another result of our research shall be emphasized: A vocabulary richness measure qualifies for intrinsic plagiarism detection only, if is has a small variance subject to the analyzed text portion's size. Our experiments revealed that the introduced averaged word frequency class outperforms other well-known measures in this respect.

References

1. Sergey Brin, James Davis, and Hector Garcia-Molina. Copy detection mechanisms for digital documents. In *SIGMOD '95*, pages 398–409, New York, NY, USA, 1995. ACM Press.
2. J.S. Chall and E. Dale. *Readability Revisited: The new Dale-Chall Readability Formula.* Brookline Books, 1995.
3. E. Dale and J.S. Chall. A formula for predicting readability. *Educ. Res. Bull.*, 27, 1948.
4. R. Flesch. A new readability yardstick. *Journal of Applied Psychology*, 32:221–233, 1948.
5. R. Garside, G. Leech, and A. McEnery. *Corpus Annotation: Linguistic Information from Computer Text Corpora.* Longman, 1997.
6. Timothy C. Hoad and Justin Zobel. Methods for Identifying Versioned and Plagiarised Documents. *JASIST*, 54(3):203–215, 2003.
7. A. Honore. Some simple measures of richness of vocabulary. *Association for Literary and Linguistic Computing Bulletin*, 7(2):172–177, 1979.
8. J. Kincaid, R.P. Fishburne, R.L. Rogers, and B.S. Chissom. Derivation of new readability formulas (automated readability index, fog count and flesch reading ease formula) for navy enlisted personnel. Research Branch Report 875 Millington TN: Naval Technical Training US Naval Air Station, 1975.
9. Moshe Koppel and Jonathan Schler. Authorship verification as a one-class classification problem. In *Proceedings of ICML 04*, Banff, Canada, 2004. ACM Press.
10. Donald McCabe. Research Report of the Center for Academic Integrity. `http://www.academicintegrity.org`, 2005.
11. Sven Meyer zu Eißen and Benno Stein. Genre Classification of Web Pages: User Study and Feasibility Analysis. In *KI 2004*, volume 3228 LNAI of *Lecture Notes in Artificial Intelligence*, pages 256–269, September 2004. Springer.
12. Jason Sorensen. A competitive analysis of automated authorship attribution techniques. `http://hbar.net/thesis.pdf`, 2005.
13. E. Stamatatos, N. Fakotakis, and G. Kokkinakis. Computer-based authorship attribution without lexical measures. *Computers and the Humanities*, 35:193–214, 2001.
14. Benno Stein. Fuzzy-Fingerprints for Text-Based Information Retrieval. In the *Proceedings of the 5th International Conference on Knowledge Management (I-KNOW 05), Graz*, J.UCS, pages 572–579. Know-Center, July 2005.
15. Benno Stein and Sven Meyer zu Eißen. Near Similarity Search and Plagiarism Analysis. In *Proc. 29th Annual Conference of the GfKl* Springer, 2006.
16. Fiona J. Tweedie and R. Harald Baayen. Lexical "constants" in stylometry and authorship studies. In *Proceedings of ACH-ALLC '97*, June 1997.
17. University of Leipzig. Wortschatz. `http://wortschatz.uni-leipzig.de`, 1995.
18. G. Yule. *The Statistical Study of Literary Vocabulary.* Cambridge University Press, 1944.