Intrinsic Plagiarism Detection

Sven Meyer zu Eissen and Benno Stein

Faculty of Media: Media Systems,
Bauhaus University Weimar, 99421 Weimar, Germany

Abstract. Current research in the field of automatic plagiarism detection for text documents focuses on algorithms that compare plagiarized documents against potential original documents. Though these approaches perform well in identifying copied or even modified passages, they assume a closed world: a reference collection must be given against which a plagiarized document can be compared.

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. We call this problem class *intrinsic plagiarism detection*. The paper is devoted to this problem class; it shows that it is possible to identify potentially plagiarized passages by analyzing a single document with respect to variations in writing style.

Our contributions are fourfold: (*i*) a taxonomy of plagiarism delicts along with detection methods, (*ii*) new features for the quantification of style aspects, (*iii*) a publicly available plagiarism corpus for benchmark comparisons, and (*iv*) promising results in non-trivial plagiarism detection settings: in our experiments we achieved recall values of 85% with a precision of 75% and better.

Keywords: plagiarism detection, style analysis, classifier, plagiarism corpus.

1 Introduction

Plagiarism refers to the use of another's information, language, or writing, when done without proper acknowledgment of the original source [10]. A recent large-scale study on 18,000 students by McCabe shows that about 50% of the students admit to plagiarize from extraneous documents [5]. Plagiarism in text documents happens in several forms: plagiarized text may be copied one-to-one, passages may be modified to a greater or lesser extent, or they may even be translated. Figure 1 shows a taxonomy of plagiarism delicts, which organizes delicts and possible detection methods.

State of the Art in Plagiarism Detection. The success of current approaches in plagiarism detection varies according to the underlying plagiarism delict. The approaches stated in [1; 3] employ cryptographic hash functions to generate digital fingerprints of so-called text chunks, which are then 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 given in [8] overcomes these limitations: unlike cryptographic fingerprints, the proposed method generates fingerprints that are robust against modifications to some extent.

M. Lalmas et al. (Eds.): ECIR 2006, LNCS 3936, pp. 565–569, 2006.

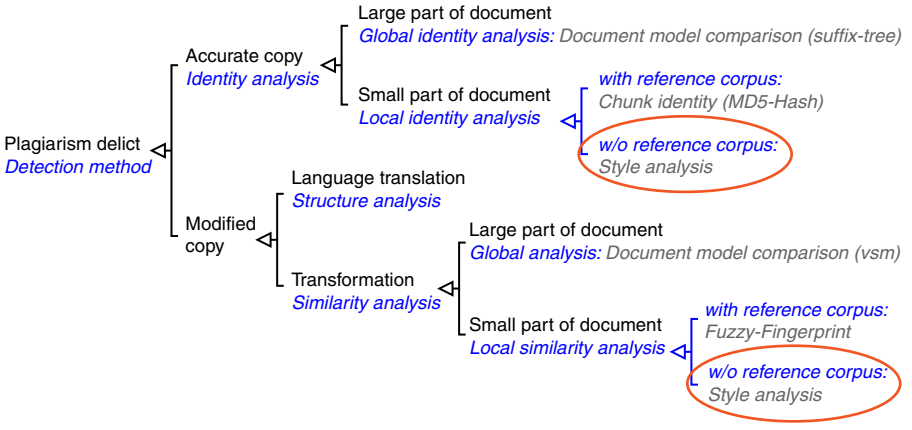

Fig. 1. A taxonomy of plagiarism delicts and analysis methods [7]. The encircled parts indicate our contributions: the detection of plagiarism delicts without having a reference corpus at hand.

Intrinsic Plagiarism Detection. The mentioned approaches have one constraint in common: they require a reference collection of potential 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.

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, which may be sentences, paragraphs, or sections, and analyzing the variance of certain style features. Within the experiments presented below 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.

Stylometric Features. Stylometric features quantify aspects of writing style, and some of them have been used successfully in the past to discriminate between books with respect to authorship [4]. Most stylometric features fall in one of the following five categories: (*i*) text statistics, which operate at the character level, (*ii*) syntactic features, which measure writing style at the sentence-level, (*iii*) part-of-speech features to

quantify the use of word classes, (*iv*) closed-class word sets to count special words, and (*v*) structural features, which reflect text organization.

In addition to these features we now introduce a new statistic, the averaged word frequency class, which turned out to be the most powerful concept with respect to intrinsic plagiarism detection that we have encountered so far.

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 [9] 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 [6].

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

Since no reference collection is available for our concern, we constructed a new corpus, oriented at the following corpus-linguistic criteria [2]: (*i*) authenticity and homogeneity, (*ii*) possibility to include many types of plagiarism, (*iii*) easy processable for both human and machine, (*iv*) clear separation of text and annotations.

We chose genuine computer science articles from the ACM digital library that we "plagiarized" with both copied as well as reformulated passages from other ACM computer science articles, contributing to criterion 1. With respect to criteria 2-4, all documents in the base corpus are represented in XML and validate against an XML schema. The schema declares a mixed content model and provides element types for plagiarism delict, plagiarism source, and other meta information.

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, containing information about the exact locations of plagiarized passages.

Experiments. For the experiments 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 these instance documents were decomposed into 50 - 100 passages from which the feature vectors were computed; the feature set included average sentence length, 18 part-of-speech features, average stopword number,

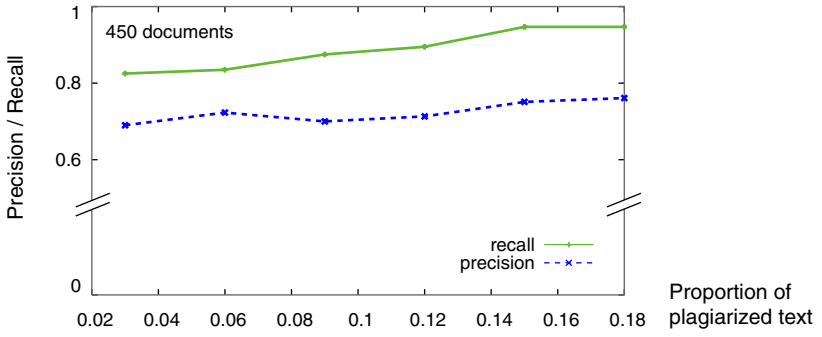

Fig. 2. Analysis performance versus severity of plagiarism delicts: The plot shows the averaged values for precision and recall of a series of experiments, where the sizes of the plagiarized passages are successively increased

Table 1. The table shows significance scores for the three best-discriminating features. Lower Lambda-values and higher F-ratios indicate better performance.

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

and the averaged word frequency class. Figure 2 illustrates good detection rates for plagiarism delicts in terms of precision and recall with respect to 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.

References

[1] S. Brin, J. Davis, and H. Garcia-Molina. Copy detection mechanisms for digital documents. In *Proc. SIGMOD '95*, pages 398–409, 1995.

[2] R. Garside, G. Leech, and A. McEnery. *Corpus Annotation: Linguistic Information from Computer Text Corpora*. Longman, 1997.

[3] T. C. Hoad and J. Zobel. Methods for Identifying Versioned and Plagiarised Documents. *JASIST*, 54(3):203–215, 2003.

[4] M. Koppel and J. Schler. Authorship verification as a one-class classification problem. In *Proc. of ICML'04*, 2004.

[5] D. McCabe. Research Report of the Center for Academic Integrity. `http://www.academicintegrity.org`, 2005.

[6] S. Meyer zu Eißen and B. Stein. Genre Classification of Web Pages: User Study and Feasibility Analysis. In *Proc. of KI'04: Advances in AI*, volume 3228 LNAI. Springer, 2005.

[7] B. Stein. Fuzzy-Fingerprints for Text-based Information Retrieval. In *Proc. of 5th Int. Conf. on Knowledge Management, Graz, Austria*. JUCS, 2005.
[8] B. Stein and S. Meyer zu Eissen. Near similarity search and plagiarism analysis. In *Proc. of GfKl '05*. Springer, 2005.
[9] University of Leipzig. Wortschatz. `http://wortschatz.uni-leipzig.de`, 1995.
[10] Wikipedia. Plagiarism. `http://en.wikipedia.org/wiki/Plagiarism`, 2005.