International Journal for Educational Integrity

State-of-the-art in detecting academic plagiarism

Norman Meuschke
University of California, Berkeley
meuschke@berkeley.edu

Bela Gipp
University of California, Berkeley and
University of Magdeburg, Department of Computer Science
gipp@berkeley.edu

Keywords: plagiarism, plagiarism detection, information retrieval, pattern recognition.

Abstract

The problem of academic plagiarism has been present for centuries. Yet, the widespread dissemination of information technology, including the internet, made plagiarising much easier. Consequently, methods and systems aiding in the detection of plagiarism have attracted much research within the last two decades. Researchers proposed a variety of solutions, which we will review comprehensively in this article. Available detection systems use sophisticated and highly efficient character-based text comparisons, which can reliably identify verbatim and moderately disguised copies. Automatically detecting more strongly disguised plagiarism, such as paraphrases, translations or idea plagiarism, is the focus of current research. Proposed approaches for this task include intrinsic, cross-lingual and citation-based plagiarism detection. Each method offers unique strengths and weaknesses; however, none is currently mature enough for practical use. In the future, plagiarism detection systems may benefit from combining traditional character-based detection methods with these emerging detection approaches.

Introduction

The advancement of information technology (IT) and especially the internet have dramatically increased the availability of information – not only for legitimate purposes. Academic plagiarism is one form of undue information use that IT has made much easier (Born, 2003; Howard, 2007).

Given the volume of available information, detecting plagiarism through manual inspection is hardly feasible (Clough, 2000, p. 9). Therefore, methods and systems capable of partially automating plagiarism detection (PD) are an active area of research. This article reviews the extensive literature on academic plagiarism detection, describes detection methods, and presents evaluations of their detection performance. We highlight the strengths and weaknesses of the various approaches and point to current research that may help in overcoming weaknesses.

© International Journal for Educational Integrity Vol. 9 No. 1 June, 2013 pp. 50–71 ISSN 1833-2595

Academic plagiarism

We define academic plagiarism as the use of ideas and/or words from sources without giving due acknowledgement as imposed by academic principles. Other researchers commonly define academic plagiarism as literary theft, i.e. stealing words or ideas from other authors (Ercegovac & Richardson Jr., 2004; Park, 2003). Theft describes the deliberate appropriation of foreign property without the permission or consent of the rightful owner (Garner, 2011, p. 125). Our definition does not necessarily characterise academic plagiarism as theft for the following three reasons.

First, academic plagiarism need not be deliberate. Authors may inadvertently fail to properly acknowledge a source, e.g. by forgetting to insert a citation, or citing a wrong source; thereby committing plagiarism unintentionally (Maurer, Kappe, & Zaka, 2006). Additionally, a psychological memory bias called cryptomnesia can cause humans to unconsciously attribute foreign ideas to themselves (Oxford University Press, 2009).

Second, academic plagiarism may not originate from other authors. We include self-plagiarism in the definition of academic plagiarism.

Third, academic plagiarists may act in consent with another author, but still commit plagiarism by not properly acknowledging the original source. The term collusion describes the behaviour of authors, who write collaboratively, or copy from one another, although they are required to work independently (Clough, 2000). We include collusion in our definition of academic plagiarism.

Observations of academic plagiarism reveal a variety of commonly found forms.

Literal plagiarism describes the undue copying of text with very little or no disguise.

- *Copy and paste (c&p)* is the most common form of literal plagiarism and is characterised by adopting text verbatim from another source (Maurer et al., 2006; Weber-Wulff, 2011).

Disguised plagiarism subsumes practices to conceal unduly copied text (Lancaster, 2003). We identified three forms of disguised plagiarism distinguished by researchers of plagiarism.

- *Shake and paste (s&p)* refers to the copying and merging of text segments with slight adjustments to form a coherent text, e.g. by changing word order, by substituting words with synonyms, or by entering or deleting filling words (Webber-Wulff, 2010).
- *Paraphrasing* is the intentional rewriting of foreign thoughts in the vocabulary and style of the plagiarist without acknowledging the source (Clough, 2000; Lancaster, 2003).
- *Technical disguise* refers to techniques that exploit weaknesses of current detection methods to make plagiarised content non-machine detectable. Examples include substituting characters with graphically identical symbols from foreign alphabets or inserting random letters in white font (Heather, 2010; Kakkonen & Mozgovoy, 2010).

Translated plagiarism is the manual or automated conversion of text from one language to another with the intention of hiding its origin (Weber-Wulff, 2010).

Idea plagiarism encompasses the use of a broader concept without due acknowledgement of the source. Examples are the appropriation of research approaches, argumentative structures, or background sources (Maurer et al., 2006).

Self-plagiarism is the partial or complete re-use of one's own writings without these being justified. Presenting updates or providing access to a larger community may justify re-publishing one's own work, but still requires appropriate acknowledgement of the previously published work (Bretag & Mahmud, 2009). Unjustified reasons include trying to artificially increase one's citation count (Collberg & Kobourov, 2005).

Plagiarism detection approaches

This section gives an overview of the generic mode of operation for all plagiarism detection systems (PDS) and presents technical descriptions of the detection methods employed by PDS.

Generic detection process

PDS are specialised computer systems supporting the identification of plagiarism incidences by implementing one of two generic detection approaches, external or intrinsic. *External PDS* compare a suspicious document with a reference collection of genuine documents (Stein, Koppel, & Stamatatos, 2007). *Intrinsic PDS* statistically examine linguistic features of a text, a process known as *stylometry*, without performing comparisons to other documents. Intrinsic PDS report changes in writing styles as indicators for potential plagiarism (Meyer zu Eissen & Stein, 2006).

Most external PDS follow a three-stage retrieval process as illustrated in Figure 1. In the first stage, PDS apply computationally inexpensive heuristic algorithms to identify a small fraction of the reference collection as candidate documents from which the suspicious text could originate.

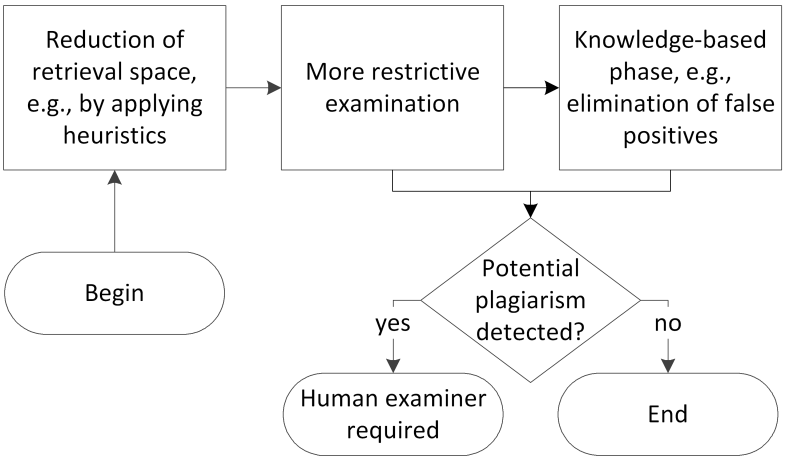

Figure 1: Generic plagiarism detection process

In the second stage, PDS perform a detailed comparison of candidate documents retrieved in the first stage using finer-grained, computationally more expensive detection methods, which we will describe in the following sections.

In the third stage, suspicious text segments retrieved in the second stage usually undergo a knowledge-based analysis. The goal of this stage is to eliminate false positives, which the specific detection procedures in the previous stages are prone to produce. Typical cases of false positives are correctly cited passages (Stein, Meyer zu Eissen, & Potthast, 2007).

The literature on academic PD emphasises that no PDS are capable of reliably identifying plagiarism without human review. An examiner is always required to check the results of the three automated retrieval stages and to verify if plagiarism is present (Lancaster, 2003; Maurer & Zaka, 2007). PDS cannot fully automate the identification of plagiarism; they are only the first step in a semi-automated detection and verification process that requires careful consideration on a case-by-case basis (Lancaster, 2003).

Overview of plagiarism detection methods

We classify plagiarism detection methods by the type of similarity assessment they commonly apply as either local or global methods, as shown in Figure 2. The leaves of the tree show the document models that the methods typically use for comparing documents.

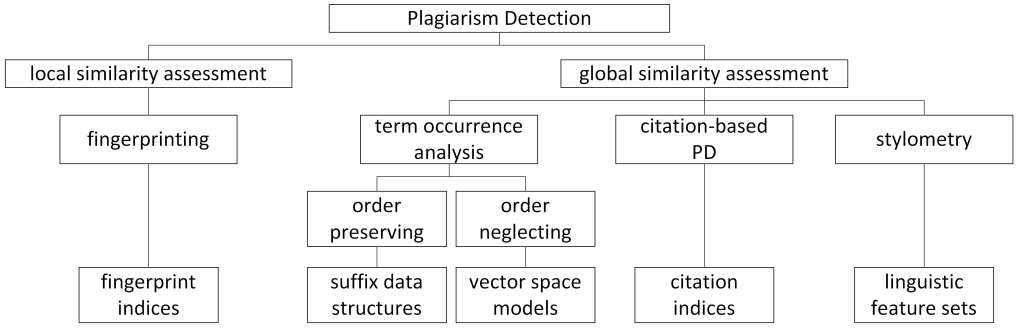

Figure 2: Classification of plagiarism detection methods

Local similarity assessment methods analyse matches of confined text segments. Fingerprinting is the most common approach in this class of detection methods. *Global similarity assessment* methods analyse characteristics of longer text sections or the full document (Stein & Meyer zu Eissen, 2006). PD methods that employ term occurrence analysis typically operate at the global level. Citation-based Plagiarism Detection (CbPD) uses the citations in academic documents to model and compare their semantic content (Gipp & Beel, 2010). Stylometry analyses stylistic differences within a document.

The classification in Figure 2 reflects the most common application of the presented detection methods, i.e. applying vector space models, string matching, or CbPD to the entire document. However, PDS can also employ the same methods to analyse fragments of a text to detect more local similarities. CbPD can detect local similarity if shorter text fragments contain sufficient citations. Figure 2 applies to the monolingual PD setting and omits cross-language PD (CLPD) for simplicity's sake. CLPD methods in part adapt building blocks from the monolingual setup and additionally use specifically designed cross-language similarity assessments.

We present all detection methods, including CLPD methods, in the next five sections. For each detection method, we will discuss typical characteristics that influence its detection capabilities.

Fingerprinting

Fingerprinting is currently the most widely applied external plagiarism detection approach (Meyer zu Eissen & Stein, 2006). Fingerprinting methods represent a document by segmenting it into substrings and selecting a subset of all the substrings formed. The substring set is the fingerprint; its elements are called minutiae (Hoad &

Zobel, 2003). PDS often apply mathematical functions to transform minutiae into computationally efficient byte strings. PDS compare a document by computing the document's fingerprint and querying each of the minutiae with a pre-computed index of fingerprints for all documents in a reference collection, as Figure 3 shows. Minutiae that match with other documents indicate shared text segments and suggest potential plagiarism when exceeding the chosen similarity threshold (Brin, Davis, & Garcia Molina, 1995).

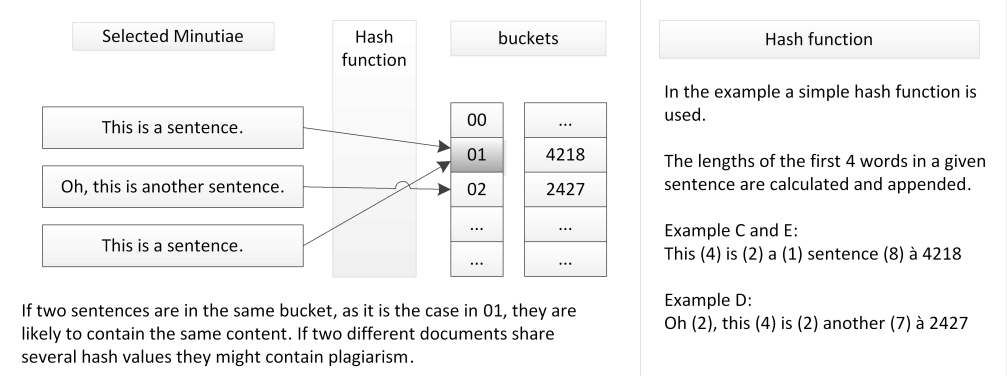

Figure 3: Concept of fingerprinting

The fingerprinting methods proposed for PD differ in the parameters explained hereafter.

The *chunking unit* defines the segments into which a fingerprinting method divides a text, and whether these segments are combined into larger composites, called chunks. Table 1 summarises chunking units proposed for fingerprinting methods.

Table 1:
Overview of chunking units proposed for fingerprinting methods

Character n-grams (*n* consecutive characters)		Barrón Cedeño & Rosso, 2009; Butakov & Scherbinin, 2009; Grozea & Popescu, 2010; Heintze, 1996; Oberreuter, L'Huillier, Rios, & Valesquez, 2011; Scherbinin & Butakovk, 2009; Zou, Long, & Ling, 2010.
Words	All words	Bernstein & Zobel, 2004, Broder, Glassman, Manasse, & Zweig, 1997; Finkel, Zaslavsky,Monostori, & Schmidt, 2002; Kasprzak & Brandejs, 2010; Lydon, Malcolm, & Dickerson, 2001.
	Stop words removed	Chowdhury, Frieder, Grossman, & McCabe, 2002; Hoad & Zobel, 2003; Kasprzak, Brandejs, & Kripac, 2009; Shivakumar & Garcia Molina, 1995.
	Stop words alone	Stamatatos, 2011
Sentences		Brin, Davis, & Garcia Molina, 1995; Pereira & Ziviani, 2004.
Hybrid terms	Word-bound n-grams	Shen, Li, Tian, & Cheng, 2009.
	Sentence-bound character n-grams	Barrón Cedeño & Rosso, 2009, Campbell, Chen, & Smith, 2000.
	Sentence-bound word n-grams	Sorokina, Gehrke, Warner & Ginsparg, 2006.

The *chunk size* determines the granularity of a fingerprint. Larger chunk sizes are more restrictive selectors and thus benefit detection accuracy, because the probability that documents share substrings decreases with increasing substring length. Larger chunks are also computationally more efficient, because fewer chunks must be stored for each document. Yet, large chunks are susceptible to failure in detecting disguised plagiarism, because changing one character alters the fingerprint of a rather long text segment. Small chunks better deal with modifications, but require higher computational effort and tend to yield false-positives when matching common substrings that documents share by chance (Heintze, 1996; Hoad & Zobel, 2003). Table 2 lists chunk sizes of fingerprinting methods found in the literature.

Table 2:
Overview of chunk sizes proposed for fingerprinting methods

Single text unit		Brin et al., 1995; Shivakumar & Garcia Molina, 1995.
Multiple text units	Without overlap	Shen et al., 2009
	With overlap	Broder et al., 1997; Lyon et al., 2001; Sorokina et al, 2006.
Favorable chunk sizes	3-5 content words	Hoad & Zobel, 2003; Kasprzak & Brandejs, 2010; Lyon et al., 2001; Shivakumar & Garcia Molina, 1996.
	8 to 11 stop words	Stamatatos, 2011.
	character 3- / 4-grams	Barrón Cedeño & Rosso, 2009.

The *resolution* is the number of minutiae a fingerprint contains and can be either fixed or variable. A fixed-resolution fingerprint encodes a decreasing percentage of text the longer the document. Fixed-resolution fingerprints are computationally more efficient, but negatively correlated to detection accuracy, especially for long documents (Heintze, 1996). When using fixed-resolution fingerprints, a book may not share enough minutiae with a paragraph copied from it to be detectable (Schleimer, Wilkerson, & Aiken, 2003).

Variable-resolution fingerprinting methods compute more minutiae the longer the document and thus encode a higher percentage of the text. Therefore, a higher fingerprint resolution benefits detection accuracy, but requires higher computational effort (Hoad & Zobel, 2003; Schleimer et al., 2003).

Full fingerprinting considers all minutiae. However, the fingerprint index for a full-resolution fingerprinting PDS requires eight or more times the hard disk space of the original document collection and significant processing time (Bernstein & Zobel, 2004; Schleimer et al., 2003). Therefore, full-resolution fingerprinting PDS are not practical for collections containing millions of documents. Table 3 lists fixed- or variable-resolution fingerprinting methods.

Table 3:
Overview of fixed-resolution and variable-resolution fingerprinting methods

fixed	Heintze, 1996
variable	Barrón Cedeño & Rosso, 2009; Bernstein & Zobel, 2004; Brin et al., 1995, Broder et al., 1997; Grozea et al., Kasprzak et al., 2009; Lyon et al., 2001; Manber, 1994; Scherbinin & Butakov, 2009; Shivakumar & Garcia Molina, 1995; Sorokina et al., 2006.

The *chunk selection strategy* determines which text sections the fingerprint encodes and thereby makes comparable to other documents. A selection of chunks is necessary, because the computational requirements of full-resolution fingerprinting are too high for most practical use cases. Table 4 summarises common chunk selection strategies.

Table 4:
Overview of common chunk selection strategies for fingerprinting methods

Start chunk at common substrings	Manber, 1994
Probabilistic selection	Brin et al., 1995; Broder et al., 1997.
Frequency-based selection	Heintze, 1996, Monostori et al., 2002; Schleimer et al., 2003.

The *similarity function* considers the minutiae that a suspicious text shares with a document in the reference collection to calculate a similarity score. Documents of the reference collection that exceed a certain threshold score represent potential plagiarism sources. The most basic similarity function, e.g. used by Kasprzak and Brandejs, defines a fixed number of matching minutiae as the threshold (Kasprzak & Brandejs, 2010). Another intuitive similarity function considers the fraction of all minutiae of a suspicious document that overlap with minutiae of a genuine document. More sophisticated similarity functions consider the length of documents (Bernstein & Zobel, 2004), relative frequencies of minutiae (Scherbinin & Butakov, 2009), or maximal differences in minutiae vectors (Zou et al., 2010).

Term occurrence analysis

Researchers frequently adopt string matching and vector space models for external PD tasks. This section explains both approaches and outlines their capabilities and limitations.

String matching

String matching refers to searching for a given character sequence in a text. PDS employing string matching commonly use suffix document models, which store each substring of a text. The PDS must compute suffix document models for the suspicious document and the entire reference collection. Because the string to search for is unknown in a PD setting, the PDS must select portions of the suspicious text and check them against all other models (Baker, 1993).

The strength of string matching PD methods is their accuracy in detecting verbatim text matches. Suffix document models encode the complete character information of a text, which distinguishes them from the document models that most fingerprinting methods employ. If two documents share substrings, suffix document models enable the detection of this overlap.

The major drawbacks of string matching in a PD context are the difficulty of detecting disguised plagiarism, which is attributable to the exact matching approach, and the high computational effort required. The most space-efficient suffix document models require about eight times as much storage space as the original document (Kurtz, 1999). Additionally, the time required for pre-computing suffix models practically prohibits the application of PDS that solely use string matching for large document sets. However, string matching becomes feasible when performed in the detailed analysis phase, after a less expensive method limits the collection size.

Vector space models

Vector space models (VSM) consider the terms of a text as unordered sets, represent the sets as vectors and compare the vectors using specialised measures. VSM consider the set *N* of all terms occurring in a collection of texts and use the *n* elements (terms) in *N* as dimensions of an *n*-dimensional space. Each text *i* of the collection is encoded as a sparse *n*-dimensional vector by recording the number of occurrences of a specific term *t* within the text *i*. Most commonly, PDS use one vector space models to encode the entire document. Some PDS employ multiple models, which encode paragraphs or sentences, to perform a more local similarity assessment. Table 5 shows papers that used global or local VSM as part of a PDS.

Table 5:
Overview of the scope of VSM proposed for plagiarism detection

entire document	Devi, Rao, Ram, & Akilandeswari, 2010; Dreher, 2007; Hoad & Zobel, 2003; Micol, Ferrandez, Llops & Munoz, 2010; Si, Leong, & Lau, 1997.
Sentences	Hariharan, Kamal, Faisal, Azharudheen, & Raman, 2010; Kang, Gelbukh, & Han, 2006; Muhr, Kern, Zechner, & Granitzer, 2010.

Most VSM consider words as terms, yet any unit of text qualifies as a *term unit*. Commonly, terms undergo *preprocessing* prior to constructing the model. Preprocessing may include stemming of words, de-capitalisation, stop word and punctuation removal, number replacement or part-of-speech tagging. Table 6 summarises term units of VSM employed for PD purposes.

Table 6:
Overview of the term units of VSM proposed for plagiarism detection

words	Dreher, 2007; Micol et al., 2010, Si et al., 1997.
word n-grams	Basile, Benedetto, Cagliotti, Cristadoro, & Esposti, 2009; Devi et al., 2010.
Sentences	Hariharan et al., 2010; Kang et al., 2006; Muhr et al., 2009.

Ranking documents by their degree of similarity requires a *similarity function* to return a numeric score. Most PDS used the cosine measure, which is a basic mathematical concept to calculate the similarity of arbitrary vectors based on their relative position in the vector space (Dreher, 2007; Hariharan et al., 2010; Muhr et al., 2009; Si et al., 1997).

VSM commonly include a *term weighting scheme* to determine the most relevant terms in a text prior to calculating a similarity score. The *tf-idf* scheme, which considers a term's frequency (*tf*) in a document and normalises it by the term's inverse frequency in all documents of the collection (*idf*), is the most widely used approach (Devi et al., 2010; Dreher, 2007; Hariharan et al., 2010; Kang et al., 2006; Si et al., 1997). The *tf-idf* scheme assigns high weights to terms that occur frequently within the analysed text, but infrequently in the entire collection. The idea is that such terms are likely specific content words that characterise a topic, which few other documents in the collection address.

VSM are well-researched and well-performing approaches for identifying verbatim text overlaps. The global similarity assessment on the document level that most VSM perform tends to be detrimental to detection accuracy in PD settings. This is because verbatim plagiarism more often encompasses smaller, confined segments of a text, which favours local similarity analysis.

Cross-language plagiarism detection

Cross-language plagiarism detection (CLPD) is an external PD approach that aims to identify documents plagiarised by translation from source documents in another language. To scale to large document collections, CLPD methods should follow the three-stage PD process composed of a heuristic retrieval, a detailed analysis and a knowledge-based post-processing phase (Potthast, Stein, & Anderka, 2008). Some prototypical PDS do not follow this guideline, but address CLPD tasks by machine translating all documents in the reference collection prior to applying monolingual PD methods (Kasprzak & Brandejs, 2010; Muhr et al., 2010; Zou et al., 2010). However, this approach is only feasible for smaller local collections (Potthast, Barrón Cedeño, Eiselt, Stein, & Rosso, 2010).

For the *heuristic retrieval* phase, a CLPD method may construct a monolingual keyword index for the reference collection, extract and machine-translate keywords from a suspicious document in another language, and query the index with the translated keywords. Alternatively, a CLPD method could machine-translate the entire suspicious document prior to extracting keywords and querying the index. In the second case, the detection method could also use a fingerprint index instead of a keyword index (Potthast, Barrón Cedeño, Stein, & Rosso, 2011; Potthast et al., 2008).

For the *detailed analysis* phase, CLPD methods can apply a number of retrieval models from cross-language information retrieval. Such models can either use pre-computed dictionaries (Ceska, 2008; Pouliquen, Steinberger, & Ignat, 2003; Steinberger, Pouliquen, & Hagman, 2002) or character similarities if the languages of the reference collection and suspicious document share sufficient syntactical similarities (Mcnamee & Mayfield, 2004).

Currently CLPD attracts less attention than monolingual PD and most research focuses on the similarity assessment in the detailed analysis stage. We found no PDS that implements the complete CLPD process. Potthast et al. view CLPD research as being “[…] still in its infancy” (Potthast, Barrón Cedeño, et al., 2011, p. 15).

Citation-based plagiarism detection

Citation-based plagiarism detection (CbPD) is an external PD method that approximates the semantic similarity of academic documents by measuring structural similarity using citation patterns. Citations, i.e. in-text pointers to the references in the bibliography of academic documents, have long been recognised as containing valuable information on semantic document relatedness (de Solla Price, 1965; Fano, 1956, Small, 1973).

In addition to offering semantic information, citations possess two characteristics valuable for plagiarism detection. First, citations are language-independent, because citing standards exist in the international academic community. Second, citations in plagiarised text are harder to alter than the language of the text, because authors of genuine works choose the sources they cite carefully and with specific goals in mind (Brooks, 1986, Garfield & Sher, 1963). Substituting or deleting citations without raising suspicion increases the effort to disguise plagiarism.

Gipp and Beel (2010) proposed exploiting the semantic information contained in citation patterns for plagiarism detection purposes. Citation patterns are sequences of citations that are shared between two documents *A* and *B*, as well as potentially intermediate non-shared citations. Figure 4 depicts the concept of CbPD.

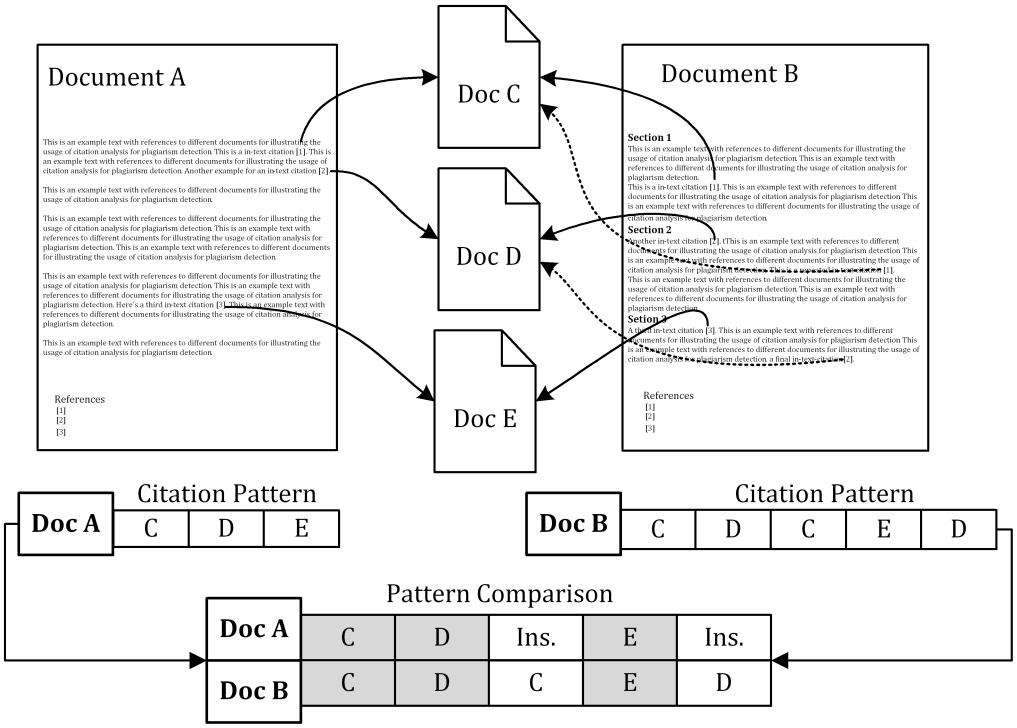

Figure 4: Concept of citation-based plagiarism detection

For identifying citation patterns, Gipp and Meuschke (2011) proposed several detection algorithms. For quantifying the patterns' similarity, the algorithms consider the order, proximity, absolute number and relative fraction of shared citations, and the probability that citations co-occur. The probability that documents share citations depends on factors such as citation counts, publication date, topic, and author connections (Meuschke, Gipp, & Breitinger, 2012).

The strength of CbPD is its ability to detect disguised plagiarism, given the documents share sufficient citations. In an analysis of a real doctoral thesis containing translated plagiarism, CbPD detected 13 of the 16 translated plagiarism instances, thus outperforming character-based PDS, which could not identify a single instance (Gipp, Meuschke, & Beel, 2011). Another advantage of CbPD compared to character-based detection methods is its lower computational effort. Only a small fraction of documents in a collection share citations. Therefore, the number of document comparisons CbPD must perform is smaller by two to three magnitudes. Moreover, citations represent only a small fraction of a document's content, thus individual document comparisons require less effort.

CbPD, however, is not a substitute, but rather a complement to character-based detection methods. CbPD tends to require longer text segments containing three or more shared citations, while character-based detection methods can identify very short instances of plagiarism regardless of whether documents share citations as long as the instances have sufficient literal text overlap.

Stylometry for intrinsic plagiarism detection

Stylometry subsumes statistical methods to quantify and analyse an author's writing style (Juola, 2008). Intrinsic PD methods employ stylometry to construct quantitative style models for segments of a text. The goal is to identify segments that are stylistically different from other segments, and thus potential indicators of plagiarism. Commonly, intrinsic methods analyse structural text segments, e.g. paragraphs or

chapters, or decompose a text into fixed-length segments based on character or word counts for analysis (Meyer zu Eissen & Stein, 2006; Stamatatos, 2009b, Suarez, Gonzalez, & Villena Roman, 2010; Uzuner, Katz, & Nahnsen, 2005).

Technically, intrinsic PD methods solve a one-class classification problem. Genuine text segments that share characteristic attributes represent the target class, while plagiarised segments form outliers with divergent attributes. A *style model* defines the attributes considered for analysis. Style models generally use a unique combination of the more than 1,000 features that researchers proposed for stylometry. Possible features can be lexical (e.g. average word lengths), syntactic (e.g. part-of-speech frequencies), or structural (e.g. average paragraph length) (Gruner & Naven, 2005). Based on the style model, a classification method must learn the characteristics of the target class and use them to reject outliers (Stein, Lipka, & Pretenhofer, 2011).

The advantage of intrinsic PD is its independence from a reference collection. Thus, in theory, intrinsic PDS can give a quick overview of document segments that need further assessment in a plagiarism investigation. The accuracy and reliability of automated stylometric analyses depends on multiple factors, including the observed linguistic attributes, genre, volume and purity of the analysed text. For instance, quoted text, headings, tables or figures can significantly skew style statistics (Juola, 2008, p. 246; Stamatatos, 2009a). Joint publications are another obstacle to text purity. Detecting writing style differences that signal potential plagiarism, and not simply multiple authorship, is a challenge for these kinds of documents (Maurer et al., 2006).

Evaluation of plagiarism detection systems

Comparing the detection performance of PDS is challenging. Authors proposing PDS prototypes often use non-standardised evaluation methods. In a review of 139 publications on PD, Potthast et al. found that 80% of the papers used individual corpora for evaluation and less than 50% offered comparisons to prior research (Potthast, Stein, Barrón Cedeño, & Rosso, 2010). For publicly available PDS, evaluations are even less objective.

We found two projects that address this lack of comparability. Both benchmark PDS using standardised collections. The first project is the annual PAN International Competition on Plagiarism Detection (PAN-PC), initiated in 2009 (Potthast, Stein, Eiselt, Barrón Cedeño, & Rosso, 2009). PAN is an acronym for “Plagiarism Analysis, Authorship Identification, and Near-Duplicate Detection”. Competitors in the PAN-PC primarily present research prototypes. The second project is a comparison of commercial and otherwise publicly available PDS, which a research group at the HTW University of Applied Sciences in Berlin performs periodically (Plagiarism Research Group HTW Berlin, 2010). We will refer to this test series as the HTW PDS Tests. We will present results from the PAN-PC in 2011 to point out the capabilities of state-of-the-art PDS prototypes and subsequently discuss the findings from the latest HTW Test for external PDS to highlight the strengths and weaknesses of PDS available to the public.

Research prototypes

The PAN-PC offers tasks for external and intrinsic plagiarism detection. The evaluation corpus of PAN-PC’11 contained 26,939 documents, of which 50% were suspicious texts, and the remainder formed the reference collection. Suspicious documents contained 61,064 artificially plagiarised sections, of which 82% were obfuscated by applying the following techniques:

- using automated or manual English translations of German and Spanish text sections

- performing random shuffles, insertions, deletions or semantic substitutions of terms
- asking humans to paraphrase sections (Potthast, Eiselt, et al., 2011).

Figure 5 illustrates the results of the PAN-PC'11. The figure shows the plagiarism detection (*plagdet*) score of the five best performing external PDS grouped by the obfuscation technique applied to the plagiarized text segments.

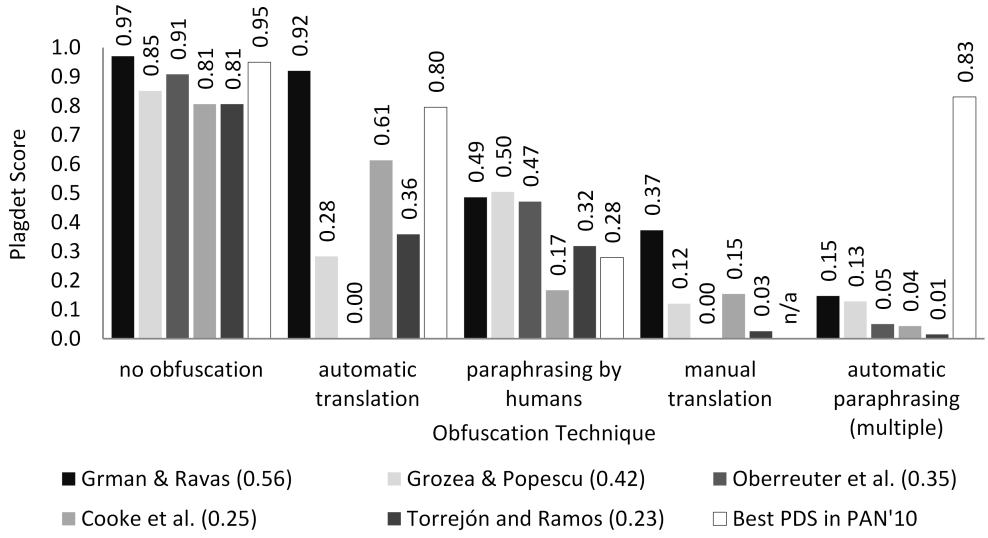

Figure 5: Plagdet scores for PDS in PAN-PC'11 by obfuscation technique
Source: (Potthast, Eiselt, et al., 2011)

The *plagdet* score considers the equally weighted harmonic mean of precision (*P*) and recall (*R*) and combines this mean with the granularity (*gran*) of a detection method. Precision denominates what percentage of all instances that a detection method reports as suspicious are plagiarism. Recall denominates what percentage of all plagiarised instances in the collection a detection method reports. The granularity reflects whether a method detects a plagiarised instance as a whole or in multiple parts. The interval of the score is [0,1]. For the exact computation of the score, see (Potthast, Barrón Cedeño, et al., 2010).

The rightmost bars with no shading in Figure 5 show the *plagdet* score of the best performing system in the competition of the previous year, PAN-PC'10. However, the bars can only provide a rough indication of the advancement of detection performance, because the evaluation corpus of PAN-PC'11 included more obfuscated segments than the corpus for PAN-PC'10. Moreover, the corpus of PAN-PC'11 included manual translations, whereas the corpora of all previous competitions included only automatic translations. Each legend entry states the overall *plagdet* score, which is the mean of the scores in the individual groups, in brackets.

Given the results, we conclude that state-of-the-art PDS can detect copies of text segments with high accuracy. Detection rates for segments plagiarised by humans are substantially lower than for non-obfuscated segments. For example, the system of Grman and Ravas (2011), which overall performed best in PAN-PC'11, achieved a recall of $R = 0.33$ for manually paraphrased segments (Potthast, Eiselt et al., 2011). In other words, the best performing system failed to identify two-thirds of the manually paraphrased plagiarism instances. There is a notable decrease in the detection performance for automatically obfuscated passages in PAN-PC'11 compared to the

earlier PAN-PC'10. We attribute this decline to the increased amount of obfuscated test cases that the organizers added to the evaluation corpus of PAN-PC'11.

The seemingly good detection performance for automatically translated text segments is misleading. The systems that performed well used automated services for translating foreign language documents in the reference collection into English. The employed services, such as Google Translate, are similar or identical to the ones used to construct the translated, plagiarised sections in the first place (Potthast, Barrón Cedeño, et al., 2011; Potthast, Eiselt, et al., 2011). The detection rates for manually translated plagiarisms are substantially lower. For instance, the best performing system of Grman and Ravas achieved a recall *R=0.26* for manually translated segments (Potthast, Eiselt, et al., 2011). We hypothesise that the translation undertaken by real authors when obfuscating their plagiarism is more complex and versatile, and hence harder to detect by the tested systems.

Figure 6 displays the *plagdet* scores of the four systems participating in the intrinsic detection track of PAN-PC'11. All systems performed significantly worse than those in the external track. The organisers attribute the good relative performance of the system presented by Oberreuter et al. (2011) to exploiting the artificial way of creating most plagiarised sections in the evaluation corpus. The procedures for generating artificial plagiarisms copy text from source documents regardless of topical relatedness. This benefits the system of Oberreuter et al. (2011), which evaluates the uniqueness of words relative to the rest of the analysed documents. This approach is unlikely to be reproducible in realistic settings (Potthast, Eiselt, et al., 2011). The performance of the remaining systems is in line with earlier PAN competitions. For comparison, a naïve baseline approach of classifying all segments as plagiarised achieved a recall *R = 0.46*, precision *P = 0.23* and plagdet score of 0.24 in 2009 (Potthast et al., 2009).

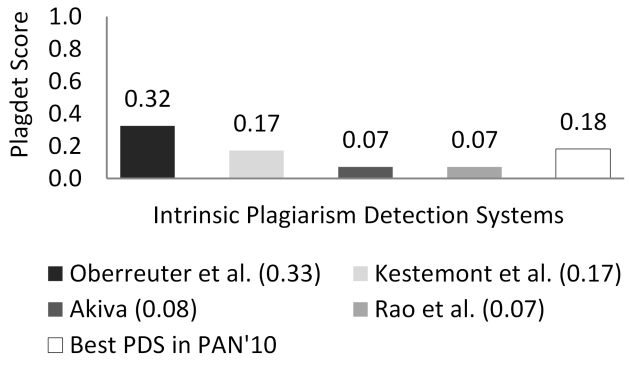

Figure 6: Plagdet scores for intrinsic PDS in PAN-PC'11
Source: (Potthast, Eiselt, et al., 2011)

Intrinsic PD appears to require longer texts to work reliably. Stein et al. analysed a subset of the PAN-PC'09 evaluation corpus. They excluded documents under 35,000 words from their evaluation for not being reliably analysable. Stein et al. report precision values ranging from 0.72 to 0.98 with corresponding recall values ranging from 0.30 to 0.60 depending on the used sub-collection (Stein et al., 2011).

Plagiarism detection systems available to the public

The latest HTW Test for external PDS in 2010 evaluated 26 publicly available detection systems using 40 manually composed essays – of which 30 were written in German and 10 in English. Most documents contained copy and paste or

shake and paste plagiarism in longer sections of the text. The sources of plagiarism are available on the internet, except for one document, which originated from a DVD encyclopedia. Five plagiarisms are manually or machine translated from English to German and one from French to English (Plagiarism Research Group HTW Berlin, 2010). If authors disguised plagiarism, they employed moderate text alterations. According to the observations of the evaluators, the obfuscation resembles the common plagiarism behaviour of students (Weber-Wulff, 2010). We view the resulting obfuscation to be comparably weaker than the manually rewritten segments contained in the PAN-PC'11.

The organisers use a three-class scale to benchmark the reliability of tested PDS. The exact scoring criteria depended on the individual test documents. For instance, the organisers judged whether a PDS could identify all sources of a plagiarism (3 points), nearly all sources (2 points), some sources (1 point) or no sources (0 points) (Weber-Wulff, 2010).

Figure 7 displays the number of test cases discovered by the top five systems in the HTW PDS Test 2010. Most undetected cases resulted from the six translations in the corpus. Due to the light obfuscation, the systems identified most other plagiarisms more or less completely.

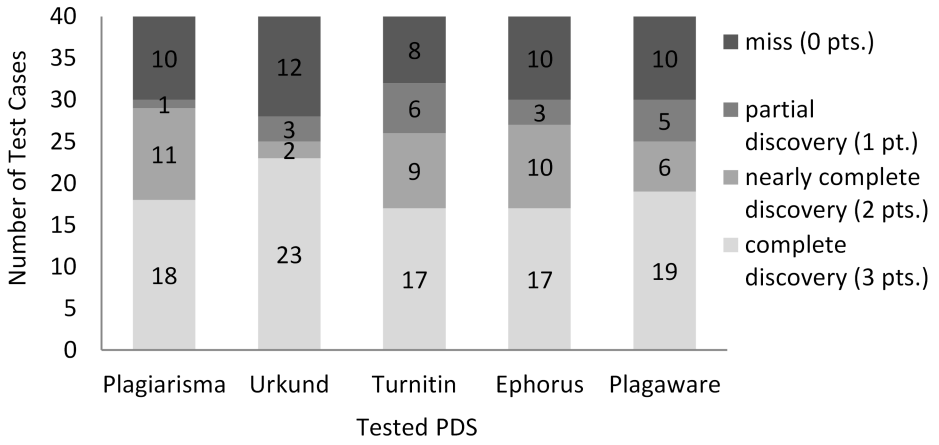

Figure 7: Performance of top five publicly available PDS
Source: (Plagiarism Research Group HTW Berlin, 2010)

Technical weaknesses of plagiarism detection systems

Technical weaknesses can significantly decrease the detection accuracy of PDS. The term technical disguise subsumes techniques to obfuscate plagiarism by exploiting technical weaknesses of PDS. Technical disguise solely affects the machine internal representation of text, which the PDS processes, while keeping the text unaltered to the human eye.

One example of technical disguise is inserting characters with font color identical to the background into plagiarised text. This renders the text as nonsense to the PDS. A similar cloak for plagiarised text is replacing letters from the original alphabet with letters from foreign alphabets that feature visually identical glyphs (Palkovskii, 2009).

Heather demonstrated three cloaks that are especially suitable for altering PDF documents (Heather, 2010). The first cloaking method is to slightly alter the PDF's character map, which assigns visible glyphs to machine-processable characters. This change renders plagiarised text meaningless to a PDS. The second cloak achieves

the same effect, but alters the font definition of plagiarised text to map, for example, the glyph representing an ‘e’ to the character ‘r’ and vice versa. Subsequently, a plagiarist would have to replace all ‘e’ and ‘r’ characters with the respective counterpart. This procedure results in visually well-formed text in which a majority of words would be uninterpretable for a PDS. The third method converts the entire plagiarized text into a graphic. To avoid triggering a warning by the PDS, because the document no longer contains any analysable text, the plagiarist can enter genuine, but unrelated text. To hide the phony text, the plagiarist may format it in a background color, place it outside the physical boundaries of the page or behind the graphics.

Conclusion

Our review of detection approaches and their performance shows that PD methods face an inevitable tradeoff between detection accuracy and computational effort. Figure 8 summarises the capabilities of current PD methods in detecting the different forms of plagiarism.

	Application				Form of plagiarism			
Detection Method	**Extrinsic PD**	**Intrinsic PD**	**Mono-lingual PD**	**Cross-lingual PD**	**(Near) copies**	**Disguised**	**Translated**	**Idea**
Character-based (Char.)	X		X					
Exact String Matching								
Approximate String Matching								
Fingerprinting								
Vector Space Models								
Semantic Enhancements								
Cross-language (CLPD)	X			X				
Citation-based (CbPD)	X		X	X				
Stylometry (Style)		X	X					
Future Perspective								
Hybrid (Char. / CLPD / CbPD / Style)	X		X	X				

Detection rate: Good | Fair | Poor | Unfit

Figure 8: Capabilities of current plagiarism detection methods

We showed that all external monolingual PD methods rely on character-based similarity between documents. Therefore, the detection accuracy of these methods decreases with increasing disguise of plagiarism. String-matching methods exhibit the strongest dependence on character-based similarity. By applying suitable term selection, fingerprinting or vector space model approaches are more stable against character alterations, but incur information loss and fail when character-based similarity falls below a certain level. The lack of textual overlap also makes translations and idea plagiarisms impossible to detect for character-based methods.

© International Journal for Educational Integrity Vol. 9 No. 1 June, 2013 pp. 50–71 ISSN 1833-2595

External, cross-language plagiarism detection is not mature or reliable at the time of writing (Potthast, Barrón Cedeño, et al., 2011). Machine translating all documents in the reference collection not written in the target language, an approach applied by some prototypes in the PAN-PC, is not scalable in practice (Potthast, Barrón Cedeño, et al., 2010).

The results of the PAN competitions, the HTW PDS Test and other studies (Hill & Page, 2009; Kakkonen & Mozgovoy, 2010; Maurer & Zaka, 2007; Scaife, 2007) prove that state-of-the-art PDS, which implement external detection methods, find incidences of verbatim and slightly modified copying with high accuracy, given the sources are accessible to the PDS. Prof. Weber-Wulff accurately assesses the current state of PDS when stating: "[…] PDS find copies, not plagiarism." (Weber-Wulff, 2010, p. 6) and: "[…] for translations or heavily edited material, the systems are powerless […]" (Plagiarism Research Group HTW Berlin, 2010).

Aside from text alterations, technical disguise can fool existing PDS. The major systems seem to not yet have implemented countermeasures. However, we expect that integrating additional checks to reveal technical disguise will present a minor challenge to future PDS.

Many researchers recognise the need to incorporate semantic information into similarity checks to allow detecting disguised plagiarism (Bao, Lyon, Lane, Wei, & Malcolm, 2007; Leung & Chan, 2007; Pera & Ng, 2011, Tsatsaronis, Varlamis, Giannakoulopoulos, & Kanellopoulos, 2010). In the experiments of Bao et al., considering synonyms increased detection performance by factor two to three. However, the processing time increased by factor 27 (Bao et al., 2007). We regard current character-based PD methods that include semantic analysis as computationally too expensive for most practical PD tasks.

Citation-based plagiarism detection is a language-independent external PD approach that considers the semantic similarity of academic documents and is computationally feasible also for large collections. In experiments, CbPD outperformed current character-based methods in detecting real-world cases of translated plagiarism (Gipp et al., 2011). To work effectively, CbPD requires sufficient shared citations, which typically implies that longer text segments have been plagiarised. A technical obstacle to CbPD is the automated acquisition of citation data, which currently works well for some, but not all citation styles (Meuschke et al., 2012).

Intrinsic plagiarism detection based on stylometry is another approach that can overcome the boundaries of character-based similarity by comparing linguistic similarity. Given that the stylistic differences between plagiarised and original text are significant, and not due to legitimate multiple authorship, stylometry is a capable aid in identifying disguised plagiarism. When a plagiarist paraphrases text to the point where it resembles the expressions of the plagiarist, stylometry fails. The results of PAN-PC 2010, PAN-PC 2011, and the experiments by Stein et al. (2011) indicate that stylometry only works reliably for document lengths of several thousand or tens of thousands of words. This restricts the applicability of this method for PD. We found no PDS in practical use that performed intrinsic PD.

Reliably detecting paraphrases, translated plagiarism and idea plagiarism requires novel approaches. Research on cross-lingual, citation-based, and intrinsic PD may provide the necessary advances to make detectable these strongly disguised forms of plagiarism in the future. Initial results show promise, although none of the three approaches is yet reliable or scalable enough for practical use. To achieve the best possible performance, future PDS could benefit from combining character-based with cross-lingual, citation-based, and intrinsic PD approaches.

References

Baker, B. S. (1993). On finding duplication in strings and software. Online Source. Retrieved June 16, 2010, from http://-cm.bell-labs.com/-cm/-cs/-doc/-93/-2-bsb-1.ps.gz

Bao, J., Lyon, C., Lane, P. C. R., Wei, J., & Malcolm, J. A. (2007). Comparing different text similarity methods. Technical report, Science and Technology Research Institute, University of Hertfordshire.

Barrón Cedeño, A., & Rosso, P. (2009). On automatic plagiarism detection based on n-grams comparison. In *Advances in Information Retrieval*: *Vol. 5478. Lecture Notes in Computer Science* (pp. 696–700). Springer.

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 PAN Workshop. Uncovering Plagiarism, Authorship and Social Software Misuse*.

Bernstein, Y., & Zobel, J. (2004). A scalable system for identifying co-derivative documents. In *String Processing and Information Retrieval: Vol. 3246. Lecture Notes in Computer Science* (pp. 1–11). Springer.

Born, A. D. (2003). How to reduce plagiarism. *Journal of Information Systems Education*, *14*(3), 223–224.

Bretag, T., & Mahmud, S. (2009). Self-plagiarism or appropriate textual re-use? *Journal of Academic Ethics*, 7, 193–205.

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

Broder, A. Z., Glassman, S. C., Manasse, M. S., & Zweig, G. (1997). Syntactic clustering of the web. *Computer Networks and ISDN Systems*, *29*(8-13), 1157–1166.

Brooks, T. (1986). Evidence of complex citer motivations. *Journal of the American Society for Information Science*, *37*(1), 34–36.

Butakov, S., & Scherbinin, V. (2009). The toolbox for local and global plagiarism detection. *Computers & Education*, *52*(4), 781–788.

Campbell, D. M., Chen, W. R., & Smith, R. D. (2000). Copy Detection Systems for Digital Documents. In *Proceedings of the Conference on Advances in Digital Libraries* (pp. 78–88). Los Alamitos, CA, USA: IEEE.

Ceska, Z. (2008). Plagiarism detection based on singular value decomposition. In *Advances in Natural Language Processing: Vol. 5221. Lecture Notes in Computer Science* (pp. 108–119). Springer.

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

Clough, P. (2000). *Plagiarism in natural and programming languages an overview of current tools and technologies*. Technical report, Department of Computer Science, University of Sheffield.

Collberg, C., & Kobourov, S. (2005). Self-plagiarism in computer science. *Communications of the ACM*, *48*(4), 88–94.

de Solla Price, D. J. (1965). Networks of scientific papers. *Science*, *149*(3683), 510–515.

Devi, S. L., Rao, P. R. K., Ram, V. S., & Akilandeswari, A. (2010). External plagiarism detection – Lab report for PAN at CLEF 2010. In *Notebook Papers of CLEF 2010 LABs and Workshops*.

Dreher, H. (2007). Automatic conceptual analysis for plagiarism detection. *Information and Beyond: The Journal of Issues in Informing Science and Information Technology*, *4*, 601–614.

Ercegovac, Z., & Richardson Jr., J. V. (2004). Academic dishonesty, plagiarism included, in the digital age: A literature review. *College and Research Libraries*, *65*(4), 301–318.

Fano, R. M. (1956). Information theory and the retrieval of recorded information. *Documentation in Action* (pp. 238–244). New York: Reinhold Publ. Co.

Finkel, R. A., Zaslavsky, A. B., Monostori, K., & Schmidt, H. W. (2002). Signature extraction for overlap detection in documents. In *Proceedings of the 25th Australasian Computer Science Conference*, volume 4 of *Conferences in Research and Practice in Information Technology* (pp. 59–64) Melbourne, Australia: Australian Computer Society Inc.

Garfield, E., & Sher, I. (1963). New factors in the evaluation of scientific literature through citation indexing. *American Documentation*, *14*(3), 195–201.

Garner, B. A. (2011). *Garner's dictionary of legal usage* (3rd ed.). Oxford University Press.

Gipp, B., & Beel, J. (2010). Citation based plagiarism detection: A new approach to identify plagiarized work language independently. In *Proceedings of the 21st ACM Conference on Hypertext and Hypermedia* (pp. 273–274). ACM. Retrieved from http://-sciplore.org/-pub.

Gipp, B., & Meuschke, N. (2011). Citation pattern matching algorithms for citation-based plagiarism detection: greedy citation tiling, citation chunking and longest common citation sequence. In *Proceedings of the 11th ACM Symposium on Document Engineering* (pp. 249–258). Mountain View, CA, USA, ACM. Retrieved from http://-sciplore.org/-pub.

Gipp, B., Meuschke, N., & Beel, J. (2011). Comparative evaluation of text- and citation-based plagiarism detection approaches using GuttenPlag. In *Proceedings of 11th ACM/IEEE-CS Joint Conference on Digital Libraries (JCDL'11)* (pp. 255–258). Ottawa, Canada: ACM. Retrieved from http://-sciplore.org/-pub.

Grman, J., & Ravas, R. (2011). Improved implementation for finding text similarities in large collections of data. In *Notebook Papers of CLEF 2011 LABs and Workshops*, Amsterdam, Netherlands.

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

Grozea, C., & Popescu, M. (2010). ENCOPLOT: Performance in the second international plagiarism detection challenge. In *Notebook Papers of CLEF 2010 LABs and Workshops*, Padua, Italy.

Gruner, S., & Naven, S. (2005). Tool support for plagiarism detection in text documents. In *Proceedings of the 2005 ACM Symposium on Applied Computing* (pp. 776–781). ACM.

Hariharan, S., Kamal, S., Faisal, A. V. M., Azharudheen, S. M., & Raman, B. (2010). Detecting plagiarism in text documents. In *Proceedings of the International Conference on Recent Trends in Business Administration and Information Processing: Vol. 70. Communications in Computer and Information Science* (pp. 497–500). Trivandrum, Kerala, India: Springer.

Heather, J. (2010). Turnitoff: Identifying and fixing a hole in current plagiarism detection software. *Assessment & Evaluation in Higher Education*, *35*(6), 647–660.

Heintze, N. (1996). Scalable document fingerprinting. In *1996 USENIX Workshop on Electronic Commerce*.

Hill, J. D., & Page, E. F. (2009). An empirical research study of the efficacy of two plagiarism-detection applications. *Journal of Web Librarianship*, *3*(3),169–181.

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

Howard, R. M. (2007). Understanding "internet plagiarism". *Computers and Composition*, *24*(1), 3–15.

Juola, P. (2008). Authorship Attribution. *Foundations and Trends Information Retrieval*, 1:233–334.

Kakkonen, T., & Mozgovoy, M. (2010). Hermetic and web plagiarism detection systems for student essays — an evaluation of the state-of-the-art. *Journal of Educational Computing Research*, *42*(2), 135–159.

Kang, N., Gelbukh, A., & Han, S. (2006). PPChecker: Plagiarism pattern checker in document copy detection. In *Text, Speech and Dialogue: Vol. 4188*. *Lecture Notes in Computer Science* (pp. 661–667). Springer.

Kasprzak, J., & Brandejs, M. (2010). Improving the reliability of the plagiarism detection system – Lab report for PAN at CLEF 2010. In *Notebook Papers of CLEF 2010 LABs and Workshops*, Padua, Italy.

Kasprzak, J., Brandejs, M., & Kripac, M. (2009). Finding plagiarism by evaluating document similarities. In *Proceedings of the 3rd PAN Workshop. Uncovering Plagiarism, Authorship and Social Software Misuse*.

Kurtz, S. (1999). Reducing the space requirement of suffix trees. *Software-Practice and Experience*, *29*(13), 1149–1171.

Lancaster, T. (2003). *Effective and efficient plagiarism detection*. PhD thesis, School of Computing, Information Systems and Mathematics South Bank University.

Leung, C.-H., & Chan, Y.-Y. (2007). A natural language processing approach to automatic plagiarism detection. In *Proceedings of the 8th ACM SIGITE Conference on Information Technology Education* (pp. 213–218). ACM.

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).

Manber, U. (1994). Finding similar files in a large file system. In *Proceedings of the USENIX Winter Technical Conference* (pp. 2–11). Berkeley, CA, USA: USENIX Association.

Maurer, H., Kappe, F., & Zaka, B. (2006). Plagiarism – A survey. *Journal of Universal Computer Science*, *12*(8), 1050–1084.

Maurer, H., & Zaka, B. (2007). Plagiarism – A problem and how to fight it. In *Proceedings of World Conference on Educational Multimedia, Hypermedia and Telecommunications* (pp. 4451–4458). Vancouver, Canada: AACE.

Mcnamee, P., & Mayfield, J. (2004). Character N-Gram tokenization for European language text retrieval. *Information Retrieval*, *7*, 73–97.

Meuschke, N., Gipp, B., & Breitinger, C. (2012). CitePlag: A citation-based plagiarism detection system prototype. In *Proceedings of the 5th International Plagiarism Conference*, Newcastle upon Tyne, UK. Retrieved from http://-sciplore.org/-pub.

Meyer zu Eissen, S., & Stein, B. (2006). Intrinsic plagiarism detection. In *Proceedings of the 28th European Conference on IR Research: Vol. 3936. Lecture Notes in Computer Science* (pp. 565–569). London, UK: Springer.

Micol, D., Ferrández, Ó., Llopis, F., & Muñoz, R. (2010). A textual-based similarity approach for efficient and scalable external plagiarism analysis – lab report for PAN at CLEF 2010. In *CLEF (Notebook Papers/LABs/Workshops)*.

Monostori, K., Finkel, R., Zaslavsky, A., Hodász, G., & Pataki, M. (2002). Comparison of overlap detection techniques. In *Proceedings of the International Conference on Computational Science*: Vol. 2329. *Lecture Notes in Computer Science* (pp. 51–60). Amsterdam, Netherlands: Springer.

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 *Notebook Papers of CLEF 2010 LABs and Workshops*, Padua, Italy.

Muhr, M., Zechner, Mario Kern, R., & Granitzer, M. (2009). External and intrinsic plagiarism detection using vector space models. In *Proceedings of the 3rd PAN Workshop. Uncovering Plagiarism, Authorship and Social Software Misuse* (pp. 47–55).

Oberreuter, G., L'Huillier, G., Ríos, S. A., & Velásquez, J. D. (2010). FastDocode: Finding approximated segments of N-Grams for document copy detection. In *Notebook Papers of CLEF 2010 LABs and Workshops*, Padua, Italy.

Oberreuter, G., L'Huillier, G., Ríos, S. A., & Velásquez, J. D. (2011). Approaches for intrinsic and external plagiarism detection. In *Notebook Papers of CLEF 2011 LABs and Workshops*, Amsterdam, Netherlands.

Oxford University Press. (2009). *A dictionary of psychology [electronic resource]*. Oxford Reference Online: Oxford University Press.

Palkovskii, Y. (2009). "Counter Plagiarism Detection Software" and "Counter Counter Plagiarism Detection" Methods. In *Proceedings of the 3rd Workshop on Uncovering Plagiarism, Authorship and Social Software Misuse and 1st International Competition on Plagiarism Detection*.

Park, C. (2003). In other peoples words: Plagiarism by university students – Literature and lessons. *Assessment Evaluation in Higher Education*, *28*(5), 471–488.

Pera, M. S., & Ng, Y.-K. (2011). SimPaD: A word-similarity sentence-based plagiarism detection tool on web documents. *Web Intelligence and Agent Systems: An International Journal (JWIAS)*, *9*(1).

Pereira, A. R. J., & Ziviani, N. (2004). Retrieving similar documents from the web. *Journal of Web Engineering*, *2*(4), 247–261.

Plagiarism Research Group HTW Berlin. (2010). Portal Plagiat - Softwaretest 2010. Online Source. Retrieved May 29, 2012, from: http://-plagiat.htw-berlin.de/-software/-2010-2/-.

Plagiarism Research Group HTW Berlin. (2012). Portal Plagiat - Softwaretest Report 2012. Online Source. Retrieved November 27, 2012, from: http://-plagiat.htw-berlin.de/-collusion-test-2012/-.

Potthast, M., Barrón Cedeño, A., Eiselt, A., Stein, B., & Rosso, P. (2010a). Overview of the 2nd international competition on plagiarism detection. In *Notebook Papers of CLEF 2010 LABs and Workshops*, Padua, Italy.

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

Potthast, M., Eiselt, A., Barrón Cedeño, A., Stein, B., & Rosso, P. (2011b). Overview of the 3rd international competition on plagiarism detection. In *Notebook Papers of CLEF 2011 LABs and Workshops*, Amsterdam, Netherlands.

Potthast, M., Stein, B., & Anderka, M. (2008). A Wikipedia-based multilingual retrieval model. In *Proceedings of the 30th European Conference on Advances in Information Retrieval* (pp. 522–530). Springer.

Potthast, M., Stein, B., Barrón Cedeño, A., & Rosso, P. (2010b). An evaluation framework for plagiarism detection. In *Proceedings of the 23rd International Conference on Computational Linguistics* (pp. 997–1005). Beijing, China: Association for 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 and 1st International Competition on Plagiarism Detection: Vol. 502.* (pp. 1–9).

Pouliquen, B., Steinberger, R., & Ignat, C. (2003). Automatic identification of document translations in large multilingual document collections. In *Proceedings of the International Conference Recent Advances in Natural Language Processing* (pp. 401–408).

Scaife, B. (2007). IT consultancy plagiarism detection software report for JISC Plagiarism advisory service. Technical report, Joint Information System Committee.

Scherbinin, V., & Butakov, S. (2009). Using Microsoft SQL server platform for plagiarism detection. In *Proceedings of the 3rd PAN Workshop. Uncovering Plagiarism, Authorship and Social Software Misuse*.

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). ACM.

Shen, Y., Li, S.-C., Tian, C.-G., & Cheng, M. (2009). Research on anti-plagiarism system and the law of plagiarism. In *Proceedings of the 1st International Workshop on Education Technology and Computer Science* (pp. 296–300).

Shivakumar, N., & Garcia Molina, H. (1995). SCAM a copy detection mechanism for digital documents. In *Proceedings of the 2nd Annual Conference on the Theory and Practice of Digital Libraries*, Austin, TX, USA.

Shivakumar, N., & Garcia Molina, H. (1996). Building a scalable and accurate copy detection mechanism. In *Proceedings of the 1st ACM International Conference on Digital Libraries* (pp. 160–168). ACM.

Si, A., Leong, Hong, V., & Lau, R. W. H. (1997). CHECK: A document plagiarism detection system. In *Proceedings of the ACM Symposium on Applied Computing* (pp. 70–77). ACM.

Small, H. (1973). Co-citation in the scientific literature: A new measure of the relationship between two documents. *Journal of the American Society for Information Science*, *24*, 265–269.

Sorokina, D., Gehrke, J., Warner, S., & Ginsparg, P. (2006). Plagiarism detection in arXiv. Technical report computer science, Cornell University TR2006-2046.

Stamatatos, E. (2009a). A survey of modern authorship attribution methods. *Journal of the American Society for Information Science and Technology*, *60*(3), 538–556.

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

Stamatatos, E. (2011). Plagiarism detection using stopword n-grams. *Journal of the American Society for Information Science and Technology*, *62*(12), 2512–2527.

Stein, B., Koppel, M., & Stamatatos, E. (Eds.). (2007a). *Plagiarism Analysis Authorship Identification, and Near Duplicate Detection: Vol. 276. CEUR Workshop Proceedings*. CEUR-WS.org. in Proceedings of the SIGIR 2007 International Workshop, held in conjunction with the 30th Annual International ACM SIGIR Conference, Amsterdam, Netherlands.

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

Stein, B., & Meyer zu Eissen, S. (2006). Near similarity search and plagiarism analysis. In *Proceedings of the 29th Annual Conference of the Gesellschaft für Klassifikation e.V.* (pp. 430–437). Magdeburg: Springer.

Stein, B., Meyer zu Eissen, S., & Potthast, M. (2007b). Strategies for retrieving plagiarized documents. In *Proceedings of the 30th Annual International ACM SIGIR Conference* (pp. 825–826). ACM.

Steinberger, R., Pouliquen, B., & Hagman, J. (2002). Document similarity calculation using the multilingual thesaurus EUROVOC. In *Proceedings of the 3rd International Conference on Computational Linguistics and Intelligent Text Processing* (pp. 415–424). London, UK: Springer.

Suárez, P., González, J. C., & Villena Román, J. (2010). A plagiarism detector for intrinsic plagiarism. In *Notebook Papers of CLEF 2010 LABs and Workshops*, Padua, Italy.

Tsatsaronis, G., Varlamis, I., Giannakoulopoulos, A., & Kanellopoulos, N. (2010). Identifying free text plagiarism based on semantic similarity. In *Proceedings of the 4th International Plagiarism Conference*, Newcastle upon Tyne, UK.

Uzuner, Ö., Katz, B., & Nahnsen, T. (2005). Using syntactic information to identify plagiarism. In *Proceedings of the 2nd Workshop on Building Educational Applications Using Natural Language Processing*, Ann Arbor, MI, USA.

Weber-Wulff, D. (2010). Test cases for plagiarism detection software. In *Proceedings of the 4th International Plagiarism Conference*, Newcastle upon Tyne, UK.

Weber-Wulff, D. (2011). Copy, Shake, and Paste – A Blog about Plagiarism written by a Professor for Media and Computing at the HTW. Online Source. Retrieved October 28, 2011, from: http://-copy-shake-paste.blogspot.com.

Zou, D., Long, W.-J., and Ling, Z. (2010). A cluster-based plagiarism detection method. In *Notebook Papers of CLEF 2010 LABs and Workshops, 22-23 September*, Padua, Italy.

About the authors

Norman Meuschke is a computer scientist researching plagiarism detection, digital document processing and document related information retrieval tasks. Norman is a member of the research project SciPlore at the University of California, Berkeley. His current work includes the design and development of an open-source hybrid PDS, which combines traditional character-based with novel semantic detection approaches.

Bela Gipp is a computer scientist working on innovative plagiarism detection approaches and scientific recommender systems at SciPlore, a research group at UC Berkeley/OvGU. Aside from theoretical research, he develops open-source software for scientists as a founder of SciPlore.org.