Citation-based Plagiarism Detection – Idea, Implementation and Evaluation

Bela Gipp

OvGU, Germany / UC Berkeley, California, USA
gipp@berkeley.edu

ABSTRACT

Currently used Plagiarism Detection Systems solely rely on text-based comparisons. They only deliver satisfying results if the plagiarized text is copied literally (copy&paste), with minor alterations (e.g. shake&paste) or is machine translated. However, if the text is paraphrased or translated by a human, the currently used methods yield a very poor performance. Using the words of Weber Wulff, who organizes regular comparisons for Plagiarism Detection Systems (PDS), the current state of available systems can be summarized as follows: "[…] PDS find copies, not plagiarism.".

In contrast to the existing approaches for Plagiarism Detection, Citation-based Plagiarism Detection compares the occurrences of citations in order to identify similarities. The most basic form is to measure the bibliographic coupling strength (citation overlap). However, this alone would lead to numerous false-positives, thus it is advisable to include further factors such as the order of citations, their proximity to each other, their chance of co-occurrence, and other more sophisticated measures. If e.g. four papers are cited in a similar order in two documents, this can be interpreted as a subtle hint that both works may not have been created independently of one another. If none of these four papers have been co-cited before in another paper or the order of citations is identical, this might indicate plagiarism.

The advantages and limitations of Citation-based Plagiarism Detection are very different from those of the currently used text-based methods. Text matching approaches continue to be suitable for detecting copy&paste plagiarism, even for short passages. They are also advantageous in that they do not require citation information; yet, they fail to identify e.g. paraphrased, translated and idea plagiarism. By applying the citation-based approach to the doctoral thesis of Guttenberg, which is a well-examined, real world plagiarism example tested by numerous conventional Plagiarism Detection Systems, we could show that the citation-based approach is able to identify 13 out of the 16 translated plagiarisms. Conventional methods failed to identify any of these sections.

However, as expected, short passages of copy&paste plagiarism can usually only be identified by text-based approaches. Therefore, Citation-based Plagiarism Detection is by no means a replacement for the currently used text-based approaches, but should be considered as a complement for identifying currently hard to find well-disguised plagiarisms. Additionally, once signs of plagiarism have been found, neither the text-based approaches, nor the citation-based approaches eliminate the need for manual examination.

Categories and Subject Descriptors

H.3.3 [**Clustering**]: INFORMATION STORAGE AND RETRIEVAL – *Information Search and Retrieval.*

General Terms

Algorithms, Experimentation, Measurement, Languages

Keywords

Plagiarism Detection, Citation Analysis, Citation Order Analysis, Citation Pattern Analysis

1. INTRODUCTION & MOTIVATION

Plagiarism describes the appropriation of another person's ideas, intellectual or creative work and passing them of as one's own [4]. The existing systems for plagiarism detection, such as fingerprinting or string matching, are generally able to identify copy&paste plagiarism; yet, they have weaknesses in that they fail to identify most other forms of plagiarism, including disguised plagiarism, translation plagiarism, idea plagiarism etc.

The purpose of this document, presented at the JCDL'11 doctoral consortium, is to summarize the author's idea on "Citation-based Plagiarism Detection". The document is based on three publications [8, 10, 9], which are co-authored with Norman Meuschke and Jöran Beel. The author's second doctoral research project, coined "Citation Proximity Analysis", is not covered in this document.

In contrast with the previously used detection approaches, this approach additionally considers citation information in identifying plagiarism. Performing citation analysis in order to identify plagiarism might sound like an oxymoron. However, experiments have shown that citation similarities often remain and offer clues of misuse.

We demonstrated this by analyzing citation patterns contained in the well-studied doctoral thesis of the former German defense Minister Karl-Theodor zu Guttenberg. The results showed that the detection rate for strongly disguised plagiarism was considerably higher (80%) than for the text-based methods (<5%). It also demonstrated, due to the lack of citation information, that the citation based approach should not be considered as a replacement, since it is unsuitable for detecting short fragments of plagiarism. In chapter 3, the detection algorithms developed for this purpose are presented and evaluated.

2. RELATED WORK & PROPOSED RESEARCH

Hundreds of papers have been published covering sophisticated approaches to detect plagiarism, and dozens of applications were developed. All of them use more or less sophisticated approaches to analyze the text, but ignore the used citations [18, 35]. These approaches deliver excellent results in detecting copied text passages, but fail if text has been paraphrased or translated - for example, from German to English. Instead of analyzing the words of a document, this paper suggests analyzing the used citations.

To our knowledge, applying citation analysis approaches to detect plagiarism has not yet been attempted. Several citation analysis approaches, however, have been developed as a measure of subject relatedness. In 1963, Kessler introduced [17] the concept of bibliographic coupling. Document A and Document B are bibliographically coupled if they cite one or more documents in common. Figure 1 illustrates this approach: Documents A and B are related because they both cite Documents 1, 2 and 3.

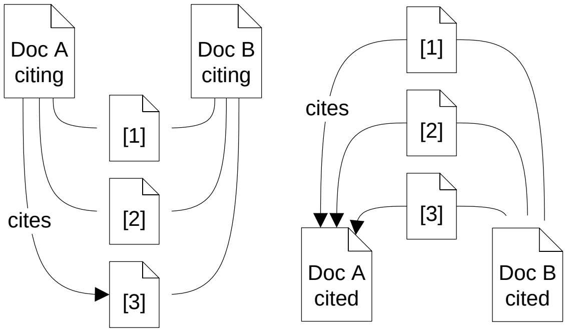


Figure 1: Bibliographic coupling (left) and co-citation (right)

A variation of this, called co-citation, was proposed by Marshakova [20] and Small [30]. Two documents are "co-cited" when at least one document cites both. This approach is illustrated on the right in Figure 1: Documents A and B are related because both are cited by Documents 1, 2 and 3. The more co-citations two documents receive, the more related they are. A further development of this approach is Citation Proximity Analysis, which identifies related documents by their co-occurrence of citations under consideration of their proximity to each other [2]. All approaches allow the calculation of the coupling strength and are used to identify related articles by academic search engines such as *SciPlore.org* and *CiteSeer*.

2.1 Forms of Plagiarism

Observations of plagiarism behavior in practice reveal a number of commonly found methods for illegitimate text usage, which can briefly be summarized as follows. Copy&Paste (c&p) plagiarism specifies the act of taking over parts or the entirety of a text verbatim from another author. Disguised plagiarism includes practices intended to mask literally copied segments. Undue paraphrasing defines the intentional rewriting of foreign thoughts, in the vocabulary and style of the plagiarist without giving due credit in order to conceal the original source [3]. Translated plagiarism is the manual or automated conversion of content from one language to another intended to cover its origin. Idea plagiarism encompasses the usage of a broader foreign concept without appropriate source acknowledgement. An Example is the appropriation of research approaches, methods, experimental setups, argumentative structures, background sources etc. [5].

2.2 Plagiarism Detection Approaches

Plagiarism Detection (PD) is a hypernym for computer-based procedures supporting the identification of plagiarism incidences. Existing PD systems (PDS) can be categorized into external and intrinsic. External PDS compare a suspicious document to a corpus of genuine works. Intrinsic PDS statistically examine linguistic features of the suspicious text, a process known as stylometry, without performing comparisons to external documents. While external PDS aim to find literally matching text segments, intrinsic PDS try to recognize changes in writing style [33].

Different comparison strategies have been proposed for external PDS. The most common ones are briefly explained. Substring matching procedures aim to identify long pairs of identical strings. Such strings are treated as indicators for potential plagiarism if their share with regard to the entire text exceeds a chosen threshold. Most commonly suffix document models, such as suffix trees or arrays, have been used for that purpose [23].

Fingerprinting methods, being the most widely used PD approach, aim at forming a representative digest of a document by selecting a set of multiple substrings from it. The set represents the fingerprint; its elements are called minutiae. Mathematical, hash-like functions can be applied on minutiae for transforming them into more space efficient byte strings [14].

More than 1.000 individual style markers have been proposed for usage in stylometry [28]. They range from lexical features, e.g. average word length, to syntactic features, e.g. part-of-speech frequencies, to structural features, e.g. frequency of punctuation. Intrinsic PD systems mostly comprise an individual combination of multiple linguistic features [34].

Citation-based Plagiarism Detection (CbPD) is a fundamentally different approach compared to text-based similarity evaluations. It is especially suitable for scientific publications, since it requires references. In a previous paper [8] we initially proposed employing citation analysis for PD and evaluated its performance using an artificially created dataset.

2.3 Strength and Weaknesses of PDS

Objective, comparative assessments of the detection performance of PD systems are difficult, since the used collections and evaluation methods differ widely. Two projects address this lack of comparability. Both attempt to benchmark PDS using standardized collections and controlled evaluation environments. The annual PAN International Competition on Plagiarism Detection (PAN-PC) was initiated in 2009, in which competitors present primarily research prototypes [25]. A periodic comparison of productive PDS is performed by a research group at the University of Applied Science Berlin (HTW) since 2004 [15].

The PAN-PC evaluation corpus mainly contains artificially plagiarized sections that were created and partially obfuscated through automated methods such as translation, random shuffles, or semantic substitutions of terms. In addition, 4000 text segments that were manually obfuscated by humans instructed to simulate a plagiarist's behavior are included [26]. In the HTW evaluations a corpus of 42 documents being manually plagiarized or original essays of approx. 1 to 1.5 pages of length is used. The original sources are known and mostly available on the internet [15, 36].

Some results of the two competitions are presented to outline the characteristic strengths and weaknesses of existing PDS. Figure 2 displays the plagiarism detection (plagdet) scores for the top 5 performing external PDS and the 2 intrinsic PDS of Muhr et al. and Suárez et al. participating in PAN-PC'10 (see [25] for further results and references regarding individual systems). The plagdet score was developed to evaluate systems participating in the PAN-PC (see [26] for details). The scores are plotted according to the obfuscation techniques applied to plagiarized text segments. The overall plagdet score for all categories is stated in brackets within each legend entry. In the legend to the figure "-I" is attached to distinguish the system of Muhr et al. participating in the intrinsic from the one in the external task.

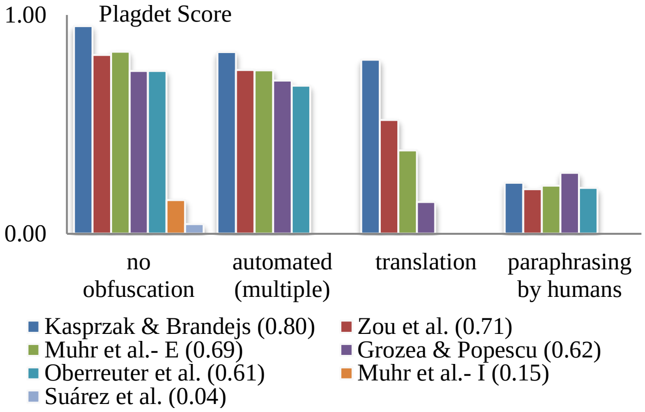


Figure 2: Results of top 5 performing PDS in PAN-PC'10 [25]

The results indicate that c&p plagiarism can be detected with high accuracy by state-of-the-art PDS. However, detection rates for disguised plagiarized segments, especially those obfuscated by humans, are substantially lower for all systems. The organizers of the competition judged the results achieved in detecting cross-lingual plagiarism to be misleading. The well-performing systems used automated services for translating foreign-language documents in the reference corpus. Those services were similar or identical to those used for constructing the plagiarized sections. It is hypothesized that the human-made translations obfuscating real-world plagiarism are much more complex and versatile, and hence less detectable by the tested PDS [25].

The findings of the HTW comparisons are in line with those of the PAN-PC. Notably, none of the tested systems was able to identify cases of translated plagiarism [15]. That supports the assumption of unrealistic detection rates for translated segments in the PAN-PC due to the laboratory-like setup of the competition.

Furthermore, it is noteworthy that the performance of any external PDS depends heavily on the reference corpus available to the individual system. Therefore, it is not surprising that tools, which use the extensive indexes of internet search providers, often achieve the best detection results [21]. The same is true for manually performed queries of suspicious keywords and fragments.

3. CITATION-BASED PD

In the academic environment, citations and references of scholarly publications have long been recognized for containing valuable information about the content of a document and its relation to other works [7]. A large volume of semantic information is contained in citation patterns because complete scientific concepts and argumentative structures are compressed into sequences of small text strings. To our knowledge the identification of plagiarism by analyzing the citations[1] and references[2] of documents has been first described and successfully applied to PD in [8, 10]. In this context, we proposed this definition:

Citation-based Plagiarism Detection (CbPD) subsumes methods that use citations and references for determining similarities between documents in order to identify plagiarism.

Citations and citation patterns offer unique features that facilitate a PD analysis. They are a comparatively easy to acquire, language independent marker, since more or less well-defined standards for using them are established in the international scientific community. This information can be exploited to detect forms of plagiarism that cannot be detected with text-based approaches.

3.1 Factors for Citation-based Text Similarity

In the following section, factors that influence a similarity assessment for documents based on citations and references are outlined for deriving a suitable document model for CbPD.

3.1.1 *Shared References*

Absolute Number

Having references in common is a well-known similarity criterion for scientific texts called bibliographic coupling [16]. The absolute number of shared references represents the coupling strength, which is used to measure the degree of relatedness.

Relative Number

The fraction that shared references represent with regard to the total number of individual references is another similarity indicator. Two texts, A and B, are more likely to be similar if they share a larger percentage of their references. This assumption is supported by results of text-based PD studies [6].

Both the amount and fraction of shared references depend on a number of factors, most importantly document length and specific document parts to be analyzed. Comprehensive documents contain on average more references than short documents, or certain document parts, e.g. related work sections in academic texts contain more citations per page than e.g. summary parts.

Considering the above factors when using reference counts for CbPD might improve their predictive value.

3.1.2 *Probability of Shared References*

The likelihood that two texts have references in common is not statistically independent. Reference co-occurrences that have a lower probability are more predictive for document similarity. The importance of shared references with regard to document similarity is dependent on a number of factors explained below.

Existing citation counts have been shown to influence future citation counts significantly. If a document is highly cited already, its likelihood of gathering additional citations from other documents increases. The phenomenon has been termed the Matthew effect in science[3]. Imagine a document C that has been widely referenced, e.g. by 500 other documents. Another

[1] Citations are short alphanumeric strings in the body of scientific texts representing sources contained in the bibliography

[2] References denote entries in the bibliography

[3] The term refers to a line in the Gospel of Matthew

document D, on the other hand, has been referenced much less frequently, e.g. by 5 other documents. In turn, document D has a smaller probability of being a shared reference of two texts A and B, which are to be analyzed. However, if document D represents a reference shared by A and B this fact is a comparably stronger indicator for similarity than in the case in which C represents a shared reference of A and B.

Time influences the likelihood of references. As citation counts tend to increase with time [24, 29], so does the probability of a document becoming a shared reference. If texts A and B have been published at different points in time, this fact should be compensated, e.g. by comparing expected citations per unit of time.

The topic of research that two documents A and B deal with also influences the likelihood that A and B share common references. They are more likely to do so if the documents address the same or very similar topics. This assumption can be derived from empirical evaluations using Co-Citation analysis to identify clusters in academic domains [11, 31]. If strong Co-Citation relations exist within a certain academic field, as has been shown, this in turn implies that a higher number of documents share common references within this domain. This is intuitive, since references are often used to illustrate prior research or origins of the ideas presented.

Proximity of authors within a social network increases the probability of respective papers to be referenced. Research showed that a text A is more likely to be referenced by a text B if the author(s) of B is/are personally more closely connected to the author(s) of A. For example, documents are referenced more frequently within texts written by former co-authors or researchers that know each other in person. This effect is sometimes referred to as cronyism [22]. The analysis of co-authorship networks might therefore increase the predictive value of reference co-occurrence assessments.

3.1.3 *Citation Pattern Similarity*

Finding similar patterns in the citations used within two scientific texts is a strong indicator for semantic text similarity and the core idea of CbPD. Patterns are subsequences in the citation tuples C_A and C_B of two texts A and B that (partially) consist of shared references and are therefore similar to each other.

The degree of similarity between patterns depends on the number of citations included in the pattern, and the extent to which their order and/or the range they cover is alike. Thus, literally matching subsequences of citations in two documents are a strong indicator for semantic similarity.

The same is true for texts containing patterns that span over similar ranges, even if the order of citations in the pattern does not necessarily correspond towards each other. The width of the covered range can be expressed with regard to sequential positions of citations in the pattern, textual ranges or combinations of both. Measuring range width in units reflecting some semantics, e.g. paragraphs or sentences, is assumed beneficial compared to considering purely syntactic character or citation counts. For example, documents containing several matching citations, one of them within a single section, the other distributed over several chapters are less likely to be similar. However, if both share identical citations e.g. within a paragraph, then their potential similarity is respectively higher. Alternatively, e.g. the document tree may be used to identify semantic clusters in the form of chapters etc.

A CbPD similarity assessment consists of two subtasks. The first is to identify matching citations and citation patterns. The second is to rate patterns with regard to their likelihood of having resulted from undue practices.

The scope of this paper is limited on presenting algorithms that tackle the first subtask of detecting citation patterns. Results of experiments with regard to the second subtask of ranking identified patterns will be presented in an upcoming paper.

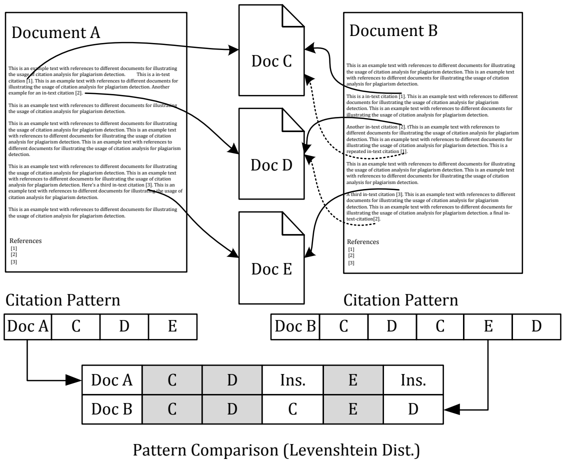


Figure 3: Identifying citation patterns for CbPD

3.1.4 *Challenge of Identifying Citation Patterns*

Detecting citation patterns is a non-trivial task due to the diverse forms of plagiarism. Copy&paste plagiarism results in different citation patterns than e.g. shake&paste plagiarism. Therefore, different algorithms are required to address the specific forms. The following challenges need to be considered.

Unknown pattern constituents – Unlike e.g. in string pattern matching the subsequences of citations to be extracted from a suspicious text and searched for within an original are initially unknown. Citations that are shared by the two documents are easily identified. However, it is unlikely that all of those shared citations represent plagiarized text passages. For instance, two documents might share 8 citations, of which 3 are contained within a plagiarized text section and 4 are distributed over the length of the text and used along with other non-shared citations without representing any form of plagiarism. The citation sequences of the two documents might therefore look like the following:

Original: `1 2 3 x x x 4 x x 5 x 6 x 7 8`
Plagiarism: `x x 5 x x x 4 x 3 x 1 x 2 x x 7 x 8`

Numbers 1-8 represent shared citations, the letter x non-shared citations. The shared citations 1-3 are supposed to represent a plagiarized passage.

Transpositions - the order of citations within text segments might be transposed compared to the corresponding original section. Possible causes can be different citation styles or sort orders of the reference list, e.g. alphabetically opposed to sorting it by publication date. Assume an original text segment contains a sentence in the form:

Studies show that <finding1>, <finding2> **[3,1,2]**.

The semantically identical content might be expressed in the form:

Studies show that <finding1>, <finding2> **[1-3]**.

Scaling - occurrences of shared citations can be used more than once, which is referred to as being scaled. Assume an original text segment in the form:

Study X showed <finding1>, <finding2> and <finding3> **[1]**. Study Y objected <finding1> **[2]**. Assessment Z proofed <finding3> **[3]**.

This segment might be plagiarized as following:

Study X showed <finding1> **[1]**, which was objected by study Y **[2]**. Study X also found <finding2> **[1]**. Assessment Z was able to proof <finding3> **[3]**, which had already been indicated by study X **[1]**.

Missing alignment - potentially plagiarized sections and their corresponding originals need not to be aligned, but can reside in very different parts of the text. For instance, the first paragraph in the first section of an original document A might be plagiarized in document B, however it may become the fifth paragraph in the third section of B. The division of corresponding text segments into paragraphs, sections or chapters might also differ significantly. For instance, a plagiarized text segment might be artificially expanded or reduced to result in a different paragraph split-up in order to conceal the plagiarism.

3.2 Citation-based Similarity Functions

Given the limited empirical knowledge base that exists for CbPD, it is intended to evaluate a balanced mixture of possible similarity functions. The goal is to include global and local similarity assessments as well as functions that focus on the order of citations opposed to functions that ignore citation order, but can handle transpositions and scaling. Besides designing new similarity functions based on the factors outlined above, testing well-proven similarity measures for their applicability to CbPD is a further objective.

The fact that citation sequences of documents can be characterized as strings has been taken as a starting point for identifying existing similarity functions. In this context, string refers to any collection of uniquely identifiable elements that are linked in a way such that each, except for exactly one leftmost and exactly one rightmost element, has one unique predecessor and one unique successor [32]. This definition is broader than the most prominent connotation of the term referring to literal character sequences in the domain of computer science. String processing is a classical and comprehensively researched domain. Thus, multiplicities of possible similarity assessments can be derived from this area see e.g. [12].

	Global Similarity Assessment	**Local Similarity Assessment**
Order preserving	Longest Common Citation Sequence	Greedy Citation Tiling
Order neglecting	Bibliographic Coupling	Citation Chunking

Figure 4: Categorization of evaluated similarity assessments

According to the objectives outlined above, similarity approaches for each category distinguishable in regard to the scope of the assessment (global vs. local) and consideration of citation order have been defined. In Figure 4, the chosen similarity assessments are outlined.

3.2.1 *Bibliographic Coupling Strength*

Bibliographic coupling is one of the first and best-known citation-based similarity assessments for academic texts. Similarity is quantified in terms of the absolute number of shared references. Order or positions of citations within the text are ignored. It can be interpreted as a raw measure of global document similarity. Solely considering bibliographic coupling strength is not a sufficient indicator for potential plagiarism and does not allow pinpointing potentially plagiarized text segments.

3.2.2 *Longest Common Citation Sequence*

The Longest Common Subsequence (LCS) of elements in a string is a traditional similarity measure. The LCS approach has been adapted to citations and comprises the maximal number of citations that can be taken from a citation sequence without changing their order, but allowing for skips over non-matching citations. For instance the sequence (3,4,5) is a subsequence of (2,3,1,4,6,8,5,9) [5], p. 4].

Intuitively, considering the LCS of two citation sequences yields high similarity scores if longer parts of the corresponding text have been adopted without altering the contained citations. Examining the Longest Common Citation Sequence has been chosen because the measure features a clear focus on order relation, opposed to bibliographic coupling. At the same time it offers flexibility for coping with slight transpositions or arbitrary sized gaps of non-matching citations.

It is capable of indicating potential cases of plagiarism in which parts of the text have been copied with no changes, or only slight alterations in the order of citations. This can be the case for copy&paste plagiarism that might have been concealed by basic rewording e.g. through synonym replacements. If significant reordering within plagiarized text segments has taken place (shake&paste plagiarism) or a different citation style has been applied that permutes the sequence of citations, the LCS approach is bound to fail.

3.2.3 *Greedy Citation Tiling*

Greedy Citation Tiling (GCT) is an adaption of a text string similarity function proposed by Wise [37]. The original procedure called Greedy String Tiling (GST) has explicitly been designed for usage in PD. It has been widely adopted and successfully applied, foremost in PDS for software source code [1, 27].

Greedy String/Citation Tiling aims to identify all matching substrings with individually longest possible size in two sequences. Individual longest matches refer to substrings that are shared by both sequences and cannot be extended to the left or right without encountering an element that is not shared by the two sequences. Corresponding individually longest matches in both sequences are permanently linked with each other and stored as a so called tile.

A tile represents a tuple $t = (s_1, s_2,$l) consisting of the starting position of a longest match in the first sequence (s_1), the starting position in the second sequence (s_2) and the length of the match (l). The tiling approach is illustrated in Figure 5. Arabic numbers represent equal elements in the sequences to be analyzed, letter x extraneous elements. Individually longest matches are indicated by boxes around elements. Roman numbers above and below the boxes identify the tiles to which the matches have been assigned. As shown in the figure, the tiling approach is able to cope with arbitrary transpositions in the order of individual substrings. A minimum size of matching substrings can be freely chosen.

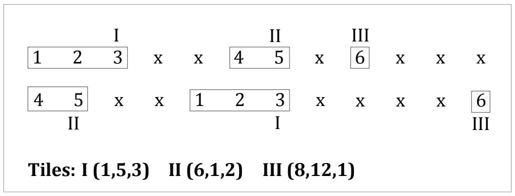


Figure 5: Citation Tiles

The principle of the tiling algorithm is illustrated in Figure 6 assuming a minimum match length of 2. The procedure strictly identifies longer tiles before shorter ones. Auxiliary arrays are used for keeping track of longest tiles and prevent elements from becoming part of multiple tiles. Elements are inserted into the auxiliary arrays at the moment they are assigned to a tile, thus they are “marked” as no longer available for matching and are ignored in future iterations.

The algorithm performs full iterations of both sequences, meaning that sequence 2 is iterated for every element of sequence 1, as long as matches longer than or equal to the specified global minimum length are found in the respective iteration. This indicates that the worst case complexity of the algorithm is $O(n^3)$.

In each iteration only maximal matches are considered for being transformed into tiles. All individual longest matches identified during the same iteration need to be equal to or longer than the maximal match found in the same iteration. If sequence 2 has been traversed for one element of sequence 1, all identified maximal matches are marked in the auxiliary arrays.

For the next iteration the current maximal match length is again set to equal the global minimum match length. This way, the “next-shorter” matches to those marked during the prior iteration are identified. One can see that this repetition continues until no more matches longer than the global minimum match length can be found, which results in the termination of the algorithm. If the minimum match length is set to 1 the GST algorithm is proven to produce the optimal coverage of matching elements with tiles [37].

The GST algorithm has been primarily designed for identifying shake&paste plagiarism. It is able to identify individually longest substrings despite potential rearrangements. Greedy Citation Tiling might serve the same purpose, but opposed to the text-based approach also identifies paraphrased shake&paste plagiarism.

The GCT approach focuses on exact equality with regard to citation order. Finding such patterns provides a strong indication for text similarity. GCT is able to deal with transpositions in the citation sequence that result from rearranging text segments, which is typical for shake&paste plagiarism. However, the approach is not capable of detecting citation scaling or transpositions resulting e.g. from the usage of different citation styles. For covering such cases, another class of detection algorithms has been designed, which is explained in the following section.

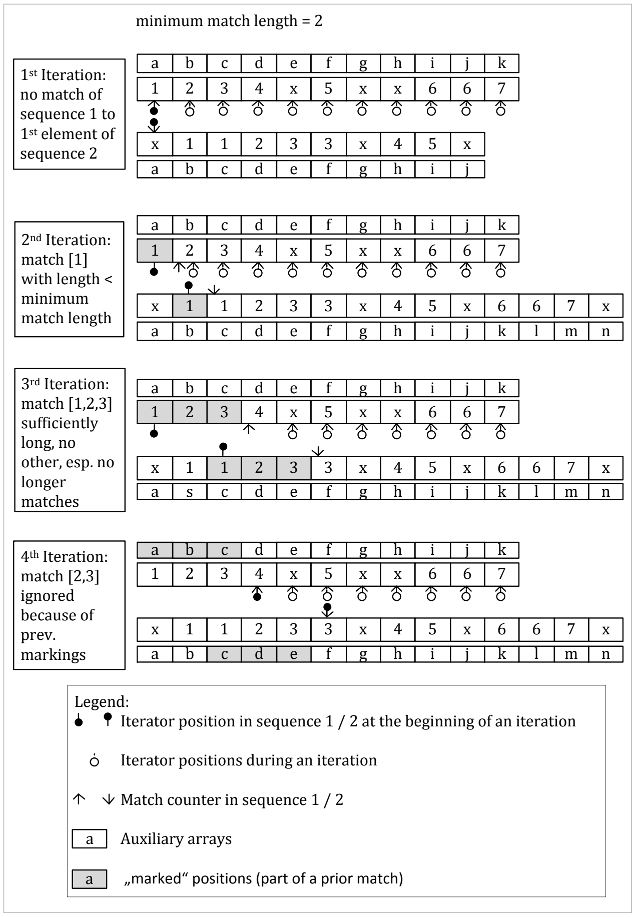

minimum match length = 2
1st Iteration: no match of sequence 1 to 1st element of sequence 2
a b c d e f g h i j k
1 2 3 4 x 5 x x 6 6 7
x 1 1 2 3 3 x 4 5 x
a b c d e f g h i j
2nd Iteration: match [1] with length < minimum match length
a b c d e f g h i j k
1 2 3 4 x 5 x x 6 6 7
x 1 1 2 3 3 x 4 5 x 6 6 7 x
a b c d e f g h i j k l m n
3rd Iteration: match [1,2,3] sufficiently long, no other, esp. no longer matches
a b c d e f g h i j k
1 2 3 4 x 5 x x 6 6 7
x 1 1 2 3 3 x 4 5 x 6 6 7 x
a s c d e f g h i j k l m n
4th Iteration: match [2,3] ignored because of prev. markings
a b c d e f g h i j k
1 2 3 4 x 5 x x 6 6 7
x 1 1 2 3 3 x 4 5 x 6 6 7 x
a b c d e f g h i j k l m n
Legend:
Iterator position in sequence 1 / 2 at the beginning of an iteration
Iterator positions during an iteration
Match counter in sequence 1 / 2
a Auxiliary arrays
a „marked“ positions (part of a prior match)

Figure 6: Example flow of the Greedy Citation Tiling algorithm.

3.2.4 *Citation Chunking*

A set of heuristic procedures that aim to identify local citation patterns regardless of potential transpositions and/or scaling have been developed for this study. The approach has been termed Citation Chunking because it is inspired by the feature selection strategies of text-based fingerprinting algorithms. A citation chunk is a variably sized substring of a document’s citation sequence.

The main idea of citation chunking is to consider shared citations as textual anchors at which local citation patterns can potentially exist. Starting from an anchor, citation chunks are constructed by dynamically increasing the considered substring of citations based on the characteristics of the chunk under construction as well as the succeeding citations.

Chunking Strategies

Strategies for forming chunks have been derived by imagining potential behaviors of a plagiarist and modeling the resulting citation patterns.

Determining the starting and ending point for a chunk is not a trivial task. There probably does not exist a best solution that fits all plagiarism scenarios. Larger chunks are believed to be better suitable for detecting overall similarities and compensate for transpositions and scaling. Smaller chunks, on the other hand, are more suitable for pinpointing specific areas of highest similarity.

In order to experiment with both tendencies, the following procedures for constructing citation chunks have been defined.

1. Only consecutive shared citations form a chunk:

```
Doc A: x,1,2,3,x,4,5,3,x,x
Doc B: x,x,3,2,1,x,5,3,4,x
```

This is the most restrictive chunking strategy. Its intention is to highlight confined text segments that have a very high citation-based similarity. It is ideal for detecting potential cases in which copy&paste plagiarism might have been concealed by rewording or translation.

2. Chunks are formed dependent on the preceding citation. A citation is included in a chunk if $n \leq 1$ or $1 > n \leq s$ non-shared citations separate it from the last preceding shared citation, with s being the number of citations in the chunk currently under construction:

```
Doc A: x,1,2,3,x,x,4,5,x,x,x,x,x,x,x,6,7
Doc B: 3,2,x,1,x,x,4,x,x,x,x,x,5,6,7,x
```

Chunking strategy 2 aims to uncover potential cases in which text segments or logical structures have been taken over from or influenced by another text. It allows for sporadic non-shared citations that may have been inserted to make the resulting text more “genuine”. It can also detect potential cases of concealed shake&paste plagiarism by allowing an increasing number of non-shared citations within a chunk, given that a certain number of shared citations have already been included. This process aims to reflect the behavior that text segments (including citations) from different sources are interwoven.

3. Citations exhibiting a textual distance below a certain threshold form a chunk.

Chunking strategy 3 aims to define a textual range inside which possible plagiarism is deemed likely. Studies have shown that plagiarism more frequently affects confined text segments, such as one or two paragraphs, rather than extended text passages or the document as a whole. Building upon this knowledge, the respective chunking strategy only considers citations within a specified range for forming chunks.

Since the split up of a plagiarized text into textual units, such as sentences or paragraphs, might be altered artificially, textual proximity might be analyzed in terms of multiple units. One possibility tested in the study has been to count the characters, words, sentences and paragraphs that separate individual citations. The respective counts have been compared to average numbers expected for a certain textual range. For instance, one paragraph might on average comprise 120 words consisting of 720 characters. If one shared citation is separated from another by 2 paragraphs, but less than 120 words, it will be included in a chunk to be formed. In this manner, even artificially created paragraph split-ups can be dealt with.

Finding a suitable maximal distance for proximity of citations in the text is highly dependent on the individual corpus analyzed. If e.g. the average length of documents is rather short, and individual documents contain smaller number of sections and paragraphs, it is believed to be harder for a plagiarist to artificially alter the textual structure. Consequently, a comparably lower maximal distance should be chosen in this scenario. In contrast, it is believed to be easier to change e.g. the paragraph split-up in longer academic texts.

The complete process of forming chunks according to the outlined chunking strategies is graphically summarized as a flow chart in figure 7. In order to experiment with larger chunk sizes, an optional merging step is tested (dashed box in figure 7).

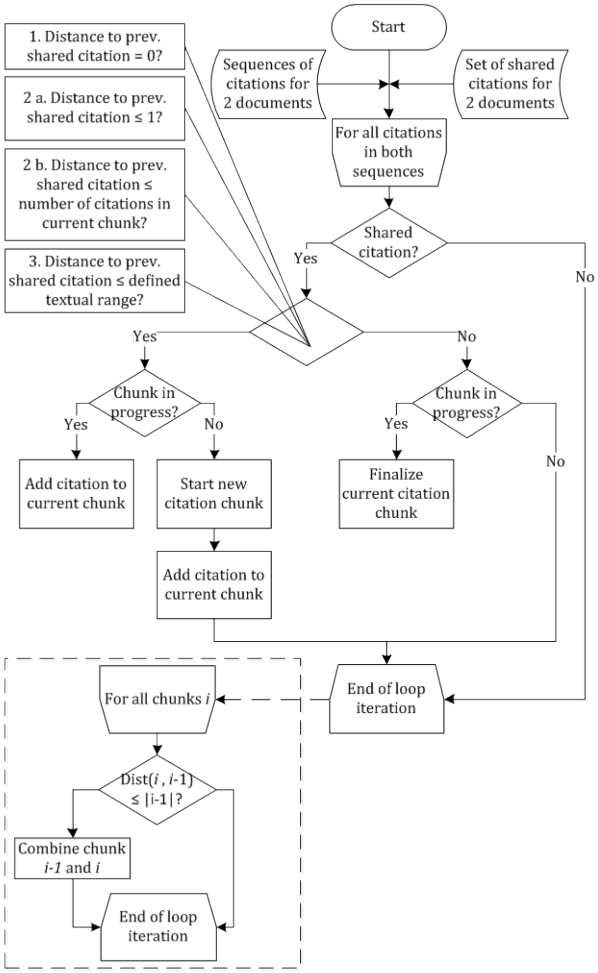

Start
Sequences of citations for 2 documents
Set of shared citations for 2 documents
For all citations in both sequences
Shared citation?
Yes
No
1. Distance to prev. shared citation = 0?
2 a. Distance to prev. shared citation ≤ 1?
2 b. Distance to prev. shared citation ≤ number of citations in current chunk?
3. Distance to prev. shared citation ≤ defined textual range?
Yes
No
Chunk in progress?
Yes
No
Add citation to current chunk
Start new citation chunk
Add citation to current chunk
Chunk in progress?
Yes
No
Finalize current citation chunk
End of loop iteration
For all chunks i
Dist(i , i-1) ≤ |i-1|?
Combine chunk i-1 and i
End of loop iteration

Figure 7: Forming of citation chunks

It is intended to combine supposedly suspicious citation patterns in order to outline longer sections of similar text e.g. as part of an idea plagiarism. Chunks are merged if they are separated by n non-shared citations, $n <= m$ with m being the number of shared citations in the preceding chunk

```
Iteration 1: XXX,x,XX,x,x,XXX,x,x,x,x,x,x,XX
Iteration 2: XXXXX,x,x,XXX,x,x,x,x,x,x,XX
Iteration 3: XXXXXXXX,x,x,x,x,x,x,XX
```

Chunk XX is not merged since its distance to preceding chunks is too large.

3.2.5 *Chunk Comparison*

Once chunks have been formed, they are considered in their entirety for comparison. That is, the order of citations within a chunk is disregarded during comparisons in order to account for potential transpositions and/or scaling. The number of shared citations within the units to be compared represents the measure of similarity.

In the following two main strategies for comparing documents based on citation chunks are described. The first is to form chunks for both documents and compare each chunk of the first document with each chunk of the second. The chunk pairs having the highest citation overlap are permanently related to each other and considered a match. If multiple chunks in the documents have an equal overlap, all combinations with maximal overlap are stored.

In the second scenario, chunks are constructed for one document only. Subsequently, each of the chunks is compared to the unaltered citation sequence of the second document by “moving” it as a sliding window over the sequence and assigning it to the position with the maximal citation overlap.

4. OVERVIEW OF PRELIMINARY RESULTS: GUTTENBERG’S DOCTORAL THESIS

4.1 Test Corpus

Artificially created evaluation corpora, such as the ones of the PAN-PC do not include citation or reference information. Moreover, they lead to unrealistically high detection rates as machines are not as creative in paraphrasing and disguising plagiarism as humans (see 2.3).

Although the scientific value of Mr. Guttenberg’s dissertation [plagiarism] is questionable, we consider it as an ideal case study for our evaluation purposes because it:

- has been thoroughly investigated by hundreds of examiners;
- was created by a human author trying to disguise plagiarism;
- provides realistic citation information.

Due to these unique characteristics, the thesis allows for comparative evaluation of commonly applied plagiarism detection and the Citation-based Plagiarism Detection approach.

Previous evaluations (as presented in 2.3) indicate that translated plagiarism is especially hard to detect by conventional text-based PDS. Therefore, the focus of our investigation has been on whether CbPD is better suitable for detecting this form of plagiarism. At the time of our investigation the GuttenPlag project had identified plagiarized passages that represented appropriations of English sources translated to German on 31 pages within the thesis. Those 31 pages were analyzed for matching citations with their identified genuine sources.

4.2 Test PD Systems

To compare citation-based with traditional text-based detection we used three popular PDS. Ephorus, which usually scores among the top 3 PDS in the HTW comparisons [15], the freely available Ferret system [19], both systems use fingerprinting detection, and WCopyFind [3], a PDS that employs substring matching. Since the two latter mentioned systems depend on local availability of possible source documents, all digitally available sources identified by the GuttenPlag project were collected and used.

For the Citation-based Plagiarism Detection we developed an Open Source software system in Java coined *CitePlag.* These steps are performed in our plagiarism detection system:

1. The document is parsed and a series of heuristics applied to process the citations, including their position within the document[4].

2. Citations are matched with their entries in the bibliography.

3. The citation-based similarity of the documents is calculated.

The developed prototype CbPDS consists of three main components. The first is a Relational Database System (RDBS) termed CbPD database storing data to be acquired from documents as well as detection results. The second is the detection software called CbPD Detector that retrieves data from the CbPD Database, runs the different analysis algorithms to be evaluated and feeds the resulting output back to the CbPD Database. The third component, the CbPD Report Generator, creates summarized reports of detection results for individual document pairs based on adjustable filter criteria. The three-tier-architecture is illustrated in the following figure.

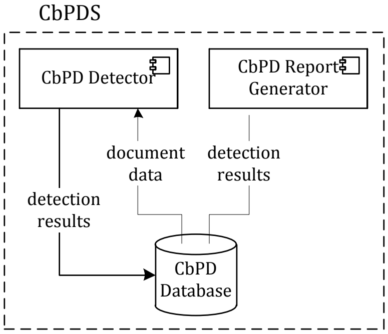


Figure 8: CbPDS system architecture

4.3 Results

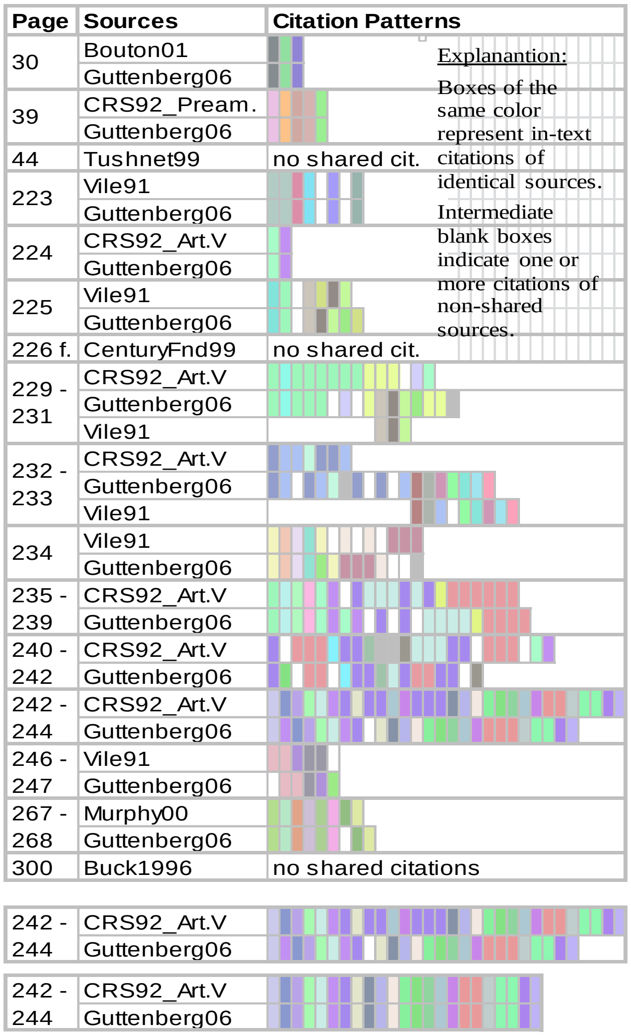
Our results obtained for text-based PD confirm earlier findings of others presented in 2.3. Manually querying search engines, such as Google, yielded high detection rates with regard to copy&paste plagiarism. Depending on the invested time and selection of keywords, even paraphrased and translated plagiarism can be found.

Plagiarism type	Text-based	Citation-based
Copy&paste	~ 70 % Good results even for short fragments	Unsuitable as short fragments cannot be detected
Disguised plagiarism	< 10 %	Depending on the fragments length ~ 30 %
Idea / structure plagiarism	0 %	Some cases could be identified
Translated plagiarism	< 5 %	~ 80 %. 13 out of 16 fragments could be identified.

Table 1: Comparison of detection results

[4] The citations were parsed using a modified version of parsCit (http://wing.comp.nus.edu.sg/parsCit) in combination with the authors' self-developed software, which is available upon request.

The text-based PDS, especially Ferret and WCopyfind, which work with local document comparisons, deliver good results for identifying copy&paste plagiarism given that the sources are available, as in our case. The performance of Ephorus in this case study was a little surprising. Only 2 % of the text in the thesis was found to match the sources of plagiarism. Given the large fraction of (almost) verbatim plagiarism in the thesis, and the fact that 77 sources of plagiarized sections, which were identified by the GuttenPlag project, are available on the internet, opposed to 63 that are not [13], this result is disappointing. Not surprisingly, all systems failed to identify almost all stronger paraphrased sections and could not detect any translated plagiarism (see Table 1 for details). However, these figures should be treated with care. Since a real thesis was used, it is uncertain whether all plagiarized fragments are known. Therefore, the stated detection rates might be too high, especially for the very hard to detect idea plagiarism.



Page	Sources	Citation Patterns
30	Bouton01	
	Guttenberg06	
39	CRS92_Pream.	
	Guttenberg06	
44	Tushnet99	no shared cit.
223	Vile91	
	Guttenberg06	
224	CRS92_Art.V	
	Guttenberg06	
225	Vile91	
	Guttenberg06	
226 f.	CenturyFnd99	no shared cit.
229 - 231	CRS92_Art.V	
	Guttenberg06	
	Vile91	
232 - 233	CRS92_Art.V	
	Guttenberg06	
	Vile91	
234	Vile91	
	Guttenberg06	
235 - 239	CRS92_Art.V	
	Guttenberg06	
240 - 242	CRS92_Art.V	
	Guttenberg06	
242 - 244	CRS92_Art.V	
	Guttenberg06	
246 - 247	Vile91	
	Guttenberg06	
267 - 268	Murphy00	
	Guttenberg06	
300	Buck1996	no shared citations

242 - 244	CRS92_Art.V	
	Guttenberg06	

242 - 244	CRS92_Art.V	
	Guttenberg06	

Explanantion:

Boxes of the same color represent in-text citations of identical sources.

Intermediate blank boxes indicate one or more citations of non-shared sources.

Figure 9: Citation Patterns for translated plagiarism

Figure 9 shows the citation patterns of all translated plagiarism fragments found by the GuttenPlag project.

The figure illustrates that the citation patterns in genuine sources and in Mr. Guttenberg's translated plagiarism are often very similar. With exception of the pages 44, 226 and 300, all other pages share the same references in a similar order in the source document and Guttenberg's translation. This becomes especially obvious after cleaning the citation sequences by removing citations that are not shared by both documents at their corresponding positions. This is exemplified at the bottom of the figure for the pages 242-244.

Whereas the currently used PDS were unable to detect a single translated fragment, the CbPD approach could identify all but three fragments. However, as with every PDS, the findings of CbPD must be carefully verified by humans, especially in cases where only a few citations form the pattern, for example in the fragments on page 30 and 224.

The evaluation indicates to a large extent that the strength of the existing PD systems are the weaknesses of the new citation based PD systems and vice versa. Whereas the strength of existing PDS lies in detecting plagiarism on the sentence level in the form of identifying similar or identical consecutive words, the strength of the citation based approach lies in identifying translation- and idea-plagiarism or disguised paraphrasing. However, since the CbPD relies on citation information, it is unable to identify short paraphrased fragments. By combining the strength of the text- and citation-based approaches the detection rate clearly outperforms currently used techniques.

5. CONCLUSION

This paper describes a new approach towards detecting plagiarism. In contrast to existing approaches, which analyze documents' words but ignore their citations, this approach is based on citation analysis and allows duplicate and plagiarism detection even if a document has been paraphrased or translated, since the relative position of citations often remain similar. Although this approach allows in many cases the detection of plagiarized work that could not be detected automatically with the currently used approaches, it should be considered as an extension rather than a substitute. Whereas the known text analysis methods can detect copied or, to a certain degree, modified passages, the proposed approach requires longer passages with at least two citations in order to create a digital fingerprint.

6. REFERENCES

[1] Ahtiainen, A., Surakka, S., and Rahikainen, M. Plaggie: GNU-licensed source code plagiarism detection engine for Java exercises. In *Proceedings of the 6th Baltic Sea conference on Computing education research: Koli Calling 2006* (New York, NY, USA, 2006), Baltic Sea '06, ACM, pp. 141–142.

[2] Bela Gipp, and Joeran Beel. Citation Proximity Analysis (CPA) - A new approach for identifying related work based on Co-Citation Analysis. In *Proceedings of the 12th International Conference on Scientometrics and Informetrics (ISSI'09)* (Rio de Janeiro (Brazil), July 2009), B. Larsen and J. Leta, Eds., vol. 2, International Society for Scientometrics and Informetrics, pp. 571–575. ISSN 2175-1935. Available at http://sciplore.org.

[3] Bloomfield, L. A. Software to detect plagiarism: WCopyfind. Online Resource, Jan. 2009. Retrieved Oct. 01, 2010 from: http://plagiarism.phys.virginia.edu/Wsoftware.html.

[4] COCEL. *Concise Oxford Companion to the English Language [electronic resource]*. Oxford Reference Online. Oxford University Press, 1998.

[5] Crochemore, M., and Rytter, W. *Jewels of Stringology*. World Scientific Publishing, 2002.

[6] Errami, M., Hicks, J. M., Fisher, W., Trusty, D., Wren, J. D., Long, T. C., and Garner, H. R. Déjà vu—A study of duplicate citations in Medline. *Bioinformatics 24*, 2 (2008), 243–249.

[7] Garfield, E. Citation Indexes for Science: A New Dimension in Documentation through Association of Ideas. *Science 122*, 3159 (July 1955), 108–111.

[8] Gipp, B., and Beel, J. Citation Based Plagiarism Detection - A New Approach to Identify Plagiarized Work Language Independently. In *Proceedings of the 21st ACM Conference on Hyptertext and Hypermedia (HT'10)* (New York, NY, USA, June 2010), ACM, pp. 273–274.

[9] Gipp, B., and Meuschke, N. 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 (DocEng2011)* (2011).

[10] Gipp, B., Meuschke, N., and Beel, J. 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)* (Ottawa, Canada, June 2011).

[11] Griffith, B. C., Small, H. G., Stonehill, J. A., and Dey, S. The Structure of Scientific Literatures II: Toward a Macro- and Microstructure for Science. *Science Studies 4*, 4 (1974), pp. 339–365.

[12] Gusfield, D. *Algorithms on Strings, Trees, and Sequences: Computer Science and Computational Biology*. Cambridge University Press, New York, NY, USA, 1997.

[13] GuttenPlag Wiki. Eine kritische Auseinandersetzung mit der Dissertation von Karl-Theodor Freiherr zu Guttenberg: Verfassung und Verfassungsvertrag. Konstitutionelle Entwicklungsstufen in den USA und der EU. Online Source, May 2011. Retrieved June 04, 2011 from: http://-de.guttenplag.wikia.com/wiki/GuttenPlag_Wiki.

[14] Hoad, T. C., and Zobel, J. Methods for Identifying Versioned and Plagiarised Documents. *Journal of the American Society for Information Science and Technology 54*, 3 (2003), 203–215.

[15] Hochschule für Technik und Wirtschaft Berlin. Portal Plagiat - Softwaretest 2010. Online Source, 2010. Retrieved Apr. 2, 2011 from: http://plagiat.htw-berlin.de/software/2010-2/.

[16] Kessler, M. M. Concerning some problems of intrascience communication. Lincoln laboratory group report, Massachusetts Institute of Technology. Lincoln Laboratory, 1958. Cited according to: B.H. Weinberg. BIBLIOGRAPHIC COUPLING: A REVIEW. Information Storage Retrieval, 10: 189-196, 1974.

[17] Kessler, M. M. Bibliographic coupling between scientific papers. *American Documentation 14* (1963), 10–25.

[18] Lukashenko, R., Graudina, V., and Grundspenkis, J. Computer-based plagiarism detection methods and tools: An overview. In *Proceedings of the 2007 international conference on Computer systems and technologies* (2007), ACM, p. 40.

[19] Lyon, C., Malcolm, J., and Dickerson, B. Detecting Short Passages of Similar Text in Large Document Collections. In *Proceedings of the 2001 Conference on Empirical Methods in Natural Language Processing* (2001), L. Lee and D. Harman, Eds., pp. 118–125.

[20] Marshakova Shaikevich, I. System of Document Connections Based on References. *Scientific and Technical Information Serial of VINITI 6*, 2 (1973), 3–8.

[21] Maurer, H., Kappe, F., and Zaka, B. Plagiarism - A Survey. *Journal of Universal Computer Science 12*, 8 (Aug. 2006), 1050–1084.

[22] Meho, L., and Yang, K. Impact of data sources on citation counts and rankings of LIS faculty: Web of Science vs. Scopus and Google Scholar. *Journal of the American Society for Information Science and Technology 58*, 13 (2007), 2105–25.

[23] Monostori, K., Zaslavsky, A., and Schmidt, H. Document Overlap Detection System for Distributed Digital Libraries. In *DL '00: Proceedings of the fifth ACM conference on Digital libraries* (New York, NY, USA, 2000), ACM, pp. 226–227.

[24] Phelan, T. A compendium of issues for citation analysis. *Scientometrics 45* (1999), 117–136. 10.1007/BF02458472.

[25] Potthast, M., Barrón Cedeño, A., Eiselt, A., Stein, B., and Rosso, P. Overview of the 2nd International Competition on Plagiarism Detection. In *Notebook Papers of CLEF 2010 LABs and Workshops, 22-23 September, Padua, Italy* (Sept. 2010), M. Braschler, D. Harman, and E. Pianta, Eds.

[26] Potthast, M., Stein, B., Barrón Cedeño, A., and Rosso, P. An Evaluation Framework for Plagiarism Detection. In *Proceedings of the 23rd International Conference on Computational Linguistics (COLING 2010)* (Block A, Xue Yan Building, Tsinghua University, Beijing 100084, China, Aug. 2010), C.-R. Huang and D. Jurafsky, Eds., Tsinghua University Press.

[27] Prechelt, L., Philippsen, M., and Malpohl, G. JPlag: Finding plagiarisms among a set of programs. Technical Report 2000-1, Universität Karlsruhe, Fakultät für Informatik, Germany, Jan. 2000.

[28] Rudman, J. The State of Authorship Attribution Studies: Some Problems and Solutions. *Computers and the Humanities 31* (1997), 351–365.

[29] Seglen, P. O. Why the impact factor of journals should not be used for evaluating research. *BMJ 314*, 7079 (1997), 497.

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

[31] Small, H., and Griffith, B. C. The Structure of Scientific Literatures I: Identifying and Graphing Specialties. *Science Studies 4*, 1 (1974), pp. 17–40.

[32] Smyth, B. *Computing Patterns in Strings*. Pearson Addison-Wesley, Harlow, England; New York, 2003.

[33] Stein, B., Koppel, M., and Stamatatos, E., Eds. *Proceedings of the SIGIR 2007 International Workshop on Plagiarism Analysis, Authorship Identification, and Near Duplicate Detection, PAN 2007, Amsterdam, Netherlands, July 27, 2007* (2007), vol. 276 of *CEUR Workshop Proceedings*, CEUR-WS.org.

[34] Stein, B., Lipka, N., and Prettenhofer, P. Intrinsic Plagiarism Analysis. *Language Resources and Evaluation [Online Resource]* (2010), 1–20.

[35] Stein, B., Rosso, P., Stamatatos, E., Koppel, M., and Eneko, A., Eds. *Proceedings of the 3rd PAN Workshop. Uncovering Plagiarism, Authorship and Social Software Misuse* (2009).

[36] Weber Wulff, D. Test cases for plagiarism detection software. In *Proceedings of the 4th International Plagiarism Conference* (Newcastle Upon Tyne, 2010).

[37] Wise, M. J. String Similarity via Greedy String Tiling and Running Karp-Rabin Matching. Online Preprint, Dec. 1993. Retrieved Oct. 13, 2010 from: http://vernix.org/marcel/share/-RKR_GST.ps.

[plagiarism] Guttenberg, K.-T. F. *Verfassung und Verfassungsvertrag : Konstitutionelle Entwicklungsstufen in den USA und der EU*. Dissertation (**Retracted as plagiarism**), Universität Bayreuth, Berlin, 2009.