@INPROCEEDINGS{Gipp11c,
 author = {{G}ipp, {B}ela and {M}euschke, {N}orman},
 title = {Citation Pattern Matching Algorithms for Citation-based Plagiarism
Detection: Greedy Citation Tiling, Citation Chunking and Longest
Common Citation Sequence},
 booktitle = {{P}roceedings of the 11th {ACM} {S}ymposium on {D}ocument {E}ngineering
({D}oc{E}ng2011)},
 year = {2011},
}

Citation Pattern Matching Algorithms for Citation-based Plagiarism Detection: Greedy Citation Tiling, Citation Chunking and Longest Common Citation Sequence

Bela Gipp
OvGU, Germany & UC Berkeley, USA
gipp@berkeley.edu

Norman Meuschke
OvGU, Germany & UC Berkeley, USA
meuschke@berkeley.edu

ABSTRACT
Plagiarism Detection Systems have been developed to locate instances of plagiarism e.g. within scientific papers. Studies have shown that the existing approaches deliver reasonable results in identifying copy&paste plagiarism, but fail to detect more sophisticated forms such as paraphrased, translated or idea plagiarism. The authors of this paper demonstrated in recent studies [4, 15] that the detection rate can be significantly improved by not only relying on text analysis, but by additionally analyzing the citations of a document. Citations are valuable language independent markers that are similar to a fingerprint. In fact, our examinations of real world cases have shown that the order of citations in a document often remains similar even if the text has been strongly paraphrased or translated in order to disguise plagiarism.

This paper introduces three algorithms and discusses their suitability for the purpose of Citation-based Plagiarism Detection. Due to the numerous ways in which plagiarism can occur, these algorithms need to be versatile. They must be capable of detecting transpositions, scaling and combinations in a local and global form. The algorithms are coined Greedy Citation Tiling, Citation Chunking and Longest Common Citation Sequence. The evaluation showed that common forms of plagiarism can be detected reliably if these algorithms are combined.

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 Systems, Citation-based, Citation Order Analysis, Citation Pattern Analysis

1. INTRODUCTION
Plagiarism describes the appropriation of other persons' ideas, intellectual or creative work and passing them of as one's own [7]. For including the act of self-plagiarism (see 2.1) we broaden the scope of the term and define academic plagiarism as *using words and/or ideas from other sources without due acknowledgement imposed by academic principles.*

DocEng'11, September 19–22, 2011, Mountain View, CA, USA.

It is a particularly common problem among college students worldwide, but also notably present among established researchers. In a self-report study among ~82,000 students about 40% of undergraduates and ~25% of graduates engaged in plagiarism within 12 months prior to the study [29]. Results of other studies range as high as ~90% of the subjects self-reporting acts of plagiarism [27].

In academia numerous cases of plagiarism have become publicly known. An automated plagiarism check of ~285,000 scientific texts of arXiv.org yielded more than 500 documents very likely to have been plagiarized. In addition, 30.000 documents (~20% of the collection) were found to be very likely duplicates or containing: "[…] excessive self-plagiarism […]" [43, p. 12].

The existing approaches for plagiarism detection have their weaknesses. Using the words of Weber-Wulff, the organizer of regular comparisons for productive Plagiarism Detection Systems (PDS), the current state of available systems can be summarized as follows: "[…] PDS find copies, not plagiarism." [50, p. 6].

The paper is structured as follows. After giving an overview of different forms of plagiarism, the detection approaches currently used and a discussion of their strength and weaknesses, the Citation-based Plagiarism Detection approach is briefly presented. Subsequently, the newly developed algorithms for Citation-based Plagiarism Detection are introduced, evaluated and their suitability for detecting different forms of plagiarism is discussed. Finally, the suitability of the presented approaches is demonstrated using real cases of plagiarism.

2. RELATED WORK

2.1 Forms of Plagiarism
Observations of plagiarism behavior in practice reveal a number of commonly found methods for illegitimate text usage, which are characterized below.

Copy&Paste (c&p) plagiarism specifies the act of taking over text verbatim from another author [49].

Disguised plagiarism subsumes practices intended to mask copied segments [26]. Four different masking techniques have been identified. These are:

- *Shake&Paste (s&p)* plagiarism is characterized by copying and merging sentences or paragraphs from different sources with slight adjustments necessary for forming a coherent text [49];
- *Expansive plagiarism* refers to the insertion of additional text into or in addition to copied segments [26];

- *Contractive plagiarism* describes the summary or trimming of copied material [26];
- *Mosaic plagiarism* encompasses the merge of text segments from different sources and obfuscating the plagiarism by changing word order, substituting words with synonyms or entering/deleting filling words [26, 49];

Technical disguise summarizes techniques for hiding plagiarized content from being automatically detected by exploiting weaknesses off current text-based analysis methods e.g. by substituting characters with graphically identical symbols from foreign alphabets or inserting letters in white font color [20].

Undue paraphrasing defines the intentional rewriting of foreign thoughts in the vocabulary and style of the plagiarist without giving due credit for concealing the original source [26].

Translated plagiarism is defined as the manual or automated conversion of content from one language to another intended to cover its origin [49].

Idea plagiarism encompasses the usage of a broader foreign concept without due source acknowledgement [28]. Examples are the appropriation of research approaches and methods, experimental setups, argumentative structures, background sources etc. [13].

Self-plagiarism characterizes the partial or complete reuse of one's own previous writings not being justified by scientific goals, e.g. for presenting updates or providing access to a larger community, but primarily serving the author, e.g. for artificially increasing citation counts [5, 11].

2.2 Existing Plagiarism Detection Approaches

Plagiarism Detection (PD) is a hypernym for computer-based procedures supporting the identification of plagiarism incidences. Existing PD methods can be categorized into external and intrinsic approaches [26, 45].

External PD methods compare a suspicious document to a collection of genuine works. Different comparison strategies have been proposed in this context.

String matching procedures [2, 32, 52] aim to identify longest pairs of identical text strings. These strings are treated as indicators for potential plagiarism if the share they represent with regard to the overall text exceeds a chosen threshold. Suffix document models, such as suffix trees or suffix arrays, have mostly been used for that purpose in the context of PD.

The strength of substring matching methods is their perfect detection accuracy with regard to literal text overlaps. Their major drawbacks are the relative difficulty of detecting disguised plagiarisms as well as the required computational effort. The former fact is intuitive when recalling the exact matching approach of the detection procedure. The later barrier results from the use of suffix data structures. The most space-efficient suffix tree [25], suffix array [24] and suffix vector [33] implementations allow searching in linear time and require on average $\sim 8n$ of storage space, with n being the number of symbols in the original document. String B-Trees allow searching in $O(\log n)$, but also require multiple times the storage space of the original document [25]. This renders them impracticable for most large document collections.

Employing *vector space retrieval* based on different term units has been proposed e.g. by [9, 40, 22]. Vector space models (VSM) are a standard, highly performance tuned Information Retrieval (IR) concept that can overcome the effort-related limitations of elaborate string matching. VSM consider a set of terms, which commonly has been extracted from the whole document or larger parts of the text, for similarity computation. Therefore, vector space retrieval methods just like string matching is classified as global similarity assessments [47].

The well-known cosine measure is a widely used similarity function in PD settings as it is for other IR tasks. More complex similarity functions tend to incorporate semantic information e.g. by considering word synonyms [21] or pre-computing semantic relations [48] between terms. The aforementioned papers show that such considerations can increase detection performance, at the cost of significantly increasing the computational effort required. In the experiments reported in [3] considering synonyms improved the F-measure of the respective detection procedures by 2-3 times. However, the runtime required for doing so increased by more than 27 times.

The detection performance of VSM based PDS is dependent on the individual plagiarism incidence to be analyzed and the parameter configuration, e.g. term unit and term selection strategy, of the specific detection method [18, p. 155]. However, the global similarity assessment of VSMs tends to be detrimental to detection accuracy in PD settings. Verbatim plagiarism is more commonly related to smaller, confined segments of a document, which favors local similarity analysis [47].

Fingerprinting methods, being the most widely used PD approach, perform a local similarity assessment. They aim to form 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 [19]. Mathematical, hash-like functions can be applied on minutiae for transforming them into more space efficient byte strings [12].

A suspicious document is checked for plagiarism by computing its fingerprint and querying each minutia with a pre-computed index of fingerprints for all documents of a reference collection. Minutiae found matching with those of other documents indicate shared text segments and suggest potential plagiarism upon exceeding a certain similarity threshold [6].

The inherent challenge of fingerprinting is finding a document representation that reduces computational effort to a suitable dimension, while limiting the information loss incurred to achieve acceptable detection accuracy [31]. A number of parameters, e.g. the chunking strategy, chunk size (granularity of the fingerprint) or number of minutiae (resolution of the fingerprint), reflect that challenge. There is no definite answer to the question of which parameter combination is the best, since this choice is strongly dependent on the nature and size of the collection as well as the amount and form of plagiarism.

Conventional fingerprinting methods implicitly encode the term order of a document in proportion to the length of the chosen text chunk. STEIN proposes an approach, termed fuzzy-fingerprinting, which disregards term order by using a VSM of document terms instead of substrings for minutia computation [44]. Fuzzy-Fingerprints are primarily targeted at reducing computational effort. Compared to fingerprinting using word-3-grams and a MD5 hash function they can be computed >5 times faster, but have been shown to be inferior in detection accuracy [47].

Intrinsic PD methods, opposed to the approaches presented so far, do not depend on the existence of a reference corpus. They

statistically examine linguistic features of the suspicious text, a process known as stylometry, without performing comparisons to external documents. They aim at recognizing changes in writing style to indicate potential plagiarism [31].

The linguistic features to be analyzed form a style model. Approximately more than 1.000 individual style markers [38] have been proposed for stylometry, most can be classified as falling into one of the following categories [46]:

- lexical features on character level, e.g. *n*-gram frequency, or word level, e.g. average word lengths or syllables per word;
- syntactic features, e.g. word or part-of-speech frequencies;
- structural features, e.g. average paragraph length or frequency of punctuation.

Style models of intrinsic PDS are generally comprised of an individual combination of multiple linguistic features [31].

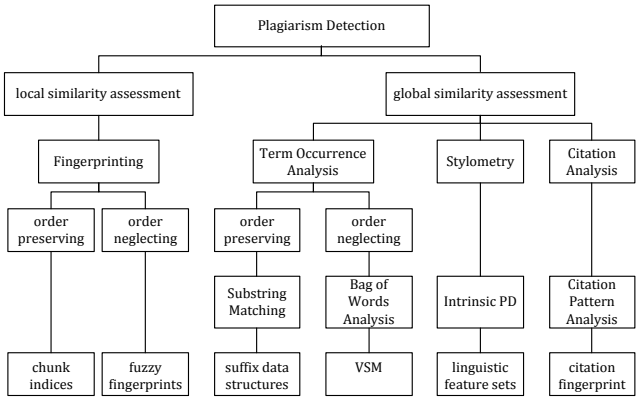

Figure 1: Classification of PD methods (inspired by: [47])

In previous papers [4, 15] we initially proposed employing citation analysis for PD and presented results of initial studies. *Citation-based Plagiarism Detection* (CbPD) is a fundamentally different approach compared to the text-based similarity evaluations described above. It is especially intended for being applied to scientific publications. Being substantially different, it is believed to be capable of overcoming some of the weaknesses of existing techniques. An overview classification of the PD approaches presented above is given in Figure 1.

2.3 Strength and Weaknesses of PDS

As described in [15] objective comparisons of the detection performance achieved by individual PDS are difficult. Authors proposing research prototypes tend to use different collections and evaluation methods. Initiated in 2009 the annual PAN International Competition on Plagiarism Detection (PAN-PC) addresses this lack of comparability. It attempts to benchmark PDS using a standardized collection and a controlled evaluation environment [36]. Results from the latest PAN-PC, held in June 2010, are presented for pointing out the capabilities of state-of-the-art PD prototypes.

Figure 2 displays the plagiarism detection (*plagdet*) scores of the top 5 performing external PDS and the 2 intrinsic PDS of MUHR ET AL. and SUÁREZ ET AL. participating in PAN-PC'10. The *plagdet* score is a measure developed for evaluating PDS in the PAN competitions. It considers the F_1 measure as well as the granularity (*gran*) of a detection method. The granularity reflects whether a plagiarized section is detected as a whole or in multiple parts: $plagdet = F_1 / log_2(1 + gran)$ [36].

In Figure 2 the scores are plotted dependent on the obfuscation techniques applied to plagiarized text segments. The overall *plagdet* score achieved in all categories is stated in brackets within each legend entry. Note that in the legend "- I" is attached to distinguish the system of MUHR ET AL. participating in the intrinsic from the one in the external task.

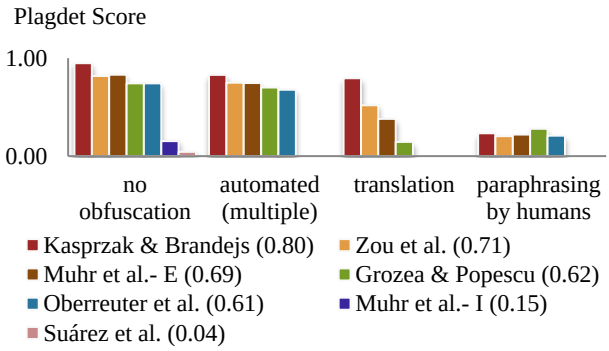

Figure 2: Results of top performing PDS in PAN-PC'10 [35]

The results indicate that unchanged copies of text segments can be detected with high accuracy by state-of-the-art PDS using fingerprinting or bag of words analyses. Detection rates for segments that were plagiarized and disguised by humans are substantially lower for all systems. On average, 76% of those realistically plagiarized segments could not be identified by the top 5 systems. The detection scores for automatically obfuscated plagiarism are 2.5 to 3.7 times higher than those for manually plagiarized sections.

The organizers of the competition judge 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 to English. The employed services, e.g. Google Translate, are similar or identical to those used for constructing the translated, plagiarized sections. It is hypothesized that human-made translations obfuscating real-world plagiarism are more complex and versatile, and hence less detectable by the tested PDS [35].

Intrinsic PDS performed significantly worse than systems using an external approach. The results are in line with those from the prior PAN competition in 2009 [36]. Intrinsic analyses seem to require larger volumes of text for working reliably [46].

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 [14]. 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

[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

documents has been first described and successfully applied to PD in [4, 15]. 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 [23]. 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 [10].

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 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 [34, 39], 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 [16, 41]. 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 [30]. 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

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

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.

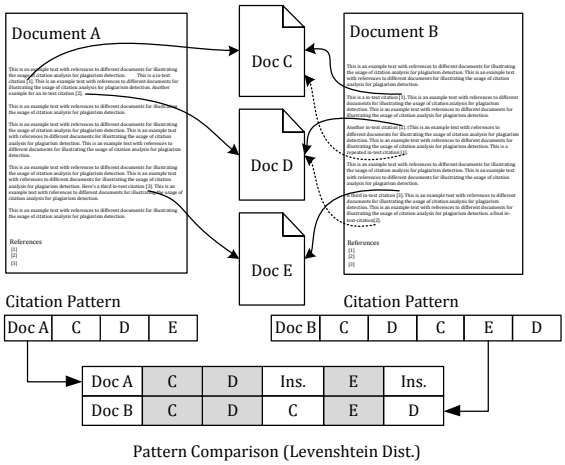

Figure 3: Identifying citation patterns for CbPD

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 [42]. 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. [17].

	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) [8, 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 [51]. 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, 37].

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 $t = (s_1, s_2, \mathrm{l})$ 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.

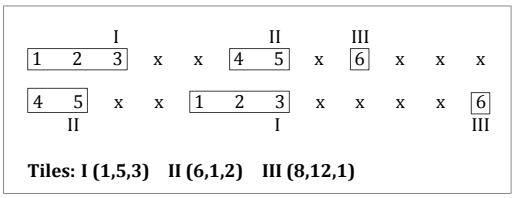

Figure 5: Citation Tiles

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.

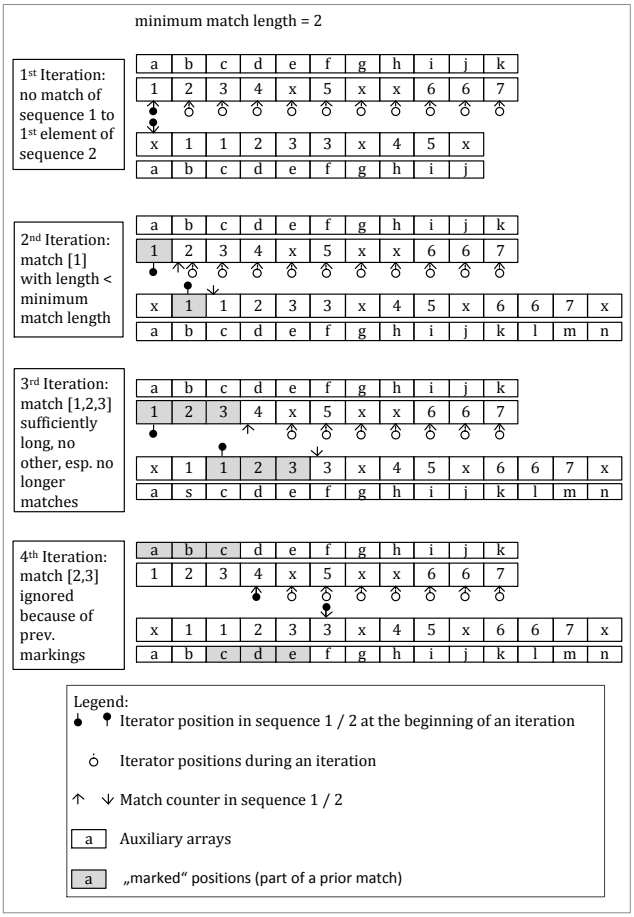

Figure 6: Example flow of the Greedy Citation Tiling algorithm.

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 [51].

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.

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,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).

   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.

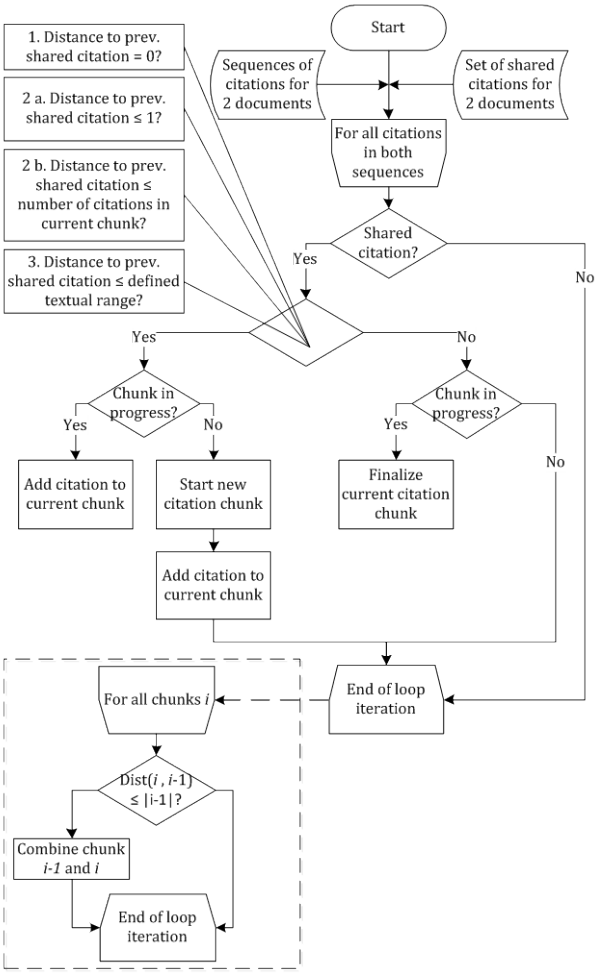

Figure 7: Forming of citation chunks

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.

3.2.6 *Strength and Weaknesses of the Algorithms*

In the following, the suitability of the presented algorithms is classified according to their ability to detect different forms of plagiarism.

	Plagiarism type	LCCS	GCT	CitChunk
Local	Identical (c&p segments, translations)	-	++	+
	Transpositions (shake&paste)	-	-	+
	Scaling	-	-	+
	Transpositions & scaling (paraphrases)	-	-	+
Global	Identical	++	++	+
	Transposition	+	-	+
	Scaling	+	-	++
	Transpositions & scaling	+	-	++

Figure 8: Overview of detection capabilities

These classifications are a generalization and should be considered with care. If, for instance, a text is translated word by word then the order of citations will not change much. This case would be classified as "identical" according to the table. In cases of free translations or the existence of several citations within one sentence varying sentence structures resulting from different languages might lead to different citation orders. In such cases a translated plagiarism would be classified as "transposition", "scaling" or even a combination of both.

Moreover, in the table it is distinguished between local and global forms of plagiarism. Local plagiarism can be observed on sentence level, whereas global forms describe document wide plagiarism.

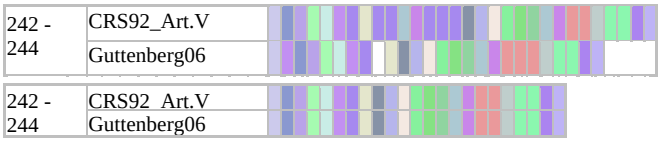

Figure 9: Example pattern identified in Guttenberg's thesis [53 plagiarism] by applying Citation Chunking

In initial experiments, the described detection procedures have been applied to prominent real world plagiarism cases in doctoral dissertations of German politicians [15]. As the table shows, Citation Chunking yielded best detection results in most cases in our tests. However, for text segments in which large portions of the contained citations were adopted unaltered, e.g. in not-freely translated plagiarisms, Greedy Citation Tiling provided clearer indications for potential plagiarism. Figure 9 shows an example of a citation pattern identified by Citation Chunking.

3.3 Prototype Citation-based PDS

For evaluating the different analysis procedures we developed an Open Source software system in Java coined *CitePlag*. 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 Figure 10.

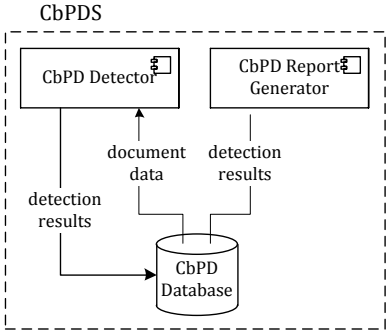

Figure 10: CbPDS system architecture

4. CONCLUSION

Previous studies have shown that CbPD is suitable for detecting forms of plagiarism that remain undetectable for the currently used text-based approaches. This paper has presented three algorithms for identifying citation patterns that have been observed in real-world plagiarism cases. The algorithms are coined Greedy Citation Tiling, Citation Chunking and Longest Common Citation Sequence.

These algorithms are able to detect citation transpositions, citation scaling and their combinations in cases of local and global plagiarism. The algorithms have been evaluated using several plagiarized documents such as the doctoral thesis of Guttenberg and by applying them to the PubMed Central Open Access Subset (PMC OAS). In [15] it was shown that the proposed algorithms could identify 13 out of the 16 sections containing translated plagiarism in the Guttenberg thesis. The tested text-based PDS were unable to detect any of them.

5. 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] BAKER, B. S. A Program for Identifying Duplicated Code. In *Computing Science and Statistics: Proceedings of the 24th Symposium on the Interface* (1992).

[3] BAO, J. P., LYON, C., LANE, P. C. R., AND JI, WEI, M. J. A. Comparing Different Text Similarity Methods. Tech. rep., Science and Technology Research Institute, University of Hertfordshire, May 2007.

[4] BELA GIPP, AND JOERAN BEEL. 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.

[5] BRETAG, T., AND MAHMUD, S. Self-Plagiarism or Appropriate Textual Re-use? *Journal of Academic Ethics 7* (2009), 193–205.

[6] BRIN, S., DAVIS, J., AND GARCIA MOLINA, H. Copy Detection Mechanisms for Digital Documents. In *Proceedings of the 1995 ACM SIGMOD international conference on Management of data* (New York, NY, USA, May 1995), M. Carey and D. Schneider, Eds., ACM, pp. 398–409.

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

[8] CROCHEMORE, M., AND RYTTER, W. *Jewels of Stringology*. World Scientific Publishing, 2002.

[9] DREHER, H. Automatic Conceptual Analysis for Plagiarism Detection. *Information and Beyond: The Journal of Issues in Informing Science and Information Technology 4* (2007).

[10] 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).

[11] ERRAMI, M., SUN, Z., LONG, T. C., GEORGE, A. C., AND GARNER, H. R. Déjà vu: a database of highly similar citations in the scientific literature. *Nucleic Acids Research 37*, suppl 1 (2009), D921–D924.

[12] FINKEL, R. A., ZASLAVSKY, A. B., MONOSTORI, K., AND SCHMIDT, H. W. Signature extraction for overlap detection in documents. In *Computer Science 2002, Twenty-Fifth Australasian Computer Science Conference (ACSC2002), Monash University, Melbourne, Victoria, January/February 2002* (Darlinghurst, Australia, 2002), M. J. Oudshoorn, Ed., vol. 4 of *Conferences in Research and Practice in Information Technology*, Australian Computer Society Inc., pp. 59–64.

[13] FRÖHLICH, G. Plagiate und unethische Autorenschaften. *Information - Wissenschaft & Praxis 57*, 2 (2006), 81–89.

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

[15] 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).

[16] 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), p 339.

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

[18] GUTBROD, M. A. *Nachhaltiges E-Learning durch sekundäre Dienste*. Dissertation, Technischen Universität Braunschweig Institut für Betriebssysteme und Rechnerverbund, Jan. 2007.

[19] 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).

[20] KAKKONEN, T., AND MOZGOVOY, M. Hermetic and Web Plagiarism Detection Systems for Student Essays—An Evaluation of the State-of-the-Art. *Journal of Educational Computing Research 42*, 2 (2010), 135–159.

[21] KANG, N., GELBUKH, A., AND HAN, S. PPChecker: Plagiarism Pattern Checker in Document Copy Detection. In *Text, Speech and Dialogue*, P. Sojka, I. Kopecek, and K. Pala, Eds., vol. 4188 of *Lecture Notes in Computer Science*. Springer Berlin / Heidelberg, 2006, pp. 661–667.

[22] KASPRZAK, J., AND BRANDEJS, M. Improving the Reliability of the Plagiarism Detection System - Lab Report for PAN at CLEF 2010. In *Notebook Papers of CLEF 2010 LABs and Workshops, 22-23 September, Padua, Italy* (2010), M. Braschler, D. Harman, and E. Pianta, Eds.

[23] 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.
[24] KO, P., AND ALURU, S. Space Efficient Linear Time Construction of Suffix Arrays. In *Combinatorial Pattern Matching* (2003), R. Baeza-Yates, E. Chávez, and M. Crochemore, Eds., vol. 2676 of *Lecture Notes in Computer Science*, Springer Berlin / Heidelberg, pp. 200–210.
[25] KURTZ, S. Reducing the Space Requirement of Suffix Trees. *Software-Practice and Experience 29*, 13 (1999), 1149–1171.
[26] LANCASTER, T. *Effective and Efficient Plagiarism Detection*. PhD thesis, School of Computing, Information Systems and Mathematics South Bank University, 2003.
[27] LIM, V. K. G., AND SEE, S. K. B. Attitudes Toward, and Intentions to Report, Academic Cheating Among Students in Singapore. *Ethics & Behavior 11*, 3 (2001), 261–274.
[28] MAURER, H., KAPPE, F., AND ZAKA, B. Plagiarism - A Survey. *Journal of Universal Computer Science 12*, 8 (Aug. 2006), 1050–1084.
[29] MCCABE, D. L. Cheating among college and university students: A North American perspective. *International Journal for Academic Integrity 1*, 1 (2005), 1–11.
[30] 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.
[31] MEYER ZU EISSEN, S., AND STEIN, B. Intrinsic Plagiarism Detection. In *Advances in Information Retrieval 28th European Conference on IR Research, ECIR 2006, London, UK, April 10-12, 2006. Proceedings* (2006), M. Lalmas, A. MacFarlane, S. M. Rüger, A. Tombros, T. Tsikrika, and A. Yavlinsky, Eds., vol. 3936 of *Lecture Notes in Computer Science*, Springer Berlin Heidelberg, pp. 565–569.
[32] 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, p. 226.
[33] MONOSTORI, K., ZASLAVSKY, A., AND SCHMIDT, H. Suffix vector: space- and time-efficient alternative to suffix trees. *Aust. Comput. Sci. Commun. 24*, 1 (2002), 157–165.
[34] PHELAN, T. A compendium of issues for citation analysis. *Scientometrics 45* (1999), 117–136. 10.1007/BF02458472.
[35] 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.
[36] POTTHAST, M., STEIN, B., EISELT, A., BARRÓN CEDEÑO, A., AND ROSSO, P. Overview of the 1st International Competition on Plagiarism Detection. In *PAN09 - 3rd Workshop on Uncovering Plagiarism, Authorship and Social Software Misuse and 1st International Competition on Plagiarism Detection* (2009), B. Stein, P. Rosso, E. Stamatatos, M. Koppel, and A. Eneko, Eds.
[37] PRECHELT, L., PHILIPPSEN, M., AND MALPOHL, G. JPlag: Finding plagiarisms among a set of programs. Technical Report 2000-1, Universität Karlsruhe, 2000.
[38] RUDMAN, J. The State of Authorship Attribution Studies: Some Problems and Solutions. *Computers and the Humanities 31* (1997), 351–365.
[39] SEGLEN, P. O. Why the impact factor of journals should not be used for evaluating research. *BMJ 314*, 7079 (1997), 497.
[40] SI, A., LEONG, H. V., AND LAU, R. W. H. CHECK: A Document Plagiarism Detection System. In *SAC '97: Proceedings of the 1997 ACM symposium on Applied computing* (New York, NY, USA, 1997), B. Bryant, J. Carroll, J. Hightower, and K. M. George, Eds., ACM, pp. 70–77.
[41] SMALL, H., AND GRIFFITH, B. C. The Structure of Scientific Literatures I: Identifying and Graphing Specialties. *Science Studies 4*, 1 (1974), pp. 17–40.
[42] SMYTH, B. *Computing Patterns in Strings*. Pearson Addison-Wesley, Harlow, England; New York, 2003.
[43] SOROKINA, D., GEHRKE, J., WARNER, S., AND GINSPARG, P. Plagiarism Detection in arXiv. Technical report computer science, Cornell University TR2006-2046, Dec. 2006.
[44] STEIN, B. Fuzzy-Fingerprints for Text-Based Information Retrieval. In *Proceedings of the I-KNOW '05, 5th International Conference on Knowledge Management, Graz, Austria* (July 2005), K. Tochtermann and H. Maurer, Eds., vol. Special Issue of *Journal of Universal Computer Science*, Springer-Verlag, Know-Center, pp. 572–579.
[45] STEIN, B., KOPPEL, M., AND STAMATATOS, E., Eds. *Proceedings of the SIGIR 2007 International Workshop on Plagiarism Analysis, Authorship Identification, and NearDuplicate Detection, PAN 2007, Amsterdam, Netherlands, July 27, 2007* (2007), vol. 276 of *CEUR Workshop Proceedings*, CEUR-WS.org.
[46] STEIN, B., LIPKA, N., AND PRETTENHOFER, P. Intrinsic Plagiarism Analysis. *Language Resources and Evaluation [Online Resource]* (2010), 1–20.
[47] STEIN, B., AND MEYER ZU EISSEN, S. Near Similarity Search and Plagiarism Analysis. In *From Data and Information Analysis to Knowledge Engineering Proceedings of the 29th Annual Conference of the Gesellschaft für Klassifikation e.V. University of Magdeburg, March 9–11, 2005* (2006), M. Spiliopoulou, R. Kruse, C. Borgelt, A. Nürnberger, and W. Gaul, Eds., Springer Berlin Heidelberg, pp. 430–437.
[48] TSATSARONIS, G., VARLAMIS, I., GIANNAKOULOPOULOS, A., AND KANELLOPOULOS, N. Identifying free text plagiarism based on semantic similarity. In *Proceedings of the 4th International Plagiarism Conference* (2010).
[49] WEBER WULFF, D. Copy, Shake, and Paste - A blog about plagiarism from a German professor, written in English. Online Source, Nov. 2010. Retrieved Nov. 28, 2010 from: http://copy-shake-paste.blogspot.com.
[50] WEBER WULFF, D. Test cases for plagiarism detection software. In *Proceedings of the 4th International Plagiarism Conference* (Newcastle Upon Tyne, 2010).
[51] WISE, M. J. String Similarity via Greedy String Tiling and Running Karp-Rabin Matching. Online Preprint, Dec. 1993. Retrieved from: http://vernix.org/marcel/share/RKR_GST.ps.
[52] ZHAN SU, BYUNG-RYUL AHN, KI-YOL EOM, MIN-KOO KANG, JIN-PYUNG KIM, AND MOON-KYUN KIM. Plagiarism Detection Using the Levenshtein Distance and Smith-Waterman Algorithm. In *Innovative Computing Information and Control, 2008. ICICIC '08. 3rd International Conference on* (June 2008), pp. 569 –569.
[53 plagiarism] GUTTENBERG, K.-T. F. *Verfassung und Verfassungsvertrag : Konstitutionelle Entwicklungsstufen in den USA und der EU*. Dissertation (**Retracted as plagiarism**), Universität Bayreuth, Berlin, 2009.