Comparing and Combining Content and Citation-Based Approaches for Plagiarism Detection

Solange de L. Pertile[1], Viviane P. Moreira[1], and Paolo Rosso[2]

[1]Instituto de Informática, UFRGS – Brazil,
{slpertile,viviane}@inf.ufrgs.br
[2]Natural Language Engineering Lab. - PRHLT Research Centre, Universitat Politècnica de València, Spain, prosso@dsic.upv.es

July 21, 2015

Abstract

The vast amount of scientific publications available online makes it easier for students and researchers reusing text from other authors and makes it harder for checking the originality of a given text. Reusing text without crediting the original authors is considered plagiarism. A number of studies report on the high prevalence of plagiarism in academia. As a consequence, numerous institutions and researchers are dedicated to devising systems to automate the process of checking for plagiarism. This work focuses on the problem of detecting text reuse in scientific papers. In this context, the contributions of this paper are twofold: (i) we survey the existing approaches for plagiarism detection based on content, based on content and structure, and based on citations and references; and (ii) we compare Content and Citation-based approaches with the goal of evaluating whether they are complementary and if their combination can improve the quality of the detection. We carried out experiments with real datasets of scientific papers and concluded that a combination of the methods can be beneficial.

Introduction

With the growing popularity of the Internet, many scientific papers are available enabling students and researchers to reuse text from other authors. Reusing text without crediting the original authors is considered plagiarism even if done unintentionally. A small fragment extracted from another paper is not considered plagiarism if it has a citation and a corresponding entry in the references/bibliography section of the paper. References must contain complete and accurate information so that readers can find the original source.

A number of studies report on the high prevalence of plagiarism in academia. Mccabe (2005) researched over 80,000 students in the U.S. and Canada and found that 36% of undergraduate students and 24% of graduate students admitted having copied or paraphrased sentences from the Internet without referencing them. Walker (2010) conducted a study in New Zealand in which 1k assignments were analysed. The author concluded that over a quarter of the submitted assignments contained plagiarism and about 10% were extensively plagiarised. Ashworth et al. (1997) and Gullifer & Tyson (2010) interviewed students and found that they had no clear definition of

plagiarism, and perhaps this was the reason for such high prevalence. Interestingly, contrary to what is expected, Walker (2010) and Youmans (2011) found that plagiarism awareness and the announced use of detection software did not reduce plagiarism. On the positive side, Park (2003) pointed out that although the Internet has made plagiarism more frequent, it has also made it easier to expose the offender.

While the aforementioned studies were based on student assignments, the literature also has some reports on cases of plagiarism in scientific publications. Y. Zhang (2010) analysed 662 papers submitted to the Journal of Zhejiang University (China) with the CrossCheck tool, of these, 22.8% presented unreasonable levels of copying or self-plagiarism, and 25.8% presented serious suspicions of plagiarism and copyright infringement. The similarity between source and suspicious documents was as high as 83% in some cases. Fang et al. (2012) analysed 2k retracted articles on the PubMed database and concluded that about 10% of them were due to plagiarism and 14% were due to duplicate publications. The most frequent cause of retraction was fraud (*i.e.*, data fabrication and falsification). High levels of text reuse have also been found by (Gupta & Rosso, 2012) who analysed the trends of text reuse in the ACL collection. This investigation focused on verbatim reuse, only. The results showed that self-reuse is more frequent than cross reuse. Recently, a database called Déjà vu[1](Garner, 2014) was populated by gathering 80K pairs of papers from Medline (a well-known repository of biomedical papers) for which a high content similarity was found by the eTBLAST search engine[2]. Not all cases are considered plagiarism, since document pairs in Déjà vu may include, for example, updated versions of a paper. García-Romero & Estrada-Lorenzo (2014) performed a bibliometric analysis of a sample of papers from Déjà vu that had been examined by curators. They concluded that plagiarism cases are published in journals with lower visibility, corroborating a finding by Fang et al. (2012), and receive fewer citations.

The enormous amount of digital documents available makes manual plagiarism analysis infeasible. As a consequence, automatic detection techniques have been proposed to deal with the various forms of plagiarism. Plagiarism analysis is generally based on the comparison of the contents of the documents. This comparison typically assigns a degree of similarity between the analysed documents which is quantified by a similarity score. Plagiarism detection systems can be classified as *intrinsic* or *external*. Intrinsic detection systems aim at identifying parts of a document which are likely to have been written by a different author, while external detection works by comparing a suspicious text with a reference collection of source documents.

While most methods for external plagiarism detection focus on comparing the textual contents of the main body of the documents (Kasprzak & Brandejs, 2010; Barrón-Cedeño & Rosso, 2009; Malcolm & Lane, 2009; Grozea et al., 2009), more recently, methods that aim at detecting plagiarism based on the analysis of references and citations have emerged (Alzahrani et al., 2012; Gipp & Beel, 2010; Gipp et al., 2014; Meuschke et al., 2012). In this work, we experiment with both types of methods.

Definitions of plagiarism

Plagiarism is one of the most serious forms of academic misconduct. It is defined as the act of appropriating of another person's ideas, words, or works without giving credit to the original source (Anderson & Steneck, 2011; Stein & zu Eissen, 2006). The literature mentions several types

[1] `http://dejavu.vbi.vt.edu/dejavu/`

[2] `http://etest.vbi.vt.edu/etblast3/`

of plagiarism which can be classified mainly into five categories (Collberg & Kobourov, 2005; Maurer et al., 2006).

- **Copy-paste Plagiarism:** verbatim copy of the original.
- **Paraphrasing Plagiarism:** changing words from the original text using synonyms, reordering, or reaffirming the same content in different words.
- **Translation Plagiarism:** translating content into different languages.
- **Self-Plagiarism:** reusing parts or entire texts from one's own previous work.
- **Idea Plagiarism:** using ideas from another author as being their own.

IEEE[3] and ACM[4] have guidelines regarding text reuse and plagiarism considering the types identified above. Regarding the severity of the offence, IEEE defines the following five levels of plagiarism:

- **Level One:** verbatim copying of a full paper, or the verbatim copying of a major portion (greater than 50%), both uncredited.
- **Level Two:** the uncredited verbatim copying of a large portion (between 20 and 50%) of the original paper.
- **Level Three:** the uncredited verbatim copying of elements, paragraphs, sentences, illustrations, etc.
- **Level Four:** the uncredited improper paraphrasing of pages or paragraphs.
- **Level Five:** the credited verbatim copying of a major portion of a paper without clear delineation (e.g., quotes or indents).

Note that levels one to three can be more easily detected. Level four requires identifying paraphrases, while level five calls for structure-based plagiarism detection. All five levels need citation analysis to decide whether the source has been adequately referenced.

A sixth level could be included in the list above to account for the cases in which the plagiarist explicitly aims at misleading plagiarism detection software. The malicious actions may consist, for example, in replacing some Latin characters by their Cyrillic lookalikes (Beall, 2013). As a result, the plagiarised text and its source would have a very low similarity.

The Committee on Publication Ethics (COPE)[5] is an important forum for journal editors and publishers. It defines guidelines regarding text recycling and recommends a course of actions (depicted as flowcharts) on how to proceed when there is a suspect of plagiarism in a submitted or published paper. The guidelines recommend that submitted papers are checked by a plagiarism detection software. When an overlap with other publications is found, if it is considered significant, the paper should be rejected. If the overlap is minor, authors could be asked to rewrite. For published papers, if the overlap is considered significant, a retraction article may be published. Still, the decision as to what extent text reuse is tolerated is left to the editor's discretion.

[3] `http://www.ieee.org/publications_standards/publications/rights/plagiarism_FAQ.html`
[4] `http://www.acm.org/publications/policies/plagiarism_policy`
[5] `http://publicationethics.org/`

The concept of self-plagiarism is quite controversial, since, as pointed out in an editorial by Cronin (2013), there are many legitimate reasons why authors may reuse previous texts of their own. For example, the paper being published may be an extended version of a workshop paper. García-Romero & Estrada-Lorenzo (2014) reported that a forum organised COPE was unable to reach a consensus on this matter. This shows even human specialists struggle to determine whether and to what level text reuse is in fact fraudulent.

Another important aspect that influences the judgement as to whether reused text can be considered plagiarism is the location (*i.e.*, the section in the paper) in which the reuse was found. Generally, recycled text in the literature review or in the methods section is more tolerated than in the results section. COPE, Garner (2014) and Alzahrani et al. (2012) mention this fact.

Authorship attribution is a related research area in which the goal is to determine the author of a given text based on characteristics such as vocabulary, syntax, and structure. This topic, however, is outside the scope of our work, since our focus is on external plagiarism detection. For an in-depth survey on author attribution, please refer to Juola (2006).

Plagiarism is a very important ethical issue, debates, tools, and efforts to address it are increasingly common. The vast number of documents to compare, the vague definition of what consists in plagiarism, and the tricks applied by the offenders to mislead detection software contribute to the complexity of this task, making it very challenging. Furthermore, one of the added difficulties in dealing with real publications is that judging whether papers which share contents are indeed cases of plagiarism is troublesome. Thus, some works prefer to call it simply "text reuse".

Aims and Scope

Our focus is on the problem of external plagiarism detection, more specifically, we are interested in plagiarism detection methods that can be applied to scientific publications. As a result, we pay particular attention to Citation-based detection.

Since the use of Citation-based metrics is recent, there is little empirical evidence about its effectiveness. An important question is whether Content-based and Citation-based detection methods are complementary or whether one subsumes the other. We have experimented with computing both types of methods on synthetic and real collections of scientific publications. If both methods are found to be complementary, a combination of them would be beneficial – this hypothesis was also evaluated.

The contributions this work include:

- an extensive survey on existing methods for plagiarism detection categorizing them as based on content, based on content and structure, and based on citation — we also report on systems that took part in plagiarism evaluation campaigns;
- a report on the available tools and their functionalities;
- an experimental comparison of Content-based and Citation-based detection methods.

Organization of the Article

This article is organized as follows: The next Section provides a description of the problem of plagiarism detection, the steps involved in the process, and the metrics used to evaluate the quality of plagiarism detection systems. Then, we survey approaches for external plagiarism found in the

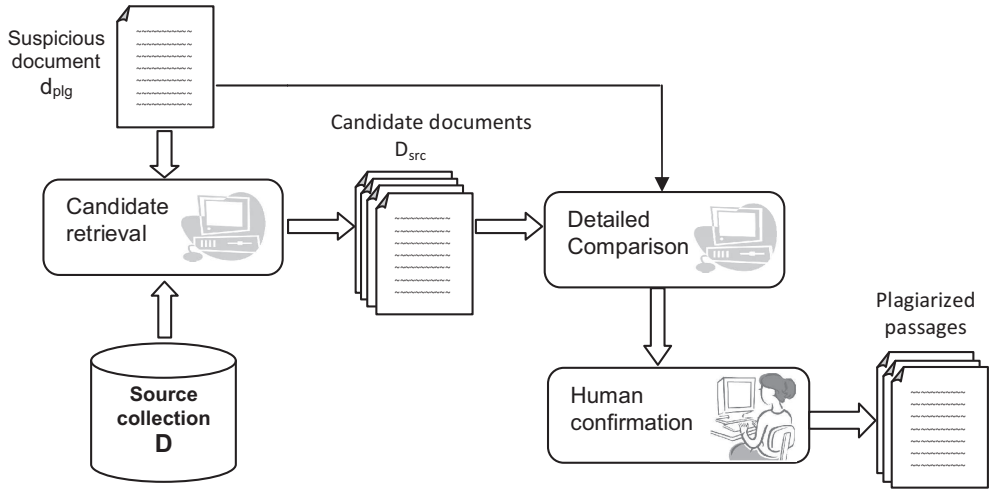


Figure 1: Three phases of the plagiarism detection process

literature classifying them into three categories: (*i*) based on content, (*ii*) based on content and structure, and (*iii*) based on citations and references. The following section is devoted to presenting existing plagiarism detection tools. Then, we present an experimental comparison between Content-based and Citation-based detection approaches, followed by the conclusion.

Problem Overview

Problem Definition

The problem of plagiarism detection is commonly defined as (Potthast et al., 2010):

Definition: Given a collection of documents D and a set of plagiarism cases S, *the task of a plagiarism detection system* is to identify a set of detections R maximizing $S \bigcap R$. Given a suspicious document d_{plg}, $D_{src} \subseteq D$ is the set of possible source documents selected from D. A case of plagiarism is represented by the tuple $s = \langle s_{plg}, d_{plg}, d_{src}, s_{src} \rangle$, where s_{plg} is a plagiarized passage from document d_{plg} and s_{src} corresponds to the original passage in the source document d_{src}. Likewise, a *plagiarism detection* for document d_{plg} associates an allegedly plagiarized passage r_{plg} in d_{plg} to r_{src} in d'_{src} and is denoted by $r = \langle r_{plg}, d_{plg}, r_{src}, d'_{src} \rangle$. We consider that r detects s if $r_{plg} \bigcap s_{plg} \neq \emptyset \wedge r_{src} \bigcap s_{src} \neq \emptyset \wedge d_{src} = d'_{src}$.

In the next sections, we discuss plagiarism detection in the light of these definitions.

Phases

The process of detecting plagiarism is usually divided into three phases (Stein et al., 2007), as depicted in Figure 1. The first step is known as *candidate retrieval*. In this step, given a suspicious document d_{plg} the aim is to identify from D the documents D_{src} which are the likely sources for the plagiarism cases in d_{plg}. This step significantly narrows down the number of candidates since typically $D_{src} \ll D$. In the *detailed comparison* step, d_{plg} is compared to each document in D_{src}

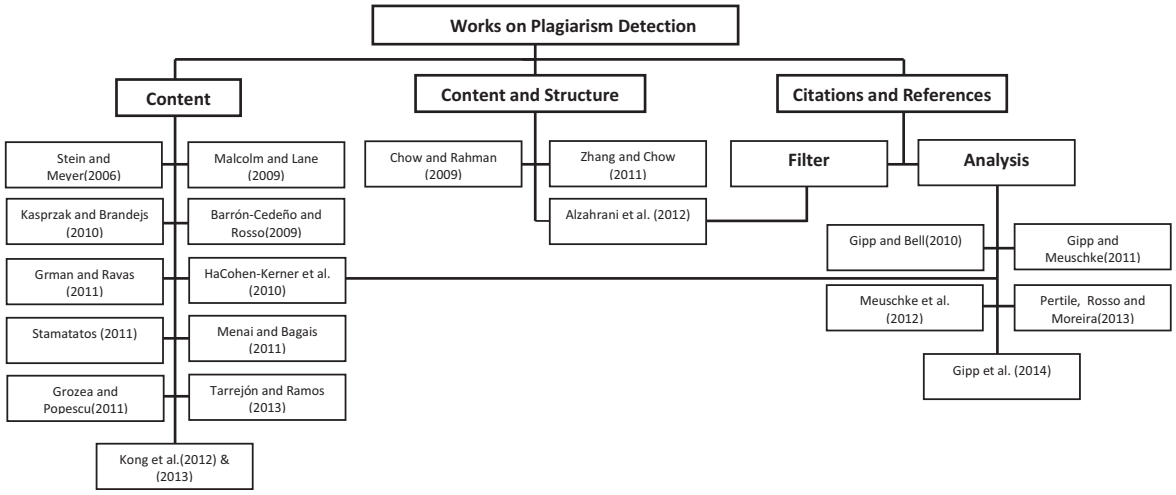


Figure 2: Taxonomy of Plagiarism Detection Methods

to enable the identification of the plagiarized passages. The third and final step requires human judgement to decide whether the suspicious passage indeed represents an instance of plagiarism.

In this section, we discuss the techniques for external plagiarism detection found in the literature. Figure 2 shows the surveyed techniques organized according to their type: (*i*) based on content, (*ii*) based on content and structure, and (*iii*) based on citations and references. Each of the following subsections is devoted to one group of techniques. Then, we present a comparison of these groups.

Works on Plagiarism Detection

Plagiarism Detection based on Content

Content-based detection is the most widely used technique for identifying plagiarism. It usually consists in contrasting text fragments (such as paragraphs, sentences, words, word n-grams, or character n-grams) from the suspicious documents against possible sources. Three main strategies are generally used for Content-based detection, namely: bag-of-words, n-grams, and fingerprints.

Bag-of-words are used mainly for the candidate retrieval step. As the name suggests, this strategy disregards word order. It consists in applying an Information Retrieval System to retrieve documents that share words with the suspicious document. Retrieval could be based on the Vector Space Model (VSM), for example. In this model, documents are represented as n-dimensional vectors containing the strength of association between the document and each of the n distinct terms in the collection. The similarity between a suspicious document and the possible sources can be computed as the cosine of their vectors. Retrieving candidates with a bag-of-words approach was used in several works (Torrejón & Ramos, 2013; Kong et al., 2012, 2013; Sanchez-Perez et al., 2014). Starting in 2012, PAN participants are requested to retrieve candidate source documents using a search engine, ChatNoir, which is made available as an API (Potthast et al., 2012).

N-gram matching is the prevalent approach in Content-based detection. A word n-gram is a

sequence of n consecutive words. The intuition is that the more n-grams in common a pair of paper has, the more similar they are. A number of studies have tried to establish the value for n that yields the best accuracy. Experiments made by Barrón-Cedeño & Rosso (2009) on the METER[6] corpus showed the best results in terms of F1 were obtained by 2-grams and 3-grams. While 2-grams achieved the best recall, 3-grams yielded the best precision. In another study, Vallés Balaguer (2009) tested values of n ranging from 4 to 6 to examine the PAN-09 corpus. The authors advocate the use of 6-grams as it achieved the best precision with a small loss in recall. 6-grams were also used by Gupta & Rosso (2012).

Grozea & Popescu (2011) used character level n-grams in their system, named Encoplot. A similarity matrix is built with scores for each suspicious-source pair. The similarity between documents included the number of times a moving window of fixed size (256 characters) contains a number of n-gram matches above a fixed threshold (64).

The CoReMo system (Torrejón & Ramos, 2010), which has taken part in PAN many times, was the winner of the 2013 competition for text alignment. Torrejón & Ramos (2013) uses *skip n-grams* detect obfuscated cases. A skip n-gram allows words to be skipped and this way more matches can be selected. This system also uses translation to detect cross-language plagiarism.

The method developed by Stamatatos (2011) is based exclusively on stopword n-grams. Documents are represented solely by the appearance of a predefined list of stopwords in the text. The aim is to find common n-grams of stopwords between the source and suspicious documents. To avoid assigning a high similarity to unrelated passages containing very frequent stopwords, a constraint is added to the candidate retrieval step to discard pairs of documents that share only very frequent stopwords. Experiments on the PAN-10 corpus presented better results compared to other approaches in cases of simulated plagiarism and artificial plagiarism with high obfuscation. Furthermore, the approach showed that the length of the plagiarism passage affects the results, cases in long passages (>10K characters) are easier to detect.

Fingerprints are compact representations of documents that aim at capturing their identity. For a fingerprinting algorithm, the chance of generating the same representation for different documents should be negligible. In order to do that, documents are separated into chunks and a function is applied to each of the chunks producing an integer value. Hoad & Zobel (2003) provide a detailed review of fingerprinting and compare it with an identity measure that they propose (and variations thereof). Their experiments showed that the proposed identity measures outperform fingerprinting methods. They have also noted that *anchor-based methods* achieved good results. An anchor is a string (or an n-gram) in the text of the document. Anchors should be chosen so that there is at least one in each document but not so common that the fingerprint becomes too large.

Stein & Meyer (2006) proposed an improvement to the MD5 algorithm based on fuzzy fingerprints with the goal of generating the same hash code to similar fragments. The MD5 hash algorithm is one of the most popular approaches for detecting copy-paste plagiarism (i.e., identical text fragments). Index terms are conflated into a small number of equivalence classes such that all terms start with the same prefix. The vector of relative frequencies for each class is computed and compared against the same vectors computed on a reference collection. This deviation is abstracted into a fuzzy deviation scheme and hash values are computed. The fuzzy fingerprint is the union of the hash values for a word n-gram. The goal is to investigate the runtime performance and the difference between the scores for fuzzy-fingerprint similarity and cosine similarity under the vector

[6] `http://nlp.shef.ac.uk/meter/`

space model. The authors concluded that fuzzy-fingerprints resembled the cosine similarity better than the MD5 fingerprint. This comparison was over all documents on the RFC collection[7], which contains 3K technical documents about the Internet.

The system by Kasprzak & Brandejs (2010) was the winner at PAN-10. It computes MD5 hashes fingerprints to documents with overlapping word 5-grams. A chunk is represented by the most significant 30 bits of the hash. The similarity between document pairs is calculated from of their common chunks. Document pairs that contain 20 or more common chunks are selected as candidates. For each pair, the method analyses whether the common chunks form one or more valid intervals, in which the gap between two neighbouring common chunks is not bigger than 50 chunks. Common chunks that satisfy this condition are reported as plagiarism cases.

HaCohen-Kerner et al. (2010) applyed a variety of methods to identify similar papers on a collection of 10,100 published academic papers from the ACL Anthology[8]. Compared methods include several variations of fingerprinting and anchor-based strategies and combinations. The paper reports how many pairs of papers were considered similar by each method. They concluded that full-fingerprints of length 3 (*i.e.*, considering all chunks of length 3) was the best method.

Dealing with Paraphrases is an important feature for identifying cases of plagiarism in which the offender tried to disguise the duplication. Systems that handle paraphrases have performed well in evaluation campaigns. The system by Grman & Ravas (2011) was the winner of the PAN-11 competition. The method is based on calculating the number of matching words for a pair of passages from the source and the suspicious documents. First, pairs of passages in which the number of matching words exceeds a certain threshold are selected. WordNet[9] is used as a source of synonyms. The use of synonyms and the disregard for word order aid the detection of paraphrased and translated plagiarism. For a deeper analysis into how paraphrase analysis impacts plagiarism detection systems, please refer to Barrón-Cedeño et al. (2013).

Plagiarism Detection based on Content and Structure

The methods discussed in the previous section are based only on the text of the documents. However, some approaches have used information about the structure of the document (Chow & Rahman, 2009; H. Zhang & Chow, 2011; El Bachir Menai & Bagais, 2011; Alzahrani et al., 2012). In some types of documents, such as scientific documents, structure plays an important role. Structural information is represented by headers, sections, paragraphs, references, etc. According to Alzahrani et al. (2012), the methods for partitioning scientific publications consider that the structure is usually presented by visual elements, such as location, position, punctuation, length, font type and size. Some methods employ keyword based strategies to label specific content, for example, using words like *chapter*, *introduction*, and *section headers* (Burget, 2007; Stoffel et al., 2010).

Representing documents as trees is an alternative for handling its structure. In the work by Chow & Rahman (2009), documents are represented as trees in which paragraphs are at the bottom layer, pages at the middle layer and whole documents at the top. Candidate retrieval is performed at the top layer while the other two layers are used for detailed analysis. The nodes

[7] `http://www.rfc-editor.org/rfc.html`
[8] `http://aclweb.org/anthology/`
[9] `http://wordnet.princeton.edu/`

contain word histograms which then go through principal component analysis to reduce dimensionality. Later, the same authors H. Zhang & Chow (2011) propose adding a weight parameter to the histogram of the documents and paragraph nodes to generate a signature representation. Documents are sorted in ascending order of distance and lower level analysis can be performed on the top matches. Experiments on a collection of 10K HTML documents[10] have been reported in both works achieving good results. A similar structure was employed by El Bachir Menai & Bagais (2011) in which documents are represented in a tree structure (document level, paragraph level, and sentence level), and then compared level by level from root to leaf. To extract the fingerprints of a document, the method uses the hash function by Kernighan & Ritchie (1988). If two documents have a common number of hashes above a fixed threshold, then the pair of documents presents a probable similarity and the analysis follows to paragraph level. The Longest Common Substring (LCS) algorithm is used to quantify the similarity between sentences. Two sentences are considered as matches if the length of the LCS is greater than a given threshold. Experiments report good results; however, they were performed over a very small set of Arabic texts.

Taking the structure of a scientific paper into consideration was central to the approach by Alzahrani et al. (2012). In order to achieve that, the text of the documents is divided into components, and a plagiarism case in the *introduction* and definitions is considered less important comparing to a case of plagiarism in the *evaluation* or *discussion* components. Thus, this proposal introduces different functions to measure component factor-weight based on their distinct terms. A case of plagiarism is identified when a component from a suspicious document gets a high similarity score in relation to a component from a source document. The experiments were performed on an artificial collection, and the results showed that the use of structural information can contribute to both candidate retrieval and plagiarism detection.

Plagiarism Detection Based on References and Citations

Strategies which analyze citations and references can be used in two different ways by plagiarism detection systems: as a filter (Alzahrani et al., 2012; Sorokina et al., 2006) (*i.e.*, checking whether the citation is giving credit to the original source); or examining similarities in citations/references across documents (HaCohen-Kerner et al., 2010; Gipp & Beel, 2010; Gipp & Meuschke, 2011; Meuschke et al., 2012; Gipp et al., 2014; Pertile et al., 2013).

Using citation analysis as a filter to discard false positives was the strategy used by Sorokina et al. (2006) and Alzahrani et al. (2012). They assume that if the author of an article A appears in the reference list of article B, then this could be a case of "mild plagiarism" since plagiarists do not cite their source. The downside is that discarding large portions of identical passages (even with proper references) means that Level 5 plagiarism cases (see Introduction) would go undetected. The authors do not analyze the gain/loss brought by such filtering.

Using citation analysis as a source of similarity Gipp & Beel (2010); Gipp & Meuschke (2011); Meuschke et al. (2012) and Gipp et al. (2014) introduced the use of citations and references as a source of similarity between documents. They propose several similarity functions based on shared references and citations. The more two documents share, the more similar they are. Documents are represented only by their citations and references, which form a pattern. Some of the proposed similarity functions take the order of the citations into consideration (*e.g.,* Longest Com-

[10] `http://www.ee.cityu.edu.hk/~twschow/Html_CityU1.rar`

mon Citation Sequence and Greedy Citation Tiling which are adaptations of traditional algorithms) while other similarity functions (such as Bibliographic Coupling and citation chunking) disregard the order of the citations. The authors report experiments on three datasets: (i) the doctoral thesis of a former German Minister of Defense, which was recognized as plagiarized from several sources; (ii) an ongoing crowd-source investigation of plagiarism allegations in German dissertations; and (iii) documents taken from the open access subset of PubMed[11]. Their results indicate that Citation-based detection outperforms Content-based detection in cases of strongly disguised plagiarism.

HaCohen-Kerner et al. (2010) also compared the titles of papers in the reference section of the two papers being analysed. The authors report that, this comparison retrieved many suspicious pairs, thus with numerous false positives. In some cases, different papers had identical references sections.

In our previous work (Pertile et al., 2013), the aim was to compare the similarity of scientific papers based on the analysis of co-occurrences in citations. Our assumption was that a high rate of inter-document co-occurrences could be an indication of plagiarism. Experimental results have shown that most of the cases with co-occurrences in citations correspond to plagiarism. However, the collection used in the experiments were artificially created using simulated cases of plagiarism thus we could not draw conclusions as to what happens in real collections of scientific documents.

Comparison of the approaches

In this section, we present a comparative analysis of the plagiarism detection methods described in the previous sections. The criteria used for this comparison are the following:

- **Approach:** indicates whether the proposal is based on content analysis (Co), on citation analysis (Ci), content and structure analysis (Co&S), or a combination of all three.
- **Candidate retrieval:** technique used in the candidate retrieval stage.
- **Detailed analysis:** technique used in the detailed analysis stage.
- **Collection:** the document collection used in the experiments.
- **Doc type:** type of document analysed (scientific paper or plain text).
- **Evaluation:** results of experimental evaluations in terms of precision (P), F1, and Plagdet. The dashes mean that the evaluation metrics were unavailable.

According to Table 1, most works are based solely on content analysis and use techniques which rely on word n-grams and fingerprints. Few authors (Alzahrani et al., 2012; Gipp et al., 2014; HaCohen-Kerner et al., 2010; Pertile et al., 2013) began to develop proposals which exploit citation analysis.

Note that the evaluation results shown on the last three columns of Table 1 are not comparable, since the experiments were done on different corpora. To enable a fair comparison, a common benchmark is needed. This is the aim of the PAN Workshops, and also, of the experiments we performed in this paper.

[11] `http://www.ncbi.nlm.nih.gov/pubmed`

Table 1: Plagiarism Detection Approaches

Work	Approach	Candidate Retrieval	Detailed Analysis	Corpus	Doc type	Evaluation	
						P	F1
Stein & Meyer (2006)	Co	fuzzy-fingerprints	N/A	RFC collection	plain	-	-
Barrón-Cedeño & Rosso (2009)	Co	n-grams in common (n=2 and 3)	classification based on a containment measure	METER	plain	0.74	0.66
Kasprzak & Brandejs (2010)	Co	n-grams in common (n=5)	gap between two neighbouring common n-grams	PAN 2010	plain	0.94	0.80
Stamatatos (2011)	Co	stopword n-grams in common (n=11)	common sequence word n-grams of stopwords (n=8)	PAN 2010	plain	0.94	0.56
Grman & Ravas (2011)	Co	matching words between a pair of passages + WordNet for synonyms	matching words between a pair of passages + WordNet for synonyms	PAN 2011	plain	0.95	0.82
Grozea & Popescu (2011)	Co	similarity matrix with the number of common n-grams within a window	common ngrams + clustering	PAN 2011	plain	0.81	0.48
Kong *et al.* (2012)	Co	ChatNoir API,TF-IDF, VSM ranking	cosine and Dice similarity	PAN 2012	plain	0.82	0.73
Gupta & Rosso (2012)	Co	word n-grams in common (n=6)	word n-grams in common (n=6)	ACL Anthology	sci	-	-
Kong *et al.* (2013)	Co	ChatNoir API, TF-IDF, PatTree and Weighted TF-IDF	cosine similarity, Bilateral Alternating Merging	PAN 2013	plain	0.83	0.82
Torrejón & Ramos (2013)	Co	IR system & Reference Monotony Pruning	Surrounding Context N-grams & Odd-Even N-grams	PAN 2013	plain	0.89	0.81
Sanchez-Perez *et al.* (2014)	Co	TF-IDF, VSM (cosine) and Dice	cosine + thresholds for allowed gaps, size &	PAN 2014	plain	0.88	0.88
Gipp (2013) Gipp (2014)	Ci	references/citations in common	references/citations in common	PubMed OAS GuttenPlag Wiki VroniPlag Wiki	sci	-	-
Pertile *et al.* (2013)	Ci	co-occurrences in citations	co-occurrences in citations	Azahrani's	sci	0.47	0.03
Chow & Rahman (2009)	Co&S	Multilayer Self Organizing Map, word histograms, PCA	Multilayer Self Organizing Map, word histograms, PCA	Html_CityU1	plain	-	-
Bachir Menai & Bagais (2011)	Co&S	fingerprints	Longest Common Substring	Arabic texts	plain	0.93	0.96
Zhang & Chow (2011)	Co&S	Multilayer Self Organizing Map, word histograms, PCA, Earth Mover's Distance	Multilayer Self Organizing Map, word histograms, PCA, Earth Mover's Distance	Html_CityU1	plain	0.74	0.74
HaCohen-Kerner *et al.* (2010)	Co&Ci	-	Fingerprints, Anchor-based methods, Titles of references	ACL Anthology	sci	-	-
Alzahrani *et al.* (2012)	Co&S&Ci	TF-IDF weighting for different sections, Cosine similarity	Jaccard similarity	Alzahrani's	sci	-	-

Plagiarism Detection Tools

Plagiarism Detection tools are computer programs that compare documents with likely sources in order to identify similarities and thus find possible cases of plagiarism. They can be used to detect and prevent plagiarism. Next, we describe various such tools. Many of them are proprietary solutions, so there are no details on the algorithms they use.

Antiplag [12] is based on the winner of the plagiarism detection competition in PAN-2011 (Grman & Ravas, 2011). It provides a web portal that queries a database comprised of academic works (BSc, MSc, doctoral dissertations etc.) and publications selected from other web sources. Slovak

[12] `http://www.svop.eu/svop/index.php/produkty/antiplag`

Table 2: Plagiarism Detection Tools

Tools	Features							
	Open Source	Free	Private	Platform	Citation Analysis	Content Analysis	Structure Analysis	Paraphrase Plagiarism
Antiplag			✓	Web		✓		
Wcopyfind	✓			Desktop		✓		
Ferret		✓		Desktop		✓		
Grammarly			✓	Web		✓		
Turnitin			✓	Web		✓		
eTBLAST		✓		Web		✓		
iThenticate			✓	Web		✓		✓
Ephorus			✓	Web		✓		
CheckForPlagiarism.net			✓	Web		✓		✓
Compilatio.net			✓	Web		✓		
DupliChecker		✓		Web		✓		
EVE2			✓	Desktop		✓		
Plagium		✓	✓	Web		✓		
PlagScan			✓	Web		✓		
PlagAware		✓	✓	Web		✓		
CoReMo			✓	Web		✓		✓
CitePlag		✓		Web	✓			

universities are obliged to send all of the publications to be compared with the contents of the publications on this database.

WCopyFind[13] is a tool developed in 2004 at the University of Virginia (USA). The input to this tool is a set of suspicious and source documents. The output presents the document pairs with similarity scores higher than a given threshold. This analysis is based on word n-grams (n is a parameter which can be set) and only exact matches are reported. As reported in the previous section, this was the tool of choice for two studies Gupta & Rosso (2012) and Vallés Balaguer (2009).

Ferret[14] is also based on word n-grams. It is a free plagiarism detection system developed at the University of Hertfordshire (UK). This similarity score is calculated as the number of shared 3-grams between two documents divided by the number of distinct 3-grams in the pair. The desktop version of the system performs pairwise comparisons with all documents from a dataset created by the user. The output highlights the similar sections between each pair.

Grammarly[15] checks a given text against a large database of web documents. It highlights the sections that have been previously published elsewhere. As a writing aid, it suggests possible references for the texts in which similarities are found.

Turnitin[16] is a web-based service which detects material copied from the web and it also does cross-checking with text documents in a database composed of documents submitted in previous analyzes.

eTBLAST[17] is also a web-based search engine which looks for literature which matches exactly the paragraph given as input. The databases searched include Medline, PubMed, and arXiv.

iThenticate[18] is a private service developed by Turnitin that aims at checking the originality of researches for educational institutions worldwide. The suspicious documents are confronted with

[13] `http://plagiarism.phys.virginia.edu/`
[14] `http://peterlane.info/software/ferret.html`
[15] `http://http://www.grammarly.com/`
[16] `http://turnitin.com/pt_br/`
[17] `http://etest.vbi.vt.edu/etblast3/`
[18] `http://www.ithenticate.com/`

CrossCheck's[19] large database.

Ephorus[20] can be integrated with digital learning environments. It also has a database of documents submitted by the participating institutions. We did not find details on how inter-document similarities are calculated.

Checkforplagiarism[21], Compilatio.net[22], DupliChecker[23], EVE[24], Plagium[25], PlagScan[26] and PlagAware[27] compare a given essay text against material freely available on the Web.

Torrejón & Ramos (2013) developed CoReMo[28] a system that allows comparing suspicious documents with a local collection or web sources. In addition, private institutions can adhere to this system and index their documents as part of the collection of source documents.

CitePlag[29] was developed by Gipp et al. (2013) and performs plagiarism detection based solely on citation patterns and references. This prototype accepts the upload of two documents which are compared against each other. Another option is to choose two publications from PubMed Open Access documents to be compared pairwise.

Table 2 shows some characteristics of plagiarism detection tools discussed in this section. Note that nearly all tools focus on exact matches and do not analyze citations and references. Another feature that is absent from most tools is paraphrase plagiarism analysis; only CoReMo, CheckForPlagiarism.net, and Ithenticate report allowing this type of analysis. None of the tools presented in Table 2 combine content and citations/references analysis.

Comparison Between Content- and Citation-Based Methods

One of the aims in this paper is to study the relationship between Content and Citation-based detection methods. Thus, in this section, we present experiments designed to answer the following questions: *Is there any correlation between Content- and Citation-based methods?* In other words: do they detect the same cases or do they complement each other? and *Is there a benefit in combining the two kinds of methods?* Our experiments computed four Citation-based metrics and contrasted with the results of a freely available tool for Content-based detection. Note that our focus was on comparing the effectiveness and not the efficiency of the methods.

The next subsections detail the experimental setup and the results.

Experimental Setup

Similarity metrics. The metrics based on references and citations we computed were:

- **Bibliographic Coupling:** computes a similarity score between a pair of documents as the absolute number of references in common between them (Meuschke et al., 2012).

[19] http://www.crossref.org/crosscheck/
[20] https://www.ephorus.com/
[21] http://www.checkforplagiarism.net/
[22] https://www.compilatio.net/
[23] http://www.duplichecker.com/
[24] http://www.canexus.com/
[25] http://www.plagium.com/
[26] http://www.plagscan.com/
[27] http://www.plagaware.com/
[28] http://www.coremodetector.com/
[29] http://citeplag.org/

- **Citing Frequency-Score (CF-Score):** computes a similarity score between a pair of documents based on matching citation patterns. It considers that rarely cited documents are more important and should receive a higher score (Meuschke et al., 2012). Here, this metric was adapted to compute a score between two documents that share matching references (rather than citation patterns). It is essentially a weighted sum of the references two documents share giving more weight to the references with fewer citations. It is computed as: $CF(d_i, d_j) = \sum_1^n \frac{N}{|r_i|}$
  where n is the number of references in common between documents d_i and d_j, N is the number of documents in the collection, and $|r_i|$ is the number of times reference r appears in the collection.

- **Jaccard similarity:** (or overlap) computes a similarity score between a pair of documents as the intersection between their references divided by their union. It is calculated as: $sim(d_i, d_j) = \frac{r_i \cap r_j}{r_i \cup r_j} = \frac{n_{i,j}}{n_i + n_j - n_{i,j}}$
  where r_i and r_j are the references in documents d_i and d_j, respectively, $n_{i,j}$ is the number of shared references between documents d_i and d_j, n_i and n_j are the number of references in documents d_i and d_j, respectively.

- **Co-occurrences in Citations:** proposed by Pertile et al. (2013), this metric calculates the co-occurrences in citations between two papers by sliding two windows of size s through the text of the documents (one for each document) and then computing the Jaccard similarity between the co-occurrences found in the documents. This is calculated as: $sim(d_i, d_j) = \frac{d_i \cap d_j}{d_i \cup d_j} = \frac{n_{i,j}}{n_i + n_j - n_{i,j}}$
  where d_i and d_j are the documents i and j, respectively, $n_{i,j}$ is the number of shared co-occurrences between documents i and j, n_i and n_j are the number of co-occurrences in documents i and j, respectively. In this experiment, s=15 was adopted as window size since it presented the best results in terms of F1 in an earlier study (Pertile et al., 2013).

For Content-based analysis of *text reuse*, we used the WCopyFind tool (described in page 12). The similarity score is based on the number of matching word n-grams between pairs of documents.

Document Collection. Evaluating plagiarism detection on scientific documents has an important limitation: the availability of test collections. Ideally, a test collection should include real cases of plagiarism, a ground truth, and references/citations to allow the detection method to establish whether the source was properly acknowledged. The collections we used here have already been used in other plagiarism detection studies (Gupta & Rosso, 2012; García-Romero & Estrada-Lorenzo, 2014). However, they do not come with a ground truth, which we had to create. The first collection, `ACL`, was used to identify trends of text reuse in Gupta & Rosso (2012). In order to build this collection, the authors collected long, short, and workshop papers published at the ACL conference in 1990 to 1997 and 2004 to 2011. In our experiments, 4686 documents between 2004 and 2011 were used as this was the period in which most cases of reuse were found by Gupta & Rosso (2012). The second collection, `PubMed`, was taken from the PubMed articles made available by the Déjà vu project[30]. This database includes about 80K pairs of articles for which the eTBLAST software has found high similarities. From those, we picked the 797 pairs for which the full-text was available. There were 1513 documents in total. The number of documents does not add up exactly to twice

[30] `http://dejavu.vbi.vt.edu/dejavu/`

Table 3: Details of the collections

Collection	# Documents	# Pooled Pairs	# Positive Pairs
ACL	4686	93	41
PubMed	1513	85	64

the number of pairs as some documents were in more than one suspicious pair. Also, it is possible that a document is treated as suspicious in one pair and then as source in another pair. Table 3 summarises the details of these text collections.

Ground Truth Generation. To generate the ground truths, we employed the *pooling method* which is widely used in Information Retrieval (Spark-Jones & Van Rijsbergen, 1975). Thus, for each collection, we computed all similarity metrics between all pairs of documents. Then, the 30 highest ranked pairs for each metric was selected to make up the pool. Pairs not in the pool were considered as not having significant reuse. Removing the duplicates (as the same pair may have been identified by more than one metric), we obtained 96 pairs for `ACL` and 85 for `PubMed`. These pairs were then evaluated by human assessors which were given the same instructions issued by Gipp et al. (2014) which asks the judges to report as `positive` the *cases with similarities, which an examiner in a check for plagiarism would find valuable to be made aware of.* With this, all five levels mentioned in the Introduction would be covered. The group had 10 evaluators, which assessed the pairs in the pool. Each pair was seen by two evaluators and a third one was requested whenever there was a disagreement[31]. The average agreement for the evaluations (not considering the third evaluation which was used for the disagreement cases) was of 84% for `ACL` and 80% for `Pubmed`, and the Fleiss Kappa (Fleiss, 1971) was 0.675 (considered *substantial*) and 0.524 (considered *moderate*) for `ACL` and `PubMed`[32], respectively. The Kappa statistic was lower for `Pubmed` as most cases were judged as positive by all judges and that increased the "chance agreement" which is taken into consideration in this metric. One important difference between the experiments reported here and the ones done by Gipp et al. (2014), is that in their study, content similarity approaches were applied only to pairs of documents which shared at least one reference. We believe this makes the test biased towards Citation-based detection since cases in which the offender copied the contents of a source document, but not the citation, would go undetected. Here, since we aim at evaluating the complementarity between both types of methods, they are run independently.

Evaluation Metrics. Since our ground truth was done at document (and not passage) level, we cannot calculate the metrics as done in the PAN evaluation campaigns Potthast et al. (2010). We then calculated precision, recall, and F1 for pairs of documents as follows:

$$precision(S, R) = \frac{R \bigcap S}{R} \tag{1}$$

$$recall(S, R) = \frac{R \bigcap S}{S} \tag{2}$$

$$F1(S, R) = 2 \times \frac{precision \times recall}{precision + recall} \tag{3}$$

[31] Annotations can be obtained from `http://www.inf.ufrgs.br/~slpertile/collections`.

[32] In the PubMed dataset, 54 of the suspicious pairs with full texts had their contents examined by curators. However, we could not use these assessments as they do not clearly state whether the similarities were considered "significant".

where R are detections identified by the similarity metric and S are the detections in the ground truth.

Since each metric outputs a ranking of document pairs, it is necessary to select a cutting point. We opted to cut the rankings at the $|S|^{th}$ position, where $|S|$ is the number of detections in the ground-truth. Essentially, this is *R-precision*, a widely used metric that assesses how many correct detections are found up to the $|S|^{th}$ rank. Cutting the list at rank $|S|$, makes precision and recall equal. We also calculated the Average Precision (AvP), which is the standard metric to evaluate ranked results and is given in Equation 4.

$$AvP = \frac{\sum_{k=1}^{|S|} P((k) \times correct(k))}{|S|} \tag{4}$$

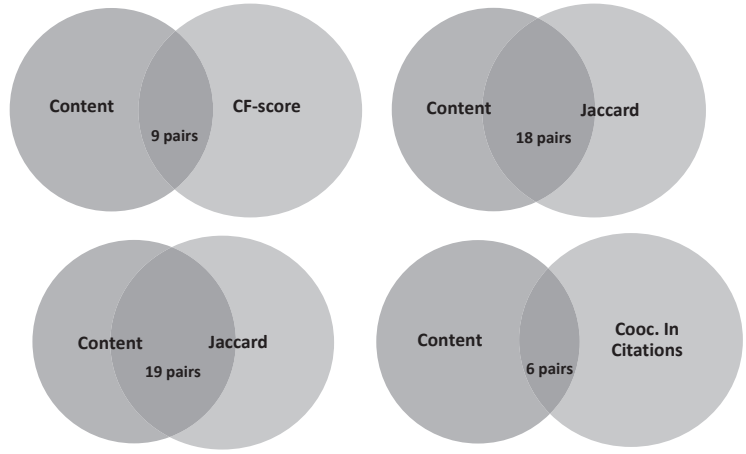
where $|S|$ is the number of detections in the ground truth, $P(k)$ is the precision at rank k, and *correct* is a function that returns 1 if the detection at rank k is correct (*i.e.*, if it is in the ground truth) and 0 otherwise.

Experimental Procedure. Recall that our first experiment aims at analysing the correlation between content and Citation-based detection methods. Thus, the first step in order to allow the calculations of the Citation-based similarity metrics was to represent the documents by their citations/references. Therefore, ParsCit[33] was used to preprocess the articles to extract this information. ParsCit applies a supervised machine learning method based on Conditional Random Fields to parse scientific papers. This tool identifies in-text citations and links them to their corresponding references in the reference list. The fields of the reference are segmented and tagged as author name, title, date, publisher, etc. In addition, the position and the size of the cited fragment are identified. The authors of ParsCit have reported on evaluation results of their approach (Councill et al., 2008). It achieved over 0.9 of F1 on Cora and Citeseer datasets on the identification of author names, paper titles, and publication date. We examined a sample of the output produced by ParsCit and conclude that the same level of accuracy was achieved in our collections. One of the difficulties we faced was that different papers may cite the same reference in different ways. Thus, computing whether two papers cite the same reference cannot be done through exact match. Our analysis focused on the following fields: author names, paper title, and publication year. First, we took the initials of all authors (e.g. if the name is "John Smith" or "J. Smith" we take "JS"). If 50% of the initials of different authors matched, we computed the edit distance on the titles of the articles. References with less than 50% matching initials were not considered matches. If the reference did not contain the year field, the threshold used for title similarity is 80%. Otherwise, the threshold was 75%, and the years are considered as matches if there is at most a one-year difference between them.

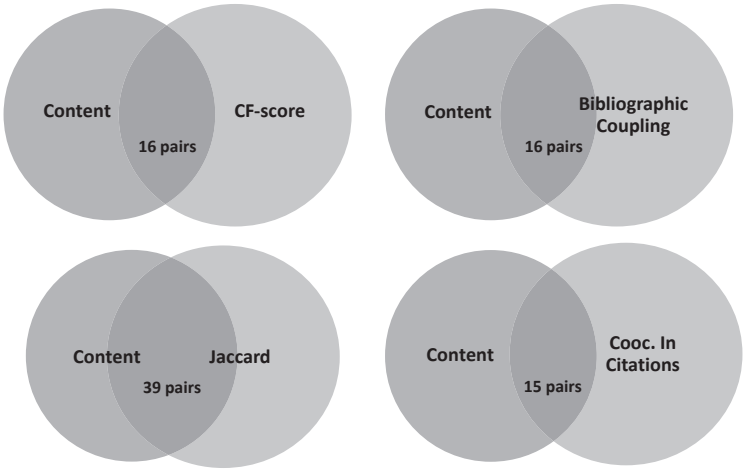
Results

Correlation between Content- and Citation-based Metrics. Once the content and the citation based similarity metrics had been calculated, we analyzed the agreement between them. Since the rankings generated by the metrics for pairwise comparisons were quite long, we limited the analysis to the top-100 pairs. Figure 3 shows the intersection between the Content-based metric and each of the four Citation-based metrics.

[33] `http://aye.comp.nus.edu.sg/parsCit/`



(a) ACL collection



(b) PubMed collection

Figure 3: Intersection between pairs found by Content and Citation-based metrics

Figure 3 shows the results for the overlap between the pairs of documents for the `ACL` and the `PubMed` collections. In the `ACL` collection, the intersection between the content and Citation-based metrics was between 15-46%. As for the `PubMed` collection, the intersection tended to be larger (23 to 61%) for all metrics. The lowest overlap being between content and co-occurrences, for both collections. Note that having a larger intersection does not necessarily mean that the methods are more correct, it just means that they have more detected cases in common.

Besides finding the intersection of the reuse cases identified by the different methods, we also calculated Spearman's rank correlation (ρ) comparing the rankings of generated by each metric. Table 4 shows the correlation coefficients. Significant correlations at α=0.05 are marked with an asterisk. The metrics that have the lowest correlations with content are Bibliographic Coupling and Jaccard, for both collections. Correlations were stronger on the `PubMed` collection.

Accuracy of Content- and Citation-based Metrics. To assess the accuracy of the similarity

Table 4: Correlation between Content and Citation-based metrics

(a) ACL collection

		Citation-based				Content-based
		Bibliographic Coupling	*CF-score*	*Jaccard*	*Co-occurrences*	
Citation-based	*Bibliographic Coupling*	1.00*				
	CF-score	0.43*	1.00*			
	Jaccard	0.36*	0.16	1.00*		
	Co-occurrences	0.10	0.38*	0.40*	1.00*	
Content-based		0.10	0.35*	0.24*	0.14	1.00*

(b) PubMed collection

		Citation-based				Content-based
		Bibliographic Coupling	*CF-score*	*Jaccard*	*Co-occurrences*	
Citation-based	*Bibliographic Coupling*	1.00*				
	CF-score	0.98*	1.00*			
	Jaccard	0.63*	0.65*	1.00*		
	Co-occurrences	-0.26*	-0.23*	0.10	1.00*	
Content-based		0.33*	0.39*	0.38*	0.04	1.00*

metrics, we computed precision, recall, F1, MAP, and the number of true positives (TP) by comparing the pairs identified by the metrics and the pairs in the ground truth. The results are in Table 5. The numbers show that content analysis yields the best results in all evaluation metrics. The best citation-based metric were Bibliographic Coupling for `ACL` and Jaccard for `PubMed`. Co-occurrences in citations was the worst metric.

Table 5: Evaluation of Content and Citation-based Metrics

Similarity Metric	ACL			PubMed		
	P, R, F	AvP	TP	P, R, F	AvP	TP
Bibliographic Coupling	0.61	0.57	25	0.44	0.41	28
CF-score	0.32	0.24	13	0.44	0.42	28
Jaccard	0.51	0.43	21	0.61	0.61	39
Co-occurrence in Citations	0.20	0.09	8	0.36	0.28	23
Content	**0.76**	**0.78**	**31**	**0.67**	**0.76**	**43**

Combining Content and Citation-Based Metrics. The fact that both groups of techniques have identified similarities in different pairs of papers could indicate that they are complementary to each other. Thus, a combination of these approaches could potentially improve the accuracy of plagiarism detection systems. To assess whether combining Content and Citation-Based Metrics can benefit plagiarism detection, we combined both types of metrics in two-ways. The first way consists is simply unioning the top-$|S|$ ranked pairs retrieved by the content and reference-based measures. The second combination was more sofisticated as we made use of machine learning algorithms. For that, each pair of papers was represented as a feature vector composed by the scores assigned by each metric to the pair. In order to train a model to classify the pairs of papers, we needed some positive and negative instances – those were taken from the ground truths. Weka (Hall et al., 2009)

Table 6: Combining Similarity Metrics

	Similarty Metric	ACL				PubMed			
		Recall	Precision	F1	TP	Recall	Precision	F1	TP
Simple	Content Only	0.76	**0.76**	**0.76**	31	0.67	**0.67**	**0.67**	43
	Bibliographic Coupling + Content	0.95	0.56	0.70	39	0.88	0.50	0.64	56
	CF-score + Content	0.85	0.44	0.58	35	0.88	0.50	0.64	56
	Jaccard + Content	0.85	0.51	0.64	35	0.75	0.54	0.63	48
	Co-occurrence in Citations + Content	0.80	0.40	0.54	33	0.83	0.47	0.60	53
	All Citation/Reference Metrics	0.76	0.26	0.38	31	0.92	0.37	0.53	59
	Combination of all Metrics	**1.00**	0.29	0.45	**41**	**1.00**	0.36	0.53	**64**
Machine Learning	Content Only	0.94	0.95	0.94	36	0.92	0.91	0.91	56
	Bibliographic Coupling + Content	0.94	0.95	0.94	36	0.91	0.91	0.91	58
	CF-score + Content	0.94	0.95	0.94	36	**0.95**	**0.95**	**0.95**	**62**
	Jaccard + Content	0.94	0.95	0.94	36	0.90	0.90	0.90	59
	Co-occurrence in Citations + Content	**0.95**	**0.96**	**0.95**	**37**	0.91	0.91	0.91	56
	All Citation/Reference Metrics	0.73	0.76	0.72	23	0.88	0.88	0.88	54
	Combination of all Metrics	**0.95**	**0.96**	**0.95**	**37**	0.93	0.93	0.93	60

was used to generate the learning models. We tested 12 algorithms available on the tool, namely, AdaBoost, Bagging, BayesNet, ConjunctiveRule, DTNB, J48, LogitBoost, Multilayer Perceptron, Naïve Bayes, RandomTree, RBFNetwork, and SMO. The rationale was to choose algorithms that use different learning strategies. The default parameters were used in all runs and 10-fold cross-validation was employed. Best results were obtained by DTNB (a hybrid decision-table/Naïve Bayes classifier) for `ACL` and J48 (a decision-tree classifier) for `Pubmed`.

Table 6 shows the results for the simple combination approach and using machine learning. The results correspond to the highest F1 achieved by the tested algorithms. The first row corresponds to the results obtained when just the score of content similarity was used. The simple combination brings improvements in recall (as more positive pairs are found), but precision decreases as the number of retrieved pairs also grows. The gain in recall does not compensate the loss in precision, reflecting in lower F1 values.

The use of machine learning produced slight improvements for both collections. The classifiers were able to filter out some of the false positives, leading to a gain in precision. At the same time, the combination provides more sources of evidence bringing a higher recall. For `PubMed`, 4 out of the 12 algorithms tested yielded gains in F1 while, for `ACL`, only 2 benefited from a combination. In many cases, the results remained the same, and in some cases, the results even degraded as the number of false positives got higher.

When comparing citation-based metrics that use an absolute number such as (Bibliographic Coupling and CF-Score) versus metrics that use proportions (Jaccard and Co-occurrences in Citations), we argue that both have advantages and drawbacks. The drawback of using absolute numbers is that two papers may have a high number of references in common (and thus a high Bibliographic Coupling) but if they have a large number of references, this may not indicate malicious reuse. This will be reflected as a low Jaccard score. On the other hand, if two papers have one co-occurrence each, and it is shared between them, then they will have a score 1 for this metric. However, if they do not share many references, they will score low in Bibliographic Coupling, which can help rule out the pair from the candidates. Thus, using a variety of metrics seems to be a good solution as one can compensate for the shortcomings of another.

Finding the source for a single suspicious document. In the previous evaluations, we

computed pairwise similarities using all metrics for all suspicious documents and used the scores to create a single ranking. Now we explore a different scenario: we assess the ranking that each metric produces for a single document. Thus, each suspicious document is compared against the entire collection and the rankings produced are evaluated in terms of Mean Average Precision (MAP). The MAP score for a set of detections is the mean of the AvP scores (Eq. 4) for each detection. It is calculated as:

$$MAP = \frac{\sum_{k=1}^{|S|} AvP(k)}{|S|} \tag{5}$$

The optimal MAP value is 1 and it means that the metric has ranked all papers with significant reuse (according to the ground truth) higher than any paper for which no significant reuse was found. The results in Table 7 show that content analysis generated the best rankings; being almost perfect for both `ACL` and `PubMed`. Citation-based metrics performed worse but still, for most suspicious papers, they assigned top rank to the corresponding source. This suggests that all metrics could potentially be used as indications of plagiarism, provided that the source documents are available for comparison.

Table 7: MAPs for rankings produced for a single document

Corpus	Content	CF-score	Bibliographic Coupling	Jaccard	Co-occurrence in Citations
ACL	1.00	0.80	0.81	0.81	0.62
PubMed	0.96	0.90	0.90	0.90	0.43

Limitations. This study focused on the types of plagiarism defined by ACM and IEEE, which considers verbatim copies or large portions of paraphrased text (entire pages or paragraphs). Thus, smaller paraphrased passages could still go unnoticed by the methods tested here. In our experiments we used the number of pairs in the ground truth as a cutting point. In practice, one would not know that number. Thus, determining a threshold to decide which pairs to keep is an important issue.

Conclusion

This paper presented a survey on plagiarism detection methods and tools. We discussed approaches based on content, on content and structure, and on references/citations. Our focus was on contrasting and combining Content and Citation-based approaches.

To evaluate whether these Content and Citation-based methods are complementary, we carried out experiments on real collections of scientific papers (ACL and PubMed). Ground truths using human judgements were produced for both corpora.

We observed that at most half of the pairs identified by the Content-based metrics were also identified by the Citation-based metrics (and vice-versa). The correlation between these metrics was weaker on the ACL collection than on PubMed. Using the ground truths, we evaluated the quality of the metrics and observed that Content-based detection yields superior results. Then, we assessed whether a combination of metrics could bring improved results. When metrics were combined through machine learning algorithms, we obtained slight gains compared to using the metrics on their own. In some cases, the machine learning algorithms were able to reduce false positives and false negatives, leading to higher precision and recall.

Despite the superior results achieved by content similarity, we believe citation analysis is promising especially in cases of cross-language plagiarism. In such a scenario, n-gram matching would not likely identify potential sources. Also, in some cases, the full text of the paper may not be available, but the references are, so metrics like Bibliographic Coupling, Jaccard, and CF-score could still be used. The downside is that they may end up identifying false positives as we found in our experiments that authors tend to reuse the same references even when the body of the works differ significantly.

Another important finding is that most cases in which significant reuse was found, the publications shared at least one author. Thus they are candidates to being considered self-plagiarism, which, as discussed in the Introduction, is quite controversial. This leads us to another important issue – what can we reasonably expect from a plagiarism detection system? Systems can only be expected to identify text reuse. The judgement as to whether the reuse consists in plagiarism cannot be done automatically. It is highly unlikely that an automatic system would rule out highly similar papers as it is the case for errata or extended versions of conference papers published as journals.

Acknowledgments This work has been funded by CNPq project 478979/2012-6. S. L. Pertile receives a grant from CAPES. We thank Parth Gupta for sharing his results with us. We are grateful to the anonymous reviewers who have made several suggestions to improve this manuscript. Finally, we thank the voluntary annotators, for identifying the significant reuse cases.

References

Alzahrani, S., Palade, V., Salim, N., & Abraham, A. (2012). Using Structural Information and Citation Evidence to Detect Significant Plagiarism Cases in Scientific Publications. *JASIST*, 286-312.

Anderson, M. S., & Steneck, N. H. (2011). The Problem of Plagiarism. *Urologic Oncology: Seminars and Original Investigations*, *29*(1), 90–94.

Ashworth, P., Bannister, P., Thorne, P., & on the Qualitative Research Methods Course Unit, S. (1997). Guilty in whose eyes? University Students's Perceptions of Cheating and Plagiarism in Academic Work and Assessment. *Studies in Higher Education*, *22*(2), 187-203.

Barrón-Cedeño, A., Vila, M., Martí, M. A., & Rosso, P. (2013). Plagiarism meets paraphrasing: Insights for the next generation in automatic plagiarism detection. *Computational Linguistics*, *39*(4), 917–947.

Barrón-Cedeño, A., & Rosso, P. (2009). On Automatic Plagiarism Detection Based on n-Grams Comparison. In *ECIR* (pp. 696–700).

Beall, J. (2013). *Five Ways to Defeat Automated Plagiarism Detection.* Retrieved from `http://scholarlyoa.com/2013/02/07/five-ways-to-defeat-automated-plagiarism-detection/` (accessed 10-Feb-2015)

Burget, R. (2007). Automatic Document Structure Detection for Data Integration. In *Proceedings of the 10th International Conference on Business Information Systems* (pp. 391–397).

Chow, T., & Rahman, M. (2009). Multilayer SOM with Tree-Structured Data for Efficient Document Retrieval and Plagiarism Detection. *Neural Networks, IEEE Transactions on*, *20*(9), 1385–1402.

Collberg, C., & Kobourov, S. (2005). Self-plagiarism in Computer Science. *Commun. ACM*, *48*(4), 88–94.

Councill, I. G., Giles, C. L., & yen Kan, M. (2008). ParsCit: An open-source CRF reference string parsing package. In *International Language Resources and Evaluation.* European Language Resources Association.

Cronin, B. (2013). Self-plagiarism: An odious oxymoron. *Journal of the American Society for Information Science and Technology*, *64*(5), 873–873.

El Bachir Menai, M., & Bagais, M. (2011). APlag: A Plagiarism Checker for Arabic Texts. In *Computer Science Education (iccse), 2011 6th International Conference on* (pp. 1379–1383).

Fang, F. C., Steen, R. G., & Casadevall, A. (2012). Misconduct accounts for the majority of retracted scientific publications. *Proceedings of the National Academy of Sciences*, *109*(42), 17028–17033.

Fleiss, J. (1971). Measuring nominal scale agreement among many raters. *Psychological Bulletin*, *76*(5), 378–382.

García-Romero, A., & Estrada-Lorenzo, J. (2014). A Bibliometric Analysis of Plagiarism and Self-plagiarism through Déjà vu. *Scientometrics*, *101*(1), 381-396.

Garner, H. (2014). The Case of the Stolen Words. *Scientific American*, *310*(3), 64– 67.

Gipp, B., & Beel, J. (2010). Citation based Plagiarism Detection: a New Approach to Identify Plagiarized Work Language Independently. In *Proceedings of the 21st ACM Conference on Hypertext and Hypermedia* (pp. 273–274).

Gipp, B., & Meuschke, N. (2011). Citation Pattern Matching Algorithms for Citation-based Plagiarism Detection: Greedy Citation Tiling, Citation Chunking and Longest Common Citation Sequence. In *Proceedings of the 11th ACM Symposium on Document Engineering (DocEng2011).*

Gipp, B., Meuschke, N., & Breitinger, C. (2014). Citation-based Plagiarism Detection: Practicability on a Large-Scale Scientific Corpus. *Journal of the Association for Information Science and Technology*, *65*(8), 1527–1540.

Gipp, B., Meuschke, N., Breitinger, C., Lipinski, M., & Nürnberger, A. (2013). Demonstration of Citation Pattern Analysis for Plagiarism Detection. In *Proceedings of the 36th International ACM SIGIR Conference on Research and Development in Information Retrieval* (pp. 1119–1120).

Grman, J., & Ravas, R. (2011). Improved Implementation for Finding Text Similarities in Large Sets of Data - Notebook for PAN at CLEF 2011. In *CLEF (Notebook Papers/Labs/Workshop).*

Grozea, C., Gehl, C., & Popescu, M. (2009). Encoplot: Pairwise Sequence Matching in Linear Time Applied to Plagiarism Detection. In *CLEF (Notebook Papers/Labs/Workshop).*

Grozea, C., & Popescu, M. (2011). The Encoplot Similarity Measure for Automatic Detection of Plagiarism - Notebook for PAN at CLEF 2011. In *CLEF (Notebook Papers/Labs/Workshop).*

Gullifer, J., & Tyson, G. A. (2010). Exploring university students perceptions of plagiarism: a focus group study. *Studies in Higher Education*, *35*(4), 463–481.

Gupta, P., & Rosso, P. (2012). Text Reuse with ACL: (upward) trends. In *Proceedings of the ACL-2012 Special Workshop on Rediscovering 50 Years of Discoveries* (pp. 76–82).

HaCohen-Kerner, Y., Tayeb, A., & Ben-Dror, N. (2010). Detection of Simple Plagiarism in Computer Science Papers. In *Proceedings of the 23rd International Conference on Computational Linguistics* (pp. 421–429).

Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., & Witten, I. H. (2009). The Weka Data Mining Software: an update. *SIGKDD Explor. Newsl.*, *11*(1).

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

Juola, P. (2006). Authorship attribution. *Foundations and Trends in Information Retrieval*, *1*(3), 233–334.

Kasprzak, J., & Brandejs, M. (2010). Improving the Reliability of the Plagiarism Detection System - Lab Report for PAN at CLEF 2010. In *Clef (notebook papers/labs/workshop)* (pp. 1–10).

Kernighan, B. W., & Ritchie, D. M. (1988). *The C Programming Language*. Prentice-Hall.

Kong, L., Qi, H., Du, C., Wang, M., & Han, Z. (2013). Approaches for Source Retrieval and Text Alignment of Plagiarism Detection - Notebook for PAN at CLEF 2013. In *CLEF (Notebook Papers/Labs/Workshop).*

Kong, L., Qi, H., Wang, S., Du, C., Wang, S., & Han, Y. (2012). Approaches for Candidate Document Retrieval and Detailed Comparison of Plagiarism Detection. In *CLEF (online working notes/labs/workshop).*

Malcolm, J. A., & Lane, P. C. R. (2009). Tackling the PAN'09 External Plagiarism Detection Corpus with a Desktop Plagiarism Detector. In *CEUR-WS.org* (Vol. 502, pp. 29–33).

Maurer, H., Kappe, F., & Zaka, B. (2006). Plagiarism - A Survey. *J-JUCS*, *12*(8), 1050–1084.

Mccabe, D. L. (2005). Cheating Among College and University Students : A North American Perspective. *International Journal for Educational Integrity*, *1*(1996).

Meuschke, N., Gipp, B., & Breitinger, C. (2012). CitePlag: A Citation-based Plagiarism Detection System Prototype. In *Proceedings of the 5th International Plagiarism Conference.* Newcastle upon Tyne, UK.

Park, C. (2003). In other (people's) words: plagiarism by university students–literature and lessons. *Assessment & Evaluation in Higher Education*, *28*(5), 471–488.

Pertile, S., Rosso, P., & Moreira, V. P. (2013). Counting Co-occurrences in Citations to Identify Plagiarised Text Fragments. In *CLEF (Notebook Papers/Labs/Workshop)* (pp. 150–154).

Potthast, M., Hagen, M., Stein, B., Grassegger, J., Michel, M., Tippmann, M., & Welsch, C. (2012). ChatNoir: A Search Engine for the ClueWeb09 Corpus. In *Proceedings of the 35th International ACM SIGIR Conference on Research and Development in Information Retrieval* (pp. 1004–1004).

Potthast, M., Stein, B., Barrón-Cedeño, A., & Rosso, P. (2010). An Evaluation Framework for Plagiarism Detection. In *Proceedings of the 23rd International Conference on Computational Linguistics: Posters* (pp. 997–1005).

Sanchez-Perez, M. A., Sidorov, G., & Gelbukh, A. F. (2014). A Winning Approach to Text Alignment for Text Reuse Detection at PAN 2014. In *Working Notes for CLEF 2014 Conference, Sheffield, UK, 2014.* (pp. 1004–1011).

Sorokina, D., Gehrke, J., Warner, S., & Ginsparg, P. (2006). Plagiarism detection in arxiv. In *Proceedings of the Sixth International Conference on Data Mining* (pp. 1070–1075). IEEE Computer Society.

Spark-Jones, K., & Van Rijsbergen, C. J. (1975). *Report on the need for and provision of an ideal information retrieval test collection.* Computer Laboratory. University of Cambridge.

Stamatatos, E. (2011). Plagiarism Detection using Stopword n-Grams. *J. Am. Soc. Inf. Sci. Technol.*, *62*(12), 2512–2527.

Stein, B., & Meyer, S. (2006). Near Similarity Search and Plagiarism Analysis. In *Gfkl* (pp. 430–437).

Stein, B., & zu Eissen, S. M. (2006). Intrinsic Plagiarism Detection. In *ECIR* (p. 565-569).

Stein, B., zu Eissen, S. M., & Potthast, M. (2007). Strategies for Retrieving Plagiarized Documents. In *Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (pp. 825–826). New York, NY, USA: ACM.

Stoffel, A., Spretke, D., Kinnemann, H., & Keim, D. A. (2010). Enhancing Document Sstructure Analysis using Visual Analytics. In *Proceedings of the 2010 ACM Symposium on Applied Computing* (pp. 8–12).

Torrejón, D. A. R., & Ramos, J. M. M. (2010). CoReMo System (Contextual Reference Monotony) - Lab Report for PAN at CLEF 2010.

Torrejón, D. A. R., & Ramos, J. M. M. (2013). Text Alignment Module in CoReMo 2.1 Plagiarism Detector - Notebook for PAN at CLEF 2013. In *CLEF (Notebook Papers/Labs/Workshop).*

Vallés Balaguer, E. (2009). Putting Ourselves in SME's shoes: Automatic Detection of Plagiarism by the Wcopyfind tool. In *Proceedings of the SEPLN'09 Workshop on Uncovering Plagiarism, Authorship and Social Software Misuse* (pp. 34–35).

Walker, J. (2010). Measuring Plagiarism: Researching What Students Do, Not What They Say They Do. *Studies in Higher Education*, *35*(1), 41–59.

Youmans, R. J. (2011). Does the Adoption of Plagiarism-detection Software in Higher Education Reduce Plagiarism?. *Studies in Higher Education*, *36*(7), 749–761.

Zhang, H., & Chow, T. W. (2011). A Coarse-to-fine Framework to Eefficiently Thwart Plagiarism. *Pattern Recognition*, *44*(2), 471–487.

Zhang, Y. (2010). CrossCheck: an Effective Tool for Detecting Plagiarism. *Learned Publishing*, *23*, 9–14.