Towards the 2nd International Competition on Plagiarism Detection and Beyond*

Alberto Barrón-Cedeño and Paolo Rosso

Natural Language Engineering Lab, ELiRF
Universidad Politécnica de Valencia, Spain
{lbarron, prosso}@dsic.upv.es

Abstract

Multiple tools that assist in the detection of text plagiarism (known as *automatic plagiarism detectors*) have been developed in recent years. However, there is still an unanswered question: Which plagiarism detection method performs best? Due to the lack of a standard evaluation framework for plagiarism detection, providing an answer to this question is still not possible nowadays.

As a response, the International Competition on Plagiarism Detection has been created (Potthast et al., 2009). In this paper an overview of the results of the first competition is provided. A description of how the second edition is being developed and its perspectives are also included. By having a competition on plagiarism detection, where different models can be objectively compared, the response to the previous question is getting closer. Additionally, an overview of what we believe to be missing from automatic plagiarism detection —cross-language plagiarism— is provided.

1 Introduction

Plagiarism can be defined in multiple ways. One of the most interesting definitions is the one of the IEEE (2008):

> *plagiarism is the reuse of someone else's prior ideas, processes, results, or words without explicitly acknowledging the original author and source.*

Whereas plagiarism may occur incidentally, due to phenomena such as cryptomnesia (Taylor, 1965), it is often the outcome of a conscious process. The definition of IEEE establishes that, independently from the vocabulary or the channel an idea is communicated with, a person that fails to provide its source is guilty of plagiarism. In recent years, due to the large amount of text available

*We would like to thank Patrick Drouin for reviewing the draft version of this paper and the conference reviewers for their valuable comments. This work was partially funded by the CONACYT-Mexico 192021 grant and the MICINN research project TEXT-ENTERPRISE 2.0 TIN2009-13391-C04-03 (Plan I+D+i).

in electronic media, plagiarism cases have increased. This is a big problem in academic, scientific and commercial circles.

People are often able to detect cases of plagiarism either by detecting text inconsistencies or by resembling previously consulted material. Nevertheless, the large amount of potential source texts available nowadays makes manual plagiarism detection infeasible. In order to assist people to uncover plagiarism, different models, grouped under the general name of *automatic plagiarism detection*, have been developed. The aim of these models is detecting potential cases of plagiarism in a suspicious document d_q and, if possible, providing the alleged source. With the information provided by the tool, a person can take the final decision: whether a text is plagiarised or not.

Standardisation on how automatic plagiarism detection should be evaluated (which implies dealing with standard text collections and evaluation measures to compare different approaches), has not been completely possible due to ethical reasons. As a response, the International Competition on Plagiarism Detection has been created. In this paper, an overview of the first competition (held in 2009), as well as the second edition (conducted nowadays) is provided. Section 2 includes a brief overview of the two main approaches to automatic plagiarism detection: intrinsic and external. Section 3 describes how the 2009 competition was organised. Some obtained results are discussed in Section 4. Section 5 overviews the new insights conducted for the organisation of this year's competition. Section 6 gives an overview of a less explored kind of plagiarism: *translated plagiarism*. Finally, Section 7 gives some final remarks.

2 Plagiarism Detection Overview

Methods for automatic plagiarism detection can be divided in two main approaches:

Intrinsic plagiarism detection. A person can identify potential cases of plagiarism by detecting unexpected irregularities through a document. Disruptive changes of style, vocabulary, or complexity are triggers of suspicion. Therefore, intrinsic models analyse different text features aiming to detect whether d_q contains text fragments written by a different author. The features considered by these models are, among others, readability and vocabulary richness, as well as average length of words and sentences (Meyer zu Eißen and Stein, 2006). Other approaches apply character n-gram[1] profiles to characterise an author's style and search for irregularities in d_q (Stamatatos, 2009).

External plagiarism detection. Researchers have paid more attention to this approach for two main reasons: (*i*) it is closely related to information retrieval; and (*ii*) obtaining the source of a potential case of plagiarism provides better evidence to make a final decision. In this approach d_q and a collection of potential source documents D are given. The task is to

[1] An n-gram is a redundant representation of a text consisting of overlapping chunks (either at character or word level) of length n. For instance, the character level 3-grams of "*example*" are $[exa, xam, amp, mlp, ple]$; and the word level 2-grams of "*this is just an example*" are $[this\ is, is\ just, just\ an, an\ example]$.

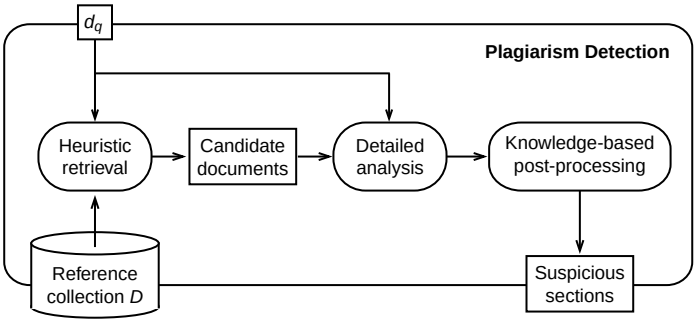

Figure 1: Retrieval process for external plagiarism detection (derived from Stein et al. (2007)).

identify the plagiarised sections in d_q (if there are any), and their respective source sections in D (Potthast et al., 2009). A prototypical process to external plagiarism detection is represented in Fig. 1. In fact, most of the participants to the 2009 edition of the competition (cf. Section 4) followed this process.

In *heuristic retrieval*, those documents $d \in D$ related to d_q are retrieved, composing the collection of candidate source documents. During the *detailed analysis*, text fragments are compared in order to determine whether the source of a text fragment in d_q is in d. Finally, during the *knowledge-based post-processing* those reused fragments in d_q which are properly referenced or quoted are discarded from the plagiarism suspects.[2]

Surveys of the research done on automatic plagiarism detection can be reviewed in (Clough, 2003) and (Maurer et al., 2006).

Commercial and non-commercial tools that assist in the detection of text plagiarism have been developed in recent years. Some systems, such as Turnitin (iParadigms, 2010), claim to be able to accurately detect cases of plagiarism, provided that their source is included in their repository. Other tools, such as DOC Cop (McCrohon, 2010), use standard search engines —for instance, those provided by *Google* or *Yahoo!*— to look up the source of a potentially plagiarised text on the Web.

Thus, the development of plagiarism detection models is not new. In fact, one of the first approaches we have track of goes back to the 1970s (Ottenstein, 1976). However, after more than 30 years, a question remains unanswered:

What automatic plagiarism detection method performs the best?

No standard evaluation framework —standard text collections with documented cases of plagiarism and evaluation measures— exists to estimate (and compare) how well the different methods perform. Consequently, researchers often use private collections of documents which cannot be freely provided to

[2]As described in Section 3.1, in the competition corpora no reused fragment includes quotation or reference to its source. Therefore, up to now, the competition only evaluates the output obtained after the *detailed analysis* stage.

others for ethical reasons.[3] Additionally, they estimate the quality of the models by considering different evaluation measures. As a result, the previous question is hard to answer.

Other information retrieval and natural language processing tasks, such as question answering, named entity recognition, or text categorisation, have benefited from different challenges and competitions, such as those of the CLEF initiative[4]. Hence, with the aim of providing a standard evaluation framework on automatic plagiarism detection, the International Competition on Plagiarism Detection was created.

3 Preparing a Competition on Plagiarism Detection

The First International Competition on Plagiarism Detection was held in 2009 (Potthast et al., 2009). Two non-exclusive tasks were proposed: intrinsic and external plagiarism detection (*cf.* Section 2). In the former case the task was, given a suspicious document d_q, to identify all plagiarised text passages. In the latter case, the task was, given d_q and a set of potential source documents D, to find all plagiarised text passages in d_q and the corresponding source passages from D.

Following, we give an overview of how the competition was created, including a description of the training and test corpora provided and the evaluation measures designed.

3.1 Creating a Corpus of *Synthetic* Plagiarism

Few people would like to be included in a collection of documents subject to plagiarism, even if it is clear that it is simulated. Additionally, text subject to copyright restrictions has to be discarded. By composing the corpus with copyright free text, it is possible to distribute it to participants (and any other interested researcher) without further agreements or permission requests. Therefore, we opted for composing the plagiarism corpus only with this kind of texts. One of the biggest, publicly available, repositories fitting these characteristics is that of Project Gutenberg[5]. Only documents from this repository were used to compose a large scale corpus of simulated plagiarism.

The resulting corpus, named *PAN-PC-09* corpus, comprises 41,223 documents and includes more than 90,000 cases of artificial plagiarism. In order to compose a realistic detection problem, the plagiarised text was automatically obfuscated. Perhaps the most drastic obfuscation was translation, as dependency between source and plagiarised text crosses languages. The main characteristics of the PAN-PC-09 are as follow:

Document Length 50% of the documents included are small (1-10 pages), 35% are medium (10-100 pages), and 15% are large (100-1000 pages).

[3] As an alternative option corpora of authorised text reuse, such as the METER corpus (Clough et al., 2002), have been used to test methods for plagiarism detection.

[4] http://www.clef-campaign.org

[5] http://www.gutenberg.org

Suspicious-to-Source Ratio 50% of the documents compose the suspicious documents collection D_q, and 50% compose the source documents collection D.

Plagiarism Percentage The percentage of plagiarism per suspicious document $d_q \in D_q$ ranges from 0% to 100%. Moreover, 50% of the suspicious documents contain no plagiarism at all.

Plagiarism Length The length of a plagiarism case is evenly distributed between 50 and 5,000 words.

Plagiarism Languages 90% of the cases are monolingual (English). The remainder of the cases are cross-language (translated). Texts written in German or Spanish were translated into English to create these cases. This process aims to resemble the fact that people whose native language is not English very often write in this language.[6]

Plagiarism Obfuscation The monolingual reused fragments created to evaluate the external approach were obfuscated. Obfuscation included, for instance, random shuffling of texts and substitution of synonyms.

3.2 Definition of Evaluation Measures

Automatic plagiarism detection can be considered a task related to Information Retrieval (IR). Thus, Recall and Precision, standard evaluation measures in IR, seem to be the option to evaluate this task. However, these measures are designed to evaluate the retrieval of entire documents. Plagiarism detection is different since its concern is retrieving specific text fragments: either a plagiarised fragment and its source (external approach) or only the plagiarised fragment (intrinsic approach). Therefore, special measures are required to accurately evaluate the output given by a plagiarism detection model.

In order to clarify how difficult evaluating this task is, let us consider the example of Fig. 2. In this case, d_q includes the plagiarised text fragments $s_{1,\ldots,3}$ (S). An analysis carried out by a given detection method considers that the plagiarised text fragments are $r_{1,\ldots,5}$ (R). The output is said to be perfect if $S \cap R = S \cup R = 1$, i.e, all the plagiarised fragments are accurately retrieved excluding any original fragment. This is not the case in the example. Some text fragments are correctly detected while some others are not. Additionally, some original fragments are wrongly detected as plagiarised.

In automatic plagiarism detection, we are interested in evaluating the following three main factors:

1. plagiarised and —if available— source fragments are retrieved;
2. original text fragments are not reported as plagiarised; and
3. plagiarised fragments are not detected over and over again.

Therefore, the output of a plagiarism detection method is evaluated by the following equation:

[6]The other side of the problem, i.e., writers that translate a source text from English into their native language, was not represented in the corpus.

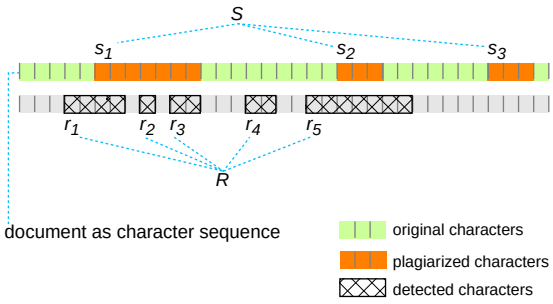

Figure 2: A suspicious document d_q as character sequence. It includes actually plagiarised text fragments (S) and detections provided by a given method (R). Squares consist of many characters.

$$overall(S, R) = \frac{F}{log_2(1 + gran)} \quad , \tag{1}$$

where F, a value in the range of $[0, 1]$, denotes the F-Measure, the harmonic mean of precision and recall, which is defined as:

$$F(S, R) = 2 \cdot \frac{prec(S, R) \cdot rec(S, R)}{prec(S, R) + rec(S, R)} \quad , \tag{2}$$

i.e., precision and recall have exactly the same relevance. Precision represents the fraction of retrieved fragments that are actually plagiarised. It is defined as:

$$prec(S, R) = \frac{1}{|R|} \sum_{r \in R} \frac{|r \sqcap \cup_{s \in S} S|}{|r|} \quad , \tag{3}$$

Recall represents the fraction of actually plagiarised fragments that were properly retrieved. It is defined as:

$$rec(S, R) = \frac{1}{|S|} \sum_{s \in S} \frac{|s \sqcap \cup_{r \in R} r|}{|s|} \quad . \tag{4}$$

These are not the original definitions of the precision and recall measures, commonly used in information retrieval (Manning et al., 2008). Instead they are tailored versions where every contiguous fragment of plagiarised characters is considered to be a basic retrieval unit. In that way, both recall and precision express whether a specific plagiarised fragment has been properly recognised and, if available, its source fragment as well. In both equations $\sqcap$ computes the positionally overlapping characters, $\cup_{x \in X}$ defines the union for every $x \in X$, and $|x|$ represents the cardinality of x.

The other factor in Eq. 1 is $gran$, a novel measure named Granularity. This measure punishes those cases where overlapping plagiarised passages are reported. The granularity of R, for a set of plagiarised sections S, is defined by the average size of the existing covers: a detection $r \in R$ belongs to the cover C_s of an $s \in S$ iff s and r overlap. Let $S_R \subseteq S$ denote the set of cases so that for each $s \in S : |Cs| > 0$. The granularity of R given S is defined as:

$$gran(S,R) = \frac{1}{|S_R|} \sum_{s \in S_R} |C_s|, \tag{5}$$

where $S_R = \{s \mid s \in S \wedge \exists r \in R : s \cap r \neq \emptyset\}$[7] and $C_s = \{r \mid r \in R \wedge s \cap r \neq \emptyset\}$. The domain of the granularity is $[1, |R|]$, where 1 marks the desirable one-to-one correspondence between R and S, and $|R|$ marks the worst case, where a single $s \in S$ is detected over an over again.

4 First International Competition on Plagiarism Detection

This competition was held in 2009. At the beginning (March, 2009), participants were provided with the developing section of the corpus. Such partition included annotations of the plagiarised text fragments and, for the case of external detection, their source. The annotations included the initial and final character of every significant (plagiarised or source) text fragment. The test corpus, provided in May 2009, did not include any annotation referring to plagiarised text. Participants had to submit their detections to the organisers in order to be evaluated.

In order to participate, teams were claimed to accomplish the following rules:

Eligibility The contest was open to any party planning to attend the PAN competition. Multiple submissions per group were allowed for each task. No feedback on the performance at the time of submission was provided: only the last submission before the deadline was evaluated and all other submissions were discarded.

Integrity The exploitation of potential flaws in the competition corpus to gain advantages was prohibited.

Text resources No other text than the one provided in the corpus could be used.

Winner Selection One winner of the "External Plagiarism Detection" task, one winner of the "Intrinsic Plagiarism Detection" task, and one overall winner were proclaimed. The winners were determined according to the following method. All participants were ranked according to their overall performance on the competition corpus for each task (cf. Section 3). Winner of a task was the participant who had the highest score on the respective part of the corpus. Winner of the competition was the participant who had the highest score on the whole competition corpus.

Award The winner of the whole competition was awarded a prise, sponsored by Yahoo! Research.

Thirteen worldwide research teams took part of the competition. The number of participants showed that the interest in automatic plagiarism detection has raised. Researchers with different backgrounds took part, such as experts

[7] This can be read as the set of elements s, given that s in S and an r in R exist, such that the intersection between s and r is not empty.

External Approach		
Heuristic Retrieval	**Detailed Analysis**	**Participants**
Character 16-grams weighted by frequency. Similarity based on the cosine measure.	Exact matches of d_q and $d \in D$ based on character 16-grams.	Grozea et al. (2009)
Word 5-grams with Boolean weights. Similarity based on the Jaccard coefficient.	Exact matches of d_q and $d \in D$ based on word 5-grams.	Kasprzak et al. (2009)
Word 8-grams weighted by frequency. Similarity based on a custom distance.	Exact matches of d_q and $d \in D$ based on word 8-grams.	Basile et al. (2009)
Intrinsic Approach		
Method Description		**Participant**
Texts characterised by character 3-grams frequencies. A window of 1000 characters is slided considering a step of 200 characters. Variations between 3-grams distributions are measured, looking for unexpected variations.		Stamatatos (2009)

Table 1: Summary of the best approaches applied in the 2009 competition (from Potthast et al. (2009)). The top three positions of the external approach and the winner of the intrinsic approach are included.

on Network Security or Mathematics. Most of the teams approached only the external detection task. Additionally, no team tried to detect the cross-language cases. In fact, cross-language plagiarism detection has drawn attention just recently (cf. Section 6). Table 1 gives a brief overview of the top methods applied in the competition.

In the external approach the most of the participants followed the schema depicted in Fig. 1 (except for the knowledge-based post-processing stage). For heuristic retrieval they opted for an n-gram based comparison (either of characters or words) on the basis of standard text similarity measures (cosine measure, Jaccard coefficient, etc.). For the detailed analysis stage, participants searched for exact matches of sequences. Greedy tiling, based on the matching of transposed sub-strings, was used as well. Just a few participants opted for fingerprinting approaches, where texts are subsampled and, frequently, text strings are mapped to short bit strings in order to improve the comparison speed (however, flexibility to editions is decreased).

Grozea et al. (2009) won this subtask. They approached the heuristic retrieval step by comparing character 16-grams. The detailed analysis consisted of a heuristic that isolated the clusters of related text. Both processes are made in linear time.

Kasprzak et al. (2009) got the second place by comparing fingerprints of word 5-grams over an inverted index. Basile et al. (2009) used a combination of word and character n-grams for the first and second steps, respectively. They opted for codifying the texts in terms of word lengths and a *T9* encoding schema. Comparing numbers instead of words allowed them to obtain the third place.

Curiously, the only difference among the approaches followed by the top three systems was computational. However, the philosophy was the same: codifying the documents into an n-gram representation and comparing them looking for common text fragments.

Less teams approached the intrinsic detection task. Their methods were based on analysing the variation of different factors such as complexity or style through the documents. Stamatatos (2009) won this subtask by quantifying

such variations on the basis of character n-grams. Given d_q, a profile of its n-grams is obtained in terms of their relative frequency. Once the document's profile is established, a sliding window is passed through the document, calculating local profiles. If the distance between local and global profile is too high, a potential case of plagiarism has been detected.

Note that both, external and intrinsic approaches, could be combined in order to improve the final results and even speed up the global process.

One of the biggest difficulties the participants argued for was the size of the suspicious and source partitions in the corpus. Exact copies can be "easily" and efficiently detected by using fingerprinting models, such as Winnowing (Schleimer et al., 2003) and COPS (Brin et al., 1995). Nevertheless, when plagiarism implies further modifications, an exhaustive comparison is necessary, thus making the process hard to accomplish in reasonable time.

The results of the competition as well as the proceedings, where the different methods applied by the participants are described, are available at `http://pan.webis.de`.

5 Towards the Second International Competition on Plagiarism Detection

After the encouraging experience of the first competition, the second edition is being held. It is one of the CLEF 2010 Labs[8], that includes some other labs, challenging different tasks, such as intellectual property in patents and multilingual question answering.[9]

The same tasks are being proposed: intrinsic and external plagiarism detection. This year the training and test corpora are larger, making the problem more challenging. In fact, the development corpus is the entire PAN-PC-09 corpus. Additionally, some drawbacks of the last year's competition corpus, such as reused text fragments in the base corpus that were unaware text reuse, are being avoided.[10]

This year, such cases of "unaware" text reuse are minimised as much as possible by discarding any anthology (that contains all the documents written by a given author) and by eliminating duplicated text fragments. This is made on the basis of fingerprinting techniques. In fact, such fingerprinting models (as those proposed by Schleimer et al. (2003) and Bernstein and Zobel (2004)), were formerly designed for automatic detection of plagiarism and co-derivation. Moreover, in order to increase the quality of the corpus, the base documents have been manually reviewed.

Another issue is the obfuscation process. More obfuscation methods are being considered. For instance, we consider the fact that the longer a plagiarised fragment is, the less modifications it will contain. This fact makes the corpus more realistic as plagiarisers use to act in that way. Once again, cross-language

[8] Refer to http://pan.webis.de/ for all the information related to this year's competition, including the new (independent) task proposed: Wikipedia vandalism detection.

[9] http://www.clef2010.org/

[10] For instance, some of the considered books from Project Gutenberg were anthologies and every single publication by the same author was used in the corpus as well. Other cases included redundant text, frequently related to religion or to Project Gutenberg annotations.

plagiarised texts are being produced by machine translation. An extra is that some handmade cases of plagiarism are being included in the test partition.

Up to now, more than twenty research teams are registered to this year's competition, including countries as Brazil, Canada, China, Germany, India, Iran, Netherlands, Singapore, Spain, Sweden, UK, and Ukraine.

6 And Beyond: Cross-Language Detection

As previously mentioned, during the first edition of the competition no team tried to detect the cross-language plagiarism cases. In fact, this is a lack in automatic plagiarism detection nowadays. Whereas some commercial tools are able to perform plagiarism analyses on different languages, to the best of our knowledge, detecting cases of translated plagiarism is not possible.

Just a few approaches already exist so far. Some of them, as the one proposed by Pouliquen et al. (2003), are aimed to detect document translations. This method is based on the exploitation of a multilingual thesaurus.

CL-ESA (cross-language explicit semantic analysis) is another interesting method for cross-language IR that has been applied to cross-language plagiarism detection (Potthast et al., 2010). CL-ESA intends to estimate, at semantic level, how similar two texts written in different languages are. This estimation is carried out on the basis of comparable corpora, such as Wikipedia.

One of the first models explicitly designed for estimation of cross-language similarity and plagiarism detection is the cross-language alignment-based similarity analysis (CL-ASA) (Barrón-Cedeño et al., 2008; Pinto et al., 2009). This approach is based on statistical machine translation techniques. It estimates the likelihood of two texts of being valid translations of each other.

7 Final remarks

In this paper we discussed the results obtained in the framework of the International Competition on Plagiarism Detection. The competition's main aim is to provide a framework for benchmarking and evaluation of automatic plagiarism detection methods in order to strength and spread the research work on this area.

After the success of the first competition, the second one is now being held as one of the 2010 CLEF labs. Research and commercial teams are participating to the challenge on intrinsic and external plagiarism detection right now. As different models for cross-language plagiarism detection have been proposed recently, we expect an increasing interest in detecting such cases.

We consider that, thanks to the competition, the answer to the question of which method performs best is closer.

References

Alberto Barrón-Cedeño, Paolo Rosso, David Pinto, and Alfons Juan. On Cross-lingual Plagiarism Analysis Using a Statistical Model. In Benno Stein, Efstathios Stamatatos, and Moshe Koppel, editors, *ECAI 2008 Workshop on*

Uncovering Plagiarism, Authorship, and Social Software Misuse (PAN 2008), pages 9–13, Patras, Greece, 2008. CEUR-WS.org.

Chiara Basile, Dario Benedetto, Giampaolo Caglioti, and Mirko Degli Esposti. A Plagiarism Detection Procedure in Three Steps: Selection, Matches and Squares. In Stein et al. (2009), pages 19–23. URL `http://ceur-ws.org/Vol-502`.

Yaniv Bernstein and Justin Zobel. A Scalable System for Identifying Co-Derivative Documents. In *Proceedings of the Symposium on String Processing and Information Retrieval*, pages 55–67. Springer, 2004.

Sergey Brin, James Davis, and Hector Garcia-Molina. Copy Detection Mechanisms for Digital Documents. In Michael J. Carey and Donovan A. Schneier, editors, *Proceedings of the 1995 ACM SIGMOD International Conference on Management of Data*, pages 398–409. ACM Press, 1995.

Paul Clough. Old and new challenges in automatic plagiarism detection. National UK Plagiarism Advisory Service, 2003. URL `http://ir.shef.ac.uk/cloughie/papers/pasplagiarism.pdf`.

Paul Clough, Robert Gaizauskas, and Scott Piao. Building and Annotating a Corpus for the Study of Journalistic Text Reuse. In *Proceedings of the 3rd International Conference on Language Resources and Evaluation (LREC 2002)*, volume V, pages 1678–1691, Las Palmas, Spain, 2002.

Cristian Grozea, Christian Gehl, and Marius Popescu. ENCOPLOT: Pairwise Sequence Matching in Linear Time Applied to Plagiarism Detection. In Stein et al. (2009), pages 10–18. URL `http://ceur-ws.org/Vol-502`.

IEEE. A plagiarism FAQ. http://www.ieee.org/web/publications/rights/plagiarism_FAQ.htm, 2008. [Online; accessed 3-March-2010].

iParadigms. Turnitin, 2010. URL `http://www.turnitin.com`. [Online; accessed 3-March-2010].

Jan Kasprzak, Michal Brandejs, and Miroslav Kriřpač. Finding Plagiarism by Evaluating Document Similarities. In Stein et al. (2009), pages 24–28. URL `http://ceur-ws.org/Vol-502`.

Christopher D. Manning, Prabhakar Raghavan, and Hinrich Schtze. *Introduction to Information Retrieval*. Cambridge University Press, New York, NY, USA, 2008. ISBN 0521865719, 9780521865715.

Hermann Maurer, Frank Kappe, and Bilal Zaka. Plagiarism - A Survey. *Journal of Universal Computer Science*, 12(8):1050–1084, 2006.

Mark McCrohon. DOC Cop, 2010. URL `http://doccop.com`. [Online; accessed 10-March-2010].

Sven Meyer zu Eißen and Benno Stein. Intrinsic plagiarism detection. *Advances in Information Retrieval: Proceedings of the 28th European Conference on IR Research (ECIR 2006)*, LNCS (3936):565–569, 2006.

KJ Ottenstein. An Algorithmic Approach to the Detection and Prevention of Plagiarism. *ACM SIGCSE Bulletin*, 8(4):30–41, 1976.

David Pinto, Jorge Civera, Alberto Barrón-Cedeño, Alfons Juan, and Paolo Rosso. A Statistical Approach to Crosslingual Natural Language Tasks. *Journal of Algorithms*, 64(1):51–60, 2009. ISSN 0196-6774. doi: DOI: 10.1016/j.jalgor.2009.02.005.

Martin Potthast, Benno Stein, Andreas Eiselt, Alberto Barrón-Cedeño, and Paolo Rosso. Overview of the 1st International Competition on Plagiarism Detection. In Stein et al. (2009), pages 1–9. URL `http://ceur-ws.org/Vol-502`.

Martin Potthast, Alberto Barrón-Cedeño, Benno Stein, and Paolo Rosso. Cross-Language Plagiarism Detection. *Language Resources and Evaluation, Special Issue on Plagiarism and Authorship Analysis*, 2010. doi: 10.1007/s10579-009-9114-z.

Bruno Pouliquen, Ralf Steinberger, and Camelia Ignat. Automatic Identification of Document Translations in Large Multilingual Document Collections. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP-2003)*, pages 401–408, Borovets, Bulgaria, 2003.

Saul Schleimer, Daniel S. Wilkerson, and Alex Aiken. Winnowing: Local Algorithms for Document Fingerprinting. In *Proceedings of the 2003 ACM SIGMOD International Conference on Management of Data*, New York, NY, 2003. ACM.

Efstathios Stamatatos. Intrinsic Plagiarism Detection Using Character n-gram Profiles. In Stein et al. (2009), pages 38–46. URL `http://ceur-ws.org/Vol-502`.

Benno Stein, Sven Meyer zu Eissen, and Martin Potthast. Strategies for Retrieving Plagiarized Documents. In Charles Clarke, Norbert Fuhr, Noriko Kando, Wessel Kraaij, and Arjen de Vries, editors, *30th Annual International ACM SIGIR Conference*, pages 825–826. ACM, 2007.

Benno Stein, Paolo Rosso, Efstathios Stamatatos, Moshe Koppel, and Eneko Agirre, editors. *SEPLN 2009 Workshop on Uncovering Plagiarism, Authorship, and Social Software Misuse (PAN 09)*, San Sebastian, Spain, 2009. CEUS-WS.org. URL `http://ceur-ws.org/Vol-502`.

Kraüpl Taylor. Cryptomnesia and Plagiarism. *The British Journal of Psychiatry*, 111:1111–1118, 1965.