Who's the Thief? Automatic Detection of the Direction of Plagiarism

Cristian Grozea[1] * and Marius Popescu[2] **

[1] Fraunhofer Institute FIRST,
Kekulestrasse 7, 12489 Berlin, Germany
cristian.grozea@first.fraunhofer.de
[2] University of Bucharest
Faculty of Mathematics and Computer Science
Academiei 14, Sect. 1,
Bucharest, Romania
popescunmarius@gmail.com

Abstract. Determining the direction of plagiarism (who plagiarized whom in a given pair of documents) is one of the most interesting problems in the field of automatic plagiarism detection. We present here an approach using an extension of the method Encoplot, which won the 1st international competition on plagiarism detection in 2009. We have tested it on a large-scale corpus of artificial plagiarism, with good results.

Key words: plagiarism detection, plagiarism direction, linguistic forensics

1 Introduction

Plagiarism is a phenomenon of increasing importance, as it is nowadays facilitated by the multitude of sources accessible through internet. It has hit also the academic world, see *dejavu* [6] for a surprisingly extensive database of plagiarized articles in the medical research, exhibiting among others cross-language plagiarism. In education, there are efforts to fight it using commercial services like Turnitin [2] and in-university developed systems [7] and [8]. A recent competition evaluated many methods of plagiarism detection [10]. The methods participating achieved fairly good results in detecting plagiarism. But detecting plagiarism is only half of the problem. The very next question is who copied after whom, who is the thief and who is the victim – although in some cases, the source is none of the two but some third party. Time stamps of some sort can easily prove the priority, but they are not always available, or are too easy to forge (file dates, timestamps on webpages) to be trusted. When two students (or two researchers) present in the same time work that is too similar, how could one know which is the original and which is the copy? There is even a case involving people as famous as Einstein and Hilbert on a subject as important as

* corresponding author
** Research supported by CNCSIS, PNII-Idei, project 228.

the theory of relativity, that was even 80 years later still a matter of debate – for details see [5], [14]. Wouldn't it be nice if the proof of originality could be found in the work itself, not in some – maybe unavailable, maybe untrusted – priority timestamp?

1.1 Related Work

Conceptually, the problem of detecting the plagiarism direction is very related to the problem of detecting stylistic changes and inconsistencies like in the intrinsic plagiarism detection and authorship attribution. If a good measure of stylistic similarity is available, this measure can be used for detecting plagiarism direction. Suppose that one is given two texts and an alleged plagiarized text fragment that belongs to both texts. Then, the stylistic similarity between the alleged plagiarized text fragment and others fragments from the two texts can be measured, and the text that is most similar (stylistically consistent) with the alleged plagiarized text fragment will be considered to be the "original", while the less similar text will be considered to be plagiarizing one.

One related research area where the problem is also the identification of which text is the copy and which text is the source is computational stemmatology "Given a collection of imperfect copies of a textual document, the aim of stemmatology is to reconstruct the history of the text, indicating for each variant the source text from it was copied." [11]

The methods used there are phylogenetic methods borrowed from evolutionary biology. Maybe it is not by chance that the only works that address the problem of plagiarism direction [13, 12] are also based on phylogenetic methods.

We are aware of no plagiarism detection methods able to identify the true source. In general, the first come is treated by the system as being the source and the second one as copying the source. Many measures developed for plagiarism detections are distances, and as such, symmetric. They consider the "effort" of going from the first text to the second text identical with the one needed to get back. Only by breaking this symmetry could one hope to obtain the information of the direction of plagiarism. Even in [12] where the developed methods target explicitly the creation of a phylogenetic tree of evolution of internet news, a complex time-space asymmetric measure is created for this, which is asymmetric, but simple timestamps (article creation time) are used in the computation that eventually decides for each direction the probability of filiation (and implicitly on which is the source and which is the copy).

2 Methods

2.1 Dataset

To approach this problem we have used the newly published plagiarism corpus [15], that has been created in order to allow for a common base of evaluation of the plagiarism detection methods in the aforementioned competition. It is a

multi-language, large-scale, public corpus of plagiarism, containing only artificial plagiarism instances. The random plagiarizing tried to mimic the attempts a human would make to hide the copying, by obfuscating to a certain degree (through reordering of the phrases, replacing words with synonyms or antonyms, deletions, insertions and changes of the words used). Also, some of the instances involve also a translation of the copied passage in the process of going from the source to the destination text, done by automatic means. The external plagiarism section of the corpus contains 14429 source documents (obtained from the Project Gutenberg [1] archive), 14428 "suspicious" documents, and 73522 plagiarized passages. The suspicious documents are also from the Project Gutenberg archive, in which random passages from the sources have been transfered with the transformations mentioned before. The documents are up to book length.

2.2 Finding the Asymmetry

Two of the methods used in the competition are employing dotplot-like analysis [4] to detect and examine the plagiarism: [3] and [9].

In the figures in both of these papers one could observe the parasitic unwanted dots that appear, in addition to the ones useful for recovering the plagiarized passages.

We have used here the second method, our own "encoplot", for which we have published the source code in [9] and which outperformed all others in the challenge. Back then we were already hinting that this method could be of use to identifying the direction of plagiarism, as we noted an asymmetry there: "it is 10% better to rank all suspicious documents for any fixed source instead of ranking all possible sources for a suspicious document (...) This asymmetry deserves more investigation, being one of the few hints of hope so far to tackling what could be the biggest open problem in automatic plagiarism detection, that is determining the direction of plagiarism in a pair of documents".

We have now found another asymmetry that is more useful than that, as it only concerns the two texts involved into a pairwise comparison. Figure 1 shows one example of "encoplot" for the source document #2400 ("Poems" by William Cullen Bryant) and the suspicious document #2 (based on "Our Churches and Chapels" by "Atticus" A. Hewitson, with changes introduced through randomly plagiarizing from two sources) in this corpus.

In Figure 2 the same pair of documents is processed, just that with twice shorter character-based n-grams (n=8 bytes). One can easier observe the parasitic clouds of dots which tend to elongate like a trace pointing to the copying document axis, parallel with that of the source text. We have used 8-grams throughout the experiment.

The apparition of these clouds is a consequence of the way encoplot works: it pairs the first instance of an n-gram in a text with the first instance of the same n-gram in the other, the second instance of it with the second one in the other text, the third with the third, and so on [9]. When the passage is copied without obfuscation and the n-grams of the passage have a single instance in the source and the copy documents, a perfect diagonal appear, without any clouds.

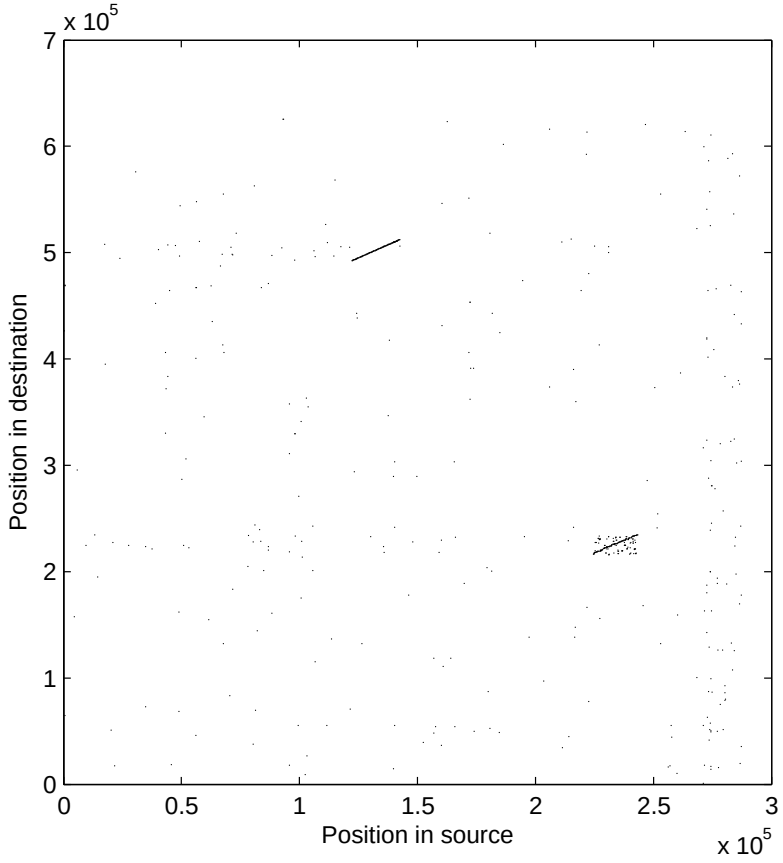

Fig. 1. Clean encoplot example, as used for plagiarism detection. Here the source #2400 and the destination #2 from the corpus, each dot is a 16-bytes n-gram that is shared by the two texts. Two copied passages can be observed as more or less clean local diagonal formations of dots.

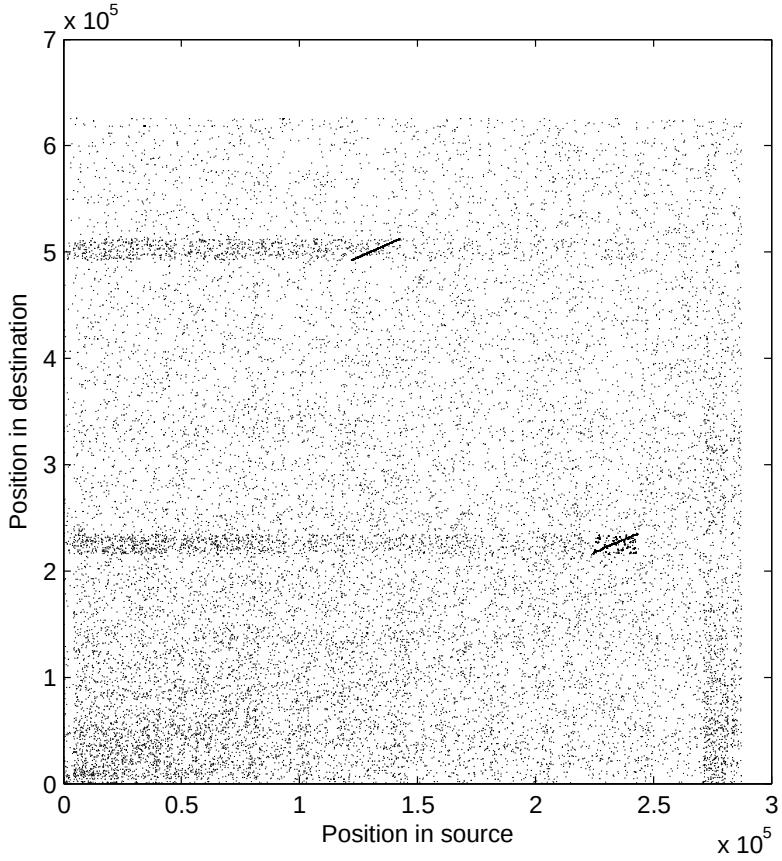

Fig. 2. Asymmetry in Encoplot. The same pair of documents (source #2400 and destination #2) from the corpus, each dot is an 8-bytes n-gram that is shared by the two texts. The shorter n-grams lead to more coincidences, "clouds" of dots that are unwanted for plagiarism detection, but useful for determining the direction of the plagiarism.

For short n-grams, the probabilities for the n-grams to appear multiple times in each document increase. For medium-size n-grams (not very short, but not very long either – n=8 bytes in the herein reported experiment) there is more probably to have multiple instances of the n-grams in the passage in the remaining text of the source document than in the remaining text of the destination document – which is the same with saying that the n-grams distribution in the copied passage matches more the one of the source text than the one of the destination one. What happens when one n-gram from the source document appears not only in the copied passage but also before and after it in the source document? Assuming for simplicity that it only appears once (in the copied passage) in the plagiarizing document, then only one match will be in the encoplot, between the first instance of that n-gram in the source (appearing before the plagiated passage) and the single instance in the destination document, in the copied passage. Effectively this means that the dot which corresponds to that match is moved forward towards the beginning of the source document. For every n-gram this offset can be different and the result is a cloud of dots moved from the diagonal towards the beginning of the source document. Of course for some n-grams the opposite could be true, to be unique in the source and have instead multiple instances in the destination, including one before the passage, which will have the effect to displace the corresponding dots from the diagonal and move those towards the beginning of the destination document. It is just that we expect this to happen less often that the former case.

We set to test how accurate a method based on this observation would be. Finding the passages in correspondence is the problem of external plagiarism detection, and as it is not our concern now, we assume we have the full details about what passages in what text correspond to what passages in what other text and the only information missing is which is the direction. To model this, we randomly permute the source and the "destination" and try to detect the correct direction using solely the asymmetry of the encoplot.

2.3 Measuring and Using the Asymmetry

Spotting the asymmetry in the encoplot graph is easy for humans. In order to do it automatically, one needs to solve a computer vision problem. The difficulty lays in the size of the encoplot data, that can be as long as one of the texts. This is still much better than the maximum length of the general dotplot sets, as those can extend up to the product of the lengths of the two documents.

Figure 3 shows the regions used for our scoring.

We have modeled the visual contrast between the horizontal trace and it's neighbor regions as the mean of the contrast to the upper band and the contrast to the lower band, each of those of the same width as the trace. The width is sometimes limited, when the trace is too close to the beginning or to the end of the vertical axis.

$$hcontrast = \frac{contrast\_up + contrast\_down}{2}. \quad (1)$$

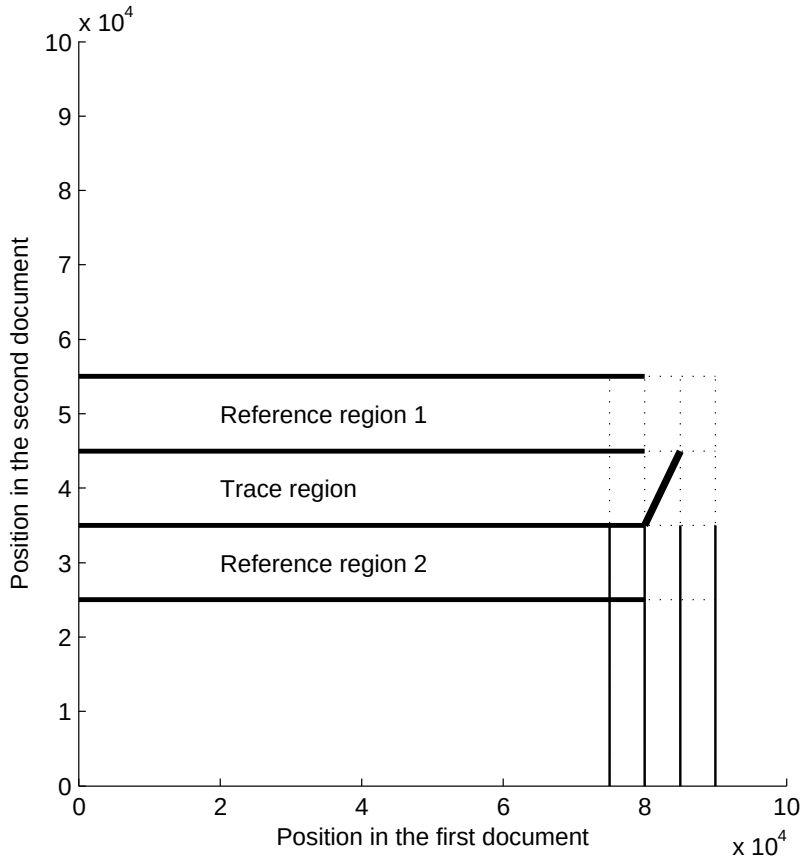

Fig. 3. Regions used to define the scoring. The density of the dots in the considered trace region (either horizontal or vertical) is compared to the density of the dots in the neighboring reference regions.

The contrast between a trace and a neighbor region is measured through the percentage of the points contained in their union that are contained in the trace, compensated for the truncation of the neighbor region (needed when the width of the neighbor region is limited by the beginning or the end of the corresponding document).

$$contrast\_up = \frac{\|Traceregion\|}{\|Traceregion\| + \|Region1\| * width_{Trace}/width_{Region1}}. \quad (2)$$

The contrast *vcontrast* between the vertical trace and the neighbor region is defined similarly, but uses the left and right neighbor regions instead of above and below ones.

We then classify each pair according to this heuristic: the higher contrast trace points to the copy and is parallel to the source.

$$encoplot\_block\_asymmetry\_indicator = hcontrast - vcontrast. \quad (3)$$

When the asymmetry indicator is positive, we predict that the source is the document on the horizontal axis, otherwise that it is the document on the vertical axis. We count the cases when the asymmetry indicator is zero as prediction errors (even though half of them could randomly match the true answer).

3 Results and Analysis

The results are summarized in Table 1.

Table 1. Results

Population layer (and proportion)	Prediction accuracy	p-Value for 1 dof χ^2
Whole population (100%)	75.417%	–
Translated (8.68%)	74.361%	0.0502
Not obfuscated (45.21%)	77.852%	$< 10^{-24}$
High obfuscation (18.58%)	69.776%	$< 10^{-52}$
Short passages (26.18%)	68.111%	$< 10^{-121}$
Long passages (73.82%)	78.008%	$< 10^{-43}$
Close to the source start (14.63%)	69.606%	$< 10^{-43}$

The global accuracy (75.417%) is surprisingly good.

It is interesting to see in what cases the method fails and why. The influence of the factors is given in Table 1, together with their statistical significance, computed using a single degree of freedom χ^2 test with the null hypothesis that the factor has no influence on the decision accuracy.

A visual inspection of those cases where the method fails show that they can be classified into one of these classes: too short passages (therefore too few

n-grams expected in the asymmetric parasitic clouds of dots); passages too close to the beginning of one of the texts (therefore again too few dots in one of the clouds); too crowded encoplots (as in Figure 4), with many closely situated passages in correspondence (decreasing thus the contrast of the trace/cloud of interest); too short texts (and again too small dot sets and too high variances of their size).

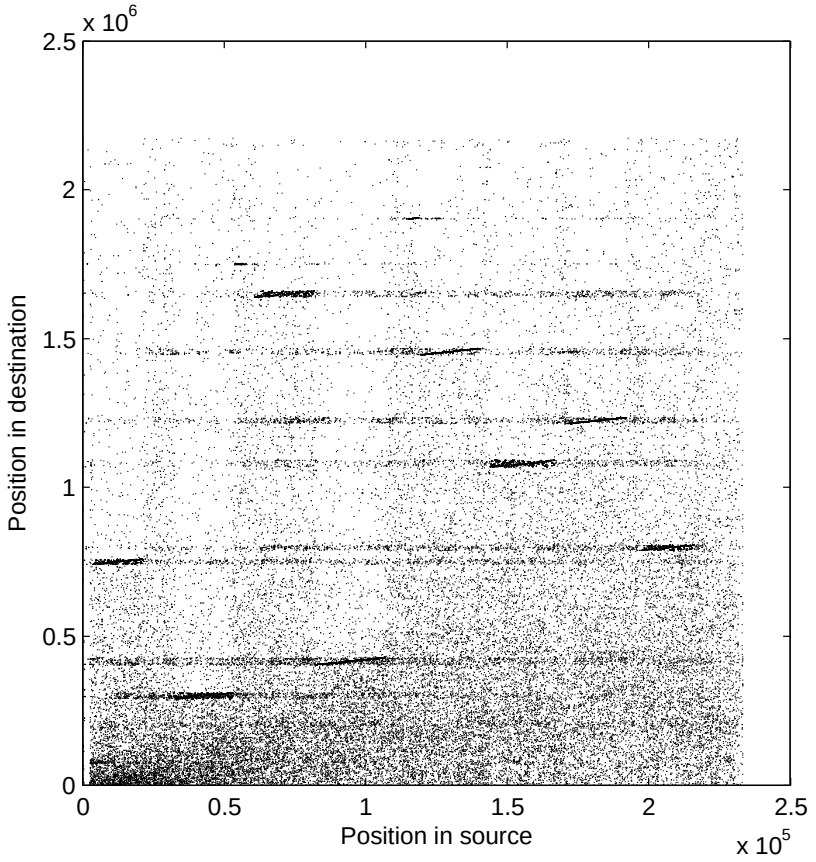

Fig. 4. Crowded encoplot, through many passages plagiarized from the same source; the horizontal traces selectively affect the density of the tested vertical bands. Here displayed source document #2225 versus suspicious document #5.

4 Discussion and Conclusion

We have to state as clearly as possible that we don't claim that we are ready to close any priority/plagiarism dispute simply by presenting the texts to our method. We are very much aware that plagiarism as blatant as many of the instances in the used corpus is maybe never to be seen in practice. This could also be said about so much lack of blending of the copied passages into the destination. All these aspects made a problem – that could be impossible to solve in real cases – solvable in 75% of the instances in this artificial plagiarism corpus.

Why does it work and what else could work? To understand that, we have to consider the meaning of those dots that help us to get back the plagiarism direction information eventually. They are shared n-grams between the two texts. Their preferential spread / higher density on a direction parallel with the axis of the source document corresponds to a better blending of that passage into the source document than into the copying document. One could say that our asymmetry indicator turned encoplot into a method for intrinsic plagiarism detection, to some extent. One could expect that other methods of intrinsic plagiarism detection can be turned into methods to determine automatically the direction of the plagiarism.

Admittedly we didn't spend too much time with tuning the asymmetry indicator or any of the other parameters. We have tuned the indicator on the first 100 plagiarism cases in a visual data exploration fashion, then we validated it by running on the remaining 73422 cases. It worked as good as it did from the first run, in the first day. Our point was mostly to have a proof of concept that this open problem of the automatic plagiarism detection, the detection of the direction of plagiarism, is solvable at least in many instances of the simulated/artificial plagiarism.

Can the performance on the population layers where the method fails more often be improved? For some of them probably yes, and here is how this could be done: for the texts too close to the beginning of the source, the encoplot can be computed on the mirrored texts. Please note that it is not enough to look for the clouds towards the end of the documents, as the encoplot procedure produces different clouds when computed on the mirrrored texts, and, as explained before, due to the way encoplot matches the texts one should only be interested in the dots displaced towards the beginnings of the texts given as input to encoplot. For the too crowded encoplots (like in Figure 4), the encoplot could be computed repeatedly for each plagiarized passage in turn, overwriting all other passages in correspondence with random text.

We found surprising that the encoplot asymmetry indicator worked so accurate on the translated passages. The encoplot for a pair of documents where the plagiarizing involved translation from Spanish to English is shown in Figure 5. In this case, the automatic translation used left untranslated all person names – as expected – and some spanish words. This was enough for the text to blend better into the original spanish context than into the english context, despite being almost in English after translation.

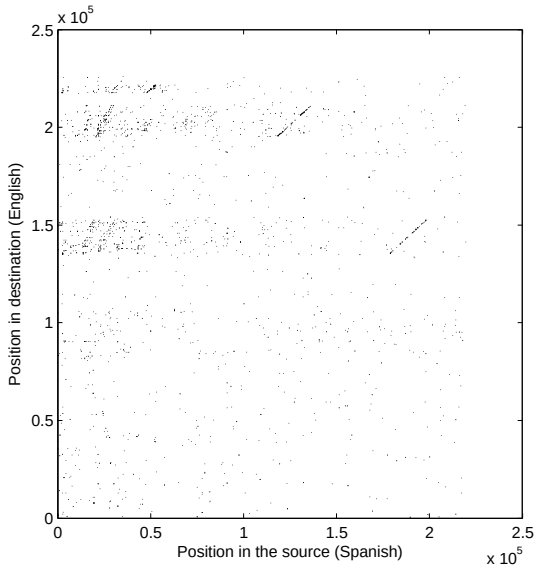

Fig. 5. Plagiarization with translation – the direction is still detectable, although the dot clouds are not very clearly delimited and not very dense. Here displayed source document #2923 (Spanish) versus suspicious document #90 (English).

Following the best practice in science, our results are fully reproducible, as both encoplot and the data used are publicly available. The corpus is available as a web resource [15] and the code for computing encoplot of two files is available in the encoplot paper [9].

To conclude, we have shown that on the largest plagiarism corpus available to date (albeit artificial) the problem of detecting the direction of the plagiarism is solvable with a fairly high accuracy (about 75%). Future work will show how well this method works on natural plagiarism. We are not aware of any publicly available corpus (even of much smaller size) that would have allowed us to test this. We are looking forward to seeing more papers on this subject, more results on the same public corpus, maybe leveraging the intrinsic plagiarism detection methods.

Acknowledgments. The authors thank Dr. Andreas Ziehe and the anonymous reviewers for the thorough review and their useful suggestions. The contribution of Filip Grozea to the development of the heuristics is also acknowledged.

References

1. Project Gutenberg `http://www.gutenberg.org`, 1971.

2. R.K. Baker, B. Thornton, and M. Adams. An Evaluation Of The Effectiveness Of Turnitin. Com As A Tool For Reducing Plagiarism In Graduate Student Term Papers. *College Teaching Methods & Styles Journal*, 4(9), 2008.
3. C. Basile, G. Cristadoro, D. Benedetto, E. Caglioti, and M. Degli Esposti. A plagiarism detection procedure in three steps: selection, matches and squares. In *3rd PAN WORKSHOP. UNCOVERING PLAGIARISM, AUTHORSHIP AND SOCIAL SOFTWARE MISUSE*, page 19.
4. P. Clough. Old and new challenges in automatic plagiarism detection. *National Plagiarism Advisory Service*, 2003.
5. L. Corry, J. Renn, and J. Stachel. Belated decision in the Hilbert-Einstein priority dispute. *Science*, 278(5341):1270, 1997.
6. M. Errami, J.M. Hicks, W. Fisher, D. Trusty, J.D. Wren, T.C. Long, and H.R. Garner. Deja vu A study of duplicate citations in Medline. *Bioinformatics*, 24(2):243, 2008.
7. M. Freire and M. Cebrian. Design of the AC Academic Plagiarism Detection System. Technical report, Tech. rep., Escuela Politecnica Superior, Universidad Autonoma de Madrid, Madrid, Spain. Nov, 2008.
8. C. Grozea. Plagiarism detection with state of the art compression programs. Report CDMTCS-247, Centre for Discrete Mathematics and Theoretical Computer Science, University of Auckland, Auckland, New Zealand, August 2004.
9. C. Grozea, C. Gehl, and M. Popescu. ENCOPLOT: Pairwise Sequence Matching in Linear Time Applied to Plagiarism Detection. In *3rd PAN WORKSHOP. UNCOVERING PLAGIARISM, AUTHORSHIP AND SOCIAL SOFTWARE MISUSE*, page 10.
10. M. Potthast, B. Stein, A. Eiselt, A. Barrón-Cedeño, and P. Rosso. Overview of the 1st International Competition on Plagiarism Detection. In *3rd PAN WORKSHOP. UNCOVERING PLAGIARISM, AUTHORSHIP AND SOCIAL SOFTWARE MISUSE*, page 1.
11. T. Roos and T. Heikkila. Evaluating methods for computer-assisted stemmatology using artificial benchmark data sets. *Literary and Linguistic Computing*, 24(4), 2009.
12. C.K. Ryu, H.J. Kim, and H.G. Cho. A detecting and tracing algorithm for unauthorized internet-news plagiarism using spatio-temporal document evolution model. In *Proceedings of the 2009 ACM symposium on Applied Computing*, pages 863–868. ACM New York, NY, USA, 2009.
13. C.K. Ryu, H.J. Kim, S.H. Ji, G. Woo, and H.G. Cho. Detecting and tracing plagiarized documents by reconstruction plagiarism-evolution tree. *CIT*, page 119, 2008.
14. T. Sauer. Einstein Equations and Hilbert Action: What is missing on page 8 of the proofs for Hilbert's First Communication on the Foundations of Physics? *Archive for history of exact sciences*, 59(6):577–590, 2005.
15. Webis at Bauhaus-Universität Weimar and NLEL at Universidad Politécnica de Valencia. PAN Plagiarism Corpus PAN-PC-09. http://www.webis.de/research/corpora, 2009. Martin Potthast, Andreas Eiselt, Benno Stein, Alberto Barrón Cedeño, and Paolo Rosso (editors).