Encoplot - Tuned for High Recall (also proposing a new plagiarism detection score)

Cristian Grozea[1] and Marius Popescu[2]

[1] Fraunhofer FOKUS, Berlin Germany
[2] University of Bucharest, Romania
cristian.grozea@brainsignals.de, popescunmarius@gmail.com

Abstract This article describes the latest changes to our plagiarism detection system Encoplot. We have sent the modified system to the PAN@CLEF 2012 automatic detection of plagiarism challenge, where it ranked 2nd by the F-measure and 3rd by the "plagdet" scoring method that we had previously shown to be flawed to some extent. The main changes have been done to the heuristic that tries to recognize the clusters of N-grams matches as matching passages in the pair of documents examined. We have aimed for high recall under difficult conditions (sparse matches) which are typical for real-life rephrasing by people. The result of the evaluation on the training and test PAN 2012 corpora shows that we have achieved our goal of improving the performance of this piece of the Encoplot plagiarism detection system. In the final part of this article we analyze the anomalies of the plagdet scoring method, show that those are not negligible, and propose a modified plagdet version that lowers those anomalies.

1 Introduction

Plagiarism detection is unfortunately a requirement of the nowadays academic life. More than once an year the press publishes about yet another politician who plagiated in his/her Ph.D. thesis. To quote from [14], "A spectre is haunting Europe, and this time it is the spectre of plagiarism and scientific misconduct. Some high-profile politicians have had to resign in the last 18 months - but the revelations are also shaking respected European universities". Scientific articles are written by appropriation of someone else's work through plagiarism, even in fields as medicine where the possible consequences of faking research and results are potentially very severe [4,3]. Given the large volume of works that should be checked for plagiarism, automatic plagiarism detection is probably the only practical way to prune until the human examiners can finally decide on whether or not the text reuse is a plagiate or not.

Our team has build a series of very competitive plagiarism detection systems, named Encoplot after the way the core algorithm used functions [5,7,8]. Those had very good performance in the previous years international competitions on plagiarism detection PAN 2009, 2010 and 2011[12,11,13].

This year we have focused only on the detailed comparison subtask, where the strengths of Encoplot lie. This task requires that given a pair of documents (one possible source and a second, suspicious document) all plagiarism instances from the source to

the suspicious document are found and reported. That is to say that each copied passage, with or without obfuscation, short or long must be found, even across languages. Given our previous bad experience with automatic translation [8] we took the easy path of using the opportunity offered by the organizers to process the texts already translated into English. Therefore, for the remaining of this paper we consider that the two texts to be compared share the same language (that doesn't have to be English, as Encoplot is language and character set independent).

2 Methods

2.1 Encoplot

Very briefly, Encoplot's core algorithm is a variant of the well-known **Dotplot** [2], faster because the set of N-gram matches (dots on the plot) is guaranteed to be linear and the run-time is also linear; also the matches lost (vs. dotplot) have good chances to bear low information, and thus be insignificant as a proof of plagiarism.

The Encoplot Core Algorithm

- Input: Sequences A and B to compare
- Output: list (x,y) of positions in A, respectively B, where there is exactly the same N-gram

1. Extract the N-grams from A and B
2. Sort these two lists of N-grams
3. Intersect these lists using a modified mergesort algorithm. Whenever the two smallest N-grams are the equal, output the position in A and the one in B.

A simple example is given in Table 1.

Encoplot pairs	Dotplot pairs
1 2 ab	1 2 ab
	4 2 ab
5 4 bd	5 4 bd

Table 1. Small Encoplot Example: A=abcabd, B=xabdy, N=2

After obtaining this (sub)set of the positions where the two texts contain the same N-grams, a heuristic process is run in order to identify the passage copying. A passage copied verbatim leads to a perfect diagonal (translated to the starting position of the passage in the source and in the suspicious document). Obfuscation leads to more diffuse clouds of dots – see Figure 1 and Figure 2.

What is new in the version for PAN 2012 is that the N-gram matches clustering heuristic part is now tuned for higher recall on the difficult cases: manual plagiarism

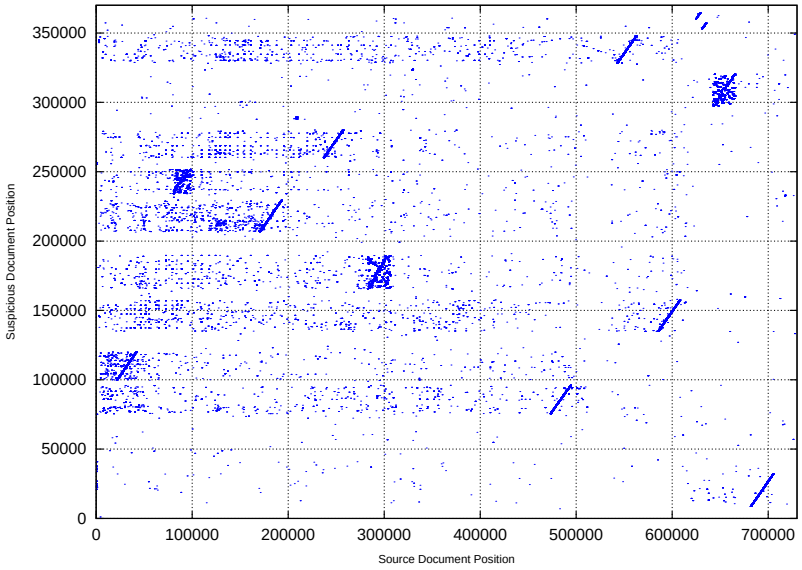

Figure 1. Plagiarism detection with Encoplot – example with several copied passages, with various degrees of obfuscation (using N=16 N-grams). Each dot is a position where the N-grams in the source and suspicious documents coincide.

and high obfuscation (either artificially induced, or occurring as a result of translation). This comes at the expense of diminishing the precision for some easier cases (like the no-obfuscation or the verbatim copying); as those are easily handled by trivially-simple methods, we have decided to ignore that. Also we have identified and fixed a bug which affected the versions we have sent to PAN from 2009 on – this bug reduced the chances of finding passage matches after the first one is found.

Tuning has been done by understanding why the heuristic failed in randomly picked cases and adjusting the parameters to improve the behavior on those cases. Also we have checked the effect of the changes on only the first 50 pairs from the lists provided in the training dataset.

In order to present the changes, we follow the full description of the heuristic used until 2011, as given in the technical report [6]. Please note that the roles of the suspicious document and source document are now reversed: the first projection is done on the suspicious document, instead of the source.

2.2 The Clustering Heuristic

The heuristic employed for clustering the "dots" produced by the encoplot core algorithm into passages consists of the following steps:

1. The dots are projected on the suspicious document's axis, then the presence bits of these projections are smoothed by convolution with a constant vector of size 256, in order to approximate their local density.

2. Within a Monte-Carlo optimization loop (100 attempts), a random starting position is selected, among the projections of the dots on the axis corresponding to the suspicious document.
3. This start is treated like the seed of a segment which is extended to the left and to the right as much as possible, while keeping the density of the projections in the segment over a certain limit (1/32).
4. If the segment is long enough (128 characters) and the projections within it are dense enough (above 1/32), the dots having projections inside the segment are isolated, their projections on the axis of the source document are pruned of outliers.
5. If the segment on the axis of the source and the segment on the axis of the suspicious document satisfies certain sanity checks (their lengths over 128 and the density of the projections of the dots above 50%), the pair of segments (passages) is selected as a candidate.
6. The best candidate (the one with longest passages found in correspondence) is reported if it satisfies the checks mentioned at the previous step, otherwise the current attempt is labeled as a failure to find a passage match.
7. Either just the dots in the box corresponding to the intersection of the two passages, or all the dots projecting on the suspicious document in the segment grown initially from the seed are removed from the set – the choice is made by evaluating the chances that the remaining dots would be enough for an equally or more dense passage correspondence. Then the Monte-Carlo loop is resumed, up to 100 times. Ten consecutive failures to find an acceptable match of passages lead to an early stop of the algorithm, such that the speed of the algorithm auto-adjusts to the size of the problem, as measured by the number of passages in correspondence.

3 Results

The results computed on all pairs from the training set (after the submission deadline) are show in Table 2. Each sub-corpus contains 1000 document pairs and could therefore provide a good prediction of the actual performance, as only 50 of them have been used for guiding our tuning.

We have shown in a previous paper [8] that the so called "plagdet" scoring formula is flawed in that it evaluates as being better detections that are obviously worse (to humans), by over-penalizing the F-measure with the granularity, where by granularity it is mean roughly in how many parts in average a single source passage is split in the reported detections. Our consequent position is that granularity barely matters, unless it's truly excessive (say, 10 or more) – what matters is the compromise between precision and recall (e.g. evaluated by F-measure). Our granularity is listed for reference, it never exceeded 1.25 – but 2 would have been as good. The previous years solutions fixing the granularity simply joined the detections that were close enough to be part of the same plagiarism instance; we focused on what we have specific and did not implement this common technique not needed in practice, but rather imposed by the way the scoring is done at PAN.

For getting a sense of how much better this year's version of Encoplot is than the one we had last year, we give in Table 3 the results on the same corpus of the old method. As

Table 2. Results on the 2012 Training Set With the 2012 Encoplot Version

Corpus	Recall	Precision	F-measure	Granularity
No obfuscation	0.87	0.74	0.80	1.02
Artificial low obfuscation	0.81	0.95	0.87	1.25
Artificial high obfuscation	0.38	0.96	0.54	1.16
Simulated paraphrase	0.56	0.85	0.67	1.00

one can see, the changes we have done to the heuristic determine significant increases in the recall performance, with minor decreases in the precision. The only sub-corpus for which the recall has diminished was surprisingly the sub-corpora of non-obfuscated plagiarism instances (more or less verbatim copies). After the analysis and the detection of the unrealistic duplicates problem presented in the next section, we have re-run the system on this sub-corpus and have obtained on it recall=0.97, precision=0.73, and thus F-measure=0.83.

Table 3. Results on the 2012 Training Set with the 2011 Encoplot Version

Corpus	Recall	Precision	F-measure	Granularity
No obfuscation	0.82	0.93	0.87	1.00
Artificial low obfuscation	0.56	0.99	0.72	1.28
Artificial high obfuscation	0.10	0.99	0.18	1.12
Simulated paraphrase	0.33	0.99	0.50	1.06

4 Discussion

4.1 What can be done to improve the realism of the PAN benchmarks in plagiarism detection

The PAN benchmark, whereas objective by design and nicely conducted, fails short sometimes. We discuss here two possible improvements.

Improving the quality of the corpus: In PAN 2010, there was an issue with the plagiarism instance duplicates: same passage from the source copied multiple times into the suspicious document, up to 17 times.

We reproduce here from [8] the description of the problem.

> "(...) some of the passages from the source were copied multiple times into the destination suspicious document – a substantial amount: out of 55723 external plagiarism instances, 10694 ($> 19\%$) had the multiplicity at least 2, 3483 multiplicity at least 3. The maximum multiplicity of a single passage was 17 (!).
>
> This probably explains our suspiciously low recall in the 2010 competition on the non-obfuscated cases (and other subcorpora). As a side effect of the

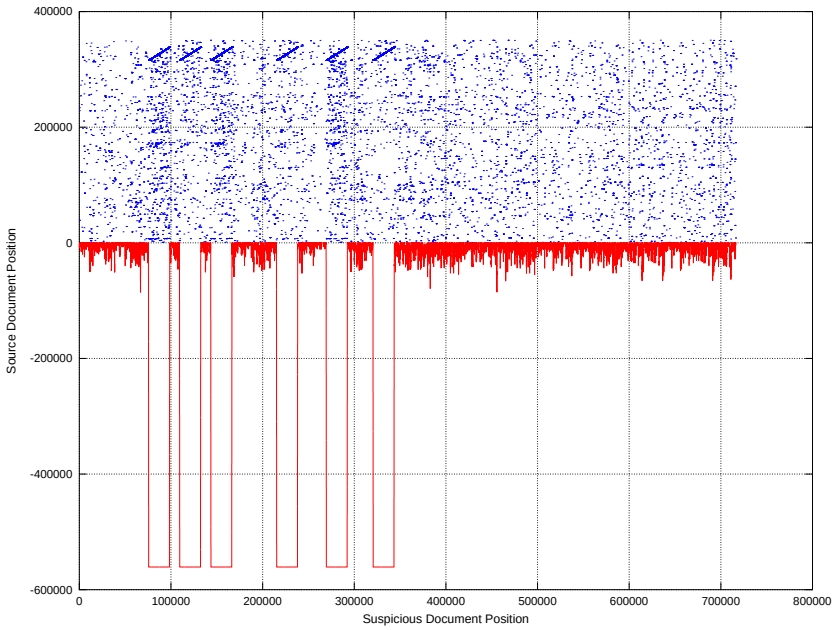

Figure 2. Unrealistic scenario: in this pair of documents in the 2012 PAN training corpus the same passage has been copied from the source 6 times into the suspicious document, at different locations. The six blue diagonals show the matches. The red curve shows their detection by the Encoplot's heuristic.

> speed and space optimizations the core encoplot algorithm offers over dotplot, for the simple case when there is no obfuscation at all and just verbatim copying multiple times, only the first copy of a passage is matched. To understand why, remember that each position in the source is paired with at most one position in the suspicious document. Therefore a full match of the source passage fully consumes it, and it cannot match any of the subsequent copies. Having a second copy of the same passage in the source would allow for a second match and so on. To cope with that, we have concatenated each source with itself 4 times before analyzing the pair in detail with our heuristic, creating thus 4 copies in the source of each passage previously there. The number 4 has been chosen as a compromise, balancing the effort and the expected increase in recall."

Although the PAN 2011 corpus was seemingly free of those problems, they have been reintroduced in 2012, judging on the training corpus, where they made a big difference in the recall achieved by Encoplot on the non-obfuscated plagiarism sub-corpus. For example, for the pair suspicious/source 1746/3773 there are no less than 5 copies of the same 19 kB long passage (about 5 pages of printed text each time). Another example, from the source 3812 to the target 1702 the same passage amounting to about 5-6 printed pages is copied no less than 6 times in different positions, as shown in Figure 2.

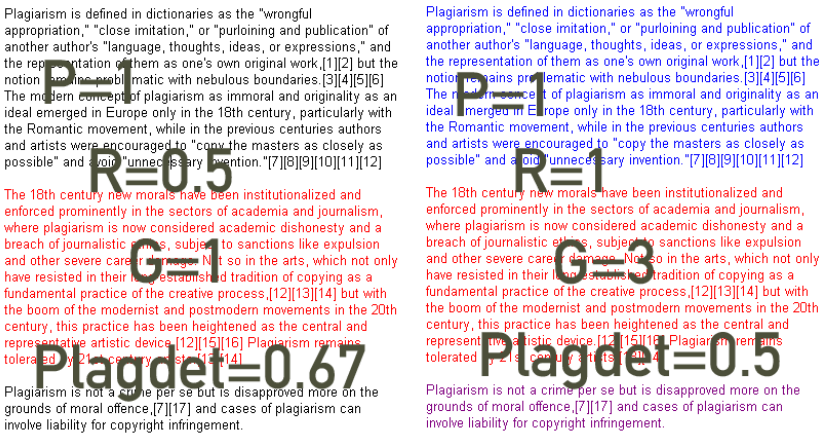

Figure 3. Left: detection preferred by the plagdet scoring, fails to find half of the copied text (in black). Right: detection preferred by humans, all copied text is found, as three parts (in red, blue and magenta). The text provenience is Wikipedia's definition of "Plagiarism", downloaded in 2011.

This is extremely unrealistic, as no real plagiator will copy multiple times the same text into his/her text; for such long passages, certainly not even twice.

Improving the scoring function: In 2011, we have shown that "plagdet" is flawed, we quote here from [8] the description of the issue with the granularity correction in the plagdet score.

> "The granularity has been introduced for plagiarism detection in [10]. It was meant to correct the standard F-score for excessive splitting of the plagiarized passages retrieved. It is an ad-hoc correction that divides the F-score by $log_2(1 + granularity)$. It exhibits unwanted behavior in certain cases. For example, let's assume we compare with plagdet two methods, one having recall 33.33%, precision 100% and granularity 1 with another method having both precision and recall 100% and granularity 3. The two methods will obtain the very same plagdet score, 0.5, as a result of applying the granularity correction, although the second method is obviously to be preferred. It has 100% recall and precision, it finds everything and nothing more and even the splitting is very far from excessive. No user will ever prefer a software that fails to find two thirds of the cases to a software that finds them all and even displays each as one block (when colouring text blocks, the adjacent parts will visually join). More thought should be spent in finding a reasonable plagiarism detection score."

The 2012 PAN retained "plagdet" with its anomalies for evaluating the plagiarism detection systems. Here is our analysis and suggestion on how to improve the scoring.

We start the analysis by giving another example in Figure 3, where a detection with recall just 50% obtains a plagdet score of 0.67 (left), whereas a detection with recall

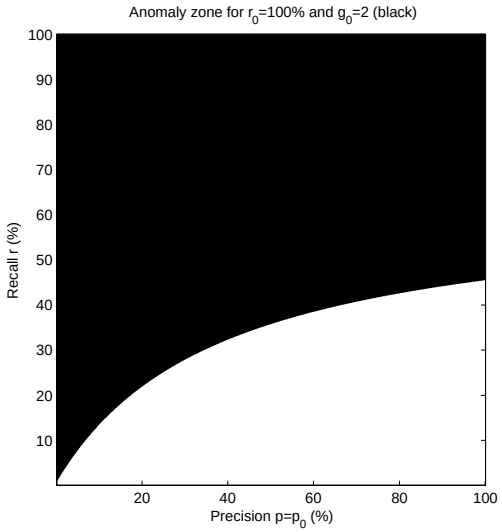

Figure 4. Anomaly area (marked black) in precision-recall space where a detection with granularity=1 and that recall<1 is better according to plagdet than a 100% recall solution with equal precision and granularity 2.

100% obtains a plagdet score of just 0.5 (on the right) - despite having both the same precision, 1. The difference is that the solution preferred by plagdet doesn't detect half of the copied text, while the obviously better one finds it all, just that as three separated parts. On this example the anomaly is clear: for the same precision, the score of one system is lower although its recall rate is much higher, due to the over-compensation for granularity.

Let us define formally the anomaly. Plagdet has this definition $plagdet(p, r, g) = F(p, r)/log_2(1 + g) = \frac{2}{\frac{1/p+1/r}{log_2(1+g)}}$ (following [10]), where p=precision, r=recall and g=granularity. Then we call formally a situation anomalous when, for two detections with precision, recall and granularity (p_0, r_0, g_0) and respectively $(p, r, g = 1)$:

$$p = p_0 \wedge r < r_0 \wedge plagdet(p, r, 1) > plagdet(p_0, r_0, g_0) \quad (1)$$

The formula 1 states that despite the detection (p_0, r_0, g_0) having the same precision and higher recall than $(p, r, g = 1)$, it has a lower plagdet score.

The first question we want to clarify in this analysis is whether or not the anomalous instances we have exemplified above are isolated cases. To this end, we fix $r_0 = 1$ and plot in Figure 4 the anomalous zone in the precision-recall space for $g_0 = 2$. It is obviously more than half of that space!

Increasing g_0 only increases the anomalous area - which is equal to the probability of reaching such an anomaly, assuming uniform distribution in the precision-recall space. In Figure 5 the areas/probabilities for the first few values of granularity are given.

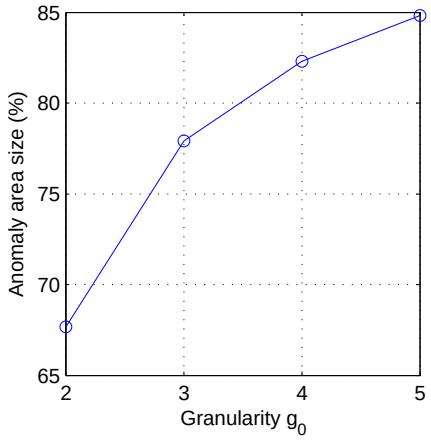

Figure 5. Probability of an anomaly with plagdet, for $r_0 = 1$.

It is important to note that the idea of penalizing the granularity is not faulty per se. It is just this particular penalization method which leads to those anomalies. What qualities an ideal plagiarism detection quality formula should have? It should favor high precision and recalls (or equivalently high F-measures), penalize the excessive granularity (e.g. 100) and should not lead to those anomalies for low granularity values g_0. In addition, for everything else being equal, it should still prefer the lower granularity detections to the higher granularity ones. Given these requirements, we propose a generalization of the plagdet formula:

$$plagdet_\beta(p, r, g) = F(p, r)/log_2(\beta + g) = \frac{\frac{2}{1/p+1/r}}{log_2(\beta + g)} \tag{2}$$

The change is minor, the 1 in the correction factor $log_2(1 + g)$ was replaced by a parameter β; therefore $plagdet_1 = plagdet$. It is our hope that by avoiding a radical departure from the old plagdet formula we increase the chances the new formula 2 is accepted by the automatic plagiarism detection community. We look now at the impact this new parameter has on the area of anomalies. In Figure 6 one can see the progressive decrease of the probability of anomalies with the increase of β; we find the levels for $\beta = 10$ more acceptable and recommend the use of $plagdet_{10}$ instead of $plagdet = plagdet_1$.

If one desires to maintain the property of plagdet that it coincides with the F-measure for granularity=1, then the normalized version of $plagdet_\beta$ can be used, which retains its desirable properties:

$$\overline{plagdet_\beta(p, r, g)} = F(p, r)\frac{log_2(\beta + 1)}{log_2(\beta + g)}$$

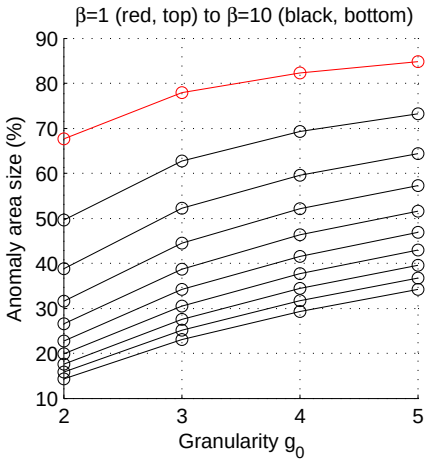

Figure 6. Progressive decrease of the anomaly level of $plagdet_\beta$ for $r_0 = 1$ and $\beta = 1..10$ (from top to down). The curve for $\beta = 1$, which corresponds to the old plagdet is in red, the other ones in black.

As a final check, in Figure 7 the granularity-depending corrections to the F-measure applied when using $\overline{plagdet_\beta}$ are displayed, in order to show that for granularity keeps being penalized and even in a very similar fashion.

5 Conclusion

We have sent Encoplot again this year to the PAN benchmark, and we have obtained a very good ranking, 2^{nd} by the F-measure. Beyond the limitations of the PAN benchmark, some of which we have explained in this paper, it remains the main benchmark in the automatic detection of plagiarism. We have also proposed here a modified $plagdet_\beta$ scoring which should present a lower level of anomalies than $plagdet$. After examining the decrease in the level of anomalies, we have proposed using either $plagdet_{10}$ or its normalized version $\overline{plagdet_{10}}$. Ever since we have won the first PAN challenge in 2009 we have managed to stay within the top few teams every year, which justifies the effort we have put into developing and maintaining Encoplot as one of the state of the art automatic plagiarism detection systems. The system can be commercially licensed – as it is, or customized for various applications – through Fraunhofer FOKUS Berlin, Germany.

References

1. Braschler, M., Harman, D., Pianta, E. (eds.): CLEF 2010 LABs and Workshops, Notebook Papers, 22-23 September 2010, Padua, Italy (2010)
2. Clough, P.: Old and new challenges in automatic plagiarism detection. National Plagiarism Advisory Service (2003)

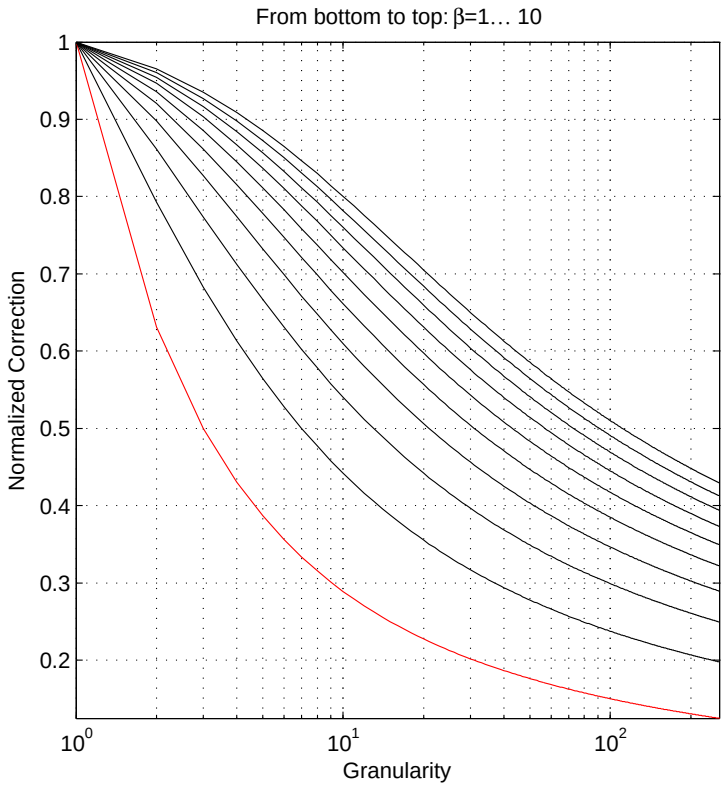

Figure 7. Granularity-dependent correction curves employed by $plagdet_\beta$ for $\beta = 1$ (bottom, red) to $\beta = 10$ (top,black).

3. Errami, M., Garner, H.: A tale of two citations. Nature 451(7177), 397–399 (2008)
4. Errami, M., Hicks, J., Fisher, W., Trusty, D., Wren, J., Long, T., Garner, H.: Déjà vuâĂŤa study of duplicate citations in medline. Bioinformatics 24(2), 243–249 (2008)
5. Grozea, C., Gehl, C., Popescu, M.: ENCOPLOT: Pairwise Sequence Matching in Linear Time Applied to Plagiarism Detection. In: 3rd PAN WORKSHOP. UNCOVERING PLAGIARISM, AUTHORSHIP AND SOCIAL SOFTWARE MISUSE. p. 10 (2009)
6. Grozea, C., Popescu, M.: The Encoplot Similarity Measure for Automatic Detection of Plagiarism - Extended Technical Report. `http://brainsignals.de/encsimTR.pdf` (Aug 2011)
7. Grozea, C., Popescu, M.: Encoplot - Performance in the Second International Plagiarism Detection Challenge - Lab Report for PAN at CLEF 2010 . In: Braschler et al. [1]
8. Grozea, C., Popescu, M.: The encoplot similarity measure for automatic detection of plagiarism - notebook for pan at clef 2011. In: Petras et al. [9]
9. Petras, V., Forner, P., Clough, P.D. (eds.): CLEF 2011 Labs and Workshop, Notebook Papers, 19-22 September 2011, Amsterdam, The Netherlands (2011)
10. Potthast, M., Stein, B., Barrón-Cedeño, A., Rosso, P.: An evaluation framework for plagiarism detection. In: Proceedings of the 23rd International Conference on

Computational Linguistics: Posters. pp. 997–1005. Association for Computational Linguistics (2010)
11. Potthast, M., Barrón-Cedeño, A., Eiselt, A., Stein, B., Rosso, P.: Overview of the 2nd International Competition on Plagiarism Detection. In: Braschler et al. [1]
12. Potthast, M., Eiselt, A., Barrón-Cedeño, A., Stein, B., Rosso, P.: Overview of the 3rd international competition on plagiarism detection. In: Petras et al. [9]
13. Potthast, M., Stein, B., Eiselt, A., universitÃd't Weimar, B., BarrÃşn-cedeÃśo, A., Rosso, P.: P.: Overview of the 1st international competition on plagiarism detection. In: In: SEPLN 2009 Workshop on Uncovering Plagiarism, Authorship, and Social Software Misuse (PAN 09), CEUR-WS.org. pp. 1–9 (2009)
14. Weber-Wulff, D.: Viewpoint: The spectre of plagiarism haunting Europe