Approaches for Intrinsic and External Plagiarism Detection

Notebook for PAN at CLEF 2011

Gabriel Oberreuter[1], Gaston L'Huillier[1], Sebastián A. Ríos[1],
and Juan D. Velásquez[1]

Department of Industrial Engineering
University of Chile
goberreu@ing.uchile.cl, {glhuilli,srios,jvelasqu}@dii.uchile.cl

Abstract Plagiarism detection has been considered as a classification problem which can be approximated with intrinsic strategies, considering self-based information from a given document, and external strategies, considering comparison techniques between a suspicious document and different sources. In this work, both intrinsic and external approaches for plagiarism detection are presented. First, the main contribution for intrinsic plagiarism detection is associated to the outlier detection approach for detecting changes in the author's style. Then, the main contribution for the proposed external plagiarism detection is the space reduction technique to reduce the complexity of this plagiarism detection task. Results shows that our approach is highly competitive with respect to the leading research teams in plagiarism detection.

1 Introduction

Plagiarism in academia is rising and multiple authors have worked to describe this phenomena [5,8]. As commented by Hunt in [5], "Internet Plagiarism" is referred sometimes as a consequence of the "Information Technology revolution", as it proves to be a big problem in academia. According to Park [8], plagiarism is analyzed from various perspectives and considered as a problem that is growing over time. To tackle this problem, the most common approach so far is to detect plagiarism using automated algorithms based on rules and string matching algorithms.

Two main strategies for plagiarism detection have been considered by researches [9,4]: Intrinsic and external plagiarism detection. Intrinsic plagiarism detection aims at discovering plagiarism by examining only the input document, deciding whether parts of the input document are not from the same author. External plagiarism detection is the approach where suspicious documents are compared against a set of possible references. From exact document copy, to paraphrasing, different levels of plagiarism techniques can been used in several contexts, according to Meyer zu Eissen [4].

The main contribution of this work is the usage of outlier detection techniques on text-based data to enhance two plagiarism detection strategies, one for intrinsic plagiarism detection using deviation parameters with respect of the writing style of a given document, and another one to reduce the search space for external plagiarism detection based on the generation of segments of n-gram for approximated plagiarism decision where unrelated documents are discarded efficiently.

This paper is structured as follows: First, in Section 2, a short summary on plagiarism detection is introduced. In Section 3 the proposed external plagiarism detection method is described. Afterwards, in Section 4, the proposed intrinsic plagiarism detection method is described. In Section 5 results are presented. Finally, in Section 6 conclusions are discussed.

2 Related Work

According to Schleimer et al. [11], copy prevention and detection methods can be combined to reduce plagiarism. While copy detection methods can only minimize it, prevention methods can fully eliminate it and decrease it. Notwithstanding this fact, prevention methods need the whole society to take part, thus its solution is non trivial. Copy plagiarism detection methods, on the other hand, are easier to implement, and tackle different levels, from simple manual comparison to complex automatic algorithms [9,10]. A short discussion on plagiarism detection strategies is presented.

2.1 Intrinsic Plagiarism Detection

When comparing texts against a reference set of possible sources, comes the complication of choosing the right set of documents to compare to. And now more than ever, with the possibilities that Internet bring to plagiarists, this task becomes more complicated to achieve. For this, the writing style can be analyzed within the document and an examination for incongruities can be done. The complexity and style of each text can be analyzed based on certain parameters such as text statistics, syntactic features, part-of-speech features, closed-class word sets, and structural features, as stated by Meyer zu Eissen [4]. The main idea is to define a criterion to determine if the style has changed enough to indicate plagiarism.

Stamatatos [13] presented a method for intrinsic plagiarism detection. As described by its author, this approach attempts to quantify the style variation within a document using character n-gram profiles and a style change function based on an appropriate dissimilarity measure originally proposed for author identification. Style profiles are first constructed using a sliding window. For the construction of those profiles the author proposed the use of character n-grams. These n-grams are used for getting information on the writer's style. The method then analyzes changes on the profiles to determine if a change is significantly enough to indicate another author style.

Other approaches have been proposed, such as the one presented by Seaward & Matwin [12]. They introduced Kolmogorov Complexity measures as a way of extracting structural information from texts for Intrinsic Plagiarism Detection. They experiment with complexity features based on the Lempel-Ziv compression algorithm for detecting style shifts within a single document, thus revealing possible plagiarized passages.

2.2 External Plagiarism Detection

In terms of external plagiarism detection algorithms, the use of n-grams have shown to give some flexibility to the detection task, as reworded text fragments could still be detected [7]. Other approaches focus on solving the plagiarism detection problem as a traditional classification problem from the machine learning community [1,3]. Bao et al. in [1], proposed to use a Semantic Sequence Kernel (SSK), and then using it into a traditional Support Vector Machines (SVMs) formulation based on the Structural Risk Minimization (SRM) principle from statistical learning theory [15], where the general objective is finding out the optimal classification hyperplane for the binary classification problem (plagiarized, not plagiarized).

Kasprzak & Brandejs [6] introduced their model for automatic external plagiarism detection. It consist of two main phases; the first is to build the index of the documents, while in the second the similarities are computed. This approach uses word n-grams, with n ranging from 4 to 6, and takes into account the number of matches of those n-grams between the suspicious documents and the source documents for computing the detections. The algorithm have the authors won the first place at the PAN@2010 competition [9].

3 Proposed Method for External Plagiarism Detection

The proposed algorithm is based on two phases; First, it executes a plagiarism search space reduction method, and then executes an exhaustive search to find plagiarized passages. The search space reduction method aims at quickly identify those pair of documents that potentially have some text in common, possibly one of them having plagiarized from the other. For this, the method's general tactics are to remove stopwords, and consider word 4-grams. If two documents have at least two word 4-grams coincidences close enough as to be in the same paragraph, the documents are given to the next phase. Otherwise the pair is discarded. For more details, please refer to [7].

For the exhaustive search, word tri-grams are used (compared to use both word bi-grams and word tri-grams in [7]), and stopwords are not removed. In Figure 1 an example of the algorithm can be seen. Two documents are being compared, where dots represent coincidences of tokens used to characterize the documents.

The system does not consider plagiarism detection between different languages. The overall mechanism for finding the plagiarized passages is described in Oberreuter et al. [7]. More details and other parameters of the algorithm are not revealed due to copyright.

4 Proposed Method for Intrinsic Plagiarism Detection

For intrinsic plagiarism detection, we first considered some ideas other authors had investigated. To characterize the writing style of an author, different details can be considered. As studied by Stein et al. [14], multiple writing style characteristics were tested in order to determine plagiarism. Likewise, Stamatatos [13] experimented with character tri-grams in combination with "n-gram profiles" for the same purpose. For this, it is fundamental to choose with precaution one or a set of language resources an author utilizes for his writing to be able to differentiate it from others.

In the following, some of the core ideas developed in this research are presented:

- To be able to distinguish different authors within the same document, one must characterize the writing style present on the text.

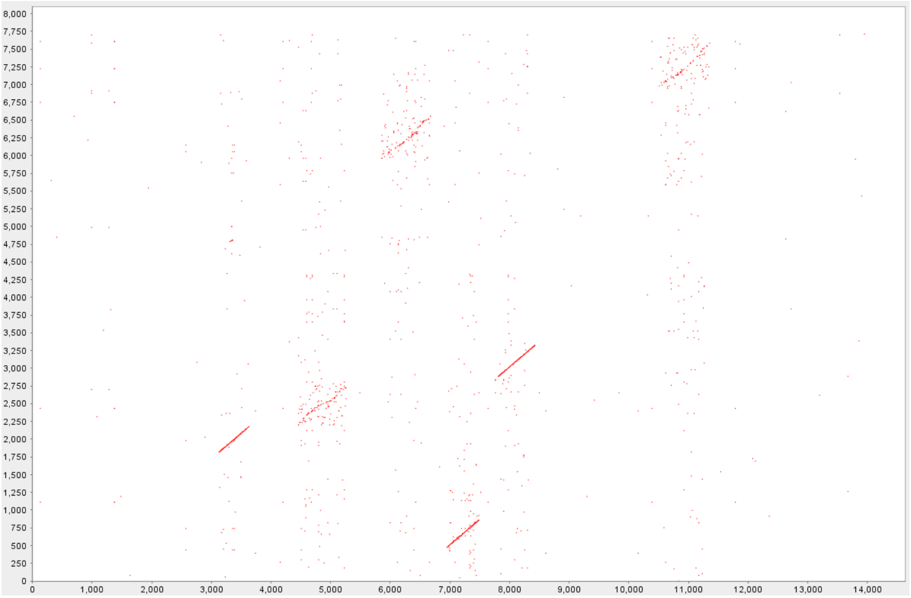

Figure 1. External plagiarism detection example: dotplot of n-grams coincidences between a pair of documents. The algorithm tries to identify those "lines" of close coincidences that represent copied passages.

- The use of "n-gram profiles" compares segments of the document against the whole document. This approach works based on the assumption that the document has a main author, who wrote the majority, if not all, the text. Therefore, it is logical that the comparison between the style of a particular segment with the whole document style could lead to detections of important variations, meaning that other authors are involved.
- Based on reading and contemplation, one of the characteristic that showed to be of interest, is the author's use of words. Different authors tend to use different words to write their ideas, be them on the same topic or not.

These ideas lead to the following intuition for the development of the algorithm: *If some of the words used on the document are author-specific, one can think that those words could be concentrated on the paragraphs (or more general, on the segments) that the mentioned author wrote.*

4.1 The method

First, the document is preprocessed removing numbers and all other Characters that don't belong to the a–z group. All Characters are considered lowercase. Second, the method uses word uni-grams and considers all non-numerical words; stopwords are not removed. Next, a frequency-based algorithm to test self-similarity of document is proposed. A hard (not normalized) frequency vector $\mathbf{v}$ is built for all words on the given document. Then, the complete document is clustered creating groups $\mathcal{C}$. As a first approach, these groups or segments $c \in \mathcal{C}$ are created using a sliding window of length m over the complete document. Afterwards, for each segment $c \in \mathcal{C}$, a new frequency vector v_c is computed, which is used in further steps to compare whether a segment is deviated with respect to the footprint of the complete document. This is performed by using the Algorithm 4.

Algorithm 1 Intrinsic plagiarism evaluation

```
Require: $\mathcal{C}, \mathbf{v}, m, \delta$
for $c \in \mathcal{C}$ do
    $d_c \leftarrow 0$
    build $v_c$ using term frequencies on segment $c$
    for word $w \in v_c$ do
        $d_c \leftarrow d_c + \frac{|freq(w,\mathbf{v})-freq(w,v_c)|}{|freq(w,\mathbf{v})+freq(w,v_c)|}$
    end for
end for
style $\leftarrow \frac{1}{|\mathcal{C}|}\sum_{c\in\mathcal{C}} d_c$
for $c \in \mathcal{C}$ do
    if $d_c < style - \delta$ then
        Mark segment $c$ as outlier and potential plagiarized passage.
    end if
end for
```

As presented in Algorithm 1, the general footprint or style of the document is represented by the average of all differences computed for each segment and the complete document. Note that every segment is compared against the whole document only in terms of the words present in the segment. Also, this algorithm takes into account the intuition; if certain words are only used on a certain segment, the comparison of that segment against the whole document would lead to a low value, because the frequency of those words would be the same in both the whole document and in the segment. Finally, all segments are classified according to its distance with respect to the document's style. As an example, in Figure 2, a graphical representation of this evaluation is presented.

In this case, the average value of the comparison of all segment with the whole document represents the document "main" style. This value is roughly computed by the difference on the frequency of words between vectors $\mathbf{v}$ and $v_c, \forall c \in \mathcal{C}$. If the variation is significant, the style function will be lower than the average value minus δ (the threshold), then the segment is classified as suspicious. In this example, real plagiarized annotations are presented along with the style function value of each segment. Five cases of plagiarism could be discovered; the value of the style function on those cases is lower than the threshold.

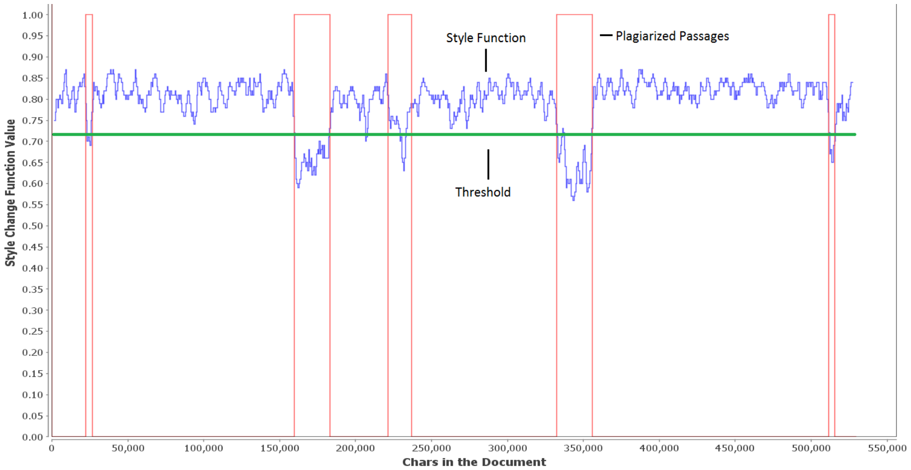

Figure 2. Intrinsic plagiarism detection example. One document is being analyzed; it's style function changes as the sliding window moves forward.

5 Evaluation

The evaluation results will now be presented. Three experiments were conducted: First, we participated with our external plagiarism detector applied to the external corpus provided by the competition. Second, our intrinsic plagiarism detector was applied to the intrinsic corpus. Last, we applied the intrinsic plagiarism detector on the external plagiarism corpus.

5.1 External Plagiarism Detection

Parameters of the algorithms were tuned considering the PAN@2010 corpus in the case of the external approach, and PAN@2009 corpus in the case of the intrinsic one.

In the PAN@2010 competition, as shown in Table 1, the best results were achieved by Kasprzak & Brandejs approach [6]. The overall score was 0.80, and their method achieved good results at the three metrics: precision, recall and granularity. The next top results show similar characteristics, being well balanced in the three metrics. Our model, in it's 2010 version, took fifth place, with an overall score of 0.61, precision of 0.85 and recall of 0.48. The granularities of the top performers were all close to 1.

Our proposed model, applied to the PAN@2010 corpus, achieves better results. The new method is more precise (0.94), thus reducing false-positive detections. The recall also improved, getting a score of 0.6. This value is acceptable considering that at the moment we do not consider detecting plagiarism between different languages, presented in the corpus. Also the corpus considers intrinsic plagiarism, which is not considered in this particular case.

In Table 2 the results from PAN@2011 competition are shown. The revised method achieves third place, obtaining high precision (0.91) but a low recall score (0.22). This could be explained as we do not consider translated plagiarism, and possibly because our plagiarism space reduction technique could be filtering lots of used source documents. The best team, Grman & Ravas, get's slightly better precision, and a recall score of 0.39.

Table 1. Results for the external proposed model using the PAN@2010 corpus. Note that this corpus considers intrinsic and external plagiarism cases.

Rank	Overall Score	F-Measure	Precision	Recall	Granularity	Lead developer
1	0.80	0.80	0.94	0.69	1.00	Kasprzak et al.
2	0.71	0.74	0.91	0.63	1.07	Zou et al.
3	0.69	0.77	0.84	0.71	1.15	Muhr et al.
4	0.62	0.63	0.91	0.48	1.02	Grozea et al.
5	0.61	0.61	0.85	0.48	1.01	Oberreuter et al.
6	0.59	0.59	0.85	0.45	1.00	Torrejón et al.
7	0.52	0.53	0.73	0.41	1.00	Pereira et al.
8	0.51	0.52	0.78	0.39	1.02	Palkovskii et al.
9	0.44	0.45	0.96	0.29	1.01	Sobha et al.
10	0.26	0.39	0.51	0.32	1.87	Gottron et al.
11	0.22	0.38	0.93	0.24	2.23	Micol et al.
12	0.21	0.23	0.18	0.30	1.07	Costa-jussá et al.
13	0.21	0.24	0.40	0.17	1.21	Nawab et al.
14	0.20	0.22	0.50	0.14	1.15	Gupta et al.
15	0.14	0.40	0.91	0.26	6.78	Vania et al.
16	0.06	0.09	0.13	0.07	2.24	Suárez et al.
17	0.02	0.09	0.35	0.05	17.31	Alzahrani et al.
18	0.00	0.00	0.60	0.00	8.68	Iftene et al.
**	**0.73**	**0.73**	**0.94**	**0.60**	**1.01**	**Proposed Model**

Table 2. Official Results for the external proposed model using the PAN@2011 corpus. This corpus considers external plagiarism cases only.

Rank	Overall Score	Recall	Precision	Granularity	Lead developer
1	0.5563430	0.3965569	0.9368736	1.0022487	Grman et al.
2	0.4153395	0.3376925	0.8119867	1.2167900	Grozea et al.
3	**0.3468605**	**0.2257937**	**0.9116530**	**1.0611984**	**Oberreuter et al.**
4	0.2467329	0.1500480	0.7106536	1.0058894	Gillam et al.
5	0.2340035	0.1612845	0.8512947	1.2328923	Torrejón et al.
6	0.1990889	0.1618067	0.4541152	1.2949292	Gupta et al.
7	0.1892155	0.1390201	0.4435687	1.1716516	Palkovskii et al.
8	0.0804139	0.0885330	0.2780244	2.1823870	Nawab et al.
9	0.0012063	0.0011829	0.0050052	2.0028818	Bandyopadhyay et al.

5.2 Intrinsic Plagiarism Detection

A sliding window of 400 words was used, and a threshold parameter $\delta = 0.075$. These were iteratively adjusted depending on text length. Sensibility analysis for all parameters was intentionally excluded by authors due to lack of space.

The results for the intrinsic task at PAN@2009 are shown in Table 3 and for PAN@2011 in Table 4. The results are based on the quality of the detection, which only considers the information on each document itself.

The winner was Stamatatos approach [13], with a recall of 0.4607, precision of 0.2321 and granularity of 1.3839. This method achieved a good combination of precision and recall, and a not top performer granularity.

The proposed method gets an overall score of 0.3457, greater than any other approach, with a positive difference of 0.0995 with the winner's approach. Our model gets the best result at F-measure, precision and granularity.

These results are confirmed with similar results on the PAN@2011 competition presented in Table 4; the proposed model gets roughly the same overall score, 0.3254, with comparable precision (0.34) and worse but not significantly different recall (0.31). We get the best results in the competition, followed by Luyckx et al. with an overall score of 0.17, almost doubling their score.

Table 3. Results for the intrinsic proposed model using the corpus PAN2009.

Rank	Overall Score	F-Measure	Precision	Recall	Granularity	Lead Developer
1	0.2462	0.3086	0.2321	0.4607	1.3839	Stamatatos (2009)
2	0.1955	0.1956	0.1091	0.9437	1.0007	Hagbi and Koppel (2009)
3	0.1766	0.2286	0.1968	0.2724	1.4524	Muhr et al. (2009)
4	0.1219	0.1750	0.1036	0.5630	1.7049	Seaward and Matwin (2009)
**	**0.3457**	**0.3458**	**0.3897**	**0.3109**	**1.0006**	**Proposed Model**

Table 4. Official Results for the intrinsic proposed model using the corpus PAN2011.

Rank	Overall Score	Precision	Recall	Granularity	Lead Developer
1	**0.3254817**	**0.3397965**	**0.3123243**	**1**	**Oberreuter et al.**
2	0.1679779	0.4279112	0.1075817	1.0329386	Luyckx et al.
3	0.0841286	0.1277831	0.0664302	1.0549085	Akiva et al.
4	0.0693820	0.1080543	0.0783903	1.4787234	Gupta et al.

5.3 Intrinsic detector with external corpus

We used the same intrinsic plagiarism detection algorithm with the same parameters on the external plagiarism corpus. The results are presented in Table 5. The recall score is the third lowest; this can be explained as the intrinsic detector provides no information on the source of the copied passages, which reduces considerably the metric itself. Also, the algorithm achieves a precision of 0.36, comparable to the precision obtained when applied to the intrinsic corpus (0.34). Overall, the intrinsic detector would have ranked 8 if participated on the external competition, out of 10 teams.

Table 5. Results using the intrinsic plagiarism detector on the external PAN@2011 corpus. This corpus considers external plagiarism cases only.

Rank	Overall Score	Recall	Precision	Granularity	Lead developer
1	0.556343	0.3965569	0.9368736	1.0022487	Grman et al.
2	0.4153395	0.3376925	0.8119867	1.21679	Grozea et al.
3	0.3468605	0.2257937	0.911653	1.0611984	Oberreuter et al.
4	0.2467329	0.150048	0.7106536	1.0058894	Gillam et al.
5	0.2340035	0.1612845	0.8512947	1.2328923	Torrejón et al.
6	0.1990889	0.1618067	0.4541152	1.2949292	Gupta et al.
7	0.1892155	0.1390201	0.4435687	1.1716516	Palkovskii et al.
8	0.0804139	0.088533	0.2780244	2.182387	Nawab et al.
9	0.0012063	0.0011829	0.0050052	2.0028818	Bandyopadhyay et al.
**	**0.1622408**	**0.1049148**	**0.3600280**	**1.0020560**	**Proposed Model**

6 Conclusions

In this lab report two approaches for plagiarism detection were described. The first method compares suspicious documents against a collection of possible sources, while the second one compares the writing style within a particular document to determine if the text was written by one or more authors.

The third place at the external plagiarism detection competition PAN@2011 was obtained, out of 9 participant teams. The precision of the proposed method, of particular importance at plagiarism detection, is close to perfect, with a score of 0.94. Future work in this task would be to integrate an automatic translator to the system, thus providing a way to detect plagiarism for cross-language tasks. Also, to investigate new ways to improve the total number detections, or recall.

The proposed intrinsic algorithm, which introduces a new variant to compute writing style differences, achieves remarkable results, obtaining the first place at the PAN@2011 competition, almost doubling the score of the second team. The method does not utilize language-dependent features such as verbs or stopwords, thus providing a starting point to experiment with other languages. Nevertheless, it is important to note that in this task, of significant difficulty, much work is still needed. The best score so far has been 0.325, indicating that in the field of writing style modeling new approaches need to be developed.

Last, the proposed intrinsic model was applied to the external corpus. This provided results for a real case scenario were one has no prior information on the suspect documents. The results indicate that the precision is still low in this case, and that a significant part of plagiarized passages are left undetected. Nevertheless, it proves that it still can be usefully as it can be the only way to get plagiarism detection done when no reference collection is available.

Acknowledgment

Authors would like to thank continuous support of "Instituto Sistemas Complejos de Ingeniería" (ICM: P-05-004- F, CONICYT: FBO16; www.isci.cl); and FONDEF project (DO8I-1015) entitled, DOCODE: Document Copy Detection (www.docode.cl). Gabriel

Oberreuter is currently “Becario CONICYT”. Finally, authors would like to thank PAN Competition Organizers for constructing such a great workshop and motivate the development of plagiarism detection techniques.

References

1. Bao, J.P., Shen, J.Y., Liu, X.D., Liu, H.Y., Zhang, X.D.: Semantic sequence kin: A method of document copy detection. In: Dai, H., Srikant, R., Zhang, C. (eds.) PAKDD. Lecture Notes in Computer Science, vol. 3056, pp. 529–538. Springer Berlin / Heidelberg (2004)
2. Braschler, M., Harman, D., Pianta, E. (eds.): CLEF 2010 LABs and Workshops, Notebook Papers, 22-23 September 2010, Padua, Italy (2010)
3. Chow, T.W.S., Rahman, M.K.M.: Multilayer som with tree-structured data for efficient document retrieval and plagiarism detection. Trans. Neur. Netw. 20(9), 1385–1402 (2009)
4. Meyer zu Eissen, S., Stein, B., Kulig, M.: Plagiarism detection without reference collections. In: Decker, R., Lenz, H.J. (eds.) GfKl. pp. 359–366. Studies in Classification, Data Analysis, and Knowledge Organization, Springer Berlin / Heidelberg (2006), http://dblp.uni-trier.de/db/conf/gfkl/gfkl2006.html
5. Hunt, R.: Let’s hear it for internet plagiarism. Teaching Learning Bridges 2(3), 2–5 (2003)
6. Kasprzak, J., Brandejs, M.: Improving the reliability of the plagiarism detection system: Lab report for pan at clef 2010. In: Braschler et al. [2]
7. Oberreuter, G., L’Huillier, G., Ríos, S.A., Velásquez, J.D.: Fastdocode: Finding approximated segments of n-grams for document copy detection: Lab report for pan at clef 2010. In: Braschler et al. [2]
8. Park, C.: In other (people’s) words: plagiarism by university students – literature and lessons. In: Assessment and Evaluation in Higher Education. pp. 471–488. No. 5, Carfax Publishing (2003)
9. Potthast, M., Barrón-Cedeño, A., Eiselt, A., Stein, B., Rosso, P.: Overview of the 2nd international competition on plagiarism detection. In: Braschler, M., Harman, D. (eds.) Notebook Papers of CLEF 2010 LABs and Workshops, 22-23 September, Padua, Italy (2010)
10. Potthast, M., Stein, B., Eiselt, A., Barrón-Cedeño, A., Rosso, P.: Overview of the 1st international competition on plagiarism detection. In: Stein, B., Rosso, P., Stamatatos, E., Koppel, M., Agirre, E. (eds.) SEPLN 2009 Workshop on Uncovering Plagiarism, Authorship, and Social Software Misuse (PAN 09). pp. 1–9. CEUR-WS.org (Sep 2009), http://ceur-ws.org/Vol-502
11. Schleimer, S., Wilkerson, D.S., Aiken, A.: Winnowing: local algorithms for document fingerprinting. In: SIGMOD ’03: Proceedings of the 2003 ACM SIGMOD international conference on Management of data. pp. 76–85. ACM, New York, NY, USA (2003)
12. Seaward, L., Matwin, S.: Intrinsic plagiarism detection using complexity analysis. In: Stein, B., Rosso, P., Stamatatos, E., Koppel, M., Agirre, E. (eds.) SEPLN 2009 Workshop on Uncovering Plagiarism, Authorship, and Social Software Misuse (PAN 09). pp. 56–61. CEUR-WS.org (Sep 2009), http://ceur-ws.org/Vol-502
13. Stamatatos, E.: Intrinsic plagiarism detection using character n-gram profiles. In: Stein, B., Rosso, P., Stamatatos, E., Koppel, M., Agirre, E. (eds.) SEPLN 2009 Workshop on Uncovering Plagiarism, Authorship, and Social Software Misuse (PAN 09). pp. 38–46. CEUR-WS.org (Sep 2009), http://ceur-ws.org/Vol-502
14. Stein, B., Lipka, N., Prettenhofer, P.: Intrinsic plagiarism analysis. Language Resources and Evaluation 45(1), 63–82 (2011)
15. Vapnik, V.N.: The Nature of Statistical Learning Theory (Information Science and Statistics). Springer Berlin / Heidelberg (1999)