# Using Hybrid Similarity Methods for Plagiarism Detection

## Notebook for PAN at CLEF 2013

Yurii Palkovskii, Alexei Belov

Zhytomyr State University, Institute of Foreign Philology
palkovskiy@yandex.ru

**Abstract.** At PAN2013 we decided to focus entirely on Text Alignment subtask. Following our previous experience at PAN2012 and CLINSS2012, we decided to put together the approaches we used in previous year to face the new challenges of PAN2013. This year competition added new way of plagiarism obfuscation via text summarization. This particular feature required represents a wide variety of typical cases of plagiarism in the wild and thus attracted our scientific interest. At this year PAN we put forward two main goals: 1) to develop a unified approach that will allow us to merge results obtained by different analysis methods and then run a unified clusterization algorithm to tackle the problem of granularity and produce clean clusters of suspected plagiarism 2) develop a new method of detecting summarization within the suspected documents. As a starting point at PAN 2013 we utilized the prototype application we developed for PAN 2012 and another application developed for FIRE 2012 (CLINSS task). Two basic approaches are - fingerprinting via 5gramm hashes with variable step as our main method and sliding window TF-IDF weighting score for similarity detection of pre-processed summarization via custom text summarizer. Euclidian distance based clusterization with additional custom filters method was used as our cluster merging technique. During the training stage we used the PAN 2012\PAN2013 provided data and performance measures scripts incorporated with genetic algorithm for best parameter tuning and overall performance. Hardware used (training\ development): 6-core Intel i7990Ex with 6GB RAM PC, Vertex3 SSD. Software used: Windows 7 x64, Visual Studio 2010, .net framework, C#, vb.net. We obtained the 6th overall score at PAN2013 with final p-det 0,6152.

## 1 Introduction

PAN 2013 has put forward a new challenges in plagiarism detection [10]. It has become a scientific occasion focused on the problem of text reuse and plagiarism detection. This year it has become even more demanding and challenging partly because of the usage of new TIRA platform and the requirement to deploy the developed application prototype in its own framework, resulting in the inability to access the test corpus thus allowing to potentially employ the cases of real life plagiarism. One more feature of this year competition we faced is the summarization obfuscation type that required the development of separate additional method of detecting similarities. We considered PAN to be one of the most valuable stimulus to push forward the development of commercially available plagiarism detection solutions in its research and development scope.

## 2 Methods

As it has become a good tradition to start off from the last code we developed for the previous PAN [8,9], we used our PAN 2012 prototype application as a basis for our this year program. As it has been already mentioned our main approach is fingerprinting via 5-gramm hashes with variable step as our main method and sliding window TF-IDF weighting score for similarity detection [3] of pre-processed summarization via custom text summarizer. We use generic .getHash method of .net platform and generic hashtable object as a hashtable to store and search hashes taken from fingerprints of 5-gramm word sequences [6]. We included such preprocessing as trimming, lowercasing, number removal, and fingerprint alphasorting [4,5] to mitigate word reordering. Additionally we utilize reference sections discarding during the preprocessing stage, thus avoiding several issues such as "multiple dots" at the document tail and several other heuristic preprocessing methods. In order to meet the new challenge of summarized obfuscation, we developed a new mechanism based on the comparison of the generated summary of document A to document B via our own summarized based on the most frequent keyword bag-of-words approach that used our comparison engine formerly developed for CLINSS competition in 2012. This method is based on a sliding window comparator model via TF-IDF comparison with ranking function working as a marker for the suspected section.

## 3 Evaluation

This year at PAN we got the 6th overall score with final p-det 0,6152 [10]. One of the issues of PAN series' cross evaluation within years', are significant changes within the plagiarism detection baseline score, due to the constant changes of the developed corpus marked by changes within its qualitative and quantitative distributions. So it is not really feasible for us to compare the results achieved during the previous PAN events. We plan to run a number of tests on previous years corpora to better understand the progress achieved by our application prototype. This is still work in progress from our side and we plan to update your evaluation in near future.

## 4 Conclusion

At PAN 2013 we continue to develop our plagiarism detection application trying to incorporate best approaches. This year we put additional effort to tackle the problem of text summarization and easily integrate the developed technique into our existing framework. We hope that at our next participation to achieve even better overall performance by further research and development.

We would like to thank the organizers of PAN for their assistance and help during our participation in PAN series.

## 5 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)
3. 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)
4. Grozea, C., Popescu, M.: The Encoplot Similarity Measure for Automatic Detection of Plagiarism - Extended Technical Report.
http://brainsignals.de/encsimTR.pdf (Aug 2011)
5. 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]
6. Grozea, C., Popescu, M.: Who's the Thief? Automatic Detection of the Direction of Plagiarism. In: Gelbukh, A.F. (ed.) CICLing. Lecture Notes in Computer Science, vol. 6008, pp. 700–710. Springer (2010)
7. Planas, J., Badia, R.M., Ayguadé, E., Labarta, J.: Hierarchical task based programming with StarSs. International Journal of High Performance Computing 23(3), 284–299 (August 2009)
8. 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)
9. 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]
10. Martin Potthast, Tim Gollub, Matthias Hagen, Martin Tippmann, Johannes Kiesel, Efstathios Stamatatos, Paolo Rosso, and Benno Stein. Overview of the 5th International Competition on Plagiarism Detection. In CLEF 2013 Evaluation Labs and Workshop – Working Notes Papers, September 2013.