Three way search engine queries with multi-feature document comparison for plagiarism detection

## Šimon Suchomel, Jan Kasprzak and Michal Brandejs

## Faculty of Informatics, Masaryk University

### {suchomel, kas, [brandejs}@fi.muni.cz](mailto:brandejs%7d@fi.muni.cz)

In this paper, we describe our approach at the PAN 2012 plagiarism detection competition. The first task was to retrieve a set of candidate source documents that may had served as an original for plagiarism. Our candidate retrieval system is based on extraction of three different types of Web queries with narrowing their execution by skipping certain passages of an input document. We have created queries based on keywords extraction, intrinsic plagiarism detection and headers extraction. We have also compared the performance of constructed queries used during the PAN 2012 test process. In PAN 2012 competition the proposed methods succeeded with competitive amount of plagiarism detected with only a small fraction of used queries compared to the other competitors.

Our detailed comparison system is based on detecting common features of several types (in the final submission, we have used two types of features: sorted word 5-grams and unsorted stop word 8- rams) in the input document pair. We have proposed a method of computing so called valid intervals from those features, represented by their offset and length attributes in both source and suspicious document. Previous works use the feature ordering as the measure of distance, which is not usable for multiple types of features, which do not have any natural ordering. From those valid intervals we compute final detections in the post-processing phase, where we merge neighbouring valid intervals and remove some types of overlapping detections. We further discuss other approaches we have explored, but which have not been used in our final submission. In the paper we also discuss the performance aspects of our program, parameter settings and the relevance of current PAN 2012 rules (including the plagdet score) to the real-world plagiarism detection systems.