Improving plagiarism detection
Notebook for PAN at CLEF 2013

Šimon Suchomel, Jan Kasprzak, and Michal Brandejs

Faculty of Informatics, Masaryk University
{suchomel,kas,brandejs}@fi.muni.cz

Abstract This paper describes approaches used for the Plagiarism Detection task in PAN 2013 international competition on uncovering plagiarism, authorship, and social software misuse.

1 Introduction

The notebooks shall contain a full write-up of your approach, including all details necessary to reproduce your results.

2 Source Retrieval

The source retrieval is a subtask in a plagiarism detection process during which only a relatively small subset of documents are retrieved from the large corpus. Those candidate documents are usually further compared in detail with the suspicious document. In the PAN 2013 source retrieval subtask the main goal was to identified web pages which have been used as a source of plagiarism for creation of the test corpus. The test corpus contained XX documents each discussing one and only one theme. Those documents were created intentionally by semiprofessional writers, thus they feature nearly realistic plagiarism cases. Such conditions are similar to a realistic plagiarism detection scenario, such as for state of the art commercial plagiarism detection systems or the anti-plagiarism service developed on and utilized at the Masaryk University. The main difference between real-world corpus of suspicious documents such as for example corpus created from theses stored in Information System of Masaryk University and the corpus of suspicious documents used during the PAN 2013 competition is that in the PAN corpus each document contains plagiarism passages. Therefore we can deepen the search during the process in certain parts of the document where no similar passage has yet been found. This is the main idea of improving recall of detected plagiarism in a suspicious document.

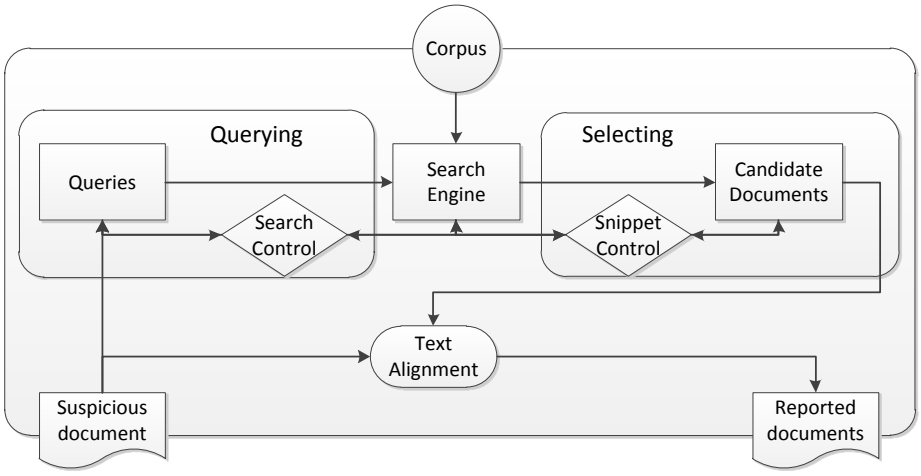

Figure 1. Source retrieval process.

An online plagiarism detection can be viewed as a reverse engineering task where we need to find original documents from which the plagiarized document was created. During the process the plagiarist locates original documents with the use of a search engine. The user decides what query the search engine to ask and which of the results from the result page to use. In real-world scenario the corpus is the whole Web and the search engine can be a contemporary commercial search engine which scales to

the size of the Web. This methodology is based on the fact that we do not possess enough resources to download and effectively process the whole corpus. In the case of PAN 2013 competition the corpus of source documents is the ClueWeb [1] corpus. As a document retrieval tool for the competition we utilized the ChatNoir [1] search engine which indexes the English subset of the ClueWeb. The reverse engineering decision process reside in creation of suitable queries on the basis of the suspicious document and in decision what to actually download and what to report as a plagiarism case from the search results.

These first two stages can be viewed in figure 1 as Querying and Selecting. Selected results from the search engine are forthwith textually aligned with the suspicious document (see section 3 for more details). This is the last decision phase – what to report. If there is any continuous passage of reused text detected, the result document is reported and the continuous passages in the suspicious document are marked as 'discovered' and no further processing of those parts is made.

2.1 Querying

Querying means to effectively utilize the search engine in order to retrieve as many relevant documents as possible with the minimum amount of queries. We consider the resulting document relevant if it shares some of text characteristics with the suspicious document.

We used 3 different types of queries [2]: i) keywords based queries, ii) intrinsic plagiarism based queries, and iii) paragraph based queries. Three main properties distinguish each type of query: i) Positional; ii) Phrasal; iii) Deterministic.

Keywords Based Queries

Intrinsic Plagiarism Based Queries

Paragraph Based Queries

2.2 Search Control

2.3 Result Selection

2.4 Snippet Control

[1] http://lemurproject.org/clueweb09.php/

[2] We used similar three-way based methodology in PAN 2012 Candidate Document Retrieval subtask. However this time we completely replaced the headers based queries with paragraph based queries, since the headers based queries did not pay off in the overall process.

3 Text Alignment

4 Conclusions

References

1. Potthast, M., Hagen, M., Stein, B., Graßegger, J., Michel, M., Tippmann, M., Welsch, C.: ChatNoir: A Search Engine for the ClueWeb09 Corpus. In: Hersh, B., Callan, J., Maarek, Y., Sanderson, M. (eds.) 35th International ACM Conference on Research and Development in Information Retrieval (SIGIR 12) (Aug 2012)