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 58 documents each discussing one and only one theme. Those documents were created intentionally by semiprofessional writers, thus they feature nearly realistic plagiarism cases [5]. 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.

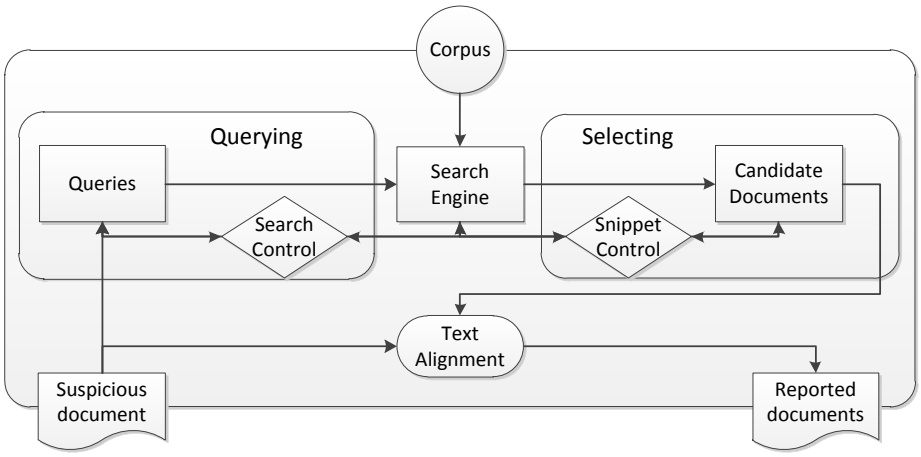

Figure 1. Source retrieval process.

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 [4] 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 done.

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. Positional queries carry extra information about a textual interval in the suspicious document which the query represents. A phrasal query aims for retrieval of documents containing the same small piece of a text. They are usually created from closely coupled words. Deterministic queries for specific suspicious document are always the same no matter how many times we run the software. On the contrary the software can create in two runs potentially different nondeterministic queries.

Keywords Based Queries. The keywords based queries compose of automatically extracted keywords from the whole suspicious document. Their purpose is to retrieve documents concerning the same theme. Two documents discussing the same theme usually share a set of overlapping keywords. Also the combination of keywords in query matters. As a method for automated keywords extraction, we used a frequency based approach described in [6]. The method combines term frequency analysis with TF-IDF score [3]. As a reference corpus we used English web corpus [1] crawled by SpiderLink [7] in 2012 which contains 4.65 billion tokens.

Each keywords based query were constructed from five top ranked keywords consecutively. Each keyword were used only in one query. Too long keywords based queries would be over-specific and it would have resulted in a low recall. On the other hand

[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.

having constructed too short (one or two tokens) queries would have resulted in a low precision and also possibly low recall since they would be too general.

In order to direct the search more at the highest ranked keywords we also extracted their most frequent two and three term long collocations. These were combined also into queries of 5 words. Resulting the 4 top ranked keywords alone can appear in two different queries, one from the keywords alone and one from the collocations. Collocation describes its keyword better than the keyword alone.

The keywords based queries are non-positional, since they represent the whole document. They are also non-phrasal since they are constructed of tokens gathered from different parts of the text. And they are deterministic, for certain input document the extractor always returns the same keywords.

Intrinsic Plagiarism Based Queries. The second type of queries purpose to retrieve pages which contain similar text detected as different, in a manner of writing style, from other parts of the suspicious document. Such a change may point out plagiarized passage which is intrinsically bound up with the text. We implemented vocabulary richness method which computes average word frequency class value for a given text part. The method is described in [2]. The problem is that generally methods based on the vocabulary statistics work better for longer texts. According to authors this method scales well for shorter texts than other text style detection methods. Still the usage is in our case limited by relatively short texts. It is also difficult to determine what parts of text to compare. Therefore we used sliding window concept for text chunking with the same settings as described in [6].

A representative sentence longer than 6 words was randomly selected among those that apply from the suspicious part of the document. An intrinsic plagiarism based query is created from the representative sentence leaving out stop words.

The intrinsic plagiarism based queries are positional. They carry the position of the representative sentence in the document. They are phrasal, since they represent a search for a specific sentence. And they are nondeterministic, because the representative sentence is selected randomly.

Paragraph Based Queries. The purpose of paragraph based queries is to check some parts of the text in more depth. Parts for which no similarity has been found during previous searches.

For this case we considered a paragraph as a minimum text chunk for plagiarism to occur. It is discussible whether a plagiarist would be persecuted for plagiarizing only one sentence in a paragraph. Also a detection of a specific sentence is very difficult if want to avoid exhaustive search approach. If someone is to reuse some peace of continuous text, it would probably be no shorter than a paragraph. Despite the fact, that paragraphs differ in length, we represent one paragraph by one query.

The paragraph based query was created from each paragraph of a suspicious document. From each paragraph we extracted the longest sentence from which the query was constructed. Ideally the extracted sentence should carry the highest information gain. The query was maximally 10 words in length which is the upper bound of ChatNoir and was constructed from the selected sentence by omitting stop words.

2.2 Search Control

For each suspicious document we prepared all three types of queries during the first phase at once. Queries were executed stepwise. After processing each query the results were evaluated (see the following subsection **??** for more details) and from all textual similarities between each result and the suspicious document, the suspicious document intervals of those similarities were marked as 'discovered'. At first the keywords based queries. All of the keywords based queries were always executed. After having all the keywords based queries processed, the intrinsic plagiarism based queries were executed according to their creation sequence. Since they carry its position not all of the intrinsic plagiarism based queries were caried out. During the execution, if any of the query position intersected with any of the 'discovered' interval, the query was dropped out. In the same way, the last paragraph based queries were processed.

This search control results in two major properties. Firstly, the source retrieval effort were increased in parts of the suspicious document, where there have not yet been found any textual similarity. Especially by the paragraph based queries. And secondly, after detection a similarity for a certain part of the text, no more intentionally retrieval attempts for that part were effectuated. Meaning that all discovered search engine results were evaluated, but there were executed no more queries regarding that passage.

2.3 Result Selection

2.4 Snippet Control

2.5 Source Retrieval Results

3 Text Alignment

4 Conclusions

References

1. Sketch Engine EnTenTen Corpus. http://trac.sketchengine.co.uk/wiki/Corpora/enTenTen (2012)
2. Eissen, S.M.Z., Stein, B.: Intrinsic plagiarism detection. In: Proceedings of the European Conference on Information Retrieval (ECIR-06) (2006)
3. Manning, C.D., Raghavan, P., Schütze, H.: Introduction to Information Retrieval. Cambridge University Press, Cambridge, UK (2008)
4. 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). p. 1004. ACM (Aug 2012)
5. Potthast, M., Hagen, M., Völske, M., Stein, B.: Crowdsourcing interaction logs to understand text reuse from the web. In: 51st Annual Meeting of the Association of Computational Linguistics (ACL 13) – (to appear). ACM (Aug 2013)
6. Suchomel, Š., Kasprzak, J., Brandejs, M.: Three way search engine queries with multi-feature document comparison for plagiarism detection. In: Forner, P., Karlgren, J., Womser-Hacker, C. (eds.) CLEF (Online Working Notes/Labs/Workshop) (2012)
7. Suchomel, V., Pomikálek, J.: Efficient web crawling for large text corpora. In: Kilgarriff, A., Sharoff, S. (eds.) Proceedings of the seventh Web as Corpus Workshop (WAC7). pp. 39–43 (2012)