Heterogeneous Queries for Synoptic and Phrasal Search

Notebook for PAN at CLEF 2014

Šimon Suchomel and Michal Brandejs

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

Abstract This paper describes an architecture of the source retrieval system used at PAN 2014 lab on uncovering plagiarism, authorship, and social software misuse. The system is based on the systems used in past years at PAN 13 [6] and PAN 12 [5]. The majority of features were adapted with some improvements described in this paper. The source retrieval subsystem forms an integral part of a modern system for plagiarism discovery.

1 Introduction

Systems which compute similarities between documents can significantly help with plagiarism detection. They automate the tedious work such as locating possible sources of plagiarism and finding similar text passages. If the suspicious passages are highlighted by the system, the supervisor only checks whether a passage is a plagiarism case or not. The state of the art anti-plagiarism systems evaluate document similarities in order to select suspicious passages of a source document. This evaluation is refered in PAN as the document alignment. Documents are algorithmically aligned to a corpus of known documents. However, if the corpus does not contain the original document the similarity between the original and the suspicious document can not be detected. Therefore a potential source documents should be retrieved from all documents prior to text alignment calculations. The corpus of all documents is usually very large, for example the entire web, and a search engine is utilized as a retrieval tool. It is ideally utilized automatically in the same manner as plagiarizing users would do manually. The global view of the system for unoriginal text detection is depicted in figure 1.

With the usage of a given search engine, the problem of source retrieval is then reduced to the problem of combining proper queries and passing them to the search engine. Selecting and downloading search engine results also influence total performance of the source retrieval system. The queries pose the most expensive piece of operation, whereas the downloads are quite cheap. In the real-world scenario, we are often limited by a number of queries executed in a given time period or by a total number of queries per document. During the operation the system should maximize the recall and precision of retrieved results and also minimize the total number of executed queries as much as possible.

The following sections describe the key parts of the system for source retrieval used at PAN 2014. More information about the task and the competition can be found in the task overview written by the lab organizers [2].

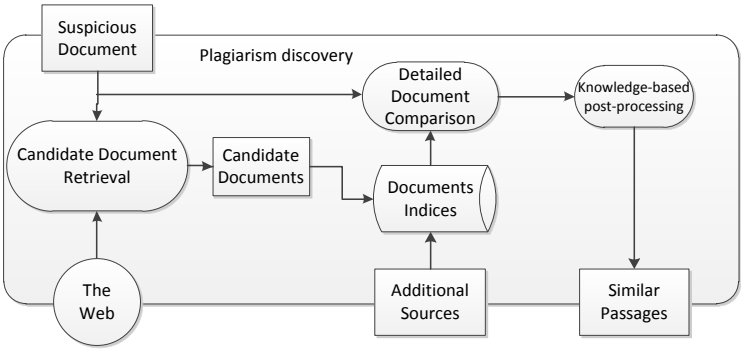

Figure 1. A global view of a modern anti-plagiarism software.

For obtaining the search results two search engines have been utilized: The Chatnoir [3] search engine for queries based on extracted keywords; and the Indri [4] search engine for combined queries and phrases. Both search engines index the ClueWeb09 corpus which served as a main external corpus for document retrieval. The software were executed and evaluated via TIRA framework [1] on a test corpus of selected english-written and plagiarized documents.

2 Building of the Queries

Several types of queries were prepared.This year we combined keywords-based queries, paragraph-based queries and headers-based queries together. Some of the prepared queries could be discarded from the execution, no query refinement was applied according to the results, but some of the top scored keywords could appear in the different combination in more than one query.

2.1 Keywords-based Queries

From the whole suspicious document, there were extracted keywords using TF-IDF scoring of lemmas created via Python NLTK lemmatizer and omitting english stopwords. Firstly we created so called pilot query from top scored six terms. This query was passed to both Chatnoir and Indri search engines with Indri setting for combine belief operator of the query.

Based on the three top-scored single-term keywords, their collocations of two more words were extracted. Those served as phrasal search queries, which were passed to Indri in order to lookup a contextual occurrence of the selected keywords. All other subsequent keywords-based queries were passed to Chatnoir only and they were created by combining collocations of top scored keywords together with the rest of the extracted keywords up to 6 tokens long. From the rest of the keywords, if any, the remaining six-term long queries were created. The total number of used keywords-based queries depended on each document characteristics. From some of the documents, there were

identified fewer keywords than from others. Usually there were prepared around 10 keywords-based queries per document.

2.2 Paragraph-based Queries

From each paragraph of the suspicious document, a single paragraph-based query was created. The longest sentence from the specific paragraph was used to build the query. From the selected sentence six subsequent terms were selected from a random position within the sentence. The query was created from those six terms, only the punctuation was removed. The resulting query was passed to the Indri search engine with proximity term number of 1 denoting the phrasal search.

2.3 Headers-based Queries

The headers-based queries were used in the form in which they appeared in the text limited to six words in length. They were passed to Indri as a phrasal query as well. For discovering the headers in the text the approach adopted from Suchomel et al. 2012 [5] was used with no modifications.

2.4 Chunking

Three types of text chunking were applied: sentence and word chunking for keywords extraction; headers detection; and paragraph chunking. For each type of query the corresponding chunking method was always applied on the whole document from the beginning.

3 Search Control

All prepared queries were processed according to their priority. Starting with the keywords-based, then the paragraph-based and last the header-based queries. Only all keywords-based queries – the pilot query, collocation queries and remaining keywords-combined queries – were executed for each suspicious document. After each query all its results were processed and positions of discovered similarities were stored. To each subsequent (not keywords-based) query its position was also attached, if that position collided with any of already found similarities, the query was omitted from the queue of prepared queries.

4 Downloading the Results

For each search engine result a snippet based on the given query can be obtained prior to the full document download. The snippet is obtained via ChatNoir API, it is considered as no operation and it provides rough information about the document content. It contains a portion of the document up to 500 characters around a given textual string. We generated the snippet for all documents from results based on specific query for each

term in that query. The all snippets from one document were concatenated together and its tuples of two terms were compared with tuples from the suspicious document. Concordance of the tuples of 20 % or more was the threshold for decision about the document download.

The downloaded results were textually aligned to the suspicious document using feature type selection for computing similarities described in Suchomel, Kasprzak et al. 2013 [6]. If any similarity were detected, the document were reported as a potential source of plagiarism. Thus all the reported documents contain at least some similarity with the suspicious document.

5 Conclusions

This paper described the key aspects and changes from our erstwhile systems for candidate document retrieval used at PAN 14 lab on uncovering plagiarism. The architecture stems from PAN 12 and PAN 13 labs and the real-world anti-plagiarism system which is in use at Masaryk University. The results of the PAN show that this approach is one of the best for a real-life adoption, since it achieved a decent recall with just a fraction of used queries. Such approach is applicable for detection of suspicious texts, which may contain plagiarism, that can then be selected for further investigation.

References

1. Gollub, T., Potthast, M., Beyer, A., Busse, M., Pardo, F.M.R., Rosso, P., Stamatatos, E., Stein, B.: Recent trends in digital text forensics and its evaluation - plagiarism detection, author identification, and author profiling. In: Forner, P., Müller, H., Paredes, R., Rosso, P., Stein, B. (eds.) CLEF. Lecture Notes in Computer Science, vol. 8138, pp. 282–302. Springer (2013)
2. Potthast, M., Gollub, T., Hagen, M., Tippmann, M., Kiesel, J., Stamatatos, E., Rosso, P., Stein, B.: Overview of the 5th international competition on plagiarism detection. In: CLEF 2013 Evaluation Labs and Workshop (September 2013)
3. 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)
4. Strohman, T., Metzler, D., Turtle, H., Croft, W.B.: Indri: a language-model based search engine for complex queries. Tech. rep., in Proceedings of the International Conference on Intelligent Analysis (2005)
5. 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). pp. 1–8 (2012)
6. Suchomel, Š., Kasprzak, J., Brandejs, M.: Diverse queries and feature type selection for plagiarism discovery. In: CLEF 2013 Evaluation Labs and Workshop. pp. 1–8 (2013)