Source Retrieval via Naïve Approach and Passage Selection Heuristics

Notebook for PAN at CLEF 2013

Ondřej Veselý, Tomáš Foltýnek, Jiří Rybička

Department of Informatics, Faculty of Business and Economics, Mendel University in Brno
xorwen@gmail.com, foltynek@pef.mendelu.cz, rybicka@mendelu.cz

Abstract. Our retrieval system tries to extract the most relevant passages from inspected text. It combines naive approach consisting of gradually increasing number of words in the search query, with simplified pre-suspiciousness index heuristics. Selected passages are used to form a search engine request queries. URLs from obtained results are then weighted and finally downloaded

1 Introduction

Potthast (2009) stated, that this given task is usually divided into following subtasks. "(1) Chunking, (2) key phrase extraction, (3) query formulation, (4) search control, and (5) download filtering."
The most specific method we used to deal with query formulation and chunking subtask is called naive approach. However, naive approach was optimized to reduce the number of search engine queries.
This paper also covers a key phrase extraction subtask, which was solved by heuristics, where the search engine results were evaluated according to given metrics to prioritize some parts of the document to be checked.

2 Optimal query size -- naive approach

The algorithm of query extraction is based on an idea of naive approach described by (Veselý, 2012) and (Veselý et. al., 2012): "The most precise results can be obtained when the phrase size is determined dynamically by querying the search engine with increasing length of selected phrase and use the last non-empty result set". That means we query the search engine with one word, two words, etc. After each query the number of results is examined. The more words in query, the less results. The query leading to the last non-zero number is then states as optimal and the set of obtained URLs is saved for further examination. After that, we start new iteration with the word following the optimal search query. However, this approach leads to

huge number of queries - equal to the number of words in the document. Hence, we have made some modifications to optimise the word count per query ratio:

- Starting query length is set to five words;
- Length of the query is increased by step of two words (not one);
- When number of results is lower than 300, this query is treated as optimal;
- In case of zero results, the previous query is considered as optimal.

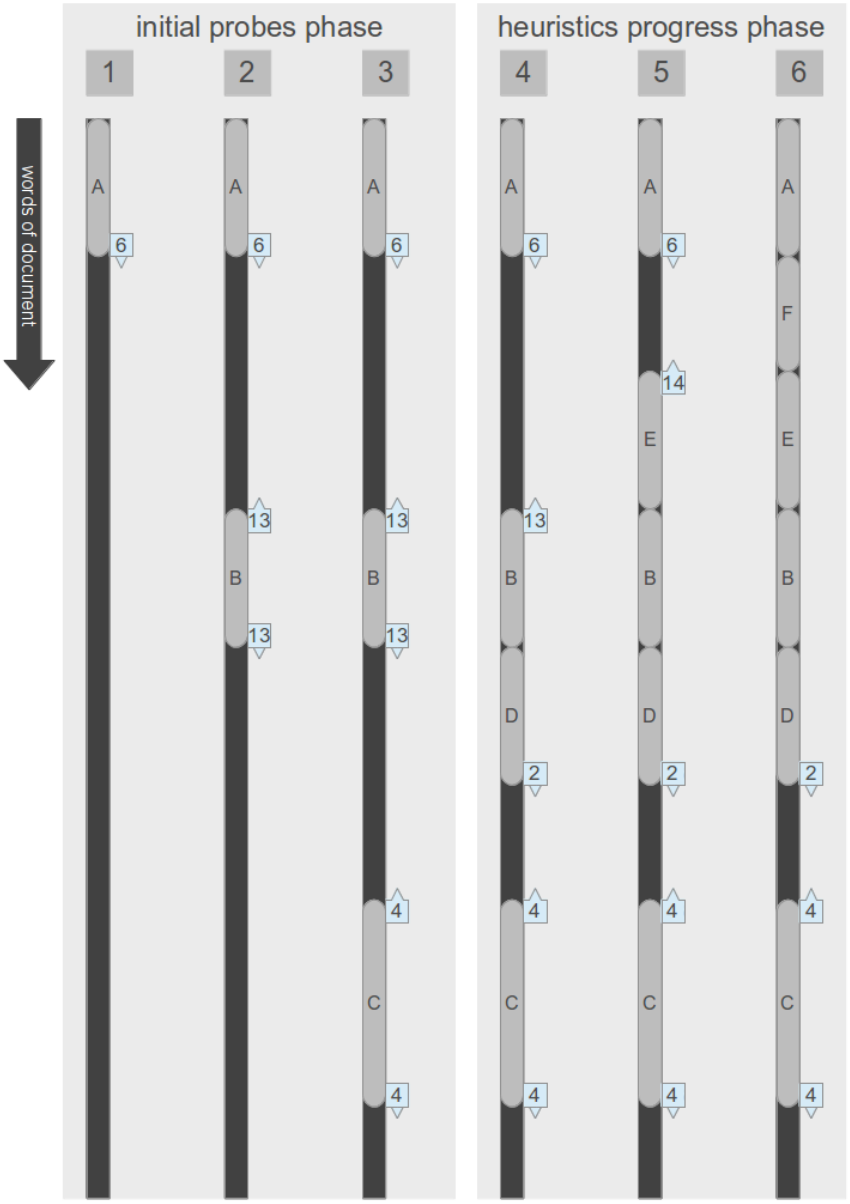

Figure 1: Heuristics for passage selection

These numbers (5 starting words, increasing by 2, boundary of 300 results) was set experimentally. However, these parameters are able to balance the number of queries, performance and precision.
ChatNoir results are provided with a lot of values accompanying obtained URLs. During the experiment, all of them were tested to find the optimal value for weighting the URLs. Some of them seemed to be negatively correlating with desired weight. In the final solution only the weight value was used.
The weight value was summarised for each URL obtained among search query results. At the end of the analysis, the content of 15 URLs with highest weight was downloaded. The value of 15 was also set experimentally and equally as previous parameters may influence final results, namely precision and recall.

Heuristics for passages selection

Previous naive approach was useful for determining the optimal query size and therefore to create set of suspicious URLs. Nonetheless, examining each part of the document would be too demanding, so we do not examine whole document, but just its relevant passages, which are identified by heuristics based on the pre-suspiciousness index. This index is also described in (Veselý, 2012) and (Veselý et. al., 2012). The examination of the documents works in two phases:

1. Initial probe phase
2. Heuristic progress phase

In initial probe phase, we make probes after the length of 100 words. Every probe consists of finding an optimal query (via the optimized naive approach described above). The pre-suspiciousness index is then equal to the length of optimal query. The length of optimal query correlates with the probability, that surrounding passage of the probe is potentially plagiarised. This index is calculated for each probe.
As every probe gives us the pre-suspiciousness index of not probed gaps between checked passages, at the second phase the gaps are probed at the order of pre-suspiciousness index value (see Figure 1). When 20% of the document's words are sent to the search engine, the algorithm starts downloading the sources. This approach allows us to skip the majority of words, where the potential plagiarism is unlikely.
There was one relevant modification against the heuristics published before (Veselý 2012) - we gave up determining the pre-suspiciousness index for every word; just for surroundings of previously checked passages.

Results

We may note that some results are directly connected with the method we have used. E.g. number of queries to the first detection is maximal in our case, because we first execute all queries and then proceed to download. We do not consider this "worst place" as serious, on contrary, we are pleased to be 4th in the number of downloads to the first detection.

Discussion

The result of the for detection plagiarism via search engines are highly dependant of the set of services provided by used search engine. For this competition, we were obliged to use ChatNoir (Potthast 2012). The main difference between ChatNoir and Seznam.cz (which was used during our previous experiments) was, that ChatNoir does not provide exact phrase searching (using quotes). Consequently, we had to omit these results, which we were used to intersect with results obtained by searching of unquoted words.
In our further research, we plan to compare different search engines (ChatNoir, Seznam.cz, Yahoo!, Google) to find out how exactly the results are influenced by using particular search engine.

Conclusion

As we can see from the results, our software is slightly above the average of all competitors. After this experience, we would like to employ other methods to improve our heuristics and possibly modify mentioned parameters to obtain better results in categories we may focus to.

Sources

Veselý, O., Kolomazník, J., Foltýnek, T.: Heuristic and AI approach to optimize plagiarism detection tool using a public search engine. [CD-ROM]. In *IADIS International conference WWW/Internet 2012*. s. 309--403. ISBN 978-989-8533-09-8.

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.

Martin Potthast, Matthias Hagen, Benno Stein, Jan Graßegger, Maximilian Michel, Martin Tippmann, and Clement Welsch. ChatNoir: A Search Engine for the ClueWeb09 Corpus. In *Bill Hersh, Jamie Callan, Yoelle Maarek, and Mark Sanderson (ed.), 35th International ACM Conference on Research and Development in Information Retrieval (SIGIR 12)*, page 1004, Portland, Oregon, August 2012

Potthast, M. et al. "Overview of the 4th International Competition on Plagiarism Detection." *Pamela Forner, Jussi Karlgren und Christa Womser—Hacker (Hg.): CLEF* 20 (2009): 72.

Veselý, O. Similarity Analysis of Theses and Online Documents. Diploma thesis. Mendel University in Brno, 2013.