Source Retrieval via Naïve Approach and Passage Selection Heuristics

Ondřej Veselý, Tomáš Foltýnek, and Jiří Rybička

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

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.

The naive approach algorithm of query extraction is based on querying the search engine with increasing length of selected phrase and using 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 stated as optimal and the set of obtained URLs is saved for further examination.

We have made some modifications to optimize 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.

Besides that, we use the Passage selection heuristics. The examination of the documents works in two phases:

* Initial probe phase
* 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. When 20% of the document’s words are sent to the search engine, the algorithm starts downloading the sources.

The result of the for detection plagiarism via search engines are highly dependant of the set of services provided by used search engine.