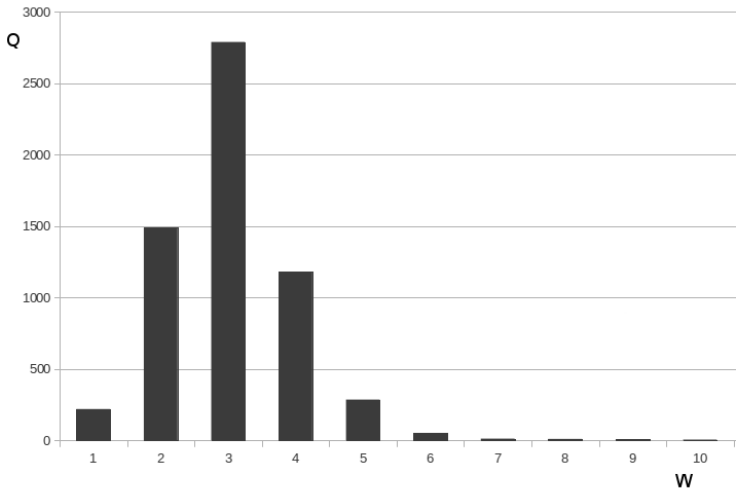
Source Retrieval via Naive Approach and Passage Selection Heuristics

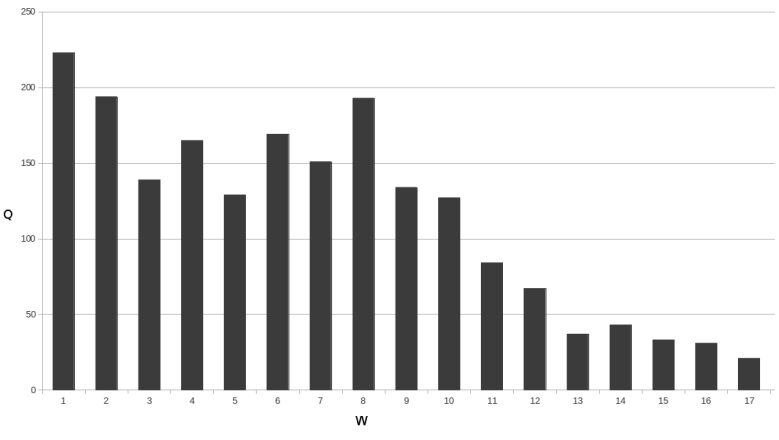
Ondřej Veselý, Tomáš Foltýnek, Jiří Rybička

Mendel University in Brno

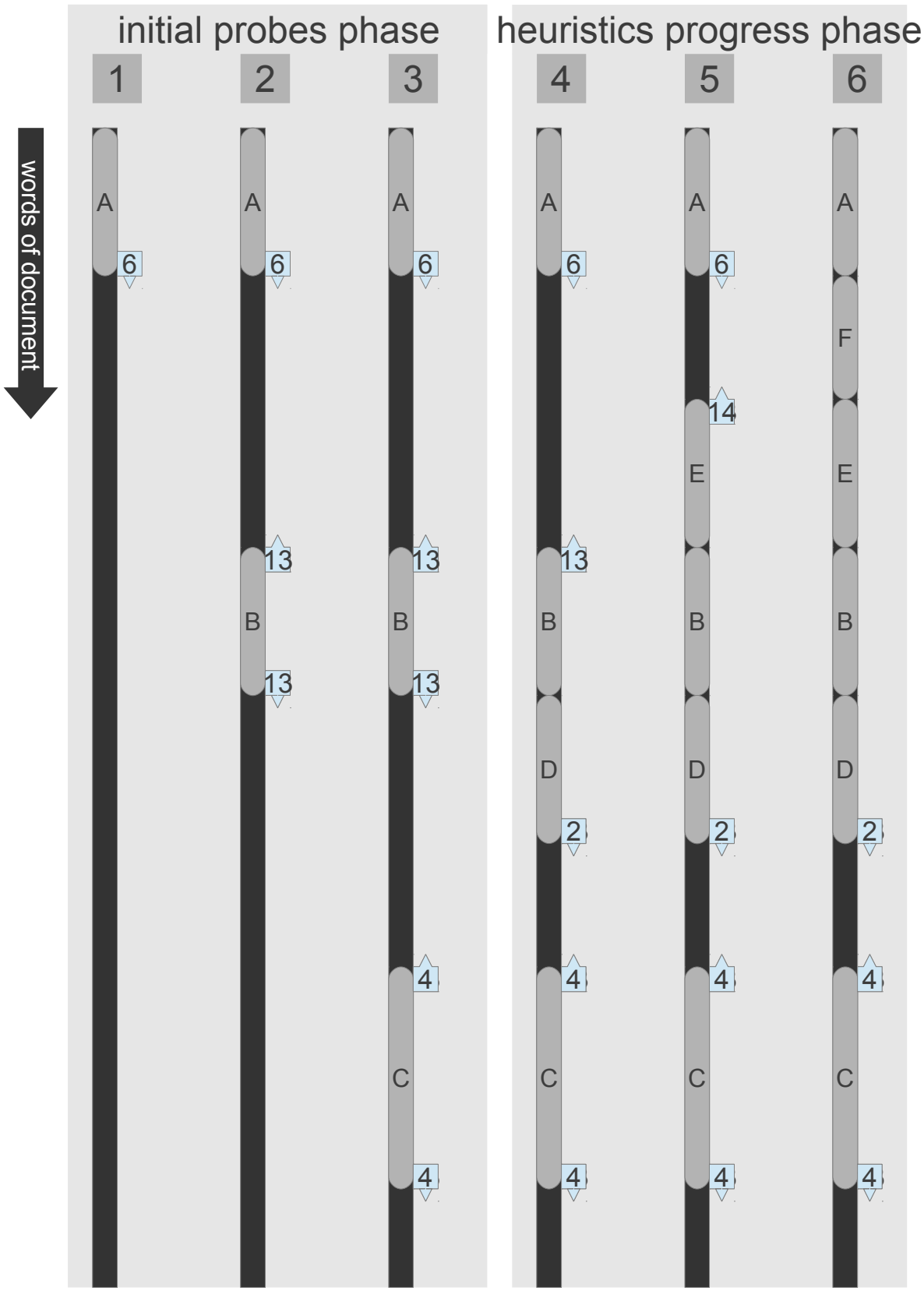
Abstract. Our retrieval system tries to extract the most relevant passages from inspected text. It combines naive approach consisting of gradually increasing number of words in the search query, with simplified pre-suspiciousness index heuristics. Selected passages are used to form a search engine request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



The first phase of algorithm we called **naive approach** is based on an idea of naive approach: "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". (Veselý 2012)
The longest phrase with non-zero result called "optimal phrase" is used for analysis.
The graph above shows the distribution optimal phrases in phrasal search.
The graph above shows the distribution optimal phrases in non-phrasal search. Both measured with Seznam.cz search engine



All connected work and sources could be found at *antiplagiator.orwen.org/en/*



Heuristics. 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.