Robust Selective Search

Yubin Kim
Language Technologies Institute
Carnegie Mellon University
yubink@cmu.edu

Abstract

Selective search is a modern distributed search architecture designed to reduce the computational cost of large-scale search. Selective search creates topical shards that are deliberately content-skewed, placing highly similar documents together in the same shard. During query time, rather than searching the entire corpus, a resource selection algorithm selects a subset of the topic shards likely to contain documents relevant to the query and search is only performed on these shards. This substantially reduces total computational costs of search while maintaining accuracy comparable to exhaustive distributed search.

Prior work has shown selective search to be effective in smaller scale, single query-at-a-time environments. However, modern practical, large-scale search and text analysis systems are often multi-stage pipeline systems where an initial, first-stage fast candidate retrieval forwards results onto downstream complex analysis components. These systems often contain other optimization components and are run in a parallel setting over multiple machines. This dissertation aims to bring selective search to wider adoption by addressing the questions related to efficiency and effectiveness in a practical implementation such as: do different instantiations of selective search have stable performance; does selective search combine well with other optimization components; can selective search deliver the high recall necessary to serve as a first-stage retrieval system? In addition, this dissertation provides tools to empower system administrators so that they can easily design and test selective search systems without full implementations.

First, this dissertation research investigates the effects of non-deterministic steps that exist in selective search on its accuracy and found the variance across system instances is acceptable. Then, selective search was combined with WAND, a common dynamic pruning algorithm and it was shown that selective search and WAND has *better*-than-additive gains due to the long posting lists of the topically focused shards. This dissertation also presents new resource selection algorithms to achieve high recall, which has been an elusive goal in prior work. A learning-to-rank based approach to resource selection can be trained without human-judged relevance data to be highly accurate and statistically equivalent to exhaustive search at deeper metrics such as MAP@1000. This result enables selective search to be used as a first-stage retrieval component in realistic multi-stage text analysis systems.

When placed in a parallel query processing environment, it was found that with judicious load-balancing to manage unequal popularity of shards, the efficiency claims of prior work remain relevant in a fully parallel processing setting, and are applicable to larger computational environments as well. A detailed simulator was built to investigate this research

question and serve as a powerful tool for administrators to test out different selective search configurations.

Finally, new evaluation metric, AUReC, is presented which can be used to easily evaluate a mapping of documents to shards without implementing a full selective search system. This allows system designers to quickly sift through many potential different document allocations to easily identify the best system configuration.

Ultimately, the dissertation aims to enable cost and energy-efficient use of large-scale data collections in not only information retrieval research, but also in other fields such as text mining and question answering, in academia and industry alike, fueling future innovation.