Language Technologies Institute

Thesis Defense

GHC 6501| Thursday, June 28| 5:00 pm

Robust Selective Search

Yubin Kim

COMMITTEE:

Jamie Callan, (chair)

Jaime Carbonell

Ralf Brown

Alistair Moffat
(Univ. of Melbourne)

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. 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.

boston.lti.cs.cmu.edu/yubink/thesis.pdf