Selective Search of Large Text Collections

Jamie Callan

Language Technologies Institute
Carnegie Mellon University, Pittsburgh, PA, USA
callan@cs.cmu.edu

ABSTRACT. *Information retrieval research has been hampered by the difficulty of conducting research on web datasets of realistic size due to the computational resources required for such datasets. When the text collection is too large for a single machine, its index is divided into 'shards' that are distributed across a computer cluster and searched in parallel, which is effective but expensive. This talk describes an alternative architecture for large-scale text search in which the corpus is decomposed into index shards that are expected to have skewed utility distributions, thus enabling most shards to be ignored for most queries. This selective search architecture is equally effective, but has lower computational costs, which makes it an attractive architecture for organizations that have modest computational resources, modest query traffic, and elastic response time requirements.*

Selective search is a new application of ideas developed originally for distributed, federated, and aggregated search, however this new application of those ideas challenges some of the assumptions and design criteria that motivated prior research. The partitioning process creates text collections, thus inviting research on what characteristics are desired or to be avoided in a text collection to enable accurate search. Most resource selection algorithms are designed to search a static number of resources; however, when the distribution of content across a set of index shards is skewed intentionally, it is important to dynamically estimate the search effort required for each query. Selective search is a cooperative search architecture, which simplifies the problem of merging results from different shards for unstructured queries; however, better methods of merging results for structured queries are still desirable.

This talk provides an introduction to the selective search architectures, and how it relates to prior research on distributed and federated search. It considers the types of problems and environments that might, or might not, benefit from the selective search architecture. It concludes by discussing open research problems and interesting research directions.

KEYWORDS: *Search engine architecture, distributed information retrieval.*

ACKNOWLEDGEMENTS. *Much of the research described in this talk was done with Anagha Kulkarni, and is part of her Ph.D. dissertation. Her research was in part supported by National Science Foundation grant IIS-0916553. Any opinions, findings, conclusions and recommendations expressed in this paper are the author's and do not necessarily reflect those of the sponsor.*