Efficient Query Processing in Distributed Search Engines

Simon Jonassen
Norwegian University of Science and Technology
simonj@idi.ntnu.no

Abstract

Web search engines have to deal with a rapidly increasing amount of information, high query loads and tight performance constraints. The success of a search engine depends on the speed with which it answers queries (efficiency) and the quality of its answers (effectiveness). These two metrics have a large impact on the operational costs of the search engine and the overall user satisfaction, which determine the revenue of the search engine. In this context, any improvement in query processing efficiency can reduce the operational costs and improve user satisfaction, hence improve the overall benefit.

In this thesis, we elaborate on query processing efficiency, address several problems within partitioned query processing, pruning and caching and propose several novel techniques:

First, we look at term-wise partitioned indexes and address the main limitations of the state-of-the-art query processing methods. Our first approach combines the advantage of pipelined and traditional (non-pipelined) query processing. This approach assumes one disk access per posting list and traditional term-at-a-time processing. For the second approach, we follow an alternative direction and look at document-at-a-time processing of sub-queries and skipping. Subsequently, we present several skipping extensions to pipelined query processing, which as we show can improve the query processing performance and/or the quality of results. Then, we extend one of these methods with intra-query parallelism, which as we show can improve the performance at low query loads.

Second, we look at skipping and pruning optimizations designed for a monolithic index. We present an efficient self-skipping inverted index designed for modern index compression methods and several query processing optimizations. We show that these optimizations can provide a significant speed-up compared to a full (non-pruned) evaluation and reduce the performance gap between disjunctive (OR) and conjunctive (AND) queries. We also propose a linear programming optimization that can further improve the I/O, decompression and computation efficiency of Max-Score.

Third, we elaborate on caching in Web search engines in two independent contributions. First, we present an analytical model that finds the optimal split in a static memory-based two-level cache. Second, we present several strategies for selecting, ordering and scheduling prefetch queries and demonstrate that these can improve the efficiency and effectiveness of Web search engines.

We carefully evaluate our ideas either using a real implementation or by simulation using real-world text collections and query logs. Most of the proposed techniques are found to

improve the state-of-the-art in the conducted empirical studies. However, the implications and applicability of these techniques in practice need further evaluation in real-life settings.

This dissertation was completed at the Department of Computer and Information Science at the Norwegian University of Science and Technology (NTNU) under advise of Prof. Svein Erik Bratsberg, Dr. Øystein Torbjørnsen and Dr. Magnus Lie Hetland. Some of the work was done in collaboration with Yahoo! Research Barcelona and mentored by Prof. Ricardo Baeza-Yates and Dr. B. Barla Cambazoglu. Prof. Alistair Moffat (University of Melbourne), Dr. Christina Lioma (University of Copenhagen) and Prof. Kjell Bratsbergsengen (NTNU) served as dissertation committee members.

Available online at: `http://www.idi.ntnu.no/research/doctor_theses/simonj.pdf`