Query Scheduling Techniques and Power-Latency Trade-off Model for Large-Scale Search Engines

Ana Freire

University of A Coruña

Spain

ana.freire@udc.es

Abstract

Web search engines have to deal with a huge increase of information, demanded by high incoming query traffic. This situation has driven companies to build large, geographically distributed data centres housing thousands of servers and consuming enormous amounts of electricity. At this scale, even minor efficiency improvements may result in large financial and power savings.

This thesis represents a novel contribution to the state-of-the-art of Query Scheduling and *Green Information Retrieval (Green IR)*, by assisting large-scale data centres to build more efficient and environmentally-friendly search engines.

The main contributions of this work are the following:

- **Query Scheduling**. We introduce query efficiency predictors as suitable estimators to improve Query Scheduling. We estimate the processing time of the queries waiting in each query server and we calculate an approximate time that a new query must spend in each queue. Based on this estimation, the fastest query server is selected.

- **Green IR**. Once we have developed new methods to improve the average response time of a search engine, we focus on reducing the power consumption of the whole system. This thesis proposes a mathematical model that establishes a trade-off between latency and power consumption. This model attempts to automatically adapt the number of active servers in the system based on the fluctuations of a daily query traffic flow.

- **Queueing Theory**. We prove the limitation of Queueing Theory models for estimating the latency in search engines. As a consequence, we develop our trade-off model by predicting the latency using historical data. Results show the good performance of this approach.

- **IR evaluation**. We attest that Simulation platforms are suitable for IR experimentation. We support this conclusion by establishing an exhaustive analysis of the current IR evaluation platforms.

Supervisor: Fidel Cacheda (University of A Coruña, Spain).
Available Online at `http://hdl.handle.net/2183/11922`.