Query Cost Prediction using Machine Learning

At Meltwater we are creating the next generation media monitoring solutions that helps our customers gain crucial insights from what is happening on the Internet. At the heart of this is a massive Elasticsearch cluster with 34B indexed documents using 450TB of storage. In order to provide the most relevant data for the customers, queries can become very complex. If not constructed properly the queries can also become very resource consuming which can affect other customers and cause unacceptable latencies or even downtime.

Our current strategy when evaluating the cost of a query is based on heuristics of a set of features from the queries, such as the number of wildcard, term and near queries and other topological properties such as the deepest nesting of the query. This helps us to detect some of the expensive queries but not all of them.

The aim of this thesis is to evaluate and implement a framework to predict the cost of executing a search query using a supervised learning approach. The work will involve all steps from the feature modelling and extraction to the implementation of a learning algorithm. The data set is composed of logs and system metrics extracted from our production environment which serve 1.5M queries per day. We hope to gain valuable insight into what drives the cost of evaluating a query. Furthermore we desire to learn how to make queries that are less expensive but deliver similar results, and automatically recommend improvements to our customer’s stored queries.

This thesis work is suitable for two students with an interest in computer science in general and machine learning in particular. Due to the nature of the problem, some knowledge of search engines, query optimization and distributed systems is preferred. Additionally, some experience working in a Linux environment using open source tools is required.

For more information, reach out to Otto Frost at otto.frost@meltwater.com