Single Query Optimisation is the Root of All Evil

J. Shane Culpepper
RMIT University
Melbourne, Australia

Overview. The demise of the age of one-shot web query optimisation is nigh. For Information Retrieval researchers and search engine engineers, this is a time to rejoice, as new opportunities to revisit old techniques are once again upon us. For years, search systems have tried to infer the intentions of a user using only a few (sometimes) carefully selected search terms. However, the classic search interface (the web browser) on a computer will soon be obsolete. Instead users will find information through mobile devices, and conversational search systems such as Alexa, Cortana, or Siri. These interfaces provide direct access to relevance feedback mechanisms from searchers, and allow new opportunities to model state instead of depending on only a single query. In this abstract, we argue that now is the time for IR researchers to once again return to building relevance models for information needs, and stop thinking in terms of one-off queries. We show that simple combinations of classic techniques along with multiple representations of a single information need can easily outperform state-of-the-art models which perform optimisations on a query-by-query basis. This is a simple first step in the right direction.

Problem. The pitfalls of over-optimising a complex multi-stage retrieval system for a single query is rarely considered by search engine designers. Recent work by Bailey et al. [1] showed that thinking in terms of queries and not the underlying information need can lead to dramatic variance in system effectiveness, but the authors do not consider the efficiency implications of query variation, or fully explore how higher level modeling of the information need might be accomplished. So, the key research challenge we set in this abstract is:

Research Challenge: How should academics and system designers model and optimise search performance based on information needs and not a single query?

<table>
<thead>
<tr>
<th>Method</th>
<th>NDCG@10</th>
<th>W/T/L</th>
</tr>
</thead>
<tbody>
<tr>
<td>BM25</td>
<td>0.212</td>
<td>-/-</td>
</tr>
<tr>
<td>SDM-Field</td>
<td>0.233</td>
<td>57/3/40</td>
</tr>
<tr>
<td>LambdaMART</td>
<td>0.225</td>
<td>59/2/39</td>
</tr>
<tr>
<td>DoubleFuse, v=all</td>
<td>0.300</td>
<td>80/1/19</td>
</tr>
</tbody>
</table>

Table 1: Effectiveness comparison of three state-of-the-art ranking methods for the most common query variation for each topic from the ClueWeb12B UQV100 collection [1]. Here $^\dagger$ means $p < 0.001$ in a Bonferroni corrected two-tailed t-test.

REFERENCES