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 underling 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?*

Method	NDCG@10	W/T/L
BM25	0.212	-/-/-
SDM-Field	0.233	57/3/40
LambdaMART	0.225	59/2/39
DoubleFuse, v=all	$0.300^{\ddagger}$	80/1/19

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 ‡ means $p < 0.001$ in a Bonferroni corrected two-tailed t-test.

DESIRES 2018, August 2018, Bertinoro, Italy

Table 1 compares three state-of-the-art search systems, with a properly tuned BM25 bag-of-words model as a baseline, using 100 adhoc queries from the ClueWeb12B UQV100 collection [1]. The three systems being compared are BM25, a field-based SDM model [9] (the exact configuration is identical to the one described by Gallagher et al. [7]), a LambdaMART learning-to-rank (LTR) model [4, 5] (here lightGBM is used with 459 features), and double unsupervised fusion [3] (RRF [6] over all UQV query variations and two systems - SDM-Field and BM25). We can see that not only does fusion make more queries better on average, it is also far less likely to make queries *worse*. This can clearly be seen when comparing Wins, Ties, and Losses (W/T/L) in the Table, where a Win or a Loss is for any query that increases or decreases the NDCG@10 score for that topic by 10% or more.

Summary. So, simple fusion over query variations is clearly effective. This has been known for some time [2, 8], particularly on "hard" queries [10]. But system designers generally still focus on learning-to-rank on single queries. How can we as a community step back and learn from over fifty years of research in Information Retrieval as we confront the radical shift from classic web search with ten blue links to interactive search through virtual assistants? Will system designers once again over-commit to optimising for the "current" query, or can we move beyond this paradigm to devise and develop entirely new approaches to search? We face many new challenges as a community – collection construction, open source stateful search systems, evaluation metrics, data privacy – in order to not be left behind by the paradigm shift in the way people search and consume information.

Acknowledgements. This work was supported by the Australian Research Council's *Discovery Projects* Scheme (DP170102231) and a grant from the Mozilla Foundation.

REFERENCES

[1] P. Bailey, A. Moffat, F. Scholer, and P. Thomas. UQV100: A test collection with query variability. In *Proc. SIGIR*, pages 725–728, 2016.
[2] N. J. Belkin, P. Kantor, E. A. Fox, and J. A. Shaw. Combining the evidence of multiple query representations for information retrieval. *Inf. Proc. & Man.*, 31(3): 431–448, 1995.
[3] R. Benham and J. S. Culpepper. Risk-reward trade-offs in rank fusion. In *Proc. ADCS*, pages 1:1–1:8, 2017.
[4] C. Burges. From ranknet to lambdarank to lambdamart: An overview. *Learning*, 11(23-581):81, 2010.
[5] R.-C. Chen, L. Gallagher, R. Blanco, and J. S. Culpepper. Efficient cost-aware cascade ranking in multi-stage retrieval. In *Proc. SIGIR*, pages 445–454, 2017.
[6] G. V. Cormack, C. L. A. Clarke, and S. Buettcher. Reciprocal rank fusion outperforms Condorcet and individual rank learning methods. In *Proc. SIGIR*, pages 758–759, 2009.
[7] L. Gallagher, J. Mackenzie, R. Benham, R.-C. Chen, F. Scholer, and J. S. Culpepper. RMIT at the NTCIR-13 We Want Web task. In *Proc. NTCIR*, 2017.
[8] K-L. Kwok, L. Grunfeld, and P. Deng. Employing web mining and data fusion to improve weak ad hoc retrieval. *Inf. Proc. & Man.*, 43(2):406–419, 2007.
[9] D. Metzler and W. B. Croft. A Markov random field model for term dependencies. In *Proc. SIGIR*, pages 472–479, 2005.
[10] E. M. Voorhees. The TREC robust retrieval track. volume 39, pages 11–20, 2005.