Evaluating Risk-Sensitive Text Retrieval

Rodger Benham
RMIT University
Melbourne, Australia

CCS CONCEPTS

• Mathematics of computing → Probabilistic inference problems; **• Information systems → Evaluation of retrieval results**; **Search engine architectures and scalability**.

KEYWORDS

Risk-Sensitive Evaluation; Probabilistic Inference

Search engines with a loyal user-base face the difficult task of improving overall effectiveness while maintaining the quality of existing work-flows. Risk-sensitive evaluation tools are designed to address that task, but, they currently do not support inference over multiple baselines. Our research objectives are to: 1) Survey and revisit risk evaluation, taking into account frequentist and Bayesian inference approaches for comparing against multiple baselines; 2) Apply that new approach, evaluating a novel web search technique that leverages previously run queries to improve the effectiveness of a new user query; and 3) Explore how risk-sensitive component interactions affect end-to-end effectiveness in a search pipeline.

Risk-Reward Trade-Offs On Multi-System Evaluations. We survey and revisit "risk" – improving semantics to avoid the double-negative of reporting a "positive-risk" score for a treatment with no risk. We extend risk measures over multiple baselines [6] to become inferential, using frequentist and Bayesian techniques. Figure 1 motivates exploring risk with Bayesian statistics. Left, we have a probability density function (PDF) built using Markov Chain Monte Carlo (MCMC) sampling, assuming the t-distribution is a good fit – as is done in TRisk [7]. Right, is the same process with a prior assumption that the original distribution is skewed. Bayes factors are used to select the model with a better fit for the data. Kass and Raftery [8] categorize the strength of these models using twice the logarithm of their Bayes factors. These values left-to-right are -50.2 and 50.2, so, the right-most PDF is a better fit for inferential purposes. We will also explore the impact of multiple comparisons corrections, such as Bonferroni and Holm-Bonferroni.

Boosting Search Performance Using Query Variations. We propose to form a *query centroid*, by clustering related queries in an offline processing step, and then fusing their rankings using a novel cost-effective CombSUM technique. Query clusters can be formed automatically with click-logs [5], but without access to these logs, we use UQV100 [2] and simulate clustering error rates by randomly assigning queries to the wrong cluster. To boost the effectiveness of subsequent queries, novel techniques are used to combine the user query ranking to its associated query centroid [3]. We use the new risk measures to explore the cost of incorrect cluster association and explore how our techniques mitigate risk.

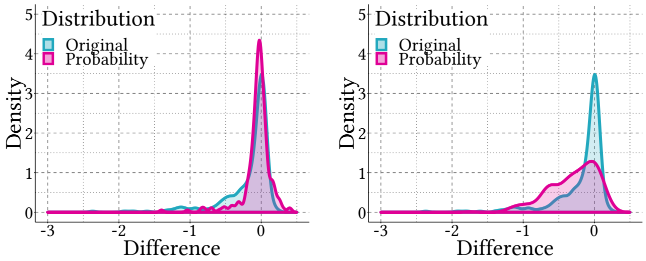

Figure 1: The differences in means, after URisk $\alpha = 10$ [4, 9] is applied to the difference in AP scores between BM25 and language modeling on Robust04. Left, the Bayesian MCMC sampled PDF is modeled assuming that the data is normal. Right, the data is assumed to be skewed-normal.

Exploring Risk-Sensitivity and Effectiveness. Armstrong et al. [1] emphasize the importance of being careful when making additivity claims. We propose exploring component interactions using risk measures. As components are typically tuned for effectiveness, that will allow us to compare the additivity of tuning components for risk. As each component has its own exposure to risk, measurement should ideally be taken group-wise. For example, a stemmer component can be measured over Porter versus Krovetz stemmers in a multi-baseline scenario.

ACKNOWLEDGMENTS

This work was supported by a SIGIR Student Travel Grant.

REFERENCES

[1] T. G. Armstrong, A. Moffat, W. Webber, and J. Zobel. Improvements that don't add up: Ad-hoc retrieval results since 1998. In *Proc. CIKM*, pages 601–610, 2009.
[2] P. Bailey, A. Moffat, F. Scholer, and P. Thomas. UQV100: A test collection with query variability. In *Proc. SIGIR*, pages 725–728, 2016.
[3] R. Benham, J. Mackenzie, A. Moffat, and J. S. Culpepper. Boosting search performance using query variations. *arXiv preprint arXiv:1811.06147*, 2018.
[4] K. Collins-Thompson, C. Macdonald, P. Bennett, F. Diaz, and E. M. Voorhees. TREC 2013 web track overview. In *Proc. TREC*, 2014.
[5] N. Craswell and M. Szummer. Random walks on the click graph. In *Proc. SIGIR*, pages 239–246, 2007.
[6] B. T. Dinçer, C. Macdonald, and I. Ounis. Risk-sensitive evaluation and learning to rank using multiple baselines. In *Proc. SIGIR*, pages 483–492, 2016.
[7] B. T. Dinçer, C. Macdonald, and I. Ounis. Hypothesis testing for the risk-sensitive evaluation of retrieval systems. In *Proc. SIGIR*, pages 23–32, 2014.
[8] R. E. Kass and A. E. Raftery. Bayes factors. *J. Am. Stat. Assoc*, 90(430):773–795, 1995.
[9] L. Wang, P. N. Bennett, and K. Collins-Thompson. Robust ranking models via risk-sensitive optimization. In *Proc. SIGIR*, pages 761–770, 2012.

SIGIR '19, July 21–25, 2019, Paris, France

ACM ISBN 978-1-4503-6172-9/19/07.
https://doi.org/10.1145/3331184.3331423