Accounting for Stability of Retrieval Algorithms using Risk-Reward Curves

Kevyn Collins-Thompson
Microsoft Research
1 Microsoft Way
Redmond, WA 98052-6399 U.S.A.
kevynct@microsoft.com

ABSTRACT

Past evaluation of information retrieval algorithms has focused largely on achieving good *average* performance, without much regard for the *stability* or variance of retrieval results across queries. In fact, two algorithms that superficially appear to have equally desirable average precision performance can have very different stability or *risk profiles*. A prime example comes from query expansion, where current techniques typically give good average improvements in mean average precision, but are also unstable and have high variance across individual queries [3]. We propose the use of *risk-reward curves* and related statistics to characterize the tradeoff an algorithm exhibits between a reward property such as mean average precision and a risk property such as the variance of the algorithm – particularly the downside variance, when the algorithm fails or makes performance worse. Such evaluation methods are broadly applicable beyond query expansion to other retrieval operations that must balance risk and reward, such as personalization, document ranking, resource selection, and others.

Categories and Subject Descriptors: H.3.3 [**Information Retrieval**]: Evaluation
General Terms: Experimentation, Measurement
Keywords: Algorithm risk, stability, query expansion

1 Risk-reward tradeoff curves

We observe that many IR scenarios have a risk-reward tradeoff. In query expansion, for example, when interpolating a feedback model with the original query model using a parameter α, giving more weight to the original query model (lower α) reduces the potential harm of a noisy expansion model, but also reduces the potential gains when the feedback model is effective, and vice versa. By plotting risk and reward jointly as α is varied from $\alpha = 0.0$ (original query only) to $\alpha = 1.0$ (all feedback), we obtain a *risk profile* in the form of a risk-reward tradeoff curve that gives a more complete picture of algorithm quality. As Fig. 1 shows, two algorithms that appear identical in terms of mean average precision (MAP) gain may have very different risk profiles.

In general, to compute a risk-reward tradeoff curve for an information retrieval algorithm, we must first decide on how to quantify risk and reward. The appropriate measures will vary with the retrieval task: a good 'reward' measure for Web search, for example, may be precision in the top-20 documents (P20); legal IR applications may focus on recall; and general IR evaluations may use mean average precision (MAP). We generally will focus on risk-reward curves using relative or absolute MAP or P20 gain as the 'reward' measure, and this is plotted on the y-axis of the chart.

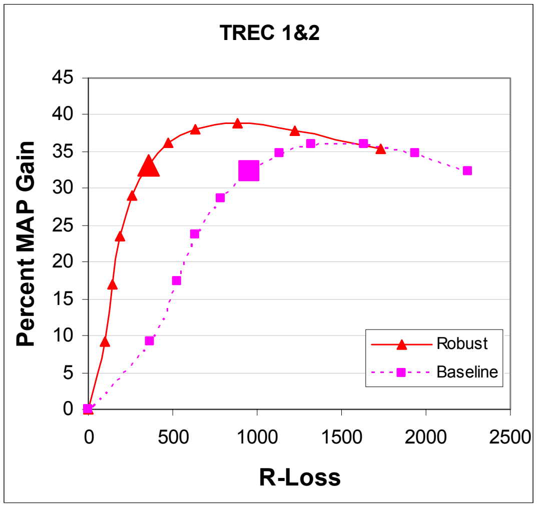

Figure 1: One form of risk-reward tradeoff curve for query expansion, showing how two algorithms that give almost identical MAP gain (33%) at a typical operational setting ($\alpha = 0.5$, shown as the enlarged points) can have very different risk profiles: the 'robust' version of the expansion algorithm is much more stable and has a much smaller net loss of relevant documents for expansion failures. The downside risk/variance (R-Loss) and MAP improvement change together as the feedback interpolation parameter α is increased from 0. (original query, no expansion) to 1.0 (all feedback model, no original query). Curves that are *higher* and *to the left* give a better tradeoff. This example is an actual experiment result (TREC 1&2 topics) taken from [2].

The key aspects of the 'risk' measure are: 1) that it captures *variance* or a related negative aspect of retrieval performance across queries, and 2) this variance/risk is based on the corresponding reward measure chosen. We are particularly interested in the *downside risk* of an expansion algorithm: the reduction in reward due to expansion failures,

SIGIR Workshop on the Future of IR Evaluation, July 23, 2009, Boston.
.

which are defined as cases where applying expansion gives worse results than the initial query. The risk measure is assigned to the x-axis of the risk-reward curve.

As one example, we can choose the reward measure to be 'percent gain in precision at k ($P@k$)' compared to using no expansion, and the risk measure, which we call *R-Loss at k* as the *net loss of relevant documents from the top k due to expansion failure*. R-Loss at k is an appropriate risk measure because it both reflects the downside variance of the reward measure and net loss in relevant documents is a concrete and important measure for users. When we use MAP gain as the reward measure instead of $P@k$, we refer to the risk measures simply as *R-Loss*, setting k to the size of the retrieved document set (typically $k = 1000$). Because R-Loss is a document count, queries with more relevant documents have greater influence on the measure. Alternatively, we could consider normalizing R-Loss over the number of relevant documents, to give each query equal weight.

A number of potentially useful concepts and extensions follow from exploiting connections to computational finance. We say that one algorithm's tradeoff curve A *dominates* another curve B if the reward achieved by A for any given risk level is always at least as high as achieved by B at the same risk level. For example, in Figure 1 the robust algorithm dominates the baseline expansion method. The *efficient frontier* on a risk-reward graph is the boundary of the convex hull of points produced by (in theory) all possible parameter settings and represents the best performance that an algorithm can achieve at any given level of risk, for any choice of parameters. Typically, the efficient frontier can be approximated, although at considerable computational cost, by broad sampling of the parameter space.

The *risk-reward ratio* $\rho(P) = G(P)/F(P)$ of a point P that achieves MAP gain $G(P)$ and R-Loss $F(P)$ is the *slope* of the line joining P to the origin. The *midpoint risk-reward tradeoff* at $\alpha = 0.5$ gives a single value that could be used to compare with other algorithms on the same collection. The *Sharpe ratio* is the optimal $\rho^\star = \rho(P^\star)$ at the point $P^\star$ of maximum slope on the (approximate) efficient frontier, identifying the *best achieved tradeoff* of an algorithm. These are just a few examples of how investigating risk-aware versions of standard retrieval statistics like MAP or P20 may be a fruitful direction for future research.

2 Related work

Risk/reward tradeoff curves were introduced by Markowitz [4] as part of his pioneering finance work on portfolio selection. Risk-aware algorithms and analysis methods are well-developed in the computational finance community but we have seen little work in IR fully exploit this connection. The downside risk of query expansion has been noted for decades [6], but only recently has this gotten more extended attention in evaluations. An early version of risk-reward curves was used by the author for query expansion robustness evaluation [3]. The connection between Markowitz-type mean-variance models and risk and reward for retrieval algorithms was first noted in a study that applied this idea to reduce the downside risk of existing query expansion methods [1]. A greatly extended exploration of risk and reward, including extensive refinement and employment of risk-reward curves for evaluation, may be found the author's doctoral dissertation [2]. Recently, a similar mean-variance paradigm was described for document ranking [7]. *Robustness histograms*[1][5], shown in Fig. 2 are another useful evaluation approach that captures variance at a single choice of risk parameter α but not the entire risk profile across all values of α. Precision-recall curves can also present a limited form of risk-reward tradeoff, but assume a binary good/bad label for the objects of interest (e.g. an expanded query), which gives only a crude approximation of variance since it ignores the magnitude of the retrieval failure or result. Risk-reward curves, in contrast, can make more effective distinctions between systems by observing the magnitude of changes in the reward measure and not merely whether gains were positive or negative.

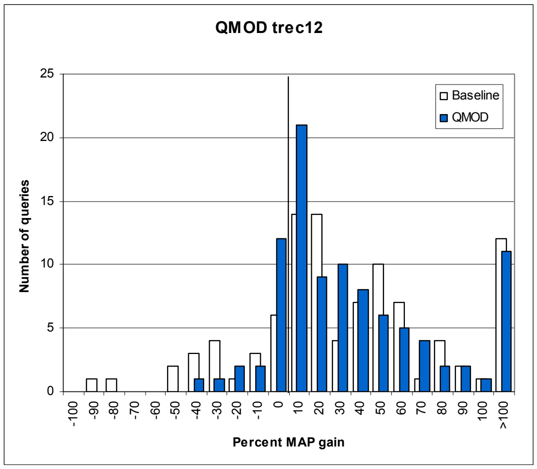

Figure 2: A *robustness histogram*, showing the variance in MAP gain/loss across queries for two different expansion algorithms at a single choice of $\alpha = 0.5$. The 'baseline' expansion method has higher downside variance than the QMOD algorithm [1], as shown by the increased left-hand tail (queries hurt by expansion).

3 Conclusion

We propose the joint analysis of risk and reward behavior for retrieval algorithms using risk-reward curves, which can capture the tradeoff between algorithm risk or variance, and a reward measure such as average-case performance. We believe risk-reward tradeoff curves are a highly useful evaluation method not only for query expansion, but also personalization, document ranking, resource selection and other risk-sensitive scenarios.

4 References

[1] K. Collins-Thompson. Estimating robust query models using convex optimization. In *Advances in NIPS 21*, 2008.

[2] K. Collins-Thompson. *Robust model estimation methods for information retrieval.* PhD thesis, Carnegie Mellon Univ., 2008.

[3] K. Collins-Thompson and J. Callan. Estimation and use of uncertainty in pseudo-relevance feedback. In *Proceedings of SIGIR 2007*, pages 303–310, 2007.

[4] H. M. Markowitz. Portfolio selection. *Journal of Finance*, 7(1):77–91, 1952.

[5] D. Metzler and W. B. Croft. Latent concept expansion using markov random fields. In *SIGIR 2007*, pages 311–318.

[6] A. Smeaton and C. J. van Rijsbergen. The retrieval effects of query expansion on a feedback document retrieval system. *The Computer Journal*, 26(3):239–246, 1983.

[7] J. Wang. Mean-variance analysis: A new document ranking theory in information retrieval. In *ECIR 2009*, pages 4–16, 2009.