Multi-Objective Optimization in Learning to Rank

Na Dai
Computer Sci. & Engr.
Lehigh University, USA
nad207@cse.lehigh.edu

Milad Shokouhi
Microsoft Research
Cambridge, UK
milads@microsoft.com

Brian D. Davison
Computer Sci. & Engr.
Lehigh University, USA
davison@cse.lehigh.edu

ABSTRACT

Supervised learning to rank algorithms typically optimize for high relevance and ignore other facets of search quality, such as freshness and diversity. Prior work on multi-objective ranking trained rankers focused on using *hybrid* labels that combine overall quality of documents, and implicitly incorporate multiple criteria into quantifying ranking risks. However, these hybrid scores are usually generated based on heuristics without considering potential correlations between individual facets (e.g., freshness versus relevance). In this poster, we empirically demonstrate that the correlation between objective facets in multi-criteria ranking optimization may significantly influence the effectiveness of trained rankers with respect to each objective.

Categories and Subject Descriptors: H.3.3 [Information Storage and Retrieval]: Information Search and Retrieval

General Terms: Algorithms, Performance

Keywords: multi-objective optimization, learning to rank

1 Introduction

To satisfy user information needs, retrieval systems consider several *facets* of search quality such as relevance, freshness and diversity in ranking documents. These facets may interact or correlate with each other in complex and query-dependent manners. For instance, previous research has demonstrated that relevance and freshness have high correlation for breaking news queries [2]. Similar scenarios exist in the information filtering and recommendation domains, where users' ratings on several aspects may correlate with each other depending on user profiles, and consequently affect the prediction models of user preferences on items [6].

Prior work that considered users' multi-criteria objectives in search or collaborative filtering have been mostly inspired by multi-criteria decision making (MCDM) theory from the operations research community [7]. The preference between different criteria is quantified by utility measures that affect optimization through preference model representation. The commonly used preference models for search or recommendation tasks include *value-focused* models [8, 9] and *outranking relations* models [3]. While these approaches exploit the search quality on each aspect (criterion-specific ratings) to enhance overall quality (measurement ratings), they ignore the inter-relationship between different objectives.

This poster explores the influence of interactions and correlations between multiple criteria for ranking optimization in the context of web search. As a preliminary step, we analyze the influence of bi-criteria inter-relationship on pairwise ranking models though the analysis can be generalized to other multi-criteria scenarios. While the definition of *measurement* utility is an open issue, we use the minimum relative ranking improvement on both criteria (denoted as *RelImp*) to measure the influence of bi-criteria optimization, emphasizing its relative benefits compared to optimizing for each single objective. We define $RelImp$ as follows:

$$RelImp = \min_{c,obj}[\text{perf}(c, bi\text{-}obj) - \text{perf}(c, obj)]/\text{perf}(c, obj) \quad (1)$$

where perf$(c, *)$ is the performance on criteria c when optimizing for objective "$*$". Our research explores the effect of correlation between two ranking criteria on the benefit of bi-objective ranking optimization, focusing on three main issues: (1) what is the correlation scale that can benefit *RelImp*? (2) how much benefit can it bring? and (3) what does a useful preference model look like under different correlation scales? We exploit a value-focused preference model implicitly for ranking optimization through minimizing bi-criteria ranking risk based on *hybrid* labels that combine the quality of documents on both aspects. We demonstrate that the correlation between multiple objectives (freshness and relevance in our case) may influence the outcome of multi-criteria ranking optimization.

2 Methodology

Given a query q and its associated documents $d_1,\ldots,d_n$, each query-document pair <q,d_k> is rated based on its quality on each facet, i.e., $y^{(1)}_{q,d_k}$ and $y^{(2)}_{q,d_k}$. By exploiting *hybrid* labels to combine the overall quality, we average the score achieved on each aspect as the hybrid label for <q,d_k>, defined as:

$$\widetilde{y}_{q,d_k} = \Big(\frac{1}{n} \cdot \sum_{i=1}^{n} \big(y^{(i)}_{q,d_k}\big)^m\Big)^{\frac{1}{m}} \quad (2)$$

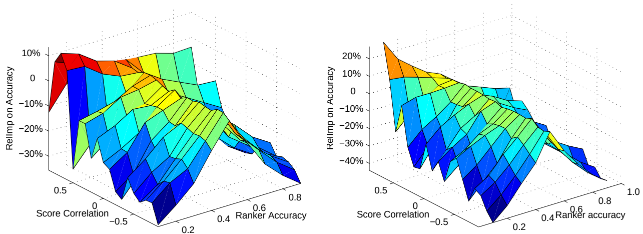
where $n = 2$ is the number of facets (e.g., freshness and relevance), and m determines the type of *hybrid* label function; quadratic mean (QM), arithmetic mean (AM), geometric mean (GM) and harmonic mean (HM) respectively for $m = 2$, $m = 1$, $m \to 0$ and $m = -1$. These variants reflect how sensitive the *hybrid* label is with respect to the lower (higher) rating scores on both aspects, assuming that the rating scores on two aspects fall into the same scale. We believe this perspective is reasonable since the criteria for judging query-document pair quality may vary from one person to another. We also include two extreme cases, i.e., MIN and MAX, representing the minimum and maximum rating scores on two aspects.

Pairwise ranking learning algorithms train a set of parameters ω by minimizing the ranking risk aggregated from loss of misclassified preferential query-document pairs based on relevance. By exploiting *hybrid* labels, we optimize model parameters by:

$$f^* = \arg\min_f \sum_{q\in\mathcal{Q}} \sum_{<d_i,d_j>\in\mathcal{D}_q} \mathcal{L}(\widehat{y}_{q,d_i}, \widehat{y}_{q,d_j}, \widetilde{y}_{q,d_i}, \widetilde{y}_{q,d_j}) \quad (3)$$

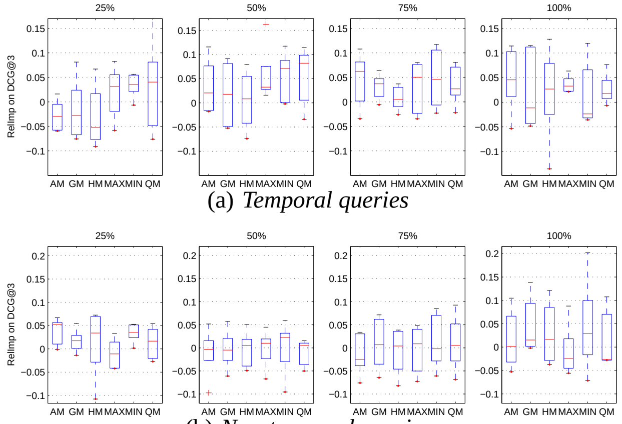
where $\mathcal{D}_q$ is the set of preferential query-document pairs for query q, and $\mathcal{L}$ is the loss function that penalizes <d_i, d_j> if its predicted

SIGIR'11, July 24–28, 2011, Beijing, China.
ACM 978-1-4503-0757-4/11/07.



(a) MIN (b) MAX

Figure 1: The minimum relative ranking improvement on accuracy based on MIN and MAX *hybrid* labels under the variance of bi-criteria correlation and ranker accuracy.



(a) *Temporal queries*

(b) *Non-temporal queries*

Figure 2: The average and standard deviation of *RelImp* on DCG@3 across five folds for the *temporal* (top) and *non-temporal* (bottom) query sets by using the top 25%, 50%, 75% and 100% (all) effective ranking features.

preferential relationship (based on $\widehat{y}_{q,d_i}$ and $\widehat{y}_{q,d_j}$) is discordant with groundtruth (based on $\widetilde{y}_{q,d_i}$ and $\widetilde{y}_{q,d_j}$). We use RankSVM [5] as our basic ranker. We note that the loss function defined on preferential query-document pairs is a linear combination of the loss on each criteria, with the coefficient depending on the actual rating scores on both criteria.[1]

3 Experiments

Our goal is to show how the gain of bi-criteria ranking optimization varies with the bi-criteria correlation and the optimization capability. To avoid data bias, we conduct experiments on two data sets.

We use syntactic data to simulate the process in which pairwise ranking models generate search results. Our dataset consists of 21 subsets, with each composed of 1000 simulated bi-criteria rating scores that have a fixed score correlation from -0.9 to 0.9 with a step size 0.1. Three pseudo-classifiers (i.e., simulated rankers) are used to generate preferential score pair relationships based on each aspect of bi-criteria (i.e., producing baseline results) and the *hybrid* labels in Section 2 respectively. To incorporate optimization capability variance, we exploit a probability threshold (ranging from [0.1,0.9] with step size 0.1) to control the chance that pseudo-classifiers generate correct pair relationships, denoted as *ranker accuracy*. For instance, if the *ranker accuracy* is 90%, the generated pair relationship has 90% chance to be consistent with the ground truth. The gain of bi-criteria ranking optimization is measured by *RelImp* based on the percentage of correctly classified preferential score pairs (i.e., *RelImp on accuracy*).

Figure 1 shows the minimum relative improvement on preferential pair classification accuracy for MIN and MAX as the bi-criteria correlation and *ranker accuracy* vary. Preliminary results demonstrate that the trends of others typically fall in between. The bi-criteria optimization brings benefits when the bi-criteria correlation is highly positive and *ranker accuracy* is low. When the ranker accuracy is high, bi-criteria optimization has negative impact on performance. It is not surprising given that it actually incorporates more inaccurate optimization objectives, and this can be mitigated with the increase of bi-criteria correlation.

To further investigate the effect of bi-criteria ranking optimization on real search scenarios, we use a learning to rank data set that is built on a large-scale archival web corpus [1]. The ranking criteria here are freshness and relevance. This data set contains 90 temporal queries manually selected from Google Trends and 90 nontemporal queries that are first randomly sampled from a 2006 MSN query log, and then filtered by a commercial temporal query classifier with high accuracy. An average of 71 documents (URLs) per query were judged by at least one worker of Amazon Mechanical Turk. Each URL is evaluated on a five point scale in terms of freshness and relevance with respect to a given query and a fixed time point (April 2007). The Pearson's correlation between freshness and relevance on temporal (nontemporal) query set is 0.912±0.004 (0.429±0.021). More details are provided in [1].

Figure 2 shows the average and standard deviation of RelImp on DCG@3 [4] across five fold cross-validation for *temporal* and *nontemporal* query sets. By using the different top k% effective ranking features that are selected by a reference model (a RankSVM model in this work) based on training data, we incorporate the influence of ranker effectiveness into the sensitivity study on the gain of bi-criteria ranking optimization. The results confirm our previous observations on simulated data and demonstrate that (1) RelImp is more sensitive to hybrid labels and ranker effectiveness when the correlation between relevance and freshness is highly positive (i.e., the *temporal* query set); and (2) bi-criteria ranking optimization can bring more benefits under highly positive bi-criteria correlation.

In summary, we showed that the gain from bi-criteria ranking optimization is sensitive to the bi-criteria correlation and the optimization capability. More benefits can be achieved when there exists a stronger positive correlation between the two criteria and the optimization capability is not strong. These observations reveal valuable insights towards better understanding multi-criteria ranking optimization and may provide hints about how we can exploit multi-criteria ranking optimization to improve search quality.

Acknowledgments

This work was supported in part by a grant from the National Science Foundation under award IIS-0803605.

4 References

[1] N. Dai, M. Shokouhi, and B. D. Davison. Learning to rank for freshness and relevance. In *Proc. 34th ACM SIGIR*, 2011.

[2] A. Dong, Y. Chang, Z. Zheng, G. Mishne, J. Bai, R. Zhang, K. Buchner, C. Liao, and F. Diaz. Towards recency ranking in web search. In *Proc. WSDM Conference*, pages 11–20, 2010.

[3] M. Farah and D. Vanderpooten. An outranking approach for rank aggregation in information retrieval. In *Proc. 30th ACM SIGIR*, pages 591–598, 2007.

[4] K. Jarvelin and J. Kekalainen. IR evaluation methods for retrieving highly relevant documents. In *Proc. 23rd ACM SIGIR*, pages 41–48, 2000.

[5] T. Joachims. Optimizing search engines using clickthrough data. In *Proc. SIGKDD*, pages 133–142, 2002.

[6] N. Manouselis and C. Costopoulou. Analysis and classification of multi-criteria recommender systems. *World Wide Web*, 10:415–441, December 2007.

[7] R. Steuer. *Multiple Criteria Optimization: Theory, Computation and Application.* John Wiley, 546 pp, 1986.

[8] L. Wang, J. Lin, and D. Metzler. Learning to efficiently rank. In *Proc. 33rd ACM SIGIR*, pages 138–145, 2010.

[9] S. R. Wolfe and Y. Zhang. User-centric multi-criteria information retrieval. In *Proc. 32nd ACM SIGIR*, pages 818–819, 2009.

[1] We omit the proof due to space limitations.