Active Query Selection for Learning Rankers

Mustafa Bilgic
Illinois Institute of Technology, Chicago, IL
mbilgic@iit.edu

Paul N. Bennett
Microsoft Research, Redmond, WA
pauben@microsoft.com

ABSTRACT

Methods that reduce the amount of labeled data needed for training have focused more on selecting which documents to label than on which queries should be labeled. One exception to this [4] uses expected loss optimization (ELO) to estimate which queries should be selected but is limited to rankers that predict absolute graded relevance. In this work, we demonstrate how to easily adapt ELO to work with any ranker and show that estimating expected loss in DCG is more robust than NDCG even when the final performance measure is NDCG.

Categories and Subject Descriptors

H.3.3 [**Information Search and Retrieval**]: Information Search and Retrieval—*active learning*

Keywords

Active learning, query selection

1. INTRODUCTION

Research in information retrieval evaluation has examined how to construct minimal test collections [2], and the balance between the number of queries judged and the depth of judging [3]. With respect to training rankers, most work has focused on document selection [1] or balancing number of queries with depth of documents judged using random query selection [5]. In this paper, we focus on selecting *queries* in order to most rapidly increase ranker retrieval performance.

In particular, we focus on the application of expected loss optimization (ELO) to query selection. Long *et al.* [4] used ELO to select queries for training but relied on having a ranker that estimated absolute graded relevance. We generalize this to work with any ranker – many of which induce a ranking but not absolute labels. To generalize to any ranker, we introduce a calibration phase over validation data.

In addition, although in theory ELO can be used with any performance measure for active learning, we show using DCG loss (as done in [4]) leads to better performance whether DCG or NDCG is used as the final evaluation of ranker performance. We provide evidence this is because ELO using DCG loss tends toward queries that have both more relevant examples *and* many degrees of relevance.

2. APPROACH

ELO suggests that, given a set of candidate queries $\mathcal{C}$, one pick the query $q \in \mathcal{C}$ for labeling where the expected loss is the greatest. Mathematically, we have:

$$\max_{q \in \mathcal{C}} \mathrm{E}_{P(Y|X_q,\mathcal{D})} \left[\max_{\pi} \mathrm{M}(\pi(X_q), y) - \mathrm{M}(\mathrm{R}(X_q), y) \right] \qquad (1)$$

where $\mathcal{D}$ is the given training data, and $P(Y \mid X_q, \mathcal{D})$ is a distribution over graded relevance labels Y for the documents, X_q, to be ranked for the query q. $\mathrm{M}(r, y)$ is a retrieval performance measure such as DCG that can evaluate the quality of a ranking, r, for a set of documents given a particular labeling of the documents, y. $\pi(X_q)$ is simply a permutation of the documents and $\mathrm{R}(X_q)$ denotes the current ranking of the documents. For most retrieval performance measures, the inner max on the left-hand side of the difference is easily found by sorting from highest relevance to the lowest.

In order to estimate the label distribution $P(Y \mid X_q, \mathcal{D})$ Long *et al.* [4] relied on training an ensemble of models to predict absolute graded relevance. We generalize ELO to work with any ranker by mapping the current ranking model to a distribution over graded labels. To do so, we introduce a calibration phase where a classification model is trained over the labels of the top k documents according to the ranker in the validation data.[1] During active learning, the classification model is used to estimate the $P(Y \mid x_q, \mathcal{D})$ for each document $x_q \in X_q$. The quantity in Eq. 1 is then estimated through sampling of the labels from this distribution.

3. EXPERIMENTS

Like most active learning evaluation settings, we start with some labeled data $\mathcal{D}$ that is randomly chosen, train the models on $\mathcal{D}$, pick a number of queries to be labeled from the candidate set $\mathcal{C}$, add those to $\mathcal{D}$, and repeat this process for a number of iterations. The performance of the active learning strategy in augmenting $\mathcal{D}$ is judged at each iteration by evaluating the current induced rankers on held-out test data. We perform 20 iterations of labeling and based on the findings reported in [5] we label 15 documents per query (or the maximum available). We repeat this process five times, each time starting with a different set of labeled data and report averages. We report both DCG@10 and NDCG@10 (one can use DCG for selection but NDCG for evaluation and so forth).

We use the publicly available Yahoo! LETOR challenge dataset that has 3 splits: we treat the train split as the candidate data $\mathcal{C}$, utilize the validation split for tuning the rankers' parameters and training the calibration models, and use the test split for evaluation. We experiment with two rankers: one that does not produce absolute graded relevance (SVMRank) and one that does (Additive Regression (AR)). SVMRank labels 750 query-document pairs per iteration where AR labels only 150. The difference is due to AR having a steeper learning curve.

We examine four query/document selection methods: (1) select queries and documents randomly (rndQ-rndD); (2) se-

[1] We use the same validation data that is used for model parameter search – this ensures our method does not require any additional labeled data.

SIGIR'12, August 12–16, 2012, Portland, Oregon, USA.
ACM 978-1-4503-1472-5/12/08.

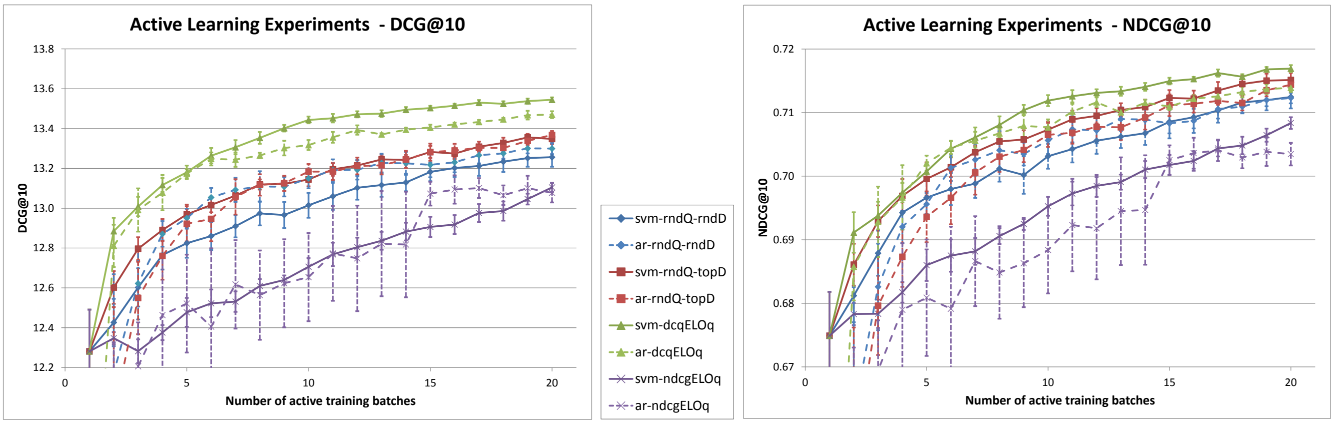

Figure 1: Comparison of query-selection strategies.

lect queries randomly and select the top documents according to the current ranker (rndQ-topD); (3) select queries according to ELO with DCG@10 as the selection measure and the top documents according to the current ranker (dcgELOq); (4) same as 3 but with NDCG@10 instead (ndcgELOq).

4. RESULTS AND DISCUSSION

Figure 1 shows the results for SVMRank (*solid*) and AR (*dashed*) when the evaluation measure is DCG@10 (*left*) and NDCG@10 (*right*).[2] Error bars are standard error about the mean over the five trials. As reported elsewhere [4], selecting the top documents performs as well or better than selecting documents randomly. Note that regardless of whether evaluating by DCG or NDCG, using NDCG for selection (ndcgELOq) leads to the worst performance. In contrast, using DCG for selection leads to the best performance across both rankers and both evaluation measures. Finally, we note that the differences between methods are less significant when evaluated by NDCG than DCG. This suggests that while the learners are more effective – finding more relevant results per query – they are contributing to the marginal relevance for each query according to NDCG. The perceived impact on user utility will likely depend on the scenario and the degree to which the task is recall-oriented.

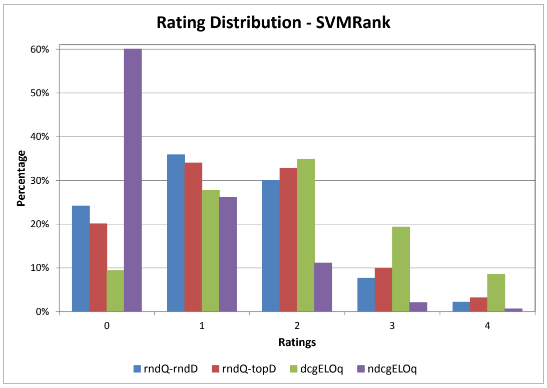

Figure 2: Rating distribution of selected queries.

Figure 2 displays the rating distribution of the training data collected at the last step of active learning as a percentage for the 15,000 labeled instances for SVMRank (the distribution trends for AR were nearly identical). Note that ndcgELOq selects far more irrelevant items than the other methods. This may seem surprising since NDCG selection is the same as DCG but normalized by the max estimate on the left-hand side of Eq. 1. However, for a poorly performing query with a single relevant document, NDCG's max will be 1 but current performance will be near zero. Thus, the selection method often selects queries with very few relevant documents. In contrast, dcgELOq not only obtains the largest percent of documents at the relevant side (labels 3,4) and fewest on the irrelevant side (label 0), it selects queries where a variety of relevance grades exist. This is consistent with the literature that biasing toward relevant documents is not sufficient in itself [1] – one also needs a variety of relevance grades present.

5. SUMMARY

We presented a method that generalizes the applicability of ELO for query selection to any ranker. Our method also has the benefit of being less of a computational burden than training ensembles at each step prior to labeling. We also demonstrated that whether one cares about DCG or NDCG for performance, using DCG provides a more stable query selection method. This is because the nature of NDCG as a ratio pushes the selection toward queries that often have few relevant documents. In contrast, using DCG in the selection mechanism promotes queries that have more relevant documents, and the expected loss component ensures that there will be a variety of relevance grades – since current performance is far below the max. These insights may be useful in developing new query selection methods.

6. REFERENCES

[1] J. Aslam, E. Kanoulas, V. Pavlu, S. Savev, and E. Yilmaz. Document selection methodologies for efficient and effective learning-to-rank. In *SIGIR '09*.

[2] B. Carterette, J. Allan, and R. Sitaraman. Minimal test collections for retrieval evaluation. In *SIGIR '06*.

[3] B. Carterette, V. Pavlu, E. Kanoulas, J. Aslam, and J. Allan. If i had a million queries. In *ECIR '09*.

[4] B. Long, O. Chapelle, Y. Zhang, Y. Chang, Z. Zheng, and B. Tseng. Active learning for ranking through expected loss optimization. In *SIGIR '10*.

[5] E. Yilmaz and S. Robertson. Deep versus shallow judgments in learning to rank. In *SIGIR '09*.

[2] SVMRank and AR are displayed together for space and to emphasize the similarity in trends. We are interested in comparisons within each and not across the two.