Unbiased Ranking Evaluation on a Budget

Tobias Schnabel
Cornell University
Ithaca, NY
tbs49@cornell.edu

Adith Swaminathan
Cornell University
Ithaca, NY
adith@cs.cornell.edu

Thorsten Joachims
Cornell University
Ithaca, NY
tj@cs.cornell.edu

1. EXTENDED ABSTRACT

We address the problem of assessing the quality of a ranking system (e.g., search engine, recommender system, review ranker) given a fixed budget for collecting expert judgments. In particular, we propose a method that selects which items to judge in order to optimize the accuracy of the quality estimate. Our method is not only efficient, but also provides estimates that are unbiased — unlike common approaches that tend to underestimate performance or that have a bias against new systems that are evaluated re-using previous relevance scores [1]. Our method is based on the insight that we can write many common performance measures as expectations, and then use Monte Carlo techniques, such as importance sampling, to estimate these expectations [1].

We compare against the traditional approach of ranking evaluation under budget constraints that is employed in the *pooling* method used in TREC [8]. Instead of judging all queries to their full depths, only the top k (e.g., $k = 100$) documents for each query are judged until the budget is exhausted. While for small document collections it is reasonable to assume that all relevant documents are within the top k documents, this working hypothesis is less valid for larger collections [3]. More complicated approaches include stratified sampling or greedy sample selection [11, 2], but usually result in algorithms that are difficult to apply for practitioners. Somewhat related to our method is the scenario in which one wants to re-use interaction logs of a system for evaluation [6, 7] or data from logged interleaving experiments [4].

Our contributions are as follows. First, we show how to get an unbiased estimator for Discounted Cumulative Gain (DCG) [5] using importance sampling. Second, we outline a simple proposal for selecting the sampling distribution. Lastly, we compare our method to two traditional approaches and show that it is vastly superior in terms of bias and accuracy.

1.1 Method

We assume that we are given a sample of rankings $\mathcal{Y} = \{(x, \boldsymbol{y})\}$ from a ranking system, each ranking $\boldsymbol{y}$ corresponding to a different context x (e.g., query). Each ranking $\boldsymbol{y}$ consists of a sequence of items d_j (e.g., documents). Denote with $\text{rank}(d, \boldsymbol{y})$ the rank of item d in ranking $\boldsymbol{y}$, or zero if d is not contained in $\boldsymbol{y}$. The relevance of item d given context x is denoted as $f_x(d) \in \mathbb{R}$, which is the relevance rating that relevance judges can provide. Let $P_{\boldsymbol{y}}(D)$ be an appropriately constructed probability distribution over items in $\boldsymbol{y}$. The DCG of a single ranking $\boldsymbol{y}$ is typically defined as

$$DCG(x, \boldsymbol{y}) = \sum_j \frac{f_x(d_j)}{\log j + 1} = \sum_{d \in \boldsymbol{y}} \frac{f_x(d)}{\text{rank}(d, \boldsymbol{y}) + 1}.$$

Using the notation above, it can now equivalently be written as the expectation

$$DCG(x, \boldsymbol{y}) = Z_{\boldsymbol{y}} \cdot \mathbb{E}_{P_{\boldsymbol{y}}}[f_x(D)] = Z_{\boldsymbol{y}} \cdot \sum_{x \in \boldsymbol{y}} f_x(d) \cdot P_{\boldsymbol{y}}(D = d),$$

where $P_{\boldsymbol{y}}(D = d) = \frac{1}{Z_{\boldsymbol{y}} \log(\text{rank}(d, \boldsymbol{y}) + 1)}$ with normalization constant $Z_{\boldsymbol{y}} = \sum_d \frac{1}{\log(\text{rank}(d, \boldsymbol{y}) + 1)}$ corresponds to the scaled DCG weights. Now, our goal is to compute the average DCG across all rankings:

$$AvgDCG(\mathcal{Y}) = \frac{1}{|\mathcal{Y}|} \sum_{(x, \boldsymbol{y}) \in \mathcal{Y}} DCG(x, \boldsymbol{y}).$$

Using importance sampling, we can estimate each single $DCG(x, \boldsymbol{y})$ as follows. Draw k samples $d_1, \ldots, d_k$ from any fixed distribution $Q_x(D)$, respecting $P_{\boldsymbol{y}}(D) > 0 \implies Q_x(D) > 0$, and get the relevance judgments $f_x(d_1), \ldots, f_x(d_k)$ of these documents. The following yields an unbiased estimator for the true DCG value of the ranking for any positive value of k:

$$\widehat{DCG}(x, \boldsymbol{y}) = \frac{Z_{\boldsymbol{y}}}{k} \sum_{j=1}^{k} f_x(d_j) \cdot \frac{P_{\boldsymbol{y}}(D = d_j)}{Q_x(D = d_j)}$$

We call this method the UnBiased Importance Sampling *UBIS-k* approach in the following.

One might ask now what distribution Q_x one should ideally pick. If we define optimality of an estimator in terms of having minimal variance, then standard results [10] tell us that the optimal distribution is

$$Q_x^*(d) = \frac{f_x(d) \cdot P_{\boldsymbol{y}}(D = d)}{\sum_{\tilde{d}} f_x\left(\tilde{d}\right) \cdot P_{\boldsymbol{y}}\left(D = \tilde{d}\right)}.$$

In other words, $Q_x^*(D)$ corresponds to a distribution that is only off by a constant factor from the true discounted relevances of a ranking. However, these were the values we were

WWW '15 Florence, Italy
ACM 978-1-4503-3473-0/15/05.
http://dx.doi.org/10.1145/2740908.2742565.

	TRUE	UBIS-1	UNI-1	DEEP-1	TOP-1	UBIS-5	UNI-5	DEEP-5	TOP-5
OPT	284.4	284.18 ± 2.85	281.98 ± 6.03	284.02 ± 3.05	5.77	284.35 ± 1.09	283.84 ± 3.20	284.24 ± 1.58	16.98
REV-75	277.63	277.90 ± 3.00	278.33 ± 4.58	278.04 ± 4.19	3.30	277.30 ± 1.02	277.78 ± 2.57	278.11 ± 1.70	10.00
SHIFT-5	274.94	274.85 ± 2.48	275.44 ± 6.25	274.54 ± 3.46	0.00	275.13 ± 1.00	275.29 ± 2.67	275.21 ± 1.50	4.72
REV-150	271.32	272.07 ± 2.00	270.93 ± 4.81	271.10 ± 3.40	2.89	271.74 ± 0.84	271.28 ± 2.32	271.50 ± 1.35	8.51
SHIFT-7	269.87	269.48 ± 2.05	269.46 ± 6.16	268.84 ± 3.42	0.00	269.67 ± 0.90	270.20 ± 2.33	268.96 ± 1.48	0.00

Table 1: Mean AvgDCG estimates for 25 runs on a synthetic dataset with 6000 queries. The error intervals correspond to the standard deviations across different runs. Error intervals for TOP-k are zero, since it is a deterministic method.

wishing to estimate in the first place. Hence, in practice, one chooses thick-tailed functions that approximate $Q^*_x(D)$ reasonably well [12]. Choosing $Q_x(D)$ also allows us to include any prior knowledge we might have about the rankings. Aggregating the individual $DCG(x, \boldsymbol{y})$ estimates into an estimate for $AvgDCG(\mathcal{Y})$, one can estimate the variance of our importance sampler using the same samples.

1.2 Experiment and Results

To verify our approach, we designed the following experiment using synthetic data. We generated a sample $\mathcal{Y}$ of 6000 rankings $\boldsymbol{y}$ with 2000 items each. The ground truth judgments $f_x(d) \in \{0, ...4\}$ for each ranking were drawn from a categorical distribution whose parameters were drawn from a Dirichlet distribution with $\alpha = (0.54, 0.25, 0.175, 0.03, 0.005)$. As the sampling distribution, we chose

$$Q_x(d) \propto \tilde{f}_x(d) \cdot P_{\boldsymbol{y}}(d) + \epsilon,$$

where $\tilde{f}_x(d) = 5 \cdot (1 - \frac{\text{rank}(d,\boldsymbol{y})}{2000})$ and $\epsilon = 0.05$. This choice of $\tilde{f}_x(d)$ reflects that more relevant documents are typically at the top of the ranking, and adding a small constant ϵ ensures that we have sufficiently heavy tails. An interesting open problem is how to pick $\tilde{f}_x(d)$ adaptively, which will probably lead to even better estimation accuracy.

Our baselines mimic two traditional approaches. Given a budget of $6000 \cdot k$ judgements, the *TOP-k* baseline gets judgments for the top k documents of each ranking and computes the average *DCG* based on these judgments. The *DEEP-k* baseline, on the other hand, gets judgments much deeper than TOP-k. To meet the same budget, it randomly selects $3 \cdot k$ of the rankings and gets them judged to full depth. Our *UBIS-k* approach samples k judgements from $Q_x(D)$ for each query and computes $\widehat{AvgDCG}$. Finally, *UNI-k* samples k judgements uniformly and also computes $\widehat{AvgDCG}$.

We simulate ranking systems of different ranking quality in the following way. Given a fixed set of true relevances $f_x(d)$ generated as described above, *OPT* denotes the perfect ranking function where each ranking $\boldsymbol{y} \in \mathcal{Y}$ is sorted according to the true relevances $f_x(d)$. The *SHIFT-m* ranking system shifts all rankings of OPT by m entries to the right; elements that get shifted beyond the last position get re-introduced at the first. *REV-m* reverses the order of the top m elements in OPT. As REV-m merely re-ranks the top few documents, its performance should degrade gracefully with m.

Table 1 shows the average across 25 runs for all estimators, using $k = 1$ and $k = 5$. The first observation to note is that – not surprisingly – TOP-k is heavily biased and systematically underestimates the true *AvgDCG*, while our UBIS-k method as well as DEEP-k are unbiased. Furthermore, the bias of TOP-k is so strong that its average estimates does not even reflect the correct ordering of the ranking functions by true AvgDCG.

Since UBIS-k and DEEP-k rely on sampling, we also need to consider the variances of these estimators. For UBIS-5, the *AvgDCG* estimates of all ranking functions are separated by at least two standard deviations, indicating that UBIS-5 (and in many cases even UBIS-1) can reliably distinguish the ranking performance of the systems. The standard deviations of DEEP-k, however, are much larger than for UBIS-k, since DEEP-k suffers from between-query variability when sampling only $3 \cdot k$ rankings. However, the DCG values of single rankings in *OPT* do not vary too much; this explains why DEEP-k has smaller variance than UNI-k on our synthetic dataset. Also, the more informative sampling prior of UBIS-k greatly reduces variance compared to the uniform sampling strategy of UNI-k.

1.3 Conclusions and Future Work

In summary, we have shown that we can obtain unbiased *DCG* estimates in a simple and practical way using importance sampling. Our experiments showed that even with a small budget of expert judgements, we can obtain DCG estimates that are substantially more accurate than conventional budgeting methods both in terms of bias and variance.

Our plan for the future is to extend this work in three directions. First, we want to look at how to best sample in order to evaluate two or more systems simultaneously. Of course, this also requires a different notion of optimality for our estimators. A natural extension to k systems would be to find the sampling function that minimizes the average variance of all individual estimators. As previous work has only followed heuristic approaches [1], we want to be able to answer this question more formally.

The second direction addresses the task of finding the best out of k systems. In particular, we may be able to adaptively sample to answer this question, where the estimator $Q_x(D)$ changes over time (in the spirit of active learning).

Lastly, it is appealing to not use manual relevance judgments for evaluation, but to use ratings that were provided by the users (e.g., Netflix). However, these ratings were not logged under a $Q_x(D)$ that was controlled by us. We therefore would need to estimate the $Q_x(D)$ that generated the data in order to de-bias the log data. Such approaches have proven successful in counterfactual contextual bandit evaluation [9]. It is an open question how effectively this can be done in real-world settings.

Acknowledgments

This research was funded in part through NSF Awards IIS-1247637 and IIS-1217686, the JTCII Cornell-Technion Research Fund, and a gift from Bloomberg.

2. REFERENCES

[1] J. A. Aslam, V. Pavlu, and E. Yilmaz. A statistical method for system evaluation using incomplete judgments. In *SIGIR*, pages 541–548, 2006.

[2] B. Carterette, J. Allan, and R. Sitaraman. Minimal test collections for retrieval evaluation. In *SIGIR*, pages 268–275, 2006.

[3] B. Carterette, V. Pavlu, E. Kanoulas, J. A. Aslam, and J. Allan. If I had a million queries. In *ECIR*, pages 288–300, 2009.

[4] K. Hofmann, S. Whiteson, and M. de Rijke. Estimating interleaved comparison outcomes from historical click data. In *CIKM: Short Papers*, pages 1779–1783, 2012.

[5] K. Järvelin and J. Kekäläinen. Cumulated gain-based evaluation of ir techniques. *TOIS*, 20(4):422–446, 2002.

[6] L. Li, W. Chu, J. Langford, and X. Wang. Unbiased offline evaluation of contextual-bandit-based news article recommendation algorithms. In *WSDM*, pages 297–306, 2011.

[7] L. Li, J. Y. Kim, and I. Zitouni. Toward predicting the outcome of an A/B experiment for search relevance. In *WSDM*, pages 37–46, 2015.

[8] R. Nuray and F. Can. Automatic ranking of retrieval systems in imperfect environments. In *SIGIR*, pages 379–380, 2003.

[9] A. Strehl, J. Langford, L. Li, and S. M. Kakade. Learning from logged implicit exploration data. In *NIPS*, pages 2217–2225. 2010.

[10] L. Wasserman. *All of statistics: a concise course in statistical inference*. Springer, 2004.

[11] E. Yilmaz, E. Kanoulas, and J. A. Aslam. A simple and efficient sampling method for estimating AP and NDCG. In *SIGIR*, pages 603–610, 2008.

[12] C. Yuan and M. J. Druzdzel. How heavy should the tails be? In *FLAIRS*, pages 799–805, 2005.