Multileaving for online evaluation of rankers

Brian Brost*
University of Copenhagen
Department of Computer Science

ABSTRACT

In online learning to rank we are faced with a tradeoff between exploring new, potentially superior rankers, and exploiting our pre-existing knowledge of what rankers have performed well in the past. Multileaving methods offer an attractive approach to this problem since they can efficiently use online feedback to simultaneously evaluate a potentially arbitrary number of rankers. In this talk we discuss some of the main challenges in multileaving, and discuss promising areas for future research.

ACM Reference format:
Brian Brost. 2017. Multileaving for online evaluation of rankers. In *Proceedings of the first Internationl Workshop on LEARning Next gEneration Rankers, Amsterdam, The Netherlands, October 1, 2017 (LEARNER'17),* 2 pages.

1 INTRODUCTION

Online evaluation of rankers is the evaluation of rankers in a *fully functioning* system based on *implicit* measurement of *real users'* experiences of the system in a *natural* usage environment [5].

By evaluating in a natural usage environment, i.e. based on how people use the system in their day-to-day lives, we can avoid a key problem of offline evaluation methods, that they can only approximate a real user's feedback, and we can avoid the possible distortions that can occur due to the less natural usage environment present in a lab-based study. Furthermore user behaviour can be easily logged with no additional effort from the user. This provides online evaluation methods with inexpensive access to large amounts of timely training data [4]. On the other hand, the implicit measurement of feedback, meaning the logging of clicks and other user behaviours, is noisy and difficult to interpret. As a result, these large amounts of training data are necessary to reliably infer quality differences between rankers. A click during a web search session can be a mistake, or even if it is not a mistake, cannot be interpreted as an absolute signal of quality. Instead, a click on a given document may only support the relative judgement that this document was more useful than the other documents inspected by the user.

One of the key drawbacks of online evaluation methods is that the outputs of new, potentially poor, rankers are presented to actual users. If a new ranker is poor, users will be presented with poor results and, in the worst case, might abandon the service [6]. Conversely, if new rankers are not presented there is a risk of overlooking better rankers in the pool of rankers. In online learning the question of determining a proper exploration level is known as the *exploration-exploitation tradeoff*. The problem of managing this tradeoff for multileaving methods was first addressed in [3].

The gold standard for online evaluation of rankers is *A/B testing* of the rankers on separate random subsets of the users or queries [8]. A/B testing allows for rankers to be compared on real users, according to the exact, specific use case that the experimenter wishes to examine, and according to the exact metric by which the experimenter measures success. The primary cost associated with A/B testing is the number of user impressions that are required to reliably distinguish performance. Since we can measure exactly what we want with A/B testing, the goal of alternative online evaluation methods should be to replicate the expected outcomes of A/B tests, while requiring fewer user impressions than A/B testing.

In online evaluation, it is often easier for users to make relative judgements, rather than absolute judgements. For example, it is easier for a user to say that document A is more relevant for a certain query than document B, than to say how relevant each document is. This intuition partly motivates the introduction of *interleaving* as a method to compare rankers. Interleaving methods have two stages, and compare pairs of rankers by first combining the ranked lists produced by each ranker into a single ranked list and displaying this list to the user. They then infer which ranker is better from implicit feedback, e.g. clicks, collected from the user. This approach has the benefit that the comparison is carried out on the same user, eliminating the between user variance which would affect a comparison between rankers A and B on separate users. Interleaving methods were found to require 1-2 orders of magnitude less interaction data than absolute metrics to detect even small differences in retrieval quality [4]. Additionally, it has been shown that the credit inference stage of interleaving methods can be tuned so that their outcomes agree well with the relative outcomes of A/B testing [8].

Multileaving is a generalisation of interleaving that allows more than two rankers to be simultaneously compared [2, 7, 9]. In this case, $K > 2$ rankers are compared by creating a new ranked results list that consists of documents selected from the documents retrieved by the K rankers and then inferring based on the user's clicks how good each ranker is. Multileaving has been shown to use click feedback more efficiently than interleaving [9].

Like interleaving, multileaving methods have two distinct stages; the first stage involves sampling the documents to be displayed to the user, and the second stage assigns credit to the rankers based on the user's clicks. The sampling stage is often a straightforward generalization of those proposed in the interleaving literature, for example one method is to randomly order the rankers and then, in turns, sample the top remaining document from each ranker. These

*Also with Spotify Research.

LEARNER'17, October 1, 2017, Amsterdam, The Netherlands

sampling strategies are not uniform, i.e. some documents are much more likely to be sampled than others. The expected outcome of the credit assignment stage is affected by the probabilities of the documents being sampled during the first stage. Specifically, the ranker quality estimates in multileaving methods are skewed by artefacts of the sampling process, and this can cause substantial errors in the accuracies of multileaving estimates of ranker quality.

Multileaving offers dramatically improved efficiency over interleaving, allowing large numbers of rankers to be compared with very little interaction data, however this comes at the price of the above described problem which can affect the accuracy of the comparisons[1].

2 FUTURE CHALLENGES

This talk identified several challenges and areas for future work in multileaving.

- An important challenge for multileaving methods is to maximize agreement with A/B testing. Ideally, the goal should be to provide theoretical guarantees that the outcome of a multileaving experiment agrees with that of A/B testing for a broad class of A/B testing metrics, and under as weak a set of assumptions as possible.
- There are tradeoffs between the efficiency with which we can learn, and the quality of the displayed list to the user. For example, if all the rankers being compared agree about a given document being relevant, it is probably a good idea from a user experience point of view to display this document to users, but we learn nothing about the relative quality of the rankers from clicks on this document. How can this tradeoff between information gain and document quality be managed in an optimal manner?
- Can document selection be done in a more intelligent manner than those currently employed by multileaving methods? In particular is it possible to aggregate the information between rankers during learning so that the multileaved list can be expected to be better than that of most of the individual rankers?
- How can counterfactual methods and multileaving methods be combined to minimize the deterioration in user experience during evaluation?

REFERENCES

[1] B. Brost. *Online Evaluation of Rankers Using Multileaving*. University of Copenhagen, 2017.
[2] B. Brost, I. J. Cox, Y. Seldin, and C. Lioma. An improved multileaving algorithm for online ranker evaluation. In *Proceedings of the 39th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 745–748, 2016.
[3] B. Brost, Y. Seldin, I. J. Cox, and C. Lioma. Multi-dueling bandits and their application to online ranker evaluation. In *Proceedings of the 25th International ACM Conference on Information and Knowledge Management*, pages 2161–2166. ACM, 2016.
[4] O. Chapelle, T. Joachims, F. Radlinski, and Y. Yue. Large-scale validation and analysis of interleaved search evaluation. *ACM Transactions on Information Systems (TOIS)*, 30(1):6, 2012.
[5] K. Hofmann, L. Li, F. Radlinski, et al. Online evaluation for information retrieval. *Foundations and Trends® in Information Retrieval*, 10(1):1–117, 2016.
[6] K. Hofmann, S. Whiteson, and M. de Rijke. Balancing exploration and exploitation in learning to rank online. In *Advances in Information Retrieval*, pages 251–263. Springer, 2011.
[7] A. Schuth, R.-J. Bruintjes, F. Buüttner, J. van Doorn, C. Groenland, H. Oosterhuis, C.-N. Tran, B. Veeling, J. van der Velde, R. Wechsler, et al. Probabilistic multileave for online retrieval evaluation. In *Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 955–958. ACM, 2015.
[8] A. Schuth, K. Hofmann, and F. Radlinski. Predicting search satisfaction metrics with interleaved comparisons. In *Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 463–472. ACM, 2015.
[9] A. Schuth, F. Sietsma, S. Whiteson, D. Lefortier, and M. de Rijke. Multileaved comparisons for fast online evaluation. In *Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management*, pages 71–80. ACM, 2014.

[1] This introduction is adapted from [1]