Contextual Bandits for Information Retrieval

Katja Hofmann **Shimon Whiteson** **Maarten de Rijke**

ISLA, University of Amsterdam
Amsterdam, The Netherlands
{K.Hofmann,S.A.Whiteson,deRijke}@uva.nl

Abstract

In this paper we give an overview of and outlook on research at the intersection of information retrieval (IR) and contextual bandit problems. A critical problem in information retrieval is online learning to rank, where a search engine strives to improve the quality of the ranked result lists it presents to users on the basis of those users' interactions with those result lists. Recently, researchers have started to model interactions between users and search engines as contextual bandit problems, and initial methods for learning in this setting have been devised. Our research focuses on two aspects: balancing exploration and exploitation and inferring preferences from implicit user interactions. This paper summarizes our recent work on online learning to rank for information retrieval and points out challenges that are characteristic of this application area.

1 Introduction

Research at the intersection of information retrieval (IR) and reinforcement learning (RL) has recently gained increasing interest. Problems like learning what news articles are most likely to be interesting to a user (news recommendation) or which ads to place on a website (ad placement) can be naturally represented as e.g., contextual bandit problems [10, 12]. These formulations allow insights and methods from the bandit literature to be applied and extended to address these problems. For research on contextual bandits, this work has opened up a realistic application area where new approaches can be evaluated on large-scale datasets [13].

This paper considers the IR problem of online learning to rank. Users submit queries to an IR system (i.e., a search engine), which matches it against a document collection to construct a result list. The result list should rank documents according to how likely they are to fulfill the information need of the user that is expressed by the query. In an online learning system, the interactions of the user with the result list can be used to infer feedback about the ranking. This feedback can then be applied to learn better rankings.

Modern search engines typically construct result rankings from hundreds of information sources (features). Currently, a ranking function based on those features is typically either tuned manually or learned via supervised methods. Both approaches require access, before deployment, to a representative sample of labeled training data specifying which documents are relevant for certain queries. This is realistic only in a small portion of application areas, e.g., large scale web search engines. In many other areas, e.g., enterprise search, personal search, etc., such data is simply not available. This problem has recently been addressed by online learning to rank approaches that can exploit feedback inferred from user interactions.

On the surface, modeling online learning to rank for IR as a contextual bandit problem seems like a natural approach. However, a number of unique challenges cannot be addressed by simply applying existing contextual bandit algorithms. For example, the retrieval system does not select individual actions, but constructs result lists from several documents, so that one result list can contain both

exploratory and exploitative elements. Also, user interactions do not provide an explicit reward signal. Instead, only noisy, relative feedback can be inferred from such interactions.[1]

In the remainder of this paper, we first formalize the problem of learning to rank as a contextual bandit problem. We then give a brief overview of our work on balancing exploration and exploitation and inferring feedback from user interactions in this setting. We conclude with an outlook on future work and the unique challenges posed by online learning to rank for IR.

2 Problem formulation

We model online learning to rank for IR as a continuous cycle of user interactions with an IR system. Because we assume that the queries are independent, a natural model for this setting is the contextual bandit problem, in which rewards depend on the observed state of the environment [1, 11].

Figure 1 shows the interaction cycle. A user submits a query to a retrieval system, which generates a document list and presents it to the user. The user interacts with the list, e.g., by clicking on links, from which the retrieval system infers feedback about the quality of the presented document list. This feedback is then used to update the ranking function, with the goal to generate better rankings in the future. This completes the cycle and the next user query can be processed. This problem formulation directly translates to an RL problem (cf., Figure 1, terminology in italics) in which the retrieval system tries, based only on implicit feedback, to maximize a hidden reward signal that corresponds to some measure of result list quality (in IR, result list quality is typically assessed using evaluation measures based on the relevance grade and rank of the documents in the result list, such as Normalized Discounted Cumulative Gain (NDCG) [8]).

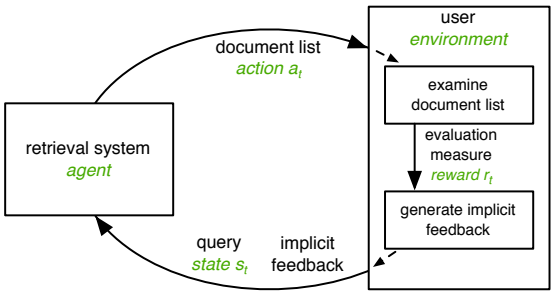

Figure 1: The IR problem modeled as a contextual bandit problem, with IR terminology in black and corresponding RL terminology in green and italics.

Because the retrieval system learns while interacting with users, its task is to optimize retrieval performance *while learning*. Previous work in learning to rank for IR has considered only final performance, i.e., performance on unseen data after training is completed [14], and, in the case of active learning, learning speed in terms of the number of required training samples [22]. Here, we measure performance in terms of cumulative reward, i.e., the sum of rewards over all queries addressed during learning [19]. Many definitions of cumulative reward are possible, depending on the modeling assumptions. We assume an infinite horizon problem, a model that is appropriate for IR learning to rank problems that run indefinitely. One issue with infinite horizon problems is the *infinitely delayed splurge*: since there are always infinitely many timesteps to go, the agent always explores, confident that enough time remains to exploit. To address the issue, infinite horizon problems typically include a *discount factor* $\gamma \in [0, 1)$ which weights immediate rewards higher than future rewards. The specific value of γ in IR problems is likely to be context-dependent, with lower values for quickly changing search environments, such as real-time search. How to set γ for a given search environment is an open problem. Given these assumptions, cumulative reward is defined as the discounted infinite sum of rewards r_i: $C = \sum_{i=1}^{\infty} \gamma^{i-1} r_i$.

The resulting problem formulation differs from those traditionally used in IR because performance depends on cumulative reward during the entire learning process, rather than just the quality of the final retrieval system produced by learning. It also differs from typical contextual bandit problems, which assume that the agent has access to the true immediate reward resulting from its actions. Typical IR evaluation measures require explicit feedback, which is not available in most realistic use cases for online learning to rank. Thus, this contextual bandit problem is distinct in that it requires the learner to cope with implicit feedback such as click behavior.

[1] In some retrieval settings, clicks are relatively reliable. For such settings, approaches optimizing for click-through rate are effective [16, 18] (these are similar to approaches for ad-placement and news recommendation). Here we focus on the more general retrieval problem where clicks can be noisy.

3 Balancing exploration and exploitation in online learning to rank for IR[2]

An online learning to rank system for IR can obtain feedback only on results that users actually examine, which typically includes only a few top-ranked documents. To obtain feedback on other documents, the system must explore by trying out variations of the current ranking that could lead to a better solution. However, the system also needs to ensure that the quality of result lists is high throughout the lifetime of the search engine, as otherwise users may be dissatisfied with the results. Thus, it has to exploit what is already known to be a good ranking. Clearly, this results in an exploration/exploitation dilemma. While this dilemma is a central challenge in research on contextual bandits, it is unclear what role, if any, it plays in practical IR settings. Our work is the first to demonstrate that balancing exploration and exploitation can substantially improve the online performance of learning to rank for IR approaches.

We developed the first online learning to rank method that could balance exploration and exploitation when learning from noisy, relative feedback. It is based on a (purely exploratory) stochastic gradient method that can learn from implicit feedback [23]. The balance between exploration and exploitation is based on the well-studied ϵ-greedy approach (cf., [20]), which we adjusted to the relative-feedback setting as follows. We maintain two document lists, one exploitative (based on the currently learned best ranking), and one exploratory (introducing variations to the current best ranking to explore potential improvements). An exploration rate k determines the relative number of documents each list contributes to the final interleaved list shown to the user. Unlike in ϵ-greedy methods, each action, i.e., interleaved list, can contain both exploratory and exploitive elements.

To compare our method to the baseline approach, we developed a new evaluation framework that can simulate user clicks on arbitrary result lists, and measure *online performance*, i.e., the quality of search results that a user would experience. We achieved this by leveraging existing learning-to-rank data sets [15] and recently developed click models [2–4] so that the performance of online learning to rank methods can be systematically explored under different assumptions, e.g., varying levels of noise in user clicks.

Table 1: Results, balancing exploration and exploitation in online learning to rank for IR. $^{\triangle}$ and $^{\blacktriangle}$ indicate statistically significant improvements of *exploit* runs ($k \in [0.1, 0.4]$) over the purely exploratory baseline ($k = 0.5$).

k	*0.5*	*0.4*	*0.3*	*0.2*	*0.1*
HP2003	102.58	109.78$^{\blacktriangle}$	**118.84**$^{\blacktriangle}$	116.38$^{\blacktriangle}$	117.52$^{\blacktriangle}$
HP2004	89.61	97.08$^{\blacktriangle}$	99.03$^{\blacktriangle}$	103.36$^{\blacktriangle}$	**105.69**$^{\blacktriangle}$
NP2003	90.32	100.94$^{\blacktriangle}$	105.03$^{\blacktriangle}$	108.15$^{\blacktriangle}$	**110.12**$^{\blacktriangle}$
NP2004	99.14	104.34$^{\triangle}$	110.16$^{\blacktriangle}$	112.05$^{\blacktriangle}$	**116.00**$^{\blacktriangle}$
TD2003	70.93	75.20$^{\blacktriangle}$	**77.64**$^{\blacktriangle}$	77.54$^{\blacktriangle}$	75.70$^{\triangle}$
TD2004	78.83	80.17	82.40$^{\triangle}$	**83.54**$^{\blacktriangle}$	80.98
OHSUMED	125.35	126.92$^{\triangle}$	127.37$^{\blacktriangle}$	**127.94**$^{\blacktriangle}$	127.21
MQ2007	95.50	94.99	95.70	**96.02**	94.94
MQ2008	89.39	90.55	91.24$^{\triangle}$	**92.36**$^{\blacktriangle}$	92.25$^{\blacktriangle}$

Table 1 shows our results when assuming a *navigational* click model, which simulates the relatively reliable click data observed in web search settings where users have a target website in mind (qualitatively similar results were obtained for both more and less reliable click models). These results show performance is highest for exploration rates $k \in [0.1, 0.3]$. Thus, on average, injecting only two exploratory documents in the top-10 result list is sufficient to optimize online performance.

These experiments validate our model of online learning to rank as a contextual bandit problem, and confirm our hypothesis that balancing exploration and exploitation is crucial for obtaining high retrieval performance while learning. In addition, they demonstrate that surprisingly little exploration is needed to achieve good learning performance.

4 A probabilistic method for inferring preferences from clicks

A critical component of systems for learning to rank from implicit feedback are methods to infer information from that feedback. Because feedback in typical IR settings is often noisy, and varies substantially across users and queries, *interleaved comparison* methods are used to infer relative judgments [17]. Interleaved comparison methods consist of two steps. One interleaving step, in which the two candidate result lists are combined into one interleaved result list that is then presented to the user, and one inference step, in which user clicks observed on the interleaved list are attributed to the candidate lists to determine the winner of the comparison.

[2]The work presented in Sections 3 and 4 is described in more detail in [6] and [7], respectively.

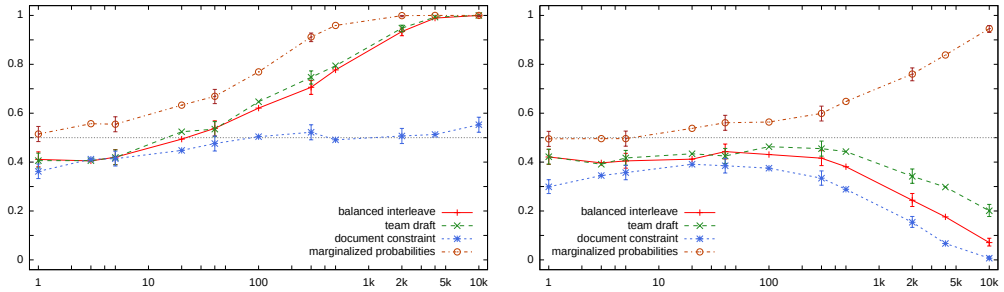

Figure 2: Assessment of interleaved comparison methods. Probability of preferring the correct ranker vs. number of interactions, under perfect (left) and noisy (right) feedback.

Previous interleaved comparison methods were shown to allow reliable comparisons but suffered from either bias or lack of sensitivity, especially when the candidate lists are similar. We developed a new probabilistic method for inferring such comparisons that addresses these problems. Our method models the interleaving step as repeated sampling without replacement from probability distributions over documents. The probability distributions are defined as softmax functions over the document rankings of the candidate lists, such that the highest-ranked documents are most likely to be drawn. After clicks are observed, comparisons are inferred by estimating the expected relative value of the two lists. Because both result lists use the same softmax function and judgments are based directly on the resulting expectation, this approach is unbiased. In addition, because a non-zero probability is assigned to every observed interleaved result list, it is possible to marginalize over click assignments, instead of relying only on the noisier observed samples of which candidate list contributed which document. Doing so improves generalization from sparse data, yielding a method that is highly sensitive even to small differences between document lists.

Figure 2 shows how our method performs on perfect and noisy click data. Our analytical and experimental results show that our approach can compare ranked result lists more accurately than previous methods, because it addresses problems of bias and sensitivity in earlier methods. In addition, our approach is robust to noise in user feedback, which makes it applicable in a much wider setting.

5 Ongoing work and challenges

Since research on online learning to rank for IR is only beginning, many directions for future research exist. Methods for comparing rankers have so far focused only on comparing individual ranker pairs. For learning methods, which require many such comparisons, methods that use available data much more efficiently can be devised. One interesting direction for research is to develop methods that can efficiently identify the best subset of a given pool of rankers. Such methods could considerably speed-up stochastic learning. Possible starting points are [9, 21].

Also, feedback can so far only be collected during live interactions with users. Many hundreds to thousands of such interactions are required for comparing a single pair of rankers and the collected data cannot currently be reused. Solutions developed for related areas, such as news recommendation, cannot be applied directly, as they require logged data that provides reasonable coverage of possible state-action pairs [13]. In IR, this form of data collection is infeasible, as the quality of the result lists during data collection would be too low. However, the probabilistic nature of our interleaved comparison method makes it possible in principle to use importance sampling to judge the relative quality of two result lists based on clicks obtained from interleaving two other lists. By reusing historical data, such an approach could greatly improve the sample efficiency of online learning to rank. This direction is the current focus of our ongoing research.

In addition, current approaches for learning from implicit feedback, and for balancing exploration and exploitation are quite simple. For example, our work is based on a stochastic gradient method that performs random local exploration of the solution space. Thus, the potential exists for large improvement using more sophisticated approaches such as estimation of distribution methods [5].

Finally, the assumption that queries are independent, which enables the application of contextual-bandit algorithms, cannot model more complex settings containing sessions of consecutive queries. Additional research is needed to formulate such settings as a subclass of (partially observable) Markov decision processes and develop efficient algorithms to exploit their unique characteristics.

Acknowledgments

This research was partially supported by the European Union's ICT Policy Support Programme as part of the Competitiveness and Innovation Framework Programme, CIP ICT-PSP under grant agreement nr 250430, the European Community's Seventh Framework Programme (FP7/2007-2013) under grant agreements nr 258191 (PROMISE Network of Excellence) and 288024 (LiMoSINe project), the Netherlands Organisation for Scientific Research (NWO) under project nrs 612.061.-814, 612.061.815, 640.004.802, 380-70-011, 727.011.005, the Center for Creation, Content and Technology (CCCT), the Hyperlocal Service Platform project funded by the Service Innovation & ICT program, the WAHSP project funded by the CLARIN-nl program, under COMMIT project Infiniti and by the ESF Research Network Program ELIAS.

References

[1] A. G. Barto, R. S. Sutton, and P. S. Brouwer. Associative search network: A reinforcement learning associative memory. *IEEE Trans. Syst., Man, and Cybern.*, 40:201–211, 1981.

[2] N. Craswell, O. Zoeter, M. Taylor, and B. Ramsey. An experimental comparison of click position-bias models. In *WSDM '08*, pages 87–94, 2008.

[3] F. Guo, L. Li, and C. Faloutsos. Tailoring click models to user goals. In *WSCD '09*, pages 88–92, 2009.

[4] F. Guo, C. Liu, and Y. M. Wang. Efficient multiple-click models in web search. In *WSDM '09*, pages 124–131, 2009.

[5] N. Hansen, S. Müller, and P. Koumoutsakos. Reducing the time complexity of the derandomized evolution strategy with covariance matrix adaptation (CMA-ES). *Evolutionary Computation*, 11(1):1–18, 2003.

[6] K. Hofmann, S. Whiteson, and M. de Rijke. Balancing exploration and exploitation in learning to rank online. In *ECIR'11*, pages 251–263, 2011.

[7] K. Hofmann, S. Whiteson, and M. de Rijke. A probabilistic method for inferring preferences from clicks. In *CIKM '11*, 2011.

[8] K. Järvelin and J. Kekäläinen. Cumulated gain-based evaluation of ir techniques. *ACM Trans. Inf. Syst.*, 20(4):422–446, 2002.

[9] S. Kalyanakrishnan and P. Stone. Efficient selection of multiple bandit arms: Theory and practice. In *ICML'10*, 2010.

[10] J. Langford, A. Strehl, and J. Wortman. Exploration scavenging. In *ICML '08*, pages 528–535, 2008.

[11] J. Langford and T. Zhang. The epoch-greedy algorithm for multi-armed bandits with side information. In *NIPS'08*, pages 817–824, 2008.

[12] L. Li, W. Chu, J. Langford, and R. E. Schapire. A contextual-bandit approach to personalized news article recommendation. In *WWW '10*, pages 661–670, 2010.

[13] L. Li, W. Chu, J. Langford, and X. Wang. Unbiased offline evaluation of contextual-bandit-based news article recommendation algorithms. In *WSDM '11*, pages 297–306, 2011.

[14] T. Y. Liu. Learning to rank for information retrieval. *Foundations and Trends in Information Retrieval*, 3(3):225–331, 2009.

[15] T.-Y. Liu, J. Xu, T. Qin, W. Xiong, and H. Li. Letor: Benchmark dataset for research on learning to rank for information retrieval. In *LR4IR '07*, 2007.

[16] T. Moon, L. Li, W. Chu, C. Liao, Z. Zheng, and Y. Chang. Online learning for recency search ranking using real-time user feedback. In *CIKM '10*, pages 1501–1504, 2010.

[17] F. Radlinski and N. Craswell. Comparing the sensitivity of information retrieval metrics. In *SIGIR '10*, pages 667–674, 2010.

[18] F. Radlinski, R. Kleinberg, and T. Joachims. Learning diverse rankings with multi-armed bandits. In *ICML '08*, pages 784–791. ACM, 2008.

[19] R. S. Sutton and A. G. Barto. *Reinforcement learning: An introduction*. MIT Press, Cambridge, MA, USA, 1998.

[20] C. Watkins. *Learning from Delayed Rewards*. PhD thesis, Cambridge University, 1989.

[21] S. Whiteson and P. Stone. On-line evolutionary computation for reinforcement learning in stochastic domains. In *GECCO '06*, pages 1577–1584, 2006.

[22] Z. Xu, K. Kersting, and T. Joachims. Fast active exploration for link-based preference learning using gaussian processes. In *ECML PKDD'10*, pages 499–514, 2010.

[23] Y. Yue and T. Joachims. Interactively optimizing information retrieval systems as a dueling bandits problem. In *ICML'09*, pages 1201–1208, 2009.