Online Learning to Rank for Information Retrieval

SIGIR 2016 Tutorial

Artem Grotov
University of Amsterdam
Amsterdam, The Netherlands
a.grotov@uva.nl

Maarten de Rijke
University of Amsterdam
Amsterdam, The Netherlands
derijke@uva.nl

ABSTRACT

During the past 10–15 years offline learning to rank has had a tremendous influence on information retrieval, both scientifically and in practice. Recently, as the limitations of offline learning to rank for information retrieval have become apparent, there is increased attention for *online* learning to rank methods for information retrieval in the community. Such methods learn from user interactions rather than from a set of labeled data that is fully available for training up front.

Below we describe why we believe that the time is right for an intermediate-level tutorial on online learning to rank, the objectives of the proposed tutorial, its relevance, as well as more practical details, such as format, schedule and support materials.

Keywords

Online learning to rank; Bandit algorithms; Exploration vs. exploitation

1. INTRODUCTION

Today's search engines have developed into complex systems that combines hundreds of ranking criteria with the aim of producing the optimal result list in response to users' queries. For automatically tuning optimal combinations of large numbers of ranking criteria, learning to rank [22, LTR] has proved invaluable. For a given query, each document is represented by a feature vector. The features may be query dependent, document dependent or capture the relationship between the query and documents. The task of the learner is to find a model that combines these features such that, when this model is used to produce a ranking for an unseen query, user satisfaction is maximized.

Traditionally, learning to rank algorithms are trained in batch mode, on a complete dataset of query and document pairs with their associated manually created relevance labels. This setting has a number of disadvantages and is impractical in many cases. First, creating such datasets is expensive and therefore infeasible for smaller search engines, such as small web-store search engines [24]. Second, it may be impossible for experts to annotate documents, as in the case of personalized search [18]. Third, the relevance of documents to queries can change over time, like in a news search engine [7].

Online learning to rank addresses all of these issues by incrementally learning from user feedback in real time [34]. Online learning is closely related to active learning, incremental learning, and counterfactual learning. However, online learning is more difficult because the agent has to balance exploration and exploitation: actions with unknown performance have to be explored to learn better solutions [11].

There is a growing body of established methods for online learning to rank for information retrieval (see the schedule below for a broad range of examples). The time is right to organize and present this material to a broad audience of interested information retrieval researchers, whether junior or senior, whether academic or industrial. The online learning to rank methods available today have been proposed by different communities, in machine learning and information retrieval. A key aim of the tutorial is to bring these together and offer a unified perspective. To achieve this we illustrate the core and state of the art methods in online learning to rank, their theoretical foundations and real-world applications, as well as existing online learning algorithms that have not been used by information retrieval community so far.

We expect the tutorial to be useful for both academic and industrial researchers who either want to develop new online learning to rank methods, use them in their own research, or apply the methods described in the tutorial to improve search and recommendation systems.

2. OBJECTIVES

Online learning to rank from user interactions is fundamentally different from currently dominant supervised learning to rank approaches for information retrieval, where training data is assumed to be randomly sampled from some underlying distribution, and where absolute and reliable labels are provided by professional annotators [15]. When learning from user interactions, a system has no control over which queries it receives, it only receives feedback on the result lists it presents to users, and it has to present high quality result lists while learning, to satisfy user expectations.

Following Hofmann et al. [11], in this tutorial we formulate online learning to rank as a reinforcement learning problem, in which an *agent*, the search engine, learns from interactions with an *environment*, the user and their interactions, by trying out actions (e.g., returning a ranked list of items) and observing rewards (e.g., interpreting user feedback as absolute or relative feedback) in multiple rounds or discrete time steps; see Figure 1.

Particularly relevant to this tutorial are methods for tackling so-called *contextual bandit problems* (also known as bandits with side information) [1]. A contextual bandit problem is a special case of

SIGIR '16, July 17 - 21, 2016, Pisa, Italy

ACM ISBN 978-1-4503-4069-4/16/07.

DOI: http://dx.doi.org/10.1145/2911451.2914798

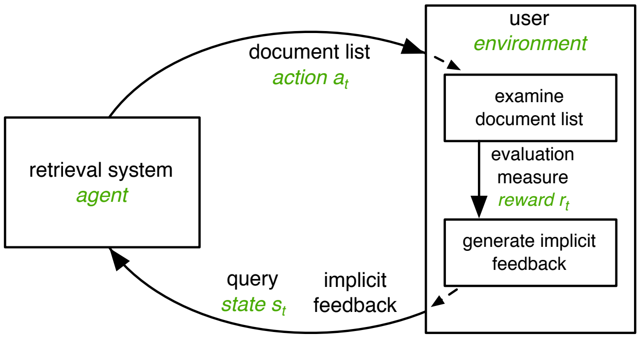

Figure 1: The information retrieval problem modeled as a contextual bandit problem, with information retrieval terminology in black and the corresponding reinforcement terminology in green and italics. (Figure taken from [11].)

a reinforcement learning problem in which states are independent of the agent's actions. In information retrieval terms, the context could consist of the user and the query and the actions are the search engine result pages. A difference between typical contextual bandit formulations and online learning to rank for information retrieval is that in information retrieval (absolute) rewards cannot be observed directly. Instead, feedback for learning is inferred from observed user interactions as noisy preference indications.

As we will demonstrate in the tutorial, an important benefit of reducing IR problems to bandit approaches is that the rich body of work on bandit approaches can be used. At the same time, information retrieval poses unique challenges that inspire additional research on bandit algorithms.

The objectives of the tutorial are as follows:

- To describe existing online LTR algorithms in a unified way, i.e., using common notation and terminology, so that different models can easily be related to each other.
- Explain the importance of balancing exploration and exploitation in an online LTR setting.
- To explain how to analyse the performance of online LTR algorithms and why it is worth the effort.
- To present appropriate experimental and evaluation methodologies for online LTR in both synthetic and real world settings.
- To describe how to deploy online LTR algorithms in an industrial setting.
- To present online learning algorithms that have not been used yet for learning to rank and indicate how IR researchers might go about using them.
- To discuss future directions of research in online LTR.

These objectives are meant to give participants a thorough understanding of existing learning to rank for information retrieval methods and to present online learning methods that have so far not been applied to learning to rank, let alone to learning to rank for information retrieval.

3. DETAILED SCHEDULE

The tutorial will be organized in two halves of 90 minutes each, each mixing theory and experiment, with formal analyses of online learning to rank methods interleaved with discussions of code and of experimental outcomes. Part I is aimed at familiarizing participants with the key concepts and algorithms. In Part II we select a small number of topics to provide a more in-depth technical treatment.

Part I

[10 minutes] Introduction, aims and historical notes

Here we discuss the context in which online LTR is applied and the most important historical milestones in its development.

[10 minutes] LTR in IR.

- The task of ranking documents, its importance, relationship to other machine learning tasks, and the unique challenges of LTR [22].
- Current approaches to LTR and argue that they all have issues that have to be addressed, such as the cost of producing labelled data and the mismatch between manually curated labels and user intent [34].

[15 minutes] Online LTR: balancing exploration and exploitation.

- How does online LTR addresses the shortcomings of offline LTR [11, 34]?
- How does online LTR relate to, and differ from, other tasks such as learning from labelled data, active learning and learning from logged interactions [30]?
- We explain the importance of balancing exploration and exploitation [13].

[5 minutes] Introduction to bandits and reinforcement learning.

- An important formalism behind many online LTR methods: bandit algorithms [1, 20]
- Connection to reinforcement learning [11, 29]
- Illustrate the importance of formal analysis and present k-armed bandits [1], contextual bandits [20] and cascading bandits [19] as ways to formalize the online LTR setting.

[10 minutes] Online signals to learn from.

- Close connection with online and logged based IR Evaluation [9, 14], because in both settings one needs to make the connection between observed user feedback and the hidden quality of the system.
- How to use observed user feedback to train the ranking models? The observed user feedback includes clicks, absence of clicks, dwell times, abandonment and many other signals [17, 21].
- Signals can be interpreted in a number of ways, for example, clicks can be interpreted in the form of absolute click through rates, or as relative preferences between documents or retrieval systems or using click models [16].

[20 minutes] Dueling bandit gradient descent.

- Dueling bandit gradient descent [34, DBGD], one of the core methods used in online LTR. We present the theory behind this method and discuss under which conditions it is guaranteed to work.
- More advanced methods that build on DBGD such as Probabilistic Multileave Gradient Decent [23, 27] and DBGD with Candidate Preselection [12].

[5 minutes] Real world applications.

- Real world applications of online LTR.
- Practical considerations such as what kind of infrastructure is required and what is important to log during online LTR.

[15 minutes] Discussion.

Part II

In this part we dig deeper into the foundations of some of the concepts introduced in Part I.

[5 minutes] Introduction.

- Focus on deepening the participants' understanding of the concepts introduced in part one.
- Connecting to online learning to rank methods not yet used in information retrieval.

[30 minutes] Online LTR in K-armed bandits setting.

- How to perform online LTR with a finite population of candidate rankers, framing it as a K-armed bandits problem [6].
- Challenges associated with deciding which ranker is the best among a population, the concept of Condorcet winner and Relative Upper Confidence Bound algorithm [35–38].

[20 minutes] Current problems: non-linear models, better exploration, safety guarantees, combining offline and online.

- State of the art in online LTR and ways to improve it [10].
- Non-linear neural network and tree ensemble-based ranking models [3], exploration strategies based on uncertainty estimations [31].

[10 minutes] Existing online algorithms not used in information retrieval.

- Online learning algorithms that have not been used in the LTR settings with the goal of inspiring researchers to adapt those algorithms for use in the IR community [4, 5]

[10 minutes] Datasets and resources.

- How to run online LTR experiments at home [25]
- CLEF LL4IR: Living Labs for IR Evaluation [26]
- TREC OpenSearch – Academic Search [32]

[15 minutes] Discussion and conclusion.

- Change the world!

4. TYPE OF SUPPORT MATERIALS TO BE SUPPLIED TO ATTENDEES

- Slides
- Draft survey on online learning to rank for information retrieval [8]
- Code and data samples to follow experimental segments of the tutoral
- Lerot – experimental environment for online learning to rank [25]

Acknowledgements

This research was supported by Ahold, Amsterdam Data Science, the Bloomberg Research Grant program, the Dutch national program COMMIT, Elsevier, the European Community's Seventh Framework Programme (FP7/2007-2013) under grant agreement nr 312827 (VOX-Pol), the ESF Research Network Program ELIAS, the Royal Dutch Academy of Sciences (KNAW) under the Elite Network Shifts project, the Microsoft Research Ph.D. program, the Netherlands eScience Center under project number 027.012.105, the Netherlands Institute for Sound and Vision, the Netherlands Organisation for Scientific Research (NWO) under project nrs 727.011.005, 612.001.116, HOR-11-10, 640.006.013, 612.066.930, CI-14-25, SH-322-15, 652.002.001, 612.001.551, the Yahoo Faculty Research and Engagement Program, and Yandex. All content represents the opinion of the authors, which is not necessarily shared or endorsed by their respective employers and/or sponsors.

REFERENCES

[1] Peter Auer. Using confidence bounds for exploitation-exploration trade-offs. *Journal of Machine Learnng Research*, 3:397–422, 2003.

[2] Alexey Borisov, Pavel Serdyukov, and Maarten de Rijke. Using metafeatures to increase the effectiveness of latent semantic models in web search. In *WWW 2016: 25th International World Wide Web Conference*. ACM, April 2016.

[3] Christopher J.C. Burges. From ranknet to lambdarank to lambdamart: An overview. Technical Report MSR-TR-2010-82, June 2010.

[4] Giuseppe Burtini, Jason Loeppky, and Ramon Lawrence. A survey of online experiment design with the stochastic multi-armed bandit. *CoRR*, abs/1510.00757, 2015. URL http://arxiv.org/abs/1510.00757.

[5] Róbert Busa-Fekete and Eyke Hüllermeier. A survey of preference-based online learning with bandit algorithms. In *Algorithmic Learning Theory: 25th International Conference, ALT 2014, Bled, Slovenia, October 8-10, 2014. Proceedings*, pages 18–39, Cham, 2014. Springer International Publishing.

[6] Róbert Busa-Fekete and Eyke Hüllermeier. A survey of preference-based online learning with bandit algorithms. In *ALT '14*, number 8776 in LNCS, pages 18–39. Springer, 2014.

[7] Susan T. Dumais. The web changes everything: Understanding and supporting people in dynamic information environments. In *Research and Advanced Technology for Digital Libraries, 14th European Conference, ECDL 2010*, 2010.

[8] Artem Grotov and Maarten de Rijke. Online learning to rank for information retrieval: A survey. *Draft*, 2016.

[9] Artem Grotov, Shimon Whiteson, and Maarten de Rijke. Bayesian ranker comparison based on historical user interactions. In *SIGIR 2015: 38th international ACM SIGIR conference on Research and development in information retrieval*. ACM, August 2015.

[10] Artem Grotov, Maarten de Rijke, and Shimon Whiteson. Online LambdaRank. In *Submitted*, 2016.

[11] Katja Hofmann, Shimon Whiteson, and Maarten de Rijke. Balancing exploration and exploitation in learning to rank online. In *ECIR 2011: 33rd European Conference on Information Retrieval*. Springer, April 2011.

[12] Katja Hofmann, Anne Schuth, Shimon Whiteson, and Maarten de Rijke. Reusing historical interaction data for faster online learning to rank for information retrieval. In

WSDM 2013: International Conference on Web Search and Data Mining. ACM, February 2013.

[13] Katja Hofmann, Shimon Whiteson, and Maarten de Rijke. Balancing exploration and exploitation in listwise and pairwise online learning to rank for information retrieval. *Information Retrieval Journal*, 16(1):63–90, February 2013.

[14] Katja Hofmann, Shimon Whiteson, and Maarten de Rijke. Fidelity, soundness, and efficiency of interleaved comparison methods. *ACM Transactions on Information Systems*, 31(3): Article 18, October 2013.

[15] Katja Hofmann, Shimon Whiteson, Anne Schuth, and Maarten de Rijke. Learning to rank for information retrieval from user interactions. *ACM SIGWEB Newsletter*, (Spring): 5:1–5:7, April 2014.

[16] Thorsten Joachims. Optimizing search engines using clickthrough data. In *Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '02, pages 133–142, New York, NY, USA, 2002. ACM.

[17] Youngho Kim, Ahmed Hassan, Ryen W. White, and Imed Zitouni. Modeling dwell time to predict click-level satisfaction. In *Proceedings of the 7th ACM International Conference on Web Search and Data Mining*, WSDM '14, pages 193–202, New York, NY, USA, 2014. ACM.

[18] Ron Kohavi, Roger Longbotham, Dan Sommerfield, and Randal M. Henne. Controlled experiments on the web: Survey and practical guide. *Data Mining and Knowledge Discovery*, 18(1):140–181, 2009.

[19] Branislav Kveton, Csaba Szepesvári, Zheng Wen, and Azin Ashkan. Cascading bandits. *CoRR*, abs/1502.02763, 2015. URL http://arxiv.org/abs/1502.02763.

[20] John Langford and Tong Zhang. The epoch-greedy algorithm for multi-armed bandits with side information. In J. C. Platt, D. Koller, Y. Singer, and S. T. Roweis, editors, *Advances in Neural Information Processing Systems 20*, pages 817–824. Curran Associates, Inc., 2008.

[21] Damien Lefortier, Pavel Serdyukov, and Maarten de Rijke. Online exploration for detecting shifts in fresh intent. In *CIKM 2014: 23rd ACM Conference on Information and Knowledge Management*. ACM, November 2014.

[22] Tie-Yan Liu. Learning to rank for information retrieval. *Found. Trends Inf. Retr.*, 3(3):225–331, March 2009.

[23] Harrie Oosterhuis, Anne Schuth, and Maarten de Rijke. Probabilistic multileave gradient descent. In *ECIR 2016: 38th European Conference on Information Retrieval*, LNCS. Springer, March 2016.

[24] Mark Sanderson. Test collection based evaluation of information retrieval systems. *Found. & Tr. Inform. Retr.*, 4(4): 247–375, 2010.

[25] Anne Schuth, Katja Hofmann, Shimon Whiteson, and Maarten de Rijke. Lerot: an online learning to rank framework. In *Living Labs for Information Retrieval Evaluation workshop at CIKM'13.*, 2013.

[26] Anne Schuth, Krisztian Balog, and Liadh Kelly. Overview of the living labs for information retrieval evaluation (ll4ir) clef lab 2015. In *Experimental IR Meets Multilinguality, Multimodality, and Interaction*, pages 484–496. Springer, 2015.

[27] Anne Schuth, Harrie Oosterhuis, Shimon Whiteson, and Maarten de Rijke. Multileave gradient descent for fast online learning to rank. In *WSDM 2016: The 9th International Conference on Web Search and Data Mining*. ACM, February 2016.

[28] Aleksandrs Slivkins, Filip Radlinski, and Sreenivas Gollapudi. Ranked bandits in metric spaces: learning optimally diverse rankings over large document collections. Technical report, arXiv preprint arXiv:1005.5197, 2010.

[29] Richard S. Sutton and Andrew G. Barto. *Reinforcement Learning: An Introduction*. MIT Press., 1998.

[30] Adith Swaminathan and Thorsten Joachims. Counterfactual risk minimization: Learning from logged bandit feedback. *CoRR*, abs/1502.02362, 2015. URL http://arxiv.org/abs/1502.02362.

[31] Aibo Tian and Matthew Lease. Active learning to maximize accuracy vs. effort in interactive information retrieval. In *Proceedings of the 34th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '11, pages 145–154, New York, NY, USA, 2011. ACM.

[32] TREC. OpenSearch track. http://trec-open-search.org, 2016.

[33] Aleksandr Vorobev, Damien Lefortier, Gleb Gusev, and Pavel Serdyukov. Gathering additional feedback on search results by multi-armed bandits with respect to production ranking. In *Proceedings of the 24th International Conference on World Wide Web*, pages 1177–1187. ACM, 2015.

[34] Yisong Yue and Thorsten Joachims. Interactively optimizing information retrieval systems as a dueling bandits problem. In *ICML '09*, 2009.

[35] Masrour Zoghi, Shimon Whiteson, Maarten de Rijke, and Remi Munos. Relative confidence sampling for efficient on-line ranker evaluation. In *7th ACM WSDM Conference (WSDM2014)*. ACM, February 2014.

[36] Masrour Zoghi, Shimon Whiteson, Remi Munos, and Maarten de Rijke. Relative upper confidence bound for the k-armed dueling bandit problem. In *ICML 2014: International Conference on Machine Learning*, June 2014.

[37] Masrour Zoghi, Shimon Whiteson, and Maarten de Rijke. Mergerucb: A method for large-scale online ranker evaluation. In *WSDM 2015: The Eighth International Conference on Web Search and Data Mining*. ACM, February 2015.

[38] Masrour Zoghi, Shimon Whiteson, Zohar Karnin, and Maarten de Rijke. Copeland dueling bandits. In *NIPS 2015*, December 2015.

[39] Masrour Zoghi, Tomáš Tunys, Lihong Li, Damien Jose, Junyan Chen, Chun Ming Chin, and Maarten de Rijke. Click-based hot fixes for underperforming torso queries. In *SIGIR 2016: 39th international ACM SIGIR conference on Research and development in information retrieval*. ACM, July 2016.