Interactive Information Retrieval with Bandit Feedback

Huazheng Wang, Yiling Jia, Hongning Wang
Department of Computer Science
University of Virginia
Charlottesville, VA, USA
{hw7ww,yj9xs,hw5x}@virginia.edu

ABSTRACT

Information retrieval (IR) in nature is a process of sequential decision making. The system repeatedly interacts with the users to refine its understanding of the users' information needs, improve its estimation of result relevance, and thus increase the utility of its returned results (e.g., the result rankings). Distinct from traditional IR solutions that rigidly execute an offline trained policy, interactive information retrieval emphasizes online policy learning. This, however, is fundamentally difficult for at least three reasons. First, the system only collects user feedback on the presented results, aka, the bandit feedback. Second, users' feedback is known to be noisy and biased. Third, as a result, the system always faces the conflicting goals of improving its policy by presenting currently underestimated results to users versus satisfying the users by ranking the currently estimated best results on top.

In this tutorial, we will first motivate the need for online policy learning in interactive IR, by highlighting its importance in several real-world IR problems where online sequential decision making is necessary, such as web search and recommendations. We will carefully address the new challenges that arose in such a solution paradigm, including sample complexity, costly and even outdated feedback, and ethical considerations in online learning (such as fairness and privacy) in interactive IR. We will prepare the technical discussions by first introducing several classical interactive learning strategies from machine learning literature, and then fully dive into the recent research developments for addressing the aforementioned fundamental challenges in interactive IR. Note that the tutorial on "*Interactive Information Retrieval: Models, Algorithms, and Evaluation*" [48] will provide a broad overview on the general conceptual framework and formal models in interactive IR, while this tutorial covers the online policy learning solutions for interactive IR with bandit feedback.

CCS CONCEPTS

• **Information systems → Retrieval models and ranking**; **Recommender systems**; • **Theory of computation → Online learning algorithms**.

SIGIR '21, July 11–15, 2021, Virtual Event, Canada

ACM ISBN 978-1-4503-8037-9/21/07…$15.00
https://doi.org/10.1145/3404835.3462810

KEYWORDS

Interactive information retrieval; bandit feedback

ACM Reference Format:
Huazheng Wang, Yiling Jia, Hongning Wang. 2021. Interactive Information Retrieval with Bandit Feedback. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '21), July 11–15, 2021, Virtual Event, Canada.* ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/3404835.3462810

1 MOTIVATION

Supervised machine learning techniques have been successfully applied in many information retrieval (IR) applications: for example learning to rank [25], matrix factorization [17], and deep neural networks [27]. Such supervised learning techniques estimate mappings from the input data to their expected labels (i.e., classification) or values (i.e., regression) based on the available annotated offline data. However, many real-world IR scenarios face the rapid appearance of new information (e.g., new users or content), together with the ever-changing nature of the users' information needs. It makes timely relevance annotation costly or even impossible. This urges us to move from this passive model learning and result serving paradigm to a more interactive or even proactive solution framework. For example, probe users for relevance feedback to estimate better mappings between the users and results. This provides a viable and principled solution for interactive IR.

However, we have to note that interactive IR cannot be simply addressed by updating an existing offline model online based on the newly observed data, for at least two reasons. First, the users can only examine the system returned results. For those not returned by the system, one cannot treat them as irrelevant. This is known as learning with bandit feedback. Second, user feedback itself is also noisy and biased. For example, the well-known position bias [14] and trust bias [2] complicate the modeling of user clicks. Repeated observations of results displayed at different positions become necessary to handle noise and biases. But as the system is learning while serving, any working interactive IR system has to carefully balance the need to focus on information that raises user interest, and simultaneously, the need to explore new information for globally improving user experience (i.e., the explore-exploit dilemma [7]). Therefore, efficient and effective techniques for online learning with bandit feedback are needed.

Significant research efforts have been spent on online learning with bandit feedback in both machine learning and information retrieval communities. And various solutions have been developed, such as multi-armed bandits [7], contextual bandits [1, 23], dueling bandits [46], and best arm identification [5]. Many algorithms have been successfully applied in interactive IR applications, such as personalized recommendation [23, 42] and online learning to rank

[36, 46]. However, beyond the standard settings of such learning solutions, new challenges arise when applying them in interactive IR scenarios, including sample complexity [36, 38, 41, 42], implicit feedback [32], costly and even outdated feedback [26, 43], and ethical considerations of exploration (such as fairness and privacy) [15, 35].

The importance and demand of the solutions for online learning with bandit feedback in interactive IR, and IR at large, is evident. We therefore structure our tutorial to provide a comprehensive introduction about the latest advances in this important direction of research, with a particular focus on covering the recently developed algorithms for those new challenges. To the best of our knowledge, none of the existing tutorials in IR community covered topics like online learning in non-stationary environments or the ethical considerations in online learning solutions. But given their importance is particularly evident to the IR community, we deeply believe in the necessity of presenting such a tutorial.

2 OBJECTIVE

The target audience of this tutorial includes researchers in IR, especially interactive IR, who are interested in research problems where interactive online learning is needed, and practitioners who may need to use online learning techniques to solve practical IR problems, for example, recommender systems, online learning to rank, or personalization problems.

The objective of this tutorial is to offer participants a comprehensive review on the recent algorithm developments in online learning with bandit feedback for interactive IR applications, and expose practitioners to a wide-range of important techniques to build and improve practical interactive IR systems. Specifically, we aim to achieve the following goals:

- Review the foundations of online learning with bandit feedback. We will provide a broad overview of existing learning techniques, covering their motivations, assumptions, application scenarios, practical implementations, and also limitations faced in interactive IR.
- Summarize and compare recent developments. We will systematically review the recent developments of bandit learning solutions in two major interactive IR application scenarios: recommender systems and learning to rank. We will categorize the challenges raised in these applications and discuss the corresponding solutions, with detailed comparisons on the existing algorithms and their theoretical guarantees.
- Facilitate future research. We will discuss the important open problems raised in recent research and highlight promising future directions for online learning with bandit feedback in interactive IR.

3 RELEVANCE TO THE IR COMMUNITY

Interactive IR is an important and well-recognized research topic in the IR community, since IR itself is an interactive process between the user and system in nature. Online learning with bandit feedback provides a principled and theoretically sound solution to address interactive IR. And there has been a growing interest in this direction in both academia and industry to develop new solutions under various new settings. And a growing number of publications have emerged in SIGIR and related conference including WWW, WSDM, KDD, and CIKM, focusing on different topics, such as recommender systems, learning to rank, and conversational systems. As a result, this tutorial is timely and of great interest to the broad attendants of SIGIR and the members of IR community.

3.1 Related tutorials

There are some recent tutorials discussing interactive information retrieval or interactive machine learning, with distinct focuses against ours. The tutors Huazheng Wang and Hongning Wang gave a tutorial on "*Learning by Exploration: New Challenges in Real-World Environments*" [44] at KDD 2020, discussed techniques in online learning with bandit feedback and their applications in recommender systems. The SIGIR 2020 tutorial on "*Interactive Information Retrieval: Models, Algorithms, and Evaluation*" [47] offered a broad overview on interactive information retrieval. The tutorial on "*Unbiased Learning to Rank: Counterfactual and Online Approaches*" [31] at WWW 2020 and SIGIR 2019 and the SIGIR 2016 tutorial on "*Online Learning to Rank for Information Retrieval*" [12] reviewed online learning to rank algorithms. The tutorial on "*Real World Interactive Learning*" at KDD2018 and ICML2017 summarized real world challenges in developing interactive machine learning systems.

Our tutorial is different from these previous one in the sense that we will systematically review the recent progress and new solutions for two important interactive IR tasks, i.e., recommender system and learning to rank. We will focus on several new challenges that arise in real-world interactive IR applications, such as more efficient exploration via exploring unique problem structures, online learning in non-stationary environments, and ethical considerations in interactive IR. Their importance is particularly evident to the information retrieval community, but none of the existing tutorials covered such topics. To satisfy the need of practitioners in interaction IR, we would also discuss the practical implementations of several popularly used bandit learning algorithms in interactive IR applications, e.g., data structures to support efficient online execution. This adds unique value to this tutorial.

4 TUTORIAL FORMAT AND SCHEDULE

Our half-day (3 hours) tutorial includes five sections, from classical bandit algorithms to solutions developed for complicated real-world environments (such as multi-user collaborative environments, and non-stationary environments) in interactive recommender systems and online learning to rank, and also ethical considerations in interactive IR.

4.1 Online learning with bandit feedback

In this section, we will introduce the role of online learning with bandit feedback in interactive IR, especially its necessity and challenges. We will motivate our discussions by the problems one typically encounters in recommender and retrieval systems, such as cold start. We will then formulate the mathematical foundation of the bandit learning solutions in various settings, such as multi-armed bandits, contextual bandits, and dueling bandits. We will also illustrate how such formal problem formulations translate to practical interactive IR problems.

4.2 Classical bandit learning algorithms

In this section, we will introduce a set of popularly used bandit learning algorithms, ranging from random exploration methods such as ϵ-greedy [21], confidence bound based methods [1, 6, 23], posterior sampling based methods [3, 16] and perturbation based methods [18, 19]. We will particularly focus on their usages in real-world interactive IR scenarios, such as in recommender systems and retrieval systems. And we will also provide sufficient coverage on their practical implementations, e.g., data structures to ensure efficient online execution. Last but not least, we will illustrate such algorithms' theoretical proprieties, such as regret bounds and convergence rates, which ensure their applicability in real world.

4.3 Interactive recommender system

In this section, we will first illustrate the need for sample efficient online learning methods in real-world interactive recommender systems, where obtaining feedback is costly. For example, user satisfaction quickly deteriorates if the system cannot provide quality results after several rounds of interactions. We will discuss recent developments on more efficient online learning algorithms under various learning scenarios, including (1) sample efficient collaborative bandit learning in multi-agent environments [8, 11, 42], where information sharing and propagation across multiple learning agents are leveraged; (2) discovering problem-dependent low rank structures of the reward function through factorization based bandit learning [24, 37, 38]; and (3) warm-start online learning using offline data through transfer learning or meta learning [49, 50]. We will then introduce the need and challenges of online learning in non-stationary environments, as the non-stationarity of users' interests naturally arises in real-world recommender system. For example, users' result preferences shift from one category to another, while the system needs to infer the underlying change and update its policy accordingly. Based on this, we will introduce the recent developments on bandit learning in non-stationary environments [10, 40, 43], especially their applications in recommender systems.

4.4 Online learning to rank

Online learning to rank eliminates offline solutions' strong dependency on manual relevance annotations, by dynamically optimizing search utility based on users' implicit (bandit) feedback about system's output. It lowers the bar for applying learning to rank in practice, and makes it possible for an online system to directly maximize the utility of result ranking through its interactions with the users. Online learning to rank studies grow rapidly in recent years in both IR and machine learning communities. In this section, we will focus on the key challenges in online learning to rank, i.e., how to effectively explore the ranking space with the biased and noisy implicit feedback while serving the users. We will categorize the existing OL2R solutions and their recent developments to improve sample efficiency, including (1) the ranker estimation in the model space formulated as a dueling bandit problem [34, 36, 46] ; and (2) the ranker estimation in the action space with respect to specific click models [20, 22, 33]. Recent works on ranker estimation based on pairwise ranking models [13, 29] will also be included. Many sample efficient online learning to rank methods have been proposed recently and its connection to offline learning to rank has attracted renewed attention recently. We will end this section with recent works on unifying online and offline (counterfactual) learning to rank [4, 30].

4.5 Ethical considerations in interactive IR with bandit learning

Ethical considerations, such as fairness and privacy-preserving, are becoming important constraints for interactive IR, especially when sensitive personal information is involved in the learning pipeline and/or the online algorithmic decisions have important consequences on people's lives. Distinct from traditional settings, where offline learnt models are used, in interactive IR with bandit learning because the system is updated based on the feedback collected from presented results, there is a higher risk of amplifying bias and disclosing private information in dynamic result serving. The concern is real, but not enough attention has been paid on these aspects. In this section, we will introduce some recent efforts in achieving fairness either with fair exploration [9, 15], or with fair ranking control [28, 45] in online setting; and the protection of potential privacy leakage during online interaction and learning [35, 39].

5 TUTORS' SHORT BIO AND EXPERTISE RELATED TO THE TUTORIAL

- **Huazheng Wang** is a Ph.D. candidate in the Department of Computer Science at the University of Virginia. His research focuses on multi-armed bandit algorithms with application to online recommendation and ranking problems. He is a recipient of Bloomberg Data Science PhD fellowship. His research has appeared in multiple top-tier venues in information retrieval and machine learning, such as SIGIR,WWW, KDD, AAAI.
- **Yiling Jia** is a Ph.D. student in the Department of Computer Science at the University of Virginia. Her research focuses on online learning to rank with a specific concern on the fairness and transparency of the system.
- **Hongning Wang** is now an Associate Professor in the Department of Computer Science at the University of Virginia. He received his PhD degree in computer science at the University of Illinois at Champaign-Urbana in 2014. His research generally lies in the intersection among machine learning, data mining and information retrieval, with a special focus on computational user behavior modeling. His work has generated over 80 research papers in top venues in data mining and information retrieval areas. He is a recipient of 2016 National Science Foundation CAREER Award.

The joint work of Huazheng Wang and Hongning Wang on online learning to rank won the Best Paper Award of SIGIR'2019. Huazheng Wang and Hongning Wang gave a tutorial on "*Learning by Exploration: New Challenges in Real-World Environments*" at KDD 2020.

6 MATERIALS

The following materials will be provided to the audience:

- Presentation slides
- A public website with related content about this tutorial, such as links to public implementations of algorithms and public benchmark datasets discussed in this tutorial
- A detailed list of references.

We will share the materials with the participants on the tutorial website before the conference.

ACKNOWLEDGMENTS

This work is supported by Bloomberg Data Science Ph.D. Fellowship and the National Science Foundation under grant IIS-1553568, IIS-2007492, and IIS-1618948.

REFERENCES

[1] Y. Abbasi-yadkori, D. Pál, and C. Szepesvári. Improved algorithms for linear stochastic bandits. In *NIPS*, pages 2312–2320. 2011.

[2] A. Agarwal, X. Wang, C. Li, M. Bendersky, and M. Najork. Addressing trust bias for unbiased learning-to-rank. In *The World Wide Web Conference*, pages 4–14, 2019.

[3] S. Agrawal and N. Goyal. Thompson sampling for contextual bandits with linear payoffs. In *International Conference on Machine Learning*, pages 127–135, 2013.

[4] Q. Ai, T. Yang, H. Wang, and J. Mao. Unbiased learning to rank: Online or offline? *ACM Transactions on Information Systems (TOIS)*, 39(2):1–29, 2021.

[5] J.-Y. Audibert, S. Bubeck, and R. Munos. Best arm identification in multi-armed bandits. In *COLT*, pages 41–53, 2010.

[6] P. Auer. Using confidence bounds for exploitation-exploration trade-offs. *Journal of Machine Learning Research*, 3(Nov):397–422, 2002.

[7] P. Auer, N. Cesa-Bianchi, and P. Fischer. Finite-time analysis of the multiarmed bandit problem. *Mach. Learn.*, 47(2-3):235–256, May 2002.

[8] N. Cesa-Bianchi, C. Gentile, and G. Zappella. A gang of bandits. *In Pro. NIPS*, 2013.

[9] Y. Chen, A. Cuellar, H. Luo, J. Modi, H. Nemlekar, and S. Nikolaidis. The fair contextual multi-armed bandit. In *Proceedings of the 19th International Conference on Autonomous Agents and MultiAgent Systems*, pages 1810–1812, 2020.

[10] A. Garivier and E. Moulines. On upper-confidence bound policies for non-stationary bandit problems. In *arXiv preprint arXiv:0805.3415 (2008)*.

[11] C. Gentile, S. Li, and G. Zappella. Online clustering of bandits. In *Pro. of the 31st International Conference on Machine Learning (ICML-14)*, pages 757–765, 2014.

[12] A. Grotov and M. de Rijke. Online learning to rank for information retrieval: Sigir 2016 tutorial. In *Proceedings of the 39th International ACM SIGIR conference on Research and Development in Information Retrieval*, pages 1215–1218, 2016.

[13] Y. Jia, H. Wang, S. Guo, and H. Wang. Pair-rank: Online pairwise learning to rank by divide-and-conquer. In *The Web Conference 2021 (WWW'21)*. ACM, 2021.

[14] T. Joachims, L. Granka, B. Pan, H. Hembrooke, and G. Gay. Accurately interpreting clickthrough data as implicit feedback. In *ACM SIGIR Forum*, volume 51, pages 4–11. Acm New York, NY, USA, 2017.

[15] M. Joseph, M. Kearns, J. H. Morgenstern, and A. Roth. Fairness in learning: Classic and contextual bandits. In *Advances in Neural Information Processing Systems*, pages 325–333, 2016.

[16] E. Kaufmann, N. Korda, and R. Munos. Thompson sampling: An asymptotically optimal finite-time analysis. In *International conference on algorithmic learning theory*, pages 199–213. Springer, 2012.

[17] Y. Koren, R. Bell, and C. Volinsky. Matrix factorization techniques for recommender systems. *Computer*, 42(8):30–37, 2009.

[18] B. Kveton, C. Szepesvari, M. Ghavamzadeh, and C. Boutilier. Perturbed-history exploration in stochastic linear bandits. *arXiv preprint arXiv:1903.09132*, 2019.

[19] B. Kveton, C. Szepesvari, S. Vaswani, Z. Wen, T. Lattimore, and M. Ghavamzadeh. Garbage in, reward out: Bootstrapping exploration in multi-armed bandits. In *International Conference on Machine Learning*, pages 3601–3610, 2019.

[20] B. Kveton, C. Szepesvari, Z. Wen, and A. Ashkan. Cascading bandits. *arXiv preprint arXiv:1502.02763*, 2015.

[21] J. Langford and T. Zhang. The epoch-greedy algorithm for contextual multi-armed bandits. In *Proceedings of the 20th International Conference on Neural Information Processing Systems*, NIPS'07, page 817–824, Red Hook, NY, USA, 2007. Curran Associates Inc.

[22] T. Lattimore, B. Kveton, S. Li, and C. Szepesvari. Toprank: A practical algorithm for online stochastic ranking. In *Advances in Neural Information Processing Systems*, pages 3945–3954, 2018.

[23] L. Li, W. Chu, J. Langford, and R. E. Schapire. A contextual-bandit approach to personalized news article recommendation. In *Proceedings of 19th WWW*, pages 661–670. ACM, 2010.

[24] S. Li, A. Karatzoglou, and C. Gentile. Collaborative filtering bandits. In *Proceedings of the 39th International ACM SIGIR conference on Research and Development in Information Retrieval*, pages 539–548, 2016.

[25] T.-Y. Liu. Learning to rank for information retrieval. 2011.

[26] H. Luo, C.-Y. Wei, A. Agarwal, and J. Langford. Efficient contextual bandits in non-stationary worlds. In *Conference On Learning Theory*, pages 1739–1776, 2018.

[27] B. Mitra, N. Craswell, et al. *An introduction to neural information retrieval*. Now Foundations and Trends, 2018.

[28] M. Morik, A. Singh, J. Hong, and T. Joachims. Controlling fairness and bias in dynamic learning-to-rank. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 429–438, 2020.

[29] H. Oosterhuis and M. de Rijke. Differentiable unbiased online learning to rank. In *Proceedings of the 27th ACM International Conference on Information and Knowledge Management*, pages 1293–1302. ACM, 2018.

[30] H. Oosterhuis and M. de Rijke. Unifying online and counterfactual learning to rank: A novel counterfactual estimator that effectively utilizes online interventions. In *Proceedings of the 14th ACM International Conference on Web Search and Data Mining*, pages 463–471, 2021.

[31] H. Oosterhuis, R. Jagerman, and M. de Rijke. Unbiased learning to rank: counterfactual and online approaches. In *Companion Proceedings of the Web Conference 2020*, pages 299–300, 2020.

[32] Y. Qi, Q. Wu, H. Wang, J. Tang, and M. Sun. Bandit learning with implicit feedback. In *Advances in Neural Information Processing Systems*, pages 7276–7286, 2018.

[33] F. Radlinski, R. Kleinberg, and T. Joachims. Learning diverse rankings with multi-armed bandits. In *25th international conference on Machine learning*, pages 784–791. ACM, 2008.

[34] A. Schuth, H. Oosterhuis, S. Whiteson, and M. de Rijke. Multileave gradient descent for fast online learning to rank. In *Proceedings of the Ninth ACM International Conference on Web Search and Data Mining*, pages 457–466. ACM, 2016.

[35] R. Shariff and O. Sheffet. Differentially private contextual linear bandits. In *Advances in Neural Information Processing Systems*, pages 4296–4306, 2018.

[36] H. Wang, S. Kim, E. McCord-Snook, Q. Wu, and H. Wang. Variance reduction in gradient exploration for online learning to rank. In *42^{nd} ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR'19, pages 835–844. ACM, 2019.

[37] H. Wang, Q. Wu, and H. Wang. Learning hidden features for contextual bandits. In *Proceedings of the 25th ACM International on Conference on Information and Knowledge Management*, pages 1633–1642, 2016.

[38] H. Wang, Q. Wu, and H. Wang. Factorization bandits for interactive recommendation. In *Thirty-First AAAI Conference on Artificial Intelligence*, 2017.

[39] H. Wang, Q. Zhao, Q. Wu, S. Chopra, A. Khaitan, and H. Wang. Global and local differential privacy for collaborative bandits. In *Fourteenth ACM Conference on Recommender Systems*, pages 150–159, 2020.

[40] Q. Wu, N. Iyer, and H. Wang. Learning contextual bandits in a collaborative environment. In *SIGIR 2018*.

[41] Q. Wu, Z. Li, H. Wang, W. Chen, and H. Wang. Factorization bandits for online influence maximization. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 636–646, 2019.

[42] Q. Wu, H. Wang, Q. Gu, and H. Wang. Contextual bandits in a collaborative environment. In *Proceedings of the 39th International ACM SIGIR conference on Research and Development in Information Retrieval*, pages 529–538, 2016.

[43] Q. Wu, H. Wang, Y. Li, and H. Wang. Dynamic ensemble of contextual bandits to satisfy users' changing interests. In *The World Wide Web Conference*, WWW '19, pages 2080–2090, New York, NY, USA, 2019. ACM.

[44] Q. Wu, H. Wang, and H. Wang. Learning by exploration: New challenges in real-world environments. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 3575–3576, 2020.

[45] T. Yang and Q. Ai. Maximizing marginal fairness for dynamic learning to rank. *arXiv preprint arXiv:2102.09670*, 2021.

[46] Y. Yue and T. Joachims. Interactively optimizing information retrieval systems as a dueling bandits problem. In *Proceedings of the 26^{th}Annual International Conference on Machine Learning*, pages 1201–1208. ACM, 2009.

[47] C. Zhai. Interactive information retrieval: Models, algorithms, and evaluation. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 2444–2447, 2020.

[48] C. Zhai. Interactive information retrieval: Models, algorithms, and evaluation. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2021.

[49] C. Zhang, A. Agarwal, H. D. Iii, J. Langford, and S. Negahban. Warm-starting contextual bandits: Robustly combining supervised and bandit feedback. In *International Conference on Machine Learning*, pages 7335–7344. PMLR, 2019.

[50] J. Zhang and E. Bareinboim. Transfer learning in multi-armed bandit: a causal approach. In *Proceedings of the 16th Conference on Autonomous Agents and MultiAgent Systems*, pages 1778–1780, 2017.