Unifying Online and Counterfactual Learning to Rank: A Novel Counterfactual Estimator that Effectively Utilizes Online Interventions (Extended Abstract)

Harrie Oosterhuis[1] and **Maarten de Rijke**[2,3]
[1]Radboud University
[2]University of Amsterdam
[3]Ahold Delhaize
harrie.oosterhuis@ru.nl, derijke@uva.nl

Abstract

State-of-the-art Learning to Rank (LTR) methods for optimizing ranking systems based on user interactions are divided into online approaches – that learn by direct interaction – and counterfactual approaches – that learn from historical interactions. We propose a novel intervention-aware estimator to bridge this online/counterfactual division. The estimator corrects for the effect of position bias, trust bias, and item-selection bias by using corrections based on the behavior of the logging policy and on online interventions: changes to the logging policy made during the gathering of click data. Our experimental results show that, unlike existing counterfactual LTR methods, the intervention-aware estimator can greatly benefit from online interventions. To the best of our knowledge, this is the first method that is shown to be highly effective in both online and counterfactual scenarios.

1 Introduction

Ranking systems form the basis for most search and recommendation applications [Liu, 2009]. As a result, the quality of such systems can greatly impact the user experience, and thus, it is important that the underlying ranking models perform well. The field of unbiased Learning to Rank (LTR) considers methods that optimize ranking models based on user clicks, while correcting for the interaction biases that are present in these interactions. User interactions are a form of implicit feedback, and therefore, generally strongly affected by other factors than user preference [Joachims *et al.*, 2005]. To be able to reliably learn from interaction data, the effect of factors other than preference has to be corrected for. In clicks on rankings three prevalent factors are well known: (i) *position bias*: users are less likely to examine, and thus click, lower ranked items [Craswell *et al.*, 2008]; (ii) *item-selection bias*: users cannot click on items that are not displayed [Ovaisi *et al.*, 2020; Oosterhuis and de Rijke, 2020a]; and (iii) *trust bias*: because users trust the ranking system, they are more likely to click on highly ranked items that they do not actually prefer [Agarwal *et al.*, 2019; Joachims *et al.*, 2005]. As a result of these forms of bias, which ranking system was used to gather clicks can have a substantial impact on the clicks that will be observed. Current unbiased LTR methods can be divided into two families: *counterfactual approaches* [Joachims *et al.*, 2017] – that learn from historical data, i.e., clicks that have been logged in the past – and *online approaches* [Yue and Joachims, 2009] – that can perform interventions, i.e., they can decide what rankings will be shown to users. Recent work has noticed that some counterfactual methods can be applied as an online method [Jagerman *et al.*, 2019], or vice versa [Zhuang and Zuccon, 2020; Ai *et al.*, 2021]. Nonetheless, every existing method was designed for either the online or counterfactual setting, never both.

In this work, we propose a novel estimator for both counterfactual and online unbiased LTR from clicks: the *intervention-aware estimator*. The intervention-aware estimator builds on ideas that underlie the latest existing counterfactual methods: the policy-aware estimator [Oosterhuis and de Rijke, 2020a] and the affine estimator [Vardasbi *et al.*, 2020]; and expands them to consider the effect of online interventions. It does so by considering how the effect of bias is changed by an intervention, and utilizes these differences in its unbiased estimation. As a result, the intervention-aware estimator is both effective when applied as a counterfactual method, i.e., when learning from historical data, and as an online method where online interventions lead to enormous increases in efficiency. In our experimental results the intervention-aware estimator is shown to reach state-of-the-art LTR performance in both online and counterfactual settings, and it is the only method that reaches top-performance in both settings.

2 Interactions with Rankings

This paper assumes that three forms of interaction bias occur: position bias, item-selection bias, and trust bias.

Position bias occurs because users only click an item after examining it, and users are more likely to examine items displayed at higher ranks [Craswell *et al.*, 2008]. Thus the rank (a.k.a. position) at which an item is displayed heavily affects the probability of it being clicked. We model this bias using $P(E = 1 \mid k)$: the probability that an item d displayed at rank k is examined by a user [Wang *et al.*, 2018].

Item-selection bias occurs when some items have a zero probability of being examined in some displayed rankings [Ovaisi *et al.*, 2020; Oosterhuis and de Rijke, 2020a]. This can happen because not all items are displayed to the user, or if the ranked list is so long that no user ever considers the entire list. We model this bias by stating: $\exists k, \forall k'$, $(k' > k \rightarrow P(E = 1 \mid k') = 0)$, i.e., there exists a rank k such that items ranked lower than k have no chance of being examined.

Finally, *trust bias* occurs because users trust the ranking system and, consequently, are more likely to perceive top ranked items as relevant even when they are not [Joachims *et al.*, 2005]. We model this bias using: $P(C = 1 \mid k, R, E)$: the probability of a click conditioned on the displayed rank k, the relevance of the item R, and examination E.

To combine these three forms of bias into a single model, we follow Agarwal *et al.* [2019] and Vardasbi *et al.* [2020]:

$$\begin{aligned}\alpha_k &= P(E=1\mid k)\big(P(C=1\mid k,R=1,E=1)\\ &\qquad\qquad - P(C=1\mid k,R=0,E=1)\big), \qquad (1)\\ \beta_k &= P(E=1\mid k)P(C=1\mid k,R=0,E=1).\end{aligned}$$

Existing work has considered how the α and β values can be inferred accurately [Agarwal *et al.*, 2019; Wang *et al.*, 2018; Fang *et al.*, 2019], we will assume they are known. With $P(R = 1 \mid d, q)$ as the probability that an item d is deemed relevant w.r.t. query q by the user, we obtain the following compact notation for the click probability:

$$P(C = 1 \mid d, k, q) = \alpha_k P(R = 1 \mid d, q) + \beta_k. \tag{2}$$

For a single ranking y, let k be the rank at which item d is displayed in y; we denote $\alpha_k = \alpha_{d,y}$ and $\beta_k = \beta_{d,y}$. Finally, let π be a ranking policy used for logging clicks, where $\pi(y \mid q)$ is the probability of π displaying ranking y for query q, then the click probability conditioned on π is:

$$P(C = 1|d, \pi, q) = \textstyle\sum_y \pi(y|q)(\alpha_{d,y} P(R = 1|d, q) + \beta_{d,y}).$$

3 Background: Counterfactual LTR

Most ranking metrics are additive w.r.t. documents; let $P(q)$ be the probability that a user-issued query is query q, then the metric reward $\mathcal{R}$ commonly has the form:

$$\mathcal{R}(\pi) = \textstyle\sum_q P(q) \sum_{d \in D_q} \lambda(d \,|\pi, q) P(R = 1|d, q). \tag{3}$$

Here, the λ function scores each item d depending on where π places d; λ can be chosen to match a desired metric, for instance, the popular Discounted Cumulative Gain (DCG) metric [Järvelin and Kekäläinen, 2002]:

$$\lambda_{\text{DCG}}(d \mid \pi, q) = \textstyle\sum_y \pi(y \mid q)(\log_2(\text{rank}(d \mid y) + 1))^{-1}. \tag{4}$$

Supervised LTR methods can optimize π to maximize $\mathcal{R}$ if relevance scores $P(R = 1 \mid d, q)$ are known [Liu, 2009].

In the counterfactual LTR setting the relevance scores are not known, instead optimization is based on historical user interactions. Let $\mathcal{D}$ be a set of collected interaction data over T timesteps; for each timestep t it contains the user issued query q_t, the logging policy $\bar{\pi}_t$ used to generate the displayed ranking $\bar{y}_t$, and the clicks c_t received on the ranking:

$$\mathcal{D} = \{(\bar{\pi}_t, q_t, \bar{y}_t, c_t)\}_{t=1}^{T}, \tag{5}$$

where $c_t(d) \in \{0, 1\}$ indicates whether item d was clicked at timestep t. While clicks are indicative of relevance they are also affected by several forms of bias, as discussed in Section 2. Counterfactual LTR methods utilize estimators that correct for such bias to unbiasedly estimate the reward of a policy π. The prevalent methods introduce a function $\hat{\Delta}$ that transforms a single click signal to correct for bias. The general estimate of the reward is:

$$\hat{\mathcal{R}}(\pi \mid \mathcal{D}) = \tfrac{1}{T} \textstyle\sum_{t=1}^{T} \sum_d \lambda(d \mid \pi, q) \hat{\Delta}(d \mid \bar{\pi}_t, q_t, \bar{y}_t, c_t). \tag{6}$$

We note the important distinction between the policy π for which we estimate the reward, and the policy $\bar{\pi}_t$ that was used to gather interactions. During optimization only π is changed in order to maximize the estimated reward.

The original Inverse-Propensity-Scoring (IPS) estimator introduced by Wang *et al.* [2016] and Joachims *et al.* [2017] weights clicks according to examination probabilities:

$$\hat{\Delta}_{\text{IPS}}(d \mid \bar{y}_t, c_t) = \frac{c_t(d)}{P(E = 1 \mid \bar{y}_t, d)}. \tag{7}$$

This estimator results in unbiased optimization if no item-selection bias or trust bias is present, thus it can only correct for position bias, for a proof we refer to previous work by Joachims *et al.* [2017] and Vardasbi *et al.* [2020]. Oosterhuis and de Rijke [2020a] introduced policy-aware propensities for the IPS estimator to correct for item-selection bias, and Vardasbi *et al.* [2020] introduced an estimator based on affine corrections that can correct for trust bias. Currently, there is no estimator to correct for all three forms of bias together.

4 The Intervention-Oblivious Estimator

Before we propose our main contribution, the intervention-aware estimator, we first introduce an estimator that simultaneously corrects for position bias, item-selection bias, and trust bias, without considering the effects of interventions.

First we note the probability of a click conditioned on a single logging policy π_t can be expressed as:

$$\begin{aligned}&P(C = 1 \mid d, \pi_t, q)\\ &\qquad = \mathbb{E}_{\bar{y}}[\alpha_d \mid \pi_t, q] P(R = 1 \mid d, q) + \mathbb{E}_{\bar{y}}[\beta_d \mid \pi_t, q].\end{aligned} \tag{8}$$

where the expected values of α and β conditioned on π_t are:

$$\begin{aligned}\mathbb{E}_{\bar{y}}[\alpha_d \mid \pi_t, q] &= \textstyle\sum_{\bar{y}} \pi_t(\bar{y} \mid q) \alpha_{d,\bar{y}},\\ \mathbb{E}_{\bar{y}}[\beta_d \mid \pi_t, q] &= \textstyle\sum_{\bar{y}} \pi_t(\bar{y} \mid q) \beta_{d,\bar{y}}.\end{aligned} \tag{9}$$

By reversing Eq. 8 the relevance probability can be obtained from the click probability. We introduce our *intervention-oblivious estimator*, which applies this transformation to correct for bias:

$$\hat{\Delta}_{\text{IO}}(d \mid q_t, c_t) = \frac{c_t(d) - \mathbb{E}_{\bar{y}}[\beta_d \mid \pi_t, q_t]}{\mathbb{E}_{\bar{y}}[\alpha_d \mid \pi_t, q_t]}. \tag{10}$$

The intervention-oblivious estimator brings together the policy-aware [Oosterhuis and de Rijke, 2020a] and affine estimators [Vardasbi *et al.*, 2020]: on every click it applies an affine transformation based on the logging policy behavior. Unlike existing estimators, we can prove that the intervention-oblivious estimator is unbiased w.r.t. our assumed click model (Section 2). For the sake of brevity and because it is extremely analogous to the proof for Theorem 5.1, we omit this proof of unbiasedness in this extended abstract.

5 The Intervention-Aware Estimator

Existing estimators for counterfactual LTR are designed for a scenario where the logging policy is static: $\forall(\pi_t, \pi_{t'}) \in \mathcal{D}$, $\pi_t = \pi_{t'}$. While they are still unbiased when interventions do take place [Jagerman *et al.*, 2019], they ignore any effect an intervention may have. In other words, any click is treated by considering how the corresponding logging policy treats the item, ignoring the treatment of all the other logging policies. Our goal is to introduce an estimator whose individual corrections are not only based on single logging policies, but instead consider the entire collection of logging policies used to gather the data $\mathcal{D}$.

For ease of notation, we use Π_T for the set of policies that gathered the data in $\mathcal{D}$: $\Pi_T = \{\pi_1, \pi_2, \ldots, \pi_T\}$. With the expected values of α and β conditioned on Π_T:

$$\begin{aligned} \mathbb{E}_{t,\bar{y}}[\alpha_d \mid \Pi_T, q] &= T^{-1} \textstyle\sum_{t=1}^{T} \sum_{\bar{y}} \pi_t(\bar{y} \mid q)\alpha_{d,\bar{y}}, \\ \mathbb{E}_{t,\bar{y}}[\beta_d \mid \Pi_T, q] &= T^{-1} \textstyle\sum_{t=1}^{T} \sum_{\bar{y}} \pi_t(\bar{y} \mid q)\beta_{d,\bar{y}}, \end{aligned} \tag{11}$$

the probability of a click can be conditioned on the Π_T set is:

$$\begin{aligned} &P(C = 1 \mid d, \Pi_T, q) \qquad (12) \\ &\quad = T^{-1} \textstyle\sum_{t=1}^{T} \sum_{\bar{y}} \pi_t(\bar{y} \mid q)(\alpha_{d,\bar{y}} P(R = 1 \mid d, q) + \beta_{d,\bar{y}}) \\ &\quad = \mathbb{E}_{t,\bar{y}}[\alpha_d \mid \Pi_T, q] P(R = 1 \mid d, q) + \mathbb{E}_{t,\bar{y}}[\beta_d \mid \Pi_T, q]. \end{aligned}$$

We propose our *intervention-aware estimator* that corrects for bias using the expectations conditioned on Π_T:

$$\hat{\Delta}_{\text{IA}}(d \mid q_t, c_t) = \frac{c_t(d) - \mathbb{E}_{t,\bar{y}}[\beta_d \mid \Pi_T, q_t]}{\mathbb{E}_{t,\bar{y}}[\alpha_d \mid \Pi_T, q_t]}. \tag{13}$$

The salient difference with the intervention-oblivious estimator is that the expectations are conditioned on Π_T, all logging policies in $\mathcal{D}$, instead of an individual logging policy π_t. While the difference with the intervention-oblivious estimator may seem small, the resulting estimates can be very different, our experimental results show that this result in a sizeable reduction in variance.

To better understand the difference, consider an example where $T = 100$ and an item was ranked very highly but not clicked for $t = 1$ up to $t = 99$ but ranked lowly and clicked at $t = 100$. The intervention-oblivious estimator would assign a high weight to the click at $t = 100$ to compensate for the low rank at which it was displayed when clicked. Conversely, the intervention-aware estimator would assign a much lower weight to this click because it compensates both for the low rank at $t = 100$ but also the high ranking of the item received the first 99 timesteps. Thus we see how online interventions can create large differences between the two estimates.

Lastly, we note that when no interventions take place the intervention-oblivious estimator and intervention-aware estimators are equivalent. Because the intervention-aware estimator is the only existing counterfactual LTR estimator whose corrections are influenced by online interventions, we consider it an important step to bridge the gap between counterfactual and online LTR.

Finally, we prove that the intervention-aware estimator is unbiased w.r.t. our assumed click model (Section 2).

Theorem 5.1. *The estimated reward $\hat{\mathcal{R}}$ (Eq. 6) using the intervention-aware estimator (Eq. 13) is unbiased w.r.t. the true reward $\mathcal{R}$ (Eq. 3) under two assumptions: (i) our click model (Eq. 2), and (ii) the click probability on every item, conditioned on Π_T, is correlated with relevance:*

$$\forall q, \forall d, \quad \mathbb{E}_{t,\bar{y}}[\alpha_d \mid \Pi_T, q] \neq 0. \tag{14}$$

Proof. Using Eq. 12 and Eq. 14 the relevance probability can be derived from the click probability by:

$$P(R = 1 \mid d, q) = \frac{P(C = 1 \mid d, \Pi_T, q) - \mathbb{E}_{t,\bar{y}}[\beta_d \mid \Pi_T, q]}{\mathbb{E}_{t,\bar{y}}[\alpha_d \mid \Pi_T, q]}. \tag{15}$$

Eq. 15 can be used to show that $\hat{\Delta}_{\text{IA}}$ is an unbiased indicator of relevance:

$$\begin{aligned} \underset{t,\bar{y},c}{\mathbb{E}}\big[\hat{\Delta}_{\text{IA}}(d|q_t, c_t)|\Pi_T\big] &= \frac{\underset{t,\bar{y},c}{\mathbb{E}}[c_t(d)|\Pi_T, q_t] - \underset{t,\bar{y}}{\mathbb{E}}[\beta_d|\Pi_T, q_t]}{\mathbb{E}_{t,\bar{y}}[\alpha_d \mid \Pi_T, q_t]} \\ &= \frac{P(C = 1|d, \Pi_T, q_t) - \underset{t,\bar{y}}{\mathbb{E}}[\beta_d|\Pi_T, q_t]}{\mathbb{E}_{t,\bar{y}}[\alpha_d \mid \Pi_T, q_t]} = P(R = 1|d, q_t). \end{aligned} \tag{16}$$

Finally, combining Eq. 16 with Eq. 6 and Eq. 3 reveals that $\hat{\mathcal{R}}$ based on $\hat{\Delta}_{\text{IA}}$ is unbiased w.r.t. $\mathcal{R}$:

$$\begin{aligned} &\mathbb{E}_{t,q,\bar{y},c}\big[\hat{\mathcal{R}}(\pi \mid \mathcal{D})\big] \qquad (17) \\ &\quad = \sum_q P(q) \sum_d \lambda(d \mid \pi, q) \mathbb{E}_{t,\bar{y},c}\Big[\hat{\Delta}_{\text{IA}}(d \mid c, q) \mid \Pi_T, q\Big] \\ &\quad = \sum_q P(q) \sum_d \lambda(d \mid \pi, q) P(R = 1 \mid d, q) = \mathcal{R}(\pi). \quad \square \end{aligned}$$

6 Experimental Setup

Our experiments aim to compare the performance of the intervention-aware estimator with existing LTR methods in both the online and counterfactual setting.[1]We use the semi-synthetic experimental setup that is common in existing work on both online LTR [Hofmann *et al.*, 2013; Oosterhuis and de Rijke, 2018; Oosterhuis and de Rijke, 2019; Zhuang and Zuccon, 2020] and counterfactual LTR [Joachims *et al.*, 2017; Vardasbi *et al.*, 2020; Ovaisi *et al.*, 2020]. At each timestep, we simulate a user-issued query by uniformly sampling from the training and validation partitions of the Yahoo Webscope dataset [Chapelle and Chang, 2011], a dataset based on real-world commercial search logs. We apply Eq. 2 with $\alpha = [0.35, 0.53, 0.55, 0.54, 0.52]$ and $\beta = [0.65, 0.26, 0.15, 0.11, 0.08]$ [Agarwal *et al.*, 2019]; the relevance probabilities are based on the five-grade annotations from the dataset: $P(R = 1 \mid d, q) = 0.25 \cdot \text{label}(d, q)$.

To obtain a production ranker policy, we apply supervised LTR on 1% of the training partition [Joachims *et al.*, 2017]. We vary the number of interventions per run; they are always evenly spread on an exponential scale, at each intervention the logging policy is replaced with the latest learned model. Ranking models are neural networks with two 32-node hidden layers, optimized using policy gradients [Oosterhuis and de Rijke, 2020b]. The propensity weights for

[1] Our experimental implementation is publicly available at https://github.com/HarrieO/2021wsdm-unifying-LTR.

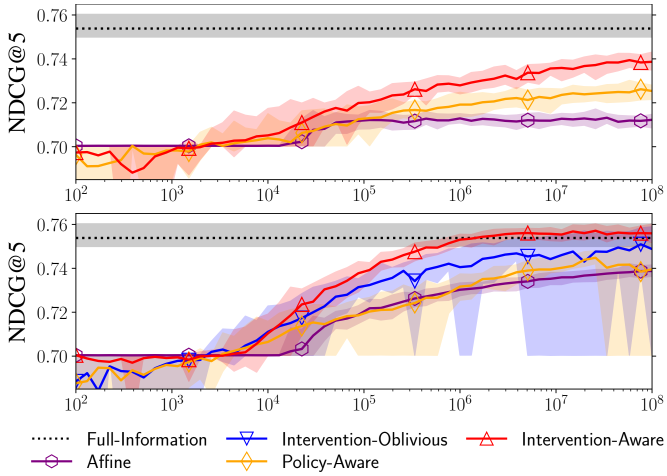

Figure 1: Comparison with counterfactual LTR estimators. Top: Counterfactual runs (no interventions); Bottom: Online runs (50 interventions). Results are average of 20 runs, shaded areas indicate the 90% confidence bounds; x-axis: number of logged queries.

training clicks are clipped at $10/\sqrt{|\mathcal{D}|}$ to reduce variance. The following baseline methods are used: (i) The policy-aware estimator [Oosterhuis and de Rijke, 2020a]. (ii) The affine estimator [Vardasbi *et al.*, 2020]. (iii) Pairwise Differentiable Gradient Descent (PDGD) [Oosterhuis and de Rijke, 2018]. (iv) Biased-PDGD, PDGD without the debiasing weights. (v) Counterfactual Online Learning to Rank (COLTR) [Zhuang and Zuccon, 2020].

7 Results and Discussion

For our comparison with existing counterfactual LTR methods, we consider Figure 1 which displays the performance of LTR using different counterfactual estimators.

The top of Figure 1 displays performance in the counterfactual setting where the logging policy is static. Very clearly the intervention-aware estimator quickly reaches a higher performance than the other methods, this is expected since it is the only unbiased estimator of the three. While the theory guarantees that intervention-aware estimator will converge at the optimal performance, we are unable to observe the number of queries it requires to do so. The bottom of Figure 1 regards an online setting with 50 online interventions. The online interventions have a clear positive effect leading to a higher mean performance for all estimators. However, this also introduces an enormous amount of variance to the policy-aware and intervention-oblivious estimators. In stark contrast, the intervention-aware estimator hardly has an increase in variance while it also learns much faster than the other estimators. Moreover, it is now able to reach optimal performance with roughly a million logged queries.

We therefore conclude that the intervention-aware estimator leads to higher performance than existing counterfactual estimators, especially when online interventions take place.

Figure 2 compares the performance of the intervention-aware estimator with online LTR methods. The top of Figure 2 shows that the intervention-aware estimator applied with 100 interventions provides performance comparable to PDGD applied fully online (an intervention at each query). We can conclude that, aside from a small difference before $2 \cdot 10^4$ queries, the intervention-aware estimator appears to match state-of-the-art online LTR performance without needing the costly overhead of constant interventions.

The bottom of Figure 2 considers the performance of PDGD when not applied fully online. We see that PDGD converges at very suboptimal performance when provided 100 or less interventions; this observation does not contradict the existing theory since PDGD is not proven to be unbiased w.r.t. ranking metrics. Its suboptimal convergence in not fully-online settings makes PDGD very unreliable in practice; we conclude that the intervention-aware estimator remains the safer and more reliable choice for these settings.

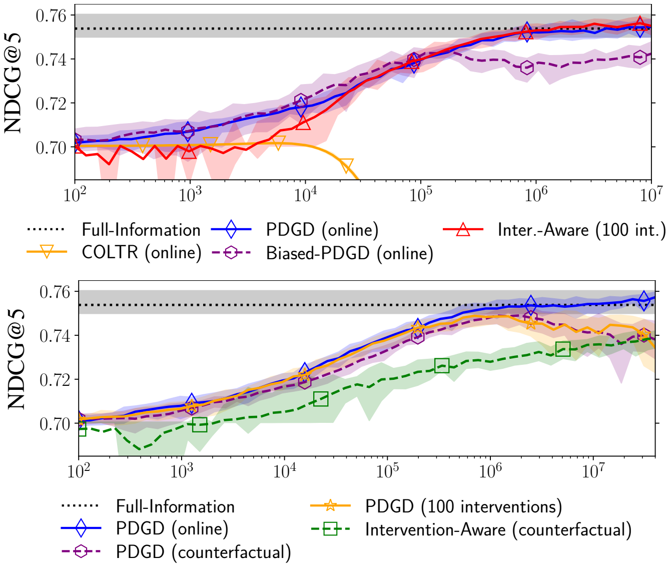

Figure 2: Comparison with online LTR methods. Results are averages of 20 runs, shaded areas indicate the 90% confidence bounds; x-axis: number of logged queries.

8 Conclusion

This paper has introduced an intervention-aware estimator which corrects for position-bias, trust-bias, and item-selection bias, and considers the effect of online interventions. Our results show that the intervention-aware estimator outperforms existing counterfactual LTR estimators, and greatly benefits from online interventions. Moreover, with only 100 interventions it can reach performance comparable to state-of-the-art online LTR methods. Because the intervention-aware estimator appears to be the most reliable method for both counterfactual and online LTR, we hope its introduction further unifies the division between these fields.

Acknowledgments

This research was partially supported by the Netherlands Organisation for Scientific Research (NWO) under project nr 612.001.551. All content represents the opinion of the authors, which is not necessarily shared or endorsed by their respective employers and/or sponsors.

References

[Agarwal *et al.*, 2019] Aman Agarwal, Xuanhui Wang, Cheng Li, Michael Bendersky, and Marc Najork. Addressing trust bias for unbiased learning-to-rank. In *The World Wide Web Conference*, pages 4–14. ACM, 2019.

[Ai *et al.*, 2021] Qingyao Ai, Tao Yang, Huazheng Wang, and Jiaxin Mao. Unbiased learning to rank: Online or offline? *ACM Transactions on Information Systems (TOIS)*, 39(2):1–29, 2021.

[Chapelle and Chang, 2011] Olivier Chapelle and Yi Chang. Yahoo! Learning to Rank Challenge Overview. *Journal of Machine Learning Research*, 14:1–24, 2011.

[Craswell *et al.*, 2008] Nick Craswell, Onno Zoeter, Michael Taylor, and Bill Ramsey. An experimental comparison of click position-bias models. In *Proceedings of the 2008 International Conference on Web Search and Data Mining*, pages 87–94. ACM, 2008.

[Fang *et al.*, 2019] Zhichong Fang, Aman Agarwal, and Thorsten Joachims. Intervention harvesting for context-dependent examination-bias estimation. In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 825–834, 2019.

[Hofmann *et al.*, 2013] Katja Hofmann, Anne Schuth, Shimon Whiteson, and Maarten de Rijke. Reusing historical interaction data for faster online learning to rank for IR. In *Proceedings of the sixth ACM International Conference on Web search and data mining*, pages 183–192. ACM, 2013.

[Jagerman *et al.*, 2019] Rolf Jagerman, Harrie Oosterhuis, and Maarten de Rijke. To model or to intervene: A comparison of counterfactual and online learning to rank from user interactions. In *Proceedings of the 42nd International ACM SIGIR Conference on Research & Development in Information Retrieval*, pages 15–24. ACM, 2019.

[Järvelin and Kekäläinen, 2002] Kalervo Järvelin and Jaana Kekäläinen. Cumulated gain-based evaluation of IR techniques. *ACM Transactions on Information Systems (TOIS)*, 20(4):422–446, 2002.

[Joachims *et al.*, 2005] Thorsten Joachims, Laura Granka, Bing Pan, Helene Hembrooke, and Geri Gay. Accurately interpreting clickthrough data as implicit feedback. In *SIGIR Forum*, pages 154–161. ACM, 2005.

[Joachims *et al.*, 2017] Thorsten Joachims, Adith Swaminathan, and Tobias Schnabel. Unbiased learning-to-rank with biased feedback. In *Proceedings of the Tenth ACM International Conference on Web Search and Data Mining*, pages 781–789, 2017.

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

[Oosterhuis and de Rijke, 2018] Harrie Oosterhuis and Maarten 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.

[Oosterhuis and de Rijke, 2019] Harrie Oosterhuis and Maarten de Rijke. Optimizing ranking models in an online setting. In *Advances in Information Retrieval*, pages 382–396, Cham, 2019. Springer International Publishing.

[Oosterhuis and de Rijke, 2020a] Harrie Oosterhuis and Maarten de Rijke. Policy-aware unbiased learning to rank for top-k rankings. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, page 489–498. ACM, 2020.

[Oosterhuis and de Rijke, 2020b] Harrie Oosterhuis and Maarten de Rijke. Taking the counterfactual online: Efficient and unbiased online evaluation for ranking. In *Proceedings of the 2020 ACM SIGIR on International Conference on Theory of Information Retrieval*, pages 137–144. ACM, 2020.

[Ovaisi *et al.*, 2020] Zohreh Ovaisi, Ragib Ahsan, Yifan Zhang, Kathryn Vasilaky, and Elena Zheleva. Correcting for selection bias in learning-to-rank systems. In *Proceedings of The Web Conference 2020*, pages 1863–1873, 2020.

[Vardasbi *et al.*, 2020] Ali Vardasbi, Harrie Oosterhuis, and Maarten de Rijke. When inverse propensity scoring does not work: Affine corrections for unbiased learning to rank. In *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*, pages 1475–1484, 2020.

[Wang *et al.*, 2016] Xuanhui Wang, Michael Bendersky, Donald Metzler, and Marc Najork. Learning to rank with selection bias in personal search. In *Proceedings of the 39th International ACM SIGIR conference on Research and Development in Information Retrieval*, pages 115–124, 2016.

[Wang *et al.*, 2018] Xuanhui Wang, Nadav Golbandi, Michael Bendersky, Donald Metzler, and Marc Najork. Position bias estimation for unbiased learning to rank in personal search. In *Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining*, pages 610–618. ACM, 2018.

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

[Zhuang and Zuccon, 2020] Shengyao Zhuang and Guido Zuccon. Counterfactual online learning to rank. In *European Conference on Information Retrieval*, pages 415–430. Springer, 2020.