Counterfactual Dimension Importance Estimation (CoDIME) for Dense Information Retrieval*

Guglielmo Faggioli[1], Nicola Ferro[1], Raffaele Perego[2] and Nicola Tonellotto[3]

[1] *University of Padova, Padova, Italy*
[2] *ISTI-CNR, Pisa, Italy*
[3] *University of Pisa, Pisa, Italy*

Abstract

Contextual dense representation models have revolutionized text processing by providing deeper semantic insights and enhancing Information Retrieval (IR) capabilities. These models represent text within a latent space, where shared underlying concepts are encoded beyond the explicit wording of the text. Nevertheless, previous studies indicate that certain dimensions within these dense embeddings can be irrelevant—or even detrimental—to retrieval success, depending on the specific information needs of the query. Limiting retrieval to a linear subspace that omits these less useful dimensions has been shown to improve performance. To tackle this issue, Dimension IMportance Estimators (DIMEs) were introduced to detect and remove harmful dimensions, thereby refining the representations of queries and documents to highlight only the valuable aspects. Current DIMEs mostly rely on pseudo-relevance feedback, which can be unreliable, or on explicit relevance judgments, which are often impractical to gather. Drawing inspiration from counterfactual analysis, we present Counterfactual DIMEs (CoDIMEs), a new technique that leverages noisy implicit feedback to assess the significance of each dimension. The CoDIME framework approximates the connection between how frequently a document is clicked and its alignment with particular query dimensions through a linear model. Empirical evidence demonstrates that CoDIME consistently outperforms traditional pseudo-relevance feedback-based DIMEs and other unsupervised counterfactual methods that make use of implicit signals.

1. Introduction

Dense text representations have demonstrated remarkable capability in capturing semantic meaning, emerging as the dominant technology across numerous text-related tasks in Information Retrieval (IR) and Natural Language Processing (NLP). These representation models are based on neural networks that project the text onto a dense representation space where semantically similar contents tend to be arranged closely. While these novel representations are more effective than traditional lexical approaches (e.g., BM25 [16] and TF-IDF [10]) in handling the semantic gap, they are far less interpretable, even if the dimensions of the representations are assumed to be associated with some latent semantic meaning. Starting from this, Faggioli et al. [3] propose the so-called *Manifold Clustering Hypothesis* which posits that it is possible to find a query-wise subspace of the dense representation space where the retrieval is more effective, i.e., where the representations of the query and its relevant documents are more aligned. to find such a subspace, Faggioli et al. define the concept of Dimension IMportance Estimator (DIME): a model explicitly meant to estimate the query-dependent importance of each dimension to preserve only the most important ones while discarding the others. In particular, Faggioli et al. propose DIMEs based on Pseudo-Relevance Feedback (PRF) or relying on explicit feedback. The former is known to have variable and not always consistent effectiveness, especially when it comes to dense models [12]. The latter, on the other hand, can be much more challenging to gather. To overcome this limitation, in this paper, we propose to employ an intermediate relevance signal, more reliable than PRF and far more available than explicit feedback: *implicit feedback*. Implicit feedback leverages the analysis of user interactions, such as clicks

IIR2025: 15th Italian Information Retrieval Workshop, 3th - 5th September 2025, Cagliari, Italy

*This is an extended abstract of [4].

guglielmo.faggioli@unipd.it (G. Faggioli); ferro@dei.unipd.it (N. Ferro); raffaele.perego@isti.cnr.it (R. Perego); nicola.tonellotto@unipi.it (N. Tonellotto)
0000-0002-5070-2049 (G. Faggioli); 0000-0001-6286-0194 (N. Ferro); 0000-0001-7189-4724 (R. Perego); 0000-0002-7427-1001 (N. Tonellotto)

and dwell times, to infer weak relevance signals for retrieved content.Akin to past efforts in the domain, we employ a set of simulated click logs to estimate the frequency of clicks on the links in a Search Engine Result Page (SERP). By relying on such click frequencies, we devise a counterfactual modelling of the click probabilities. This model is then used as a source of implicit feedback information, and we exploit it to determine the importance of the dimensions in the dense representation space, thus instantiating a set of novel Counterfactual DIMEs (CoDIMEs). In particular, we design a set of *linear* CoDIMEs that quantifies the importance of a dimension by considering the characteristics of a linear model that regresses the documents' click frequency on the interaction between the query and the documents on such a dimension.

Compared to CoRocchio [22], a state-of-the-art counterfactual approach, the CoDIME framework achieves up to +0.235 nDCG@10 points, moving from 0.404 to 0.639 (+58%) (Dragon and Robust '04) and +0.117 nDCG@100 points, moving from 0.356 to 0.473 (+33%) (Dragon and Robust '04).

2. The CoDIME Framework

We refer to the list of the top k documents retrieved in response to a query q as $\mathcal{R}=\{d_1,...,d_k\}$ where d_i is the i-th retrieved document. All users $\mathcal{U}$ who submitted the query q interact with $\mathcal{R}$, each one generating a click log $\mathcal{C}_u$ such that $c_{u,d}$ is either 1 or 0, depending on whether the user clicked on document $d\in\mathcal{R}$ or not. Based on this historical information, we can compute the observed frequency of clicks of $\mathcal{U}$ for q and d as: $\hat{f}_d=\frac{1}{|\mathcal{U}|}\sum_{u\in\mathcal{U}}c_{u,d}$. In other terms, $\hat{f}_d$ describes the proportion of clicks received by a document $d\in\mathcal{R}$ retrieved in response to q. We can expect this frequency to be somewhat correlated with the relevance of the top k documents retrieved in response to q, but also with the position at which the document is shown in the $\mathcal{R}$. Akin to CoRocchio, we debias click frequencies using the Inverse Propensity Score (IPS) [6, 9]. The debiased click frequency is thus defined as: $f_d = \hat{f}_d \cdot (1/k)^{-\eta}$ Where k is the position at which the document d was observed in $\mathcal{R}$ and η is the propensity parameter. The debiased click frequencies describe how likely it is that a document is clicked, regardless of where it is placed in the SERP. Finally, we define $(f_{d_1},...,f_{d_k})$ as the list of debiased click frequencies of the top k documents for q, included in $\mathcal{R}$.

Given a query q and a document d and their respective dense representations $\mathbf{q}$ and $\mathbf{d}$, we can define the interaction H_d between them as the Hadamard product between their representations. Assuming each dimension corresponds to a latent concept, observing strong interaction $H_{d,i}$ between the query and the document on the i-th dimension indicates that the concept is prime for the query and the document.

Linear CoDIME estimate the importance of a dimension for a query by examining the linear correlation between the dimension's interaction (H) and the debiased click frequency (f) on the documents.

Correlation CoDIME The first linear CoDIME is inspired by the *Oracle DIME* as proposed by Faggioli et al. [3]. More in detail, we define $\bar{f}$ and $\bar{H}_i$ as the mean debiased click frequency and the interaction on the i-th component for a given query and the corresponding retrieved documents, respectively. Called ρ the Pearson's correlation, the Correlation CoDIME, or CoDIME$_{corr}$, is defined as follows: $CoDIME_{corr}(i)=\rho((f_{d_1},...,f_{d_k}),(H_{d_1,i},...,H_{d_k,i}))$.

This CoDIME quantifies the linear correlation between the interactions on a given dimension and the debiased click frequencies as their Pearson's ρ correlation. If the interaction on a dimension between the query and the documents aligns with the debiased click frequencies on such documents, the importance will be 1, and the dimension likely belongs to the optimal subspace. If the interaction and the debiased click probability are uncorrelated, the importance of the dimension will be zero. Finally, when the interaction and the debiased click probability have a negative relation, the importance is negative, and the dimension will likely be discarded.

Slope CoDIME One of the major limitations of the Correlation CoDIME is that it cannot consider how fast the interactions and the click frequencies tend to vary. In fact, the linear model that best fits the points might be more or less steep. A steeper linear model indicates that the dimension is better at separating the good and the bad documents. Vice-versa, if the linear model grows slowly, it is harder to separate documents clicked often from rarely clicked documents. The value of the Correlation CoDIME does not depend on such steepness, but only on how well a linear model fits the data. Therefore, we propose a second linear CoDIME that explicitly quantifies the dimension's importance based on the slope of the linear model that best fits the data according to the Ordinary Least Square (OLS) approach.

In more detail, let us call $\mathbf{H}_i \in \mathbb{R}^{k\times 2}$ a matrix such that its first column contains k 1s and the second column contains the values $H_{d_1,i},\ldots,H_{d_k,i}$. This is the regressor matrix, while we treat $\mathbf{f} = [f_{d_1},\ldots,f_{d_k}]^\top$ as the response variable. We fit a linear model using the OLS approach by computing $\mathbf{b}_i \in \mathbb{R}^2$: $\mathbf{b}_i = (\mathbf{H}_i^\top \mathbf{H}_i)^{-1}\mathbf{H}_i^\top \mathbf{f}$. Since we added a column of ones to the regressor matrix, the first element $b_{i,1}$ of $\mathbf{b}_i$ is the intercept of the OLS linear model while the second element $b_{i,2}$ of $\mathbf{b}_i$ is the slope.[1] The CoDIME$_{slope}$ is defined as: $CoDIME_{slope}(i) = b_{i,2}$.

3. Experimental Evaluation

Experimental Setup and Query Logs Simulation To assess the proposed counterfactual strategy we consider three well-known state-of-the-art dense encoders: Contriever [7], TAS-B [5], and Dragon [12], fine-tuned on MSMARCO.[2] Experiments are conducted on three well-known TREC collections: TREC Robust 2004 (Robust '04) [19], TREC Deep Learning 2019 (DL '19) [2], and TREC Deep Learning 2020 (DL '20) [1][3]. The parameter η, describing the user's patience in clicking a document of the SERP, is set to 1, while the maximum depth of inspection is set to 20 documents unless specified differently. Furthermore, the experiments are conducted by repeating 1000 times for each topic the simulation of the click log. Differently from Faggioli et al. [3], our CoDIMEs strategies choose the fraction α of representation dimension retained by applying a 5-fold cross-validation on the validation set. The code and the data are publicly released.[4]

The CoDIME approach is based on historical user feedback needed to instantiate the counterfactual framework and estimate the click probabilities. In a real-world deployment with a consistent user base, click logs are easy to collect and use for our purposes. Following the previous literature [22, 8, 13, 15, 23, 21, 14, 9, 20] on counterfactual implicit feedback and learning to rank, we simulate the interaction of the users with the documents to generate a set of synthetic click logs. To do so, we need to simulate i) the selection bias, ii) the position bias, and iii) the relevance bias. The selection bias is implemented by assuming that every user interacts and inspects the SERP up to the document in position k'. To simulate the click propensity, we model $\hat{p}_{e,i}$, the probability of examination, as inversely proportional to the position, i.e., $\hat{p}_e(k) = \left(\frac{1}{k}\right)^\eta$. To simulate the relevance bias, we model $\hat{p}_r(q,d)$, the probability that the user will click on a document d, given its relevance to q. More in detail, we consider three ideal user models: the *perfect user* (P) whose click probability is directly proportional to the relevance of the document; the *binarized user* (B) that, clicks on a non-relevant or partially relevant document with probability 0.1 and clicks on a relevant or highly relevant document with probability 1; the *near random user* (R) that clicks on a non-relevant document with probability 0.4 and clicks on a highly relevant document with probability 0.6. For the perfect and near-random users, the probabilities are a linear spacing between the minimum and maximum probabilities, with as many steps as the relevance grades. For the binarized user, the click probability of a document with relevance within the lowest half of the grades is set to 0.1; otherwise, it is set to 1. The simulated click probability is computed as: $\hat{p}_c(q,d,k) = \hat{p}_r(q,d)\cdot\hat{p}_e(k)$. In other terms, to simulate the click of a user on a document d retrieved in position k in response to the query q, we combine, by multiplying, the probability that the user will click on such document given its relevance to the query (i.e., the relevance bias) and the probability that the user will click on a document in position k, regardless of its relevance (i.e., the position bias). We consider the following baselines Vector PRF (VPRF) [11], LLM DIME [3], PRF DIME [3], and CoRocchio [22].

Performance We report the comparison in terms of effectiveness between the different approaches. Table 1 report the effectiveness of our solution and the competitors on DL '19. In bold, we report the

[1] We also experimented with a linear model without the intercept, obtaining slightly inferior empirical results.

[2] We use the model weights publicly available on https://huggingface.co/

[3] For space reasons, we report here only the results for DL '19. The interested reader can find the results for other collections, for which we observe substantially similar patterns, in the original paper.

[4] https://github.com/guglielmof/25-SIGIR-FFPT

Table 1
Performance comparison on DL ‘19. In bold the most effective approach, underlined the runner-up. The top-tier according to an ANOVA is marked with *. P, B, and R are the Perfect, Binarized, and Near Random user models.

	Contriever			Dragon			TAS-B			Contriever			Dragon			TAS-B		
	P	B	R	P	B	R	P	B	R	P	B	R	P	B	R	P	B	R
	nDCG@10									nDCG@20								
retrieval only	.674	.674	.674	.740	.740	.740	.717	.717	.717	.655	.655	.655	.726	.726	.726	.679	.679	.679
VPRF-5	.664	.664	.664	.752	.752	.752	.721	.721	.721	.656	.656	.656	.743	.743	.743*	.701	.701	.701
VPRF-20	.636	.636	.636	.732	.732	.732	.667	.667	.667	.643	.643	.643	.717	.717	.717	.651	.651	.651
DIME$_{LLM}$	.742	.742	.742*	.767	.767	.767	.749	.749	.749	.710	.710	.710	.746	.746	.746*	.724	.724	.724*
DIME$_{PRF}$	.668	.668	.668	.740	.740	.740	.717	.717	.717	.655	.655	.655	.726	.726	.726	.682	.682	.682
CoRocchio	.804*	.766*	.632	.830	.824*	.724	.810*	.780*	.665	.761*	.729*	.634	.805*	.796*	.721	.771*	.746*	.642
CoDIME$_{corr}$	.851*	.828*	**.810***	.891*	**.854***	.831*	.856*	.835*	.804*	**.796***	**.786***	**.777***	.830*	.815*	.792*	.807*	.781*	.760*
CoDIME$_{slope}$	**.855***	**.829***	.809*	**.897***	.854*	**.842***	**.863***	**.839***	**.821***	.796*	.774*	.770*	**.838***	**.817***	**.793***	**.821***	**.785***	**.775***
	nDCG@50									nDCG@100								
retrieval only	.642	.642	.642	.686	.686	.686	.650	.650	.650	.634	.634	.634	.673	.673	.673	.637	.637	.637
VPRF-5	.644	.644	.644	.718	.718	.718*	.677	.677*	.677*	.647	.647	.647	.703	.703	.703*	.667	.667*	.667*
VPRF-20	.637	.637	.637	.698	.698	.698	.641	.641	.641	.636	.636	.636	.692	.692	.692*	.636	.636	.636
DIME$_{LLM}$	.686*	.686*	.686*	.709	.709	.709*	.693*	.693*	.693*	.676*	.676*	.676*	.700	.700	.700*	.684*	.684*	.684*
DIME$_{PRF}$	.647	.647	.647	.686	.686	.686	.653	.653	.653	.643	.643	.643	.673	.673	.673	.647	.647	.647
CoRocchio	.737*	.708*	.625	.772*	**.767***	.698	.741*	.723*	.626	.733*	.703*	.628	**.757***	**.750***	.689*	**.726***	.710*	.621
CoDIME$_{corr}$	**.743***	**.735***	**.721***	**.778***	.764*	.751*	.745*	.725*	.712*	.734*	.716*	**.708***	.747*	.741*	**.729***	.725*	**.711***	.703*
CoDIME$_{slope}$	.741*	.721*	.715*	.766*	.763*	**.754***	**.747***	**.729***	**.733***	**.735***	**.723***	.699*	.739*	.739*	.726*	.723*	**.711***	**.712***

highest performance achieved, underlined the runner-up. At the same time, the symbol * denotes the top tier of systems (i.e., the set of systems deemed statistically not distinguishable) according to an ANalysis Of the VAriance (ANOVA) [17] with Tukey's Honestly Significant Differences (HSD) [18] pairwise multiple comparison test and significance level of 0.05. The columns (P, B, and R) correspond respectively to *Perfect*, *Binarized*, and *Near Random* users. Within each measure, the first line reports the performance of the base encoder. Notice that some of the baselines (the encoder itself, VPRF, DIME$_{LLM}$ and DIME$_{PRF}$) are not based on users' clicks, and thus, they are not affected by the user model, and their performance is the same across all user models. We notice that the most effective approaches for nDCG@10 and nDCG@20 are those based on the Linear CoDIMEs (CoDIME$_{corr}$ and CoDIME$_{slope}$). Both approaches have comparable performance, with no clear dominance between the two: in all the cases, the two approaches are statistically equivalent according to the chosen statistical testing procedure. In general, the approaches based on the weighted magnitude of the dimensions (CoDIME$_{wavg}$ and CoDIME$_{wmax}$) tend to be far less effective and are generally surpassed by the CoRocchio baseline. Comparing the Linear CoDIMEs with the most effective baseline, CoRocchio, we notice that the CoDIMEs are almost always more effective than CoRocchio for cutoffs lower than 100 (the only exception is Dragon with nDCG@50 and the binarized model). If we consider nDCG@100, CoRocchio is more effective than the CoDIMEs in the case of TAS-B and Dragon with a perfect or binarized user model but the difference is not statistically significant and small (0.01 or less nDCG@100 points). Compared to CoRocchio, all CoDIMEs are less vulnerable to changes in the user model considered. In fact, for CoRocchio, when moving from the perfect to the near-random user model, we notice a performance drop which is in the range of 10 to 15 points, depending on the considered measure or cutoff. Vice versa, the CoDIMEs tends to be stable, with variations between the perfect and near-random users in the 1-3 points range, up to 5 points in the worst scenarios. This is a desirable property: in a real-world scenario, where the clicks are far more affected by noise than in a simulated environment, a more stable solution as the Linear CoDIMEs offers better guarantees of a good performance.

4. Conclusion

In this work, we introduced CoDIME, a novel counterfactual framework for dimension importance estimation in dense text representations, leveraging implicit user feedback to address challenges in existing DIME approaches. By incorporating counterfactual modelling of click probabilities in various dimension importance estimation strategies, our CoDIME approaches achieved state-of-the-art performance in multiple dense IR testbeds. Compared to CoRocchio [22], a state-of-the-art counterfactual approach, the CoDIME framework achieves up to +0.235 nDCG@10 points, moving from 0.404 to 0.639 (+58%) (Dragon and Robust ‘04) and +0.117 nDCG@100 points, moving from 0.356 to 0.473 (+33%) (Dragon and Robust ‘04). These findings highlight the efficacy of counterfactual techniques and DIME approaches in adapting dense representations and improving retrieval effectiveness.

Acknowledgments

This work is supported, in part, by the Spoke “FutureHPC & BigData” of the ICSC – Centro Nazionale di Ricerca in High-Performance Computing, Big Data and Quantum Computing, the Spoke “Human-centered AI” of the M4C2 - Investimento 1.3, Partenariato Esteso PE00000013 - "FAIR - Future Artificial Intelligence Research", the “Extreme Food Risk Analytics” (EFRA) project, Grant no. 101093026, funded by European Union – NextGenerationEU, the FoReLab project (Departments of Excellence), the NEREO PRIN project funded by the Italian Ministry of Education and Research Grant no. 2022AEFHAZ and the CAMEO PRIN 2022 Project Grant no. 2022ZLL7MW.

Declaration on Generative AI

During the preparation of this work, the author did not use any AI tool.

References

[1] Nick Craswell, Bhaskar Mitra, Emine Yilmaz, and Daniel Campos. Overview of the TREC 2020 deep learning track. In Ellen M. Voorhees and Angela Ellis, editors, *Proceedings of the Twenty-Ninth Text REtrieval Conference, TREC 2020, Virtual Event [Gaithersburg, Maryland, USA], November 16-20, 2020*, volume 1266 of *NIST Special Publication*. National Institute of Standards and Technology (NIST), 2020. URL https://trec.nist.gov/pubs/trec29/papers/OVERVIEW.DL.pdf.

[2] Nick Craswell, Bhaskar Mitra, Emine Yilmaz, Daniel Campos, and Ellen M. Voorhees. Overview of the TREC 2019 deep learning track. *CoRR*, abs/2003.07820, 2020. URL https://arxiv.org/abs/2003.07820.

[3] Guglielmo Faggioli, Nicola Ferro, Raffaele Perego, and Nicola Tonellotto. Dimension importance estimation for dense information retrieval. In Grace Hui Yang, Hongning Wang, Sam Han, Claudia Hauff, Guido Zuccon, and Yi Zhang, editors, *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2024, Washington DC, USA, July 14-18, 2024*, pages 1318–1328. ACM, 2024. doi: 10.1145/3626772.3657691. URL https://doi.org/10.1145/3626772.3657691.

[4] Guglielmo Faggioli, Nicola Ferro, Raffaele Perego, and Nicola Tonellotto. Codime: a counterfactual approach for dimension importance estimation through click logs. In *SIGIR ’25: The 48th International ACM SIGIR Conference on Research and Development in Information Retrieval July 13-18, 2025*. ACM, 2025.

[5] Sebastian Hofstätter, Sheng-Chieh Lin, Jheng-Hong Yang, Jimmy Lin, and Allan Hanbury. Efficiently teaching an effective dense retriever with balanced topic aware sampling. In Fernando Diaz, Chirag Shah, Torsten Suel, Pablo Castells, Rosie Jones, and Tetsuya Sakai, editors, *SIGIR ’21: The 44th International ACM SIGIR Conference on Research and Development in Information Retrieval, Virtual Event, Canada, July 11-15, 2021*, pages 113–122. ACM, 2021. doi: 10.1145/3404835.3462891. URL https://doi.org/10.1145/3404835.3462891.

[6] D. G. Horvitz and D. J. Thompson. A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association*, 47(260):663–685, 1952. ISSN 01621459, 1537274X. URL http://www.jstor.org/stable/2280784.

[7] Gautier Izacard, Mathilde Caron, Lucas Hosseini, Sebastian Riedel, Piotr Bojanowski, Armand Joulin, and Edouard Grave. Towards unsupervised dense information retrieval with contrastive learning. *CoRR*, abs/2112.09118, 2021. URL https://arxiv.org/abs/2112.09118.

[8] 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 Benjamin Piwowarski, Max Chevalier, Éric Gaussier, Yoelle Maarek, Jian-Yun Nie, and Falk Scholer, editors, *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2019, Paris, France, July 21-25, 2019*, pages 15–24. ACM, 2019. doi: 10.1145/3331184.3331269. URL https://doi.org/10.1145/3331184.3331269.

[9] Thorsten Joachims, Adith Swaminathan, and Tobias Schnabel. Unbiased learning-to-rank with biased feedback. In Maarten de Rijke, Milad Shokouhi, Andrew Tomkins, and Min Zhang, editors, *Proceedings of the Tenth ACM International Conference on Web Search and Data Mining, WSDM 2017, Cambridge, United Kingdom, February 6-10, 2017*, pages 781–789. ACM, 2017. doi: 10.1145/3018661.3018699. URL https://doi.org/10.1145/3018661.3018699.

[10] Karen Spärck Jones. A statistical interpretation of term specificity and its application in retrieval. *J. Documentation*, 28(1):11–21, 1972. doi: 10.1108/00220410410560573. URL https://doi.org/10.1108/00220410410560573.

[11] Hang Li, Ahmed Mourad, Shengyao Zhuang, Bevan Koopman, and Guido Zuccon. Pseudo relevance feedback with deep language models and dense retrievers: Successes and pitfalls. *ACM Trans. Inf. Syst.*, 41(3):62:1–62:40, 2023. doi: 10.1145/3570724. URL https://doi.org/10.1145/3570724.

[12] Sheng-Chieh Lin, Akari Asai, Minghan Li, Barlas Oguz, Jimmy Lin, Yashar Mehdad, Wen-tau Yih, and Xilun Chen. How to train your dragon: Diverse augmentation towards generalizable dense retrieval. In Houda Bouamor, Juan Pino, and Kalika Bali, editors, *Findings of the Association for Computational Linguistics: EMNLP 2023, Singapore, December 6-10, 2023*, pages 6385–6400. Association for Computational Linguistics, 2023. doi: 10.18653/V1/2023.FINDINGS-EMNLP.423. URL https://doi.org/10.18653/v1/2023.findings-emnlp.423.

[13] Harrie Oosterhuis and Maarten de Rijke. Differentiable unbiased online learning to rank. In Alfredo Cuzzocrea, James Allan, Norman W. Paton, Divesh Srivastava, Rakesh Agrawal, Andrei Z. Broder, Mohammed J. Zaki, K. Selçuk Candan, Alexandros Labrinidis, Assaf Schuster, and Haixun Wang, editors, *Proceedings of the 27th ACM International Conference on Information and Knowledge Management, CIKM 2018, Torino, Italy, October 22-26, 2018*, pages 1293–1302. ACM, 2018. doi: 10.1145/3269206.3271686. URL https://doi.org/10.1145/3269206.3271686.

[14] Harrie Oosterhuis and Maarten de Rijke. Unifying online and counterfactual learning to rank: A novel counterfactual estimator that effectively utilizes online interventions. In Liane Lewin-Eytan, David Carmel, Elad Yom-Tov, Eugene Agichtein, and Evgeniy Gabrilovich, editors, *WSDM '21, The Fourteenth ACM International Conference on Web Search and Data Mining, Virtual Event, Israel, March 8-12, 2021*, pages 463–471. ACM, 2021. doi: 10.1145/3437963.3441794. URL https://doi.org/10.1145/3437963.3441794.

[15] Zohreh Ovaisi, Ragib Ahsan, Yifan Zhang, Kathryn Vasilaky, and Elena Zheleva. Correcting for selection bias in learning-to-rank systems. In Yennun Huang, Irwin King, Tie-Yan Liu, and Maarten van Steen, editors, *WWW '20: The Web Conference 2020, Taipei, Taiwan, April 20-24, 2020*, pages 1863–1873. ACM / IW3C2, 2020. doi: 10.1145/3366423.3380255. URL https://doi.org/10.1145/3366423.3380255.

[16] Stephen E. Robertson, Steve Walker, Susan Jones, Micheline Hancock-Beaulieu, and Mike Gatford. Okapi at TREC-3. In Donna K. Harman, editor, *Proceedings of The Third Text REtrieval Conference, TREC 1994, Gaithersburg, Maryland, USA, November 2-4, 1994*, volume 500-225 of *NIST Special Publication*, pages 109–126. National Institute of Standards and Technology (NIST), 1994. URL http://trec.nist.gov/pubs/trec3/papers/city.ps.gz.

[17] Henry Scheffe. *The analysis of variance.* John Wiley & Sons, 1959.

[18] John W. Tukey. Comparing individual means in the analysis of variance. *Biometrics*, 5(2):99–114, 1949. ISSN 0006341X, 15410420.

[19] Ellen M. Voorhees. Overview of the TREC 2004 robust track. In Ellen M. Voorhees and Lori P. Buckland, editors, *Proceedings of the Thirteenth Text REtrieval Conference, TREC 2004, Gaithersburg, Maryland, USA, November 16-19, 2004*, volume 500-261 of *NIST Special Publication*. National Institute of Standards and Technology (NIST), 2004. URL http://trec.nist.gov/pubs/trec13/papers/ROBUST.OVERVIEW.pdf.

[20] Xuanhui Wang, Michael Bendersky, Donald Metzler, and Marc Najork. Learning to rank with selection bias in personal search. In Raffaele Perego, Fabrizio Sebastiani, Javed A. Aslam, Ian Ruthven, and Justin Zobel, editors, *Proceedings of the 39th International ACM SIGIR conference on Research and Development in Information Retrieval, SIGIR 2016, Pisa, Italy, July 17-21, 2016*, pages 115–124. ACM, 2016. doi: 10.1145/2911451.2911537. URL https://doi.org/10.1145/2911451.2911537.

[21] Shengyao Zhuang and Guido Zuccon. Counterfactual online learning to rank. In Joemon M.

Jose, Emine Yilmaz, João Magalhães, Pablo Castells, Nicola Ferro, Mário J. Silva, and Flávio Martins, editors, *Advances in Information Retrieval - 42nd European Conference on IR Research, ECIR 2020, Lisbon, Portugal, April 14-17, 2020, Proceedings, Part I*, volume 12035 of *Lecture Notes in Computer Science*, pages 415–430. Springer, 2020. doi: 10.1007/978-3-030-45439-5_28. URL https://doi.org/10.1007/978-3-030-45439-5_28.

[22] Shengyao Zhuang, Hang Li, and Guido Zuccon. Implicit feedback for dense passage retrieval: A counterfactual approach. In Enrique Amigó, Pablo Castells, Julio Gonzalo, Ben Carterette, J. Shane Culpepper, and Gabriella Kazai, editors, *SIGIR '22: The 45th International ACM SIGIR Conference on Research and Development in Information Retrieval, Madrid, Spain, July 11 - 15, 2022*, pages 18–28. ACM, 2022. doi: 10.1145/3477495.3531994. URL https://doi.org/10.1145/3477495.3531994.

[23] Shengyao Zhuang, Zhihao Qiao, and Guido Zuccon. Reinforcement online learning to rank with unbiased reward shaping. *Inf. Retr. J.*, 25(4):386–413, 2022. doi: 10.1007/S10791-022-09413-Y. URL https://doi.org/10.1007/s10791-022-09413-y.