On the Orthogonality of Bias and Utility in Ad hoc Retrieval

Amin Bigdeli
abigdeli@ryerson.ca
Ryerson University

Negar Arabzadeh
narabzad@uwaterloo.ca
University of Waterloo

Shirin Seyedsalehi
shirin.seyedsalehi@ryerson.ca
Ryerson University

Morteza Zihayat
mzihayat@ryerson.ca
Ryerson University

Ebrahim Bagheri
bagheri@ryerson.ca
Ryerson University

ABSTRACT

Various researchers have recently explored the impact of different types of biases on information retrieval tasks such as ad hoc retrieval and question answering. While the impact of bias needs to be controlled in order to avoid increased prejudices, the literature has often viewed the relationship between increased retrieval utility (effectiveness) and reduced bias as a tradeoff where one can suffer from the other. In this paper, we empirically study this tradeoff and explore whether it would be possible to reduce bias while maintaining similar retrieval utility. We show this would be possible by revising the input query through a bias-aware pseudo-relevance feedback framework. We report our findings based on four widely used TREC corpora namely Robust04, Gov2, ClueWeb09 and ClueWeb12 and using two classes of bias metrics. The findings of this paper are significant as they are among the first to show that decrease in bias does not necessarily need to come at the cost of reduced utility.

CCS CONCEPTS

• **Information systems** → **Information Retrieval**; • **Evaluation of retrieval results** → *Presentation of retrieval results.*

KEYWORDS

Bias, Fairness, Query Expansion, Ad hoc Retrieval

ACM Reference Format:
Amin Bigdeli, Negar Arabzadeh, Shirin Seyedsalehi, Morteza Zihayat, and Ebrahim Bagheri . 2021. On the Orthogonality of Bias and Utility in Ad hoc Retrieval. 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, 5 pages. https://doi.org/10.1145/3404835.3463110

1 INTRODUCTION

There has been growing awareness on how various forms of biases can be introduced as a result of representational and algorithmic aspects of computational models in information retrieval (IR) [8, 18, 20, 27]. More specifically, the IR literature has recognized how biases, such as those associated with gender and ethnicity, can impact the outcomes of tasks including ad hoc retrieval and conversational search [4, 11, 22, 23]. Unidentified biases can leak and impact the outcome of retrieval systems and potentially impact users by exposing them to biased information that can in turn lead to stereotypical biases that reinforce known yet unfair prejudices.

Recent literature has observed a host of computational methods that attempt to identify, explain or potentially reduce biases, e.g., from neural representations [3, 7, 21, 35]. The recent work by Rekabsaz et al. [22] specifically shows that the state-of-the-art neural ranking methods are more inclined towards retrieving male-dominated documents when applied over gender-neutral queries. In line with the work by Rekabsaz, Bigdeli et al. [2] further explored whether stereotypical gender biases can be systematically observed within IR relevance judgements. As a result of their experiments on the MS MARCO relevance judgements, the authors found that gender biases are prevalent among relevance judgments, which can be learnt by neural models that are trained based on them.

One of the important practical considerations of dealing with biases is its potential impact on retrieval utility. Several researchers have already argued that reducing biases can often be achieved at the cost of reduced utility, pointing to a *tradeoff* between reducing bias and increasing utility [6, 9, 10, 17, 25]. In other words, such perspective identifies an orthogonal relation between fairness (reduced bias) and utility. While such position may sound reasonable, to the best of our knowledge, it has not yet been empirically explored within the context of information retrieval. For this reason, the **main objective of this paper** is to systematically study the tradeoff between bias and utility in ad hoc document retrieval. The ad hoc retrieval task is defined as a process of retrieving a ranked list of relevant documents D_q to a query q using an efficient retrieval method M that can estimate the relevance of each document in a large document collection C to q. In this context, there are at least two elements that can be controlled to reduce bias, namely (1) retrieval method M, and (2) query q. In this paper, we focus on the second element, i.e., query q, and empirically study whether: **(a)** a tradeoff necessarily exists between bias and utility, and **(b)** there exists a possible revision to q that will lead to reduced bias while maintaining (or possibly increasing) utility. We hypothesize that one way to avoid a tradeoff between bias and utility is to identify a revised query for q, i.e., q', such that q' has at least the same utility as q but results in a less biased ranked list of documents.

To explore whether a revised query such as q' exists, we show how q can be revised using a bias-aware pseudo-relevance feedback method that explicitly considers bias when revising the original query. While it has already been shown that it is possible to revise a query to increase utility [5, 12, 15, 26, 28, 29, 34], to the best of

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

ACM ISBN 978-1-4503-8037-9/21/07…$15.00
https://doi.org/10.1145/3404835.3463110

our knowledge, there are no works that revise a query to reduce bias and maintain utility. On this basis, we answer three Research Questions (RQ) in this paper: **(RQ1)** Is it possible to revise an initial query such that it maintains at least the same degree of utility while significantly reducing bias? **(RQ2)** Are potential reductions in bias as a result of revising q consistent across different quantitative measures for bias? **(RQ3)** Do the characteristics of a revised query substantially differ from that of the original query when optimized for lower bias even if the same levels of utility are maintained?

We conduct our experiments on four well-known TREC corpora namely Robust04, Gov2, ClueWeb09 and Clueweb12 and their associated topics. We find that it is possible to systematically revise an input query so that it maintains the same utility while substantially reducing bias in the retrieved list of documents. We believe the findings of this paper are **impactful** as it shows: (1) the widely discussed hypothesis about a tradeoff between bias and utility does not necessarily always hold, and (2) it is possible to use simple yet effective methods to reduce the degree of bias of the documents that are retrieved and presented to the users.

2 BIAS-AWARE PSEUDO-RELEVANCE FEEDBACK

The main hypothesis behind our work is that one could potentially show that the tradeoff between bias and utility does not necessarily always hold, if there is some revised version of the input query q, such as q', that maintains the same degree of utility but significantly decreases bias in the retrieved ranked list of documents. In this paper, we propose an effective strategy for revising a query based on **bias-aware pseudo-relevance feedback**. In essence, a Pseudo-Relevance Feedback (PRF) strategy revises the original query by expanding it using the most informative terms extracted from the top-k retrieved documents by the original query [14, 24, 30, 31]. We argue that it is possible to reformulate the query revision strategy based on PRF such that it is not solely dependent on the relevance of the top-k retrieved documents to the original query but also cognizant of the degree of bias exposed by each of these documents.

We hypothesize that by considering the degree of bias of the documents when choosing the top-k documents to be considered as the pseudo-relevant feedback set, we reduce the likelihood of choosing terms that have bias. Simply put, the inclusion of less biased documents in the pseudo-relevant feedback set will decrease the likelihood of choosing biased terms to be included in the revised query q'. As a consequence, a less biased q' is likely to show a higher degree of relevance to less biased documents, and therefore, lead to a less biased ranked list of documents.

Let us now formalize our bias-aware pseudo-relevance feedback strategy for generating a revised query q'. We assume that a retrieval method M retrieves a ranked list of documents from the collection C based on their relevance to a query q. We refer to the set of retrieved documents for query q as D_q. A pseudo-relevance based strategy would select terms from D_q to be used for developing q'. However, in our work, we revise the rank order of the documents in D_q such that the revised ranking explicitly considers the degree of bias exposed by each document in D_q. We rerank D_q such that the relevance of each document to q and the bias of each document is taken into consideration in tandem. More specifically:

$$Rel_{debiased}(d) = (1 - \lambda)Rel(d) - \lambda Bias(d) \tag{1}$$

where $d \in D$, $Rel(d)$ is the relevance of document d to query q, $Bias(d)$ is some measure of bias computed for document d and λ is a linear interpolation coefficient. Since lower values of $Bias(d)$ are desirable, they are subtracted from $Rel(d)$. Based on Equation 1, the initial ranked list of document D_q is re-ranked based on $Rel_{debiased}(d)$ producing a new ranking for the top-k documents, which we refer to as $D_q^{debiased}$. Given the re-ranked list of documents, to develop the revised query q', we adopt the RM3 strategy, which is a pseudo-relevance feedback framework for query expansion [1, 13] and has shown outstanding performance across various corpora and queries [16, 28, 33]. In order to develop q', we select and expand q with those top-n terms that have the highest score as follows:

$$Score_t = \sum_{d \in D_q^{debiased}} (P(t|d) log \frac{P(t|d)}{P(t|C)}) \tag{2}$$

Now, each term w in the original query and the selected top-n expansion terms are weighted as follows:

$$W_{debiased}(w, q) = \alpha P(w|q) + (1 - \alpha)P(w|D_q^{debiased}) \tag{3}$$

where $\alpha \in [0, 1]$ and $P(w|D_q^{debiased})$ is defined as:

$$P(w|D_q^{debiased}) = \sum_{d \in D_q^{debiased}} P(w|d) \prod_{t \in q} P(t|d) \tag{4}$$

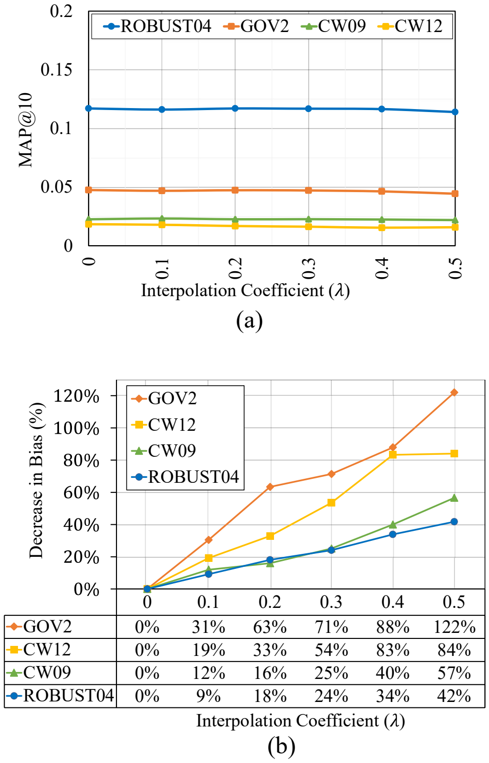
The proposed strategy for developing the revised query q' is likely to reduce bias in the final list of ranked documents for two main reasons: (1) it retrieves terms from a list that has been re-ranked by considering the bias of each document; therefore, reducing the chances of including biased terms in the top-n expansion terms, and (2) it weighs the terms in the query and the top-n expansion terms based on their likelihood to appear in $D_q^{debiased}$ as captured in Equation 4. As such, even if biased terms do appear in the final composition of the query, it is unlikely they would receive a higher weight compared to less biased terms in q'.

3 EXPERIMENTS

3.1 Setup

3.1.1 Datasets. We employ four different corpora, namely, Robust04, Gov2, ClueWeb09 (i.e., CW09), and ClueWeb12 (i.e., CW12) and their TREC topics: 301-450 and 601-700 for Robust04, 701-850 for Gov2, 1-200 for ClueWeb09, and 201-300 for ClueWeb12.

3.1.2 Bias Metrics. In order to measure the degree of bias for each document as required by $Bias(d)$ in Equation 1, we employ two strategies to show that our findings are not prejudiced towards a certain definition of bias. In the first strategy, we employ the metrics proposed by Rekabsaz et al. for measuring gender biases [22]. The authors propose two classes of metrics for measuring gender bias based on the (i) presence (boolean) and (ii) term frequency of gendered terms within each single document. On this basis, the authors extend the measurement of bias from a single document to a ranked list by proposing the Average Ranking Bias (ARaB) metric that considers the bias of each document and the ranking of that document in the list. In the second strategy, we adopt the method proposed by Bigdeli et al. [2] to measure stereotypical biases within documents. Their approach is based on measuring document gender inclinations based on LIWC's male and female references [19].



	0	0.1	0.2	0.3	0.4	0.5
GOV2	0%	31%	63%	71%	88%	122%
CW12	0%	19%	33%	54%	83%	84%
CW09	0%	12%	16%	25%	40%	57%
ROBUST04	0%	9%	18%	24%	34%	42%

Figure 1: The impact of our proposed approach on utility (map) and bias (ARaB) metrics.

3.1.3 Implementation Details. We implement our work based on Anserini [32]. We adopt the tuned BM25 implementation for each corpus from Anserini. The base RM3 query expander is also adopted from this library with top-k and top-n set to 10, and $\alpha = 0.5$. We implement and incorporate both measures of bias for our bias-aware pseudo-relevance feedback in Anserini. All our code and the results of our runs on all four corpora are publicly available[1]. The values of the interpolation coefficient (λ) in Equation 1 are selected from [0,1] with 0.1 increments. We note that while our results are consistent for all values of λ, we resort to reporting the results for [0,0.5] due to space limitation, all other results are available online[2].

3.2 Findings

We structure our findings based on our three research questions introduced earlier. We note statistical significance is measured based on a paired t-test at 95% confidence.

3.2.1 RQ1: Utility-Bias Tradeoff. In the first research question, we empirically explore whether it would be possible to revise an initial query such that utility is maintained while significantly reducing bias. We adopt mean average precision (map) as the measure for utility and term-frequency version of the ARaB metric proposed by Rekabsaz et al. [22] as the measure of bias of a ranked list. We report our findings at rank 10 of the retrieved ranked list of documents. Figure 1 depicts the impact of our work on both utility (Figure 1a) and bias (Figure 1b). In the sub-figures, the x-axis represents the impact of the interpolation coefficient λ such that λ=0 is equivalent to the base pseudo-relevance feedback method since the impact of bias measurement in Equation 1 is canceled out in this case. As seen in the figure, regardless of the value of λ, the value of map does not experience any statistically significant changes as a result of the inclusion of the bias term in Equation 1. However, the degree of bias of the retrieved ranked list of documents reduces significantly with the increase of λ starting from λ=0.1. In other words, the figure shows that it is possible to revise the initial query q such that the utility of retrieval is maintained while significantly reducing the bias of the retrieved list of documents. For instance, while the least degree of decrease in bias was observed on the Robust04 dataset with 29.48%, the best reduction on bias was observed on Gov2 with 55.01%. The impact of our proposed approach is further visualized in Figure 2 where map has been maintained while the degree of bias has been significantly reduced on all corpora.

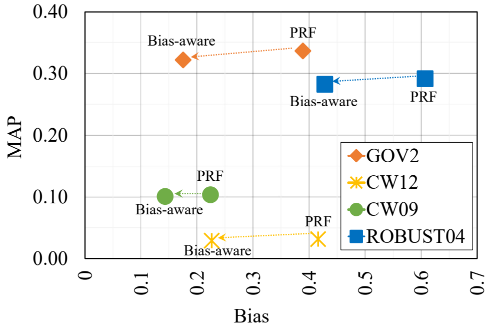


Figure 2: The impact of our approach on utility and bias.

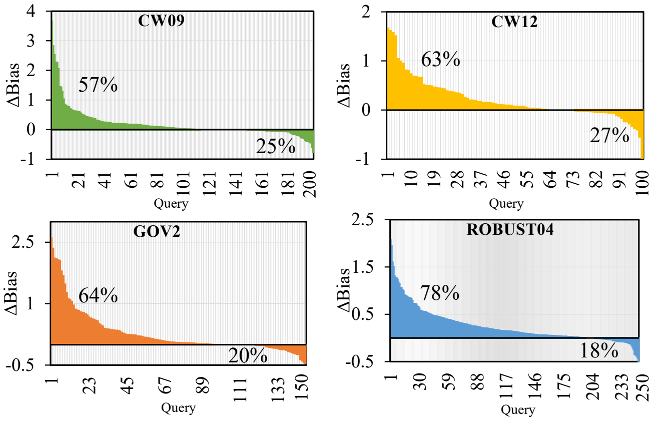


Figure 3: Reduction in bias per query-basis compared to PRF.

Finally, to show that the proposed approach has been able to consistently reduce bias on a wide range of queries, we report the degree to which bias was reduced on a per-query basis in Figure 3. This figure shows how much bias was positively or negatively impacted by the proposed approach when compared to the baseline PRF method. Positive values show how much the proposed approach was able to reduce bias compared to PRF and negative values show the inverse. The four figures show that in the majority of the cases, our approach was able to improve bias compared to PRF. For instance, in ClueWeb09, 57% of the queries were improved while only 25% were negatively impacted and 18% were tied. The

[1] https://github.com/aminbigdeli/bias-aware-PRF
[2] https://github.com/aminbigdeli/bias-aware-PRF/tree/main/results

Table 1: Bias measurements using ARaB and LIWC-based metrics.

Method	Robust04			Gov2			CW09			CW12		
	ARaB		LIWC	ARaB		LIWC	ARaB		LIWC	ARaB		LIWC
	TF	Boolean		TF	Boolean		TF	Boolean		TF	Boolean	
BM25	0.61	0.35	0.48	0.33	0.14	0.07	0.23	0.08	0.05	0.40	0.14	0.19
PRF	0.61	0.34	0.45	0.39	0.11	0.07	0.22	0.07	0.07	0.42	0.10	0.20
Our Approach	0.43	0.27	0.34	0.18	0.07	0.05	0.14	0.06	0.04	0.23	0.05	0.13
Decrease in Bias (%)	29.50	20.58	24.44	53.84	36.36	28.57	36.36	14.28	42.85	45.23	50.00	35.00

Table 2: The number and weight of biased terms in the revised queries.

	Bias metric	Robust04			Gov2			CW09			CW12		
		PRF	Ours	Δ%	PRF	Ours	Δ%	PRF	Ours	Δ%	PRF	Ours	Δ%
Number of Biased Terms	ARaB	65	49	-25%	16	5	-69%	17	14	-17%	11	8	-27%
	LIWC	69	54	-22%	24	10	-58%	24	19	-21%	15	11	-27%
Sum of Biased Term Weights	ARaB	3.84	2.72	-29%	1.42	0.79	-44%	2.20	1.85	-16%	0.99	0.84	-15%
	LIWC	4.22	3.08	-27%	1.80	1.02	-43%	2.34	1.96	-16%	1.71	1.54	-9%

Table 3: Qualitative comparison of the revised queries. Darker colors denote higher weight of biased term.

Original Query	PRF	Our Approach
anorexia nervosa bulimia	spell anorexia movi nervosa opiat teen she her bulimia tobacco diet seri	anorexia japan nervosa opiat your teen research studi eat bulimia tobacco diet
Cult Lifestyles	student krishna kim chilton lifestyl she mother cult her car pension	student di krishna mambro lifestyl she cult her campu car pension
History of Physicians in America	societi bibliographi, black my america york medicin physician histori press women	medicar black ohio my americai Insur journal african physician w1 histori
american muslim mosques schools	mosqu muslim american wahhabi my saudi america religion islam school he	mosqu muslim american my percent america religion countri islam school religi he
symptoms of heart attack	pain chest medic diseas blood heart panic symptom attack women	treatment pain pressur diseas blood stroke heart panic symptom attack sudden
weather strip	door bottom girl cheroke strip caulk jeep weather cme replac	door channel seal presen chanc strip stop weather cme bogen

most number of improved queries were seen on Robust04 were 78% of queries were improved compared to only 18% that were hurt.

3.2.2 RQ2: Bias Evaluation using Different Metrics. To confirm that the decrease in bias is not due to the fact that it has been indirectly controlled for in Equation 1, we measure the degree of bias using another completely independent measure of bias proposed by Bigdeli et al. [2] that measures bias as the degree to which male and female affiliations are observed within a document based on psychometric properties offered in Linguistic Inquiry and Word Count (LIWC) [19]. We also report the boolean version of the ARaB metric in addition to its term-frequency variation that was used in RQ1. As shown in Table 1, the percentage of decrease in bias is consistent across all metrics and always statistically significant. This shows that our approach systematically reduces bias even when measured on a different bias metric than the one considered in Equation 1.

3.2.3 RQ3. Revised Query Characteristics. We analyze how the characteristics of a query undergo changes from both quantitative and qualitative perspectives. From quantitative point of view, we compare the revised queries based on our approach with the ones from PRF from two angles: (1) the set of terms that appear in the revised queries: we report the number of biased terms appeared in each set of expanded terms that are added to the original query, and (2) the weights assigned to query terms: we compute the sum of the weights assigned to the biased terms that are used to expand the original query. The list of biased terms are those suggested by Rekabsaz et al[3]. Table 2 shows that both the number and the weight of biased terms have significantly decreased in the revised queries developed by our proposed approach in all four corpora. This supports our initial hypothesis that a reduction in the bias of the revised query will lead to a reduction in the bias of the retrieved list of documents. This also shows that the use of the biased terms did not necessarily contribute to an improved utility (Figure 2) but did lead to an increasingly biased retrieval.

To provide a qualitative insight, Table 3 shows several sample queries that represent how our revised queries differ from the baseline in two aspects: (1) there are cases where the baseline revised queries have introduced additional biased terms to the query, which do not appear in our revised queries. In such cases, the biased terms do not seem to be necessary for retrieval. For instance, for the 'symptoms of heart attack', the baseline method adds the biased word 'women' to the query, which does not appear in our revised query. (2) there are other cases where biased terms do appear in both forms of the revised query but the weights of these terms are lower in our approach. We indicate the weight differences based on a color encoding. For instance, for the 'cult lifestyles' query, the baseline adds the biased terms 'she', 'mother' and 'her' but our method does not include the term 'mother' and only includes two biased terms 'her' and 'she' with much lower weights.

4 CONCLUDING REMARKS

This paper investigates the widely assumed tradeoff between utility and bias, which asserts that reducing bias (higher fairness) can come at the cost of reduced utility (lower retrieval effectiveness). In this paper we show that this hypothesis does not necessarily always hold and one can potentially find cases where bias can be systematically reduced while maintaining utility. To this end, we have shown that it is possible to effectively revise a user query that would lead to a less biased ranked list of documents. Based on our experiments, a less biased revised query can maintain utility and at the same time reduce bias. We believe that this work lays the foundation for considering fairness and utility as two cooperating measures as opposed to being viewed as competing aspects.

[3] https://github.com/navid-rekabsaz/GenderBias_IR

REFERENCES

[1] Nasreen Abdul-Jaleel, James Allan, W Bruce Croft, Fernando Diaz, Leah Larkey, Xiaoyan Li, Mark D Smucker, and Courtney Wade. 2004. UMass at TREC 2004: Novelty and HARD. *Computer Science Department Faculty Publication Series* (2004), 189.

[2] Amin Bigdeli, Negar Arabzadeh, Morteza Zihayat, and Ebrahim Bagheri. 2021. Exploring Gender Biases in Information Retrieval Relevance Judgement Datasets. In *43rd European Conference on IR Research (ECIR 2021)*. Springer, 216–224.

[3] Tolga Bolukbasi, Kai-Wei Chang, James Zou, Venkatesh Saligrama, and Adam Kalai. 2016. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. *arXiv preprint arXiv:1607.06520* (2016).

[4] Aylin Caliskan, Joanna J Bryson, and Arvind Narayanan. 2017. Semantics derived automatically from language corpora contain human-like biases. *Science* 356, 6334 (2017), 183–186.

[5] Mostafa Dehghani, Sascha Rothe, Enrique Alfonseca, and Pascal Fleury. 2017. Learning to Attend, Copy, and Generate for Session-Based Query Suggestion. In *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management, CIKM 2017, Singapore, November 06 - 10, 2017*, Ee-Peng Lim, Marianne Winslett, Mark Sanderson, Ada Wai-Chee Fu, Jimeng Sun, J. Shane Culpepper, Eric Lo, Joyce C. Ho, Debora Donato, Rakesh Agrawal, Yu Zheng, Carlos Castillo, Aixin Sun, Vincent S. Tseng, and Chenliang Li (Eds.). ACM, 1747–1756. https://doi.org/10.1145/3132847.3133010

[6] Fernando Diaz, Bhaskar Mitra, Michael D Ekstrand, Asia J Biega, and Ben Carterette. 2020. Evaluating stochastic rankings with expected exposure. In *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*. 275–284.

[7] Alessandro Fabris, Alberto Purpura, Gianmaria Silvello, and Gian Antonio Susto. 2020. Gender stereotype reinforcement: Measuring the gender bias conveyed by ranking algorithms. *Information Processing & Management* 57, 6 (2020), 102377.

[8] Joseph Fisher, Dave Palfrey, Christos Christodoulopoulos, and Arpit Mittal. 2019. Measuring social bias in knowledge graph embeddings. *arXiv preprint arXiv:1912.02761* (2019).

[9] Ruoyuan Gao and Chirag Shah. 2019. How fair can we go: Detecting the boundaries of fairness optimization in information retrieval. In *Proceedings of the 2019 ACM SIGIR International Conference on Theory of Information Retrieval*. 229–236.

[10] Ruoyuan Gao and Chirag Shah. 2020. Toward creating a fairer ranking in search engine results. *Information Processing & Management* 57, 1 (2020), 102138.

[11] Emma J Gerritse, Faegheh Hasibi, and Arjen P de Vries. 2020. Bias in Conversational Search: The Double-Edged Sword of the Personalized Knowledge Graph. In *Proceedings of the 2020 ACM SIGIR on International Conference on Theory of Information Retrieval*. 133–136.

[12] Andisheh Keikha, Faezeh Ensan, and Ebrahim Bagheri. 2018. Query expansion using pseudo relevance feedback on wikipedia. *Journal of Intelligent Information Systems* 50, 3 (2018), 455–478.

[13] Victor Lavrenko and W Bruce Croft. 2017. Relevance-based language models. In *ACM SIGIR Forum*, Vol. 51. ACM New York, NY, USA, 260–267.

[14] Kyung Soon Lee, W Bruce Croft, and James Allan. 2008. A cluster-based resampling method for pseudo-relevance feedback. In *Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval*. 235–242.

[15] Thang Luong, Hieu Pham, and Christopher D. Manning. 2015. Effective Approaches to Attention-based Neural Machine Translation. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, EMNLP 2015, Lisbon, Portugal, September 17-21, 2015*, Lluís Màrquez, Chris Callison-Burch, Jian Su, Daniele Pighin, and Yuval Marton (Eds.). The Association for Computational Linguistics, 1412–1421. https://doi.org/10.18653/v1/d15-1166

[16] Yuanhua Lv and ChengXiang Zhai. 2010. Positional relevance model for pseudo-relevance feedback. In *Proceedings of the 33rd international ACM SIGIR conference on Research and development in information retrieval*. 579–586.

[17] Rishabh Mehrotra, James McInerney, Hugues Bouchard, Mounia Lalmas, and Fernando Diaz. 2018. Towards a fair marketplace: Counterfactual evaluation of the trade-off between relevance, fairness & satisfaction in recommendation systems. In *Proceedings of the 27th acm international conference on information and knowledge management*. 2243–2251.

[18] Abbe Mowshowitz and Akira Kawaguchi. 2005. Measuring search engine bias. *Information processing & management* 41, 5 (2005), 1193–1205.

[19] James W Pennebaker, Martha E Francis, and Roger J Booth. 2001. Linguistic inquiry and word count: LIWC 2001. *Mahway: Lawrence Erlbaum Associates* 71, 2001 (2001), 2001.

[20] Evaggelia Pitoura, Panayiotis Tsaparas, Giorgos Flouris, Irini Fundulaki, Panagiotis Papadakos, Serge Abiteboul, and Gerhard Weikum. 2018. On measuring bias in online information. *ACM SIGMOD Record* 46, 4 (2018), 16–21.

[21] Navid Rekabsaz, James Henderson, Robert West, and Allan Hanbury. 2018. Measuring Societal Biases in Text Corpora via First-Order Co-occurrence. *arXiv preprint arXiv:1812.10424* (2018).

[22] Navid Rekabsaz and Markus Schedl. 2020. Do Neural Ranking Models Intensify Gender Bias?. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*. 2065–2068.

[23] A Roegiest, A Lipani, A Beutel, A Olteanu, A Lucic, A Stoica, A Das, A Biega, Bart Voorn, C Hauff, et al. 2019. FACTS-IR: Fairness, Accountability, Confidentiality, Transparency, and Safety in Information Retrieval. (2019).

[24] Dwaipayan Roy, Sumit Bhatia, and Mandar Mitra. 2019. Selecting Discriminative Terms for Relevance Model. In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval*. 1253–1256.

[25] Ashudeep Singh and Thorsten Joachims. 2018. Fairness of exposure in rankings. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. 2219–2228.

[26] Alessandro Sordoni, Yoshua Bengio, Hossein Vahabi, Christina Lioma, Jakob Grue Simonsen, and Jian-Yun Nie. 2015. A Hierarchical Recurrent Encoder-Decoder for Generative Context-Aware Query Suggestion. In *Proceedings of the 24th ACM International Conference on Information and Knowledge Management, CIKM 2015, Melbourne, VIC, Australia, October 19 - 23, 2015*, James Bailey, Alistair Moffat, Charu C. Aggarwal, Maarten de Rijke, Ravi Kumar, Vanessa Murdock, Timos K. Sellis, and Jeffrey Xu Yu (Eds.). ACM, 553–562. https://doi.org/10.1145/2806416.2806493

[27] Tony Sun, Andrew Gaut, Shirlyn Tang, Yuxin Huang, Mai ElSherief, Jieyu Zhao, Diba Mirza, Elizabeth Belding, Kai-Wei Chang, and William Yang Wang. 2019. Mitigating gender bias in natural language processing: Literature review. *arXiv preprint arXiv:1906.08976* (2019).

[28] Mahtab Tamannaee, Hossein Fani, Fattane Zarrinkalam, Jamil Samouh, Samad Paydar, and Ebrahim Bagheri. 2020. ReQue: A Configurable Workflow and Dataset Collection for Query Refinement. In *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*. 3165–3172.

[29] Xiao Wang, Craig Macdonald, and Iadh Ounis. 2020. Deep Reinforced Query Reformulation for Information Retrieval. *arXiv preprint arXiv:2007.07987* (2020).

[30] Jinxi Xu and W Bruce Croft. 2017. Quary expansion using local and global document analysis. In *Acm sigir forum*, Vol. 51. ACM New York, NY, USA, 168–175.

[31] Yang Xu, Gareth JF Jones, and Bin Wang. 2009. Query dependent pseudo-relevance feedback based on wikipedia. In *Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval*. 59–66.

[32] Peilin Yang, Hui Fang, and Jimmy Lin. 2017. Anserini: Enabling the use of Lucene for information retrieval research. In *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 1253–1256.

[33] Ruifan Yu, Yuhao Xie, and Jimmy Lin. 2019. Simple techniques for cross-collection relevance feedback. In *European Conference on Information Retrieval*. Springer, 397–409.

[34] Hamed Zamani and W Bruce Croft. 2017. Relevance-based word embedding. In *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 505–514.

[35] Jieyu Zhao, Yichao Zhou, Zeyu Li, Wei Wang, and Kai-Wei Chang. 2018. Learning gender-neutral word embeddings. *arXiv preprint arXiv:1809.01496* (2018).