University of Glasgow Terrier Team at the TREC 2022 Fair Ranking Track

Thomas Jaenich
University of Glasgow, UK
t.jaenich.1@research.gla.ac.uk

Graham McDonald
University of Glasgow, UK
Graham.McDonald@Glasgow.ac.uk

Iadh Ounis
University of Glasgow, UK
Iadh.Ounis@glasgow.ac.uk

ABSTRACT

In our participation in the TREC 2022 Fair Ranking Track, we investigated several methods for ensuring a fair distribution of exposure over groups. For Task 1 (single ranking) we investigated search results diversification as well as query expansion techniques for generating fair rankings. For Task 2 (multiple rankings) we model the problem as a Multi Armed Bandit task and investigate different exploration strategies as well as different trade-offs in the relationship between fairness and relevance. Our results show that our submitted runs are competitive. For Task 1, our runs achieve the highest NDCG*AWRF scores out of all the submitted systems for almost half of the queries. For Task 2, our best performing Multi-Armed Bandit approach performs better than the TREC-median for 44 out of the 47 evaluation queries. From our submitted runs, our top-3 best performing systems are above the TREC-median for at least 91% of the evaluation queries.

1 INTRODUCTION

In our participation in the TREC 2022 Fair Ranking Track, the University of Glasgow Terrier Team aimed to build on their Terrier.org Information Retrieval platform [4, 6] and their 2021 Fair Ranking Track [1] participation, to further investigate fair ranking strategies. For Task 1 (single ranking) we investigate search result diversification and query expansion strategies to enhance the fair distribution of exposure among groups. For Task 2 (multiple rankings) we considered the creation of a sequence of rankings as a Multi Armed Bandit problem. We investigate whether an agent that is presented with a pool of candidate rankings at each iteration can find a useful strategy to add one of those rankings to the sequence, to ensure a fair distribution of exposure.

The remainder of this paper is structured as follows: In Section 2, we briefly describe our indexing and the relevance components used by our fairness approaches in the submitted runs. We then present our proposed fairness components in Section 3, before presenting our runs and results in Section 4. We present concluding remarks in Section 5.

2 INDEXING & RETRIEVAL

This section describes the indexing and retrieval of the provided data. The TREC 2022 Fair Ranking Track provided the participants with a corpus of Wikipedia documents, training topics, and metadata corresponding to the documents. We first parsed the document collection to remove the Wiki-Markup. To ensure efficient parsing, we used PyAutoCorpus.[1] PyAutoCorpus results in roughly 40x speed-up compared to other existing Wiki-Parsers [1]. The 2022 Fair Ranking Track test collection with removed Wiki-Markup was also integrated into the well known ir_datasets package [3]. Using ir_datasets allows to conveniently conduct our experiments within PyTerrier [5]. For our experiments, we built a ColBERT index for our submissions that use ColBERT [2] as their underlying retrieval model. We also created a traditional inverted index with porter stemming and removed stopwords for a run using BM25 [7] retrieval.

3 FAIRNESS COMPONENTS

We investigated the suitability of three different techniques for fair ranking that we discuss in this section.

Search results diversification: To develop a fairness component for Task 1, we leverage a well known search results diversification approach that is based on proportional representation, i.e., a percentage of the available positions in a ranking are allocated to particular query aspects based on the distribution of a background population. To leverage this approach for generating fair rankings, we allocate positions in the ranking to different fairness attributes according to their exposure targets by considering the attention an item gets at an allocated position. Our exposure targets were calculated using the page-alignments from the publicly available metric[2] as well as the predicted relevance from an initial retrieval.

Query Expanison: To investigate whether query expansion can be used to enhance the fair distribution of exposure, we applied two query expansion strategies that have been shown to be successful in improving the relevance of a search result set. We explore the potential of a traditional sparse query expansion technique as well as dense retrieval query expansion.

Multi Armed Bandit: For the second task we modelled the creation of the sequence of rankings as a multi-armed bandit problem. We present an agent with the task of filling the sequence with rankings in an iterative fashion so that the exposure is fairly distributed over all fairness characteristics. The agent can select from a pool of rankings with different distributions of exposure and fairness-relevance relationships. We reward our agent if the overall distribution of exposure gets fairer and penalise it if it adds a ranking that decreases fairness. Specifically, a single ranking in the pool is optimised for a fair distribution of exposure relating to only one fairness characteristic. For each fairness characteristic, there is at least one ranking present in the pool. All of the rankings are considered potential candidates to be added to the sequence by the agent. The number of rankings in the pool as well as the fairness-relevance relationship in the single rankings are defined by different weighting parameters. Moreover, the selection strategy of the rankings differs in the submitted runs. For the reward function, we used information from the publicly available metric.

[1] https://github.com/seanmacavaney/pyautocorpus

[2] https://github.com/fair-trec/trec2022-fair-public

Runs	NDCG↑	AWRF↑	NDCG*AWRF↑
UoGTrT1ColPRF	0.6044	0.5245	0.3253
UoGTrQE	0.5367	0.4983	0.2734
UoGTrExpE2	0.5977	0.5243	0.3229
UoGTrExpE1	0.5176	0.5121	0.2716
UoGRelvOnlyT1	0.6044	0.5245	0.3253

Table 1: The table shows the average results over all 47 queries for every submitted run in Task 1. We report the NDCG, the AWRF and the combined metric score. For every metric, the ideal value is 1.

Runs	>Minimum	>=Median	= Maximum
UoGTrT1ColPRF	46	28	10
UoGTrQE	40	23	2
UoGTrExpE2	45	29	8
UoGTrExpE1	38	22	2
UoGRelvOnlyT1	46	28	10

Table 2: The number of queries for which a run is the minimum, ≥ to the median, or equal to the maximum. Maximum represents the best performing submitted system.

	EE-L	EE-D	EE-R
Relevance Only	2.2230	2.2397	0.0788
UoGTrMabWeSA	1.1230	1.0790	0.0484
UoGTrMabSAED	1.2227	1.1934	0.0558
UoGTrMabSaWR	1.1847	1.1461	0.0511
UoGTrMabSANr	2.1397	2.0486	0.0249

Table 3: The table shows the results for Task 2 averaged over 47 queries. Lower values are better for EE-L & EE-D, higher values are better for EE-R.

4 SUBMITTED RUNS AND RESULTS

In this section, we present our submitted runs for Task 1 in Section 4.1 and Task 2 in Section 4.2.

4.1 Task 1: Single Ranking

We submitted five runs for Task 1. Of our five runs, two investigate our diversification approach and two investigate query expansion. As a baseline, we also submitted a relevance only approach that contains no explicit fairness component.

4.1.1 Submitted runs.

- UoGRelvOnlyT1: A relevance only baseline with no fairness intervention for Task 1. The retrieval was done with ColBERT-PRF. (Note that this run was meant to be Colbert-E2E retrieval but we discovered after submission that a duplicate version of UoGTrT1ColPRF was submitted.)
- UoGTrExpE1 uses a heuristic approach to rerank an initial retrieval set using our expected exposure targets that are computed from the page alignments and our predicted relevance scores. The initial retrieval was done using ColBERT-E2E [2]. The expected exposure targets evenly distribute the available exposure between fairness attributes. We deploy a diversification strategy to rerank the documents with respect to our exposure targets.
- UoGTrExpE2: Similar to UoGTrExpE1 this run uses a heuristic approach to rerank an initial retrieval set using our expected exposure targets and deploy diversification to rerank the documents. This run also uses ColBERT-E2E for the relevance component.
- UoGTrT1ColPRF: This run uses Colbert-PRF [9] to expand a user's query to enhance the performance on relevance.
- UoGTrQE: A traditional query expansion method to expand a query with the goal to improve the distribution of documents in a fair manner.

4.1.2 Results. Table 1 shows our results for Task 1. The table reports the average NCDG score as well as the average attention-weighted ranked fairness [8] (AWRF) and the product of both, which is the official metric provided by the organizers. The averages are calculated over the 47 evaluation queries per submitted run.

From Table 1 we can see that UoGTrT1ColPRF and UoGRelvOnlyT1 perform the best out of all systems. They have the highest scores on nDCG as well on the fairness metric AWRF. Analysing our runs that are based on search results diversification we note that our run with specific exposure targets (UoGTrExpE2) achieves scores only marginally different to our best-performing relevance only approach. For UoGTrExpE1 we notice a drop-off in AWRF which could be expected since it optimises for equal exposure distribution and not specific targets. Investigating the Query Expansion run UoGTrQE we also note a decrease in the AWRF metric as well as on nDCG compared to the best performing systems UoGTrT1ColPRF and UoGRelvOnlyT1.

To gain further insights into the performance of our approaches, Table 2 reports a per-query analysis of the number of queries for which our runs perform better than the TREC-Minimum or TREC-Median, or are equal to the TREC-Maximum performance. We can see from Table 2 that all of our runs are better than the TREC-Minimum on the majority of the queries. Following the trend from Table 1 UoGTrExpE1 has the lowest scores on nine of the queries while also achieving the highest score on two of the queries. Our relevance only baseline has the lowest score on one query but achieves the maximum score on ten queries. Considering the TREC-Median UoGRelvOnlyT1 & UoGTrExpE2 perform better than the median system on more than 50% of the queries. UoGTrExpE1 & UoGTrQE are better than the TREC-median on just under half of the queries (48%&47%). Next we analyse the results of our Multi-Armed Bandit approach for Task 2.

4.2 Task 2: Multiple Rankings

We submitted five runs for Task 2. Four of our runs use a Multi-Armed Bandit strategy with different parameters and strategies, while the fifth run is a relevance only baseline. For all runs, an agent tries to find the optimal strategy when adding rankings to the sequence. The rankings are selected from a pool of rankings. For every protected attribute, there are multiple different rankings

Runs	= Minimum	<=Median	=Maximum
UogTRelvOnlyT2	5	21	5
UoGTrMabWeSA	6	44	9
UoGTrMabSAED	3	43	0
UoGTrMabSaWR	7	47	0
UoGTrMabSANr	2	18	4

Table 4: The number of queries a run achieves the minimum, ≤ the median or the maximum EE-L score. Lower values of EE-L are better, i.e., minimum is the best performing system.

with different fairness-relevance relationships. The initial retrieval before the creation of the pools was done with Colbert-E2E.

4.2.1 Submitted runs.

- UoGTrMabWeSA: An approach that uses an epsilon-greedy strategy and fairness weights in the creation of a single ranking.
- UoGTrMabSAED: This approach uses a variation of an epsilon-decay strategy. No weights are used in the creation of the pool of the rankings.
- UoGTrMabSaNR: This approach does not use any randomisation or weighting in the creation of the rankings.
- UoGTrMabSaWR: This run does not use weighting in the creation of the rankings. However, it does use randomisation in the MAB component.
- UogTRelvOnlyT2: A run designed to be used as a baseline with no explicit fairness component. Every position in the sequence is filled with the same ranking retrieved through our ColBERT-E2E Ranker based only on relevance.

4.2.2 Results. Table 3 shows our results for Task 2. We report the expected exposure loss (EE-L), the Expected Exposure Disparity (EE-D) and the Expected Exposure Relevance (EE-R) averaged over all 47 evaluation queries. Lower values are better for EE-L and EE-D. Higher values are better for EE-R. It is notable that all of our approaches outperform the relevance only baseline for EE-L. As expected the relevance only baseline achieves the highest scores on EE-R. UoGTrMabWeSA achieves the best scores on EE-L as well as on EE-D. Our approach UoGTrMabSANr that does not use randomisation in the exploring phase and no weights in the creation of the rankings, achieves the highest EE-L and EE-D scores. Notably there is only an 8% difference in EE-L scores between our best performing MAB approaches (UoGTrMabWeSA, UoGTrMabSaWR and UoGTrMabSaWR). For each of our submitted runs improvements in fairness results in some decrease in relevance. To further analyse the performance of our approaches we investigate the per-query performance.

Table 4 reports the per-query analysis of the number of queries for which our runs achieve the TREC-Minium EE-L, are ≤ the TREC-Median, or are equal to the TREC-Maximum for Task 2. From Table 4, we can see that three of our runs that include a fairness component, i.e., UoGTrMabSAED, UoGTrMabSaWR, UoGTrMabWeSA achieve scores lower than the TREC-Median on a majority of the queries (lower is better), with UoGTrMabSaWR achieving a better performance than the median on all of the queries. UoGTrMabSANr performs better than the TREC-Median on 18 out of 47 approaches. Each of our submitted approaches, including the relevance baseline, achieves the overall best score (minimum) for some of the queries. Our best performing approach in terms of EE-L, i.e. UoGTrMabWeSA scores lower than the median on 44 out of the 47 queries and achieves the minimum EE-L score for six queries. This is the highest number of maximum scores for a run. UoGTrMabSAED as well as UoGTrMabSaWR have no queries on which they achieve the maximum scores, while UoGTrMabSANr and the relevance only baseline score highest and therefore worst on 4 and 5 queries.

5 CONCLUSIONS

In this year's TREC participation we explored the potential of Search Results diversification and Query Expansion for distributing exposure fairly over a single ranking in Task 1. For Task 2 we introduced an optimisation strategy for selecting a ranking from a pool of possible candidate rankings to add to the sequence. In Task 1 we found that our relevance only baseline performed the best. This provides us with an interesting starting point for further research. Specifically, we plan to further investigate how to optimise our fairness components so they do not decrease the relevance of rankings while ensuring the fairness of exposure. Our results for Task 2 show that our optimisation strategy can be effective in ensuring fairness in a sequence of rankings. We tested different exploration strategies and weighting parameters and found valuable insights about how to optimise for a target exposure.

REFERENCES

[1] Thomas Jaenich, Graham McDonald, and Iadh Ounis. 2021. University of Glasgow Terrier Team at the TREC 2021 Fair Ranking Track.

[2] Omar Khattab and Matei Zaharia. 2020. Colbert: Efficient and effective passage search via contextualized late interaction over bert. In *Proceedings of the 43rd International ACM SIGIR conference on research and development in Information Retrieval*. 39–48.

[3] Sean MacAvaney, Andrew Yates, Sergey Feldman, Doug Downey, Arman Cohan, and Nazli Goharian. 2021. Simplified Data Wrangling with ir_datasets. In *SIGIR*.

[4] Craig Macdonald and Nicola Tonellotto. 2020. Declarative experimentation in information retrieval using pyterrier. In *Proc. of ICTIR*.

[5] Craig Macdonald and Nicola Tonellotto. 2020. Declarative experimentation in information retrieval using pyterrier. In *Proceedings of the 2020 ACM SIGIR on International Conference on Theory of Information Retrieval*. 161–168.

[6] Iadh Ounis, Gianni Amati, Vassilis Plachouras, Ben He, Craig Macdonald, and Christina Lioma. 2006. Terrier: A High Performance and Scalable Information Retrieval Platform. In *Proc. OSIR at SIGIR*.

[7] Stephen Robertson, Hugo Zaragoza, et al. 2009. The probabilistic relevance framework: BM25 and beyond. *Foundations and Trends® in Information Retrieval* 3, 4 (2009), 333–389.

[8] Piotr Sapiezynski, Wesley Zeng, Ronald E Robertson, Alan Mislove, and Christo Wilson. 2019. Quantifying the Impact of User Attentionon Fair Group Representation in Ranked Lists. In *Companion Proceedings of The 2019 World Wide Web Conference*. 553–562.

[9] Xiao Wang, Craig Macdonald, Nicola Tonellotto, and Iadh Ounis. 2021. Pseudo-relevance feedback for multiple representation dense retrieval. In *Proc. of ICTIR*.