Overview of the TREC 2019 Fair Ranking Track[*]

Asia J. Biega
Microsoft Research Montréal
asia.biega@microsoft.com

Fernando Diaz
Microsoft Research Montréal
diazf@acm.org

Michael D. Ekstrand
Boise State University
michaelekstrand@boisestate.edu

Sebastian Kohlmeier
Allen Institute for Artificial Intelligence
sebastiank@allenai.org

Abstract

The goal of the TREC Fair Ranking track was to develop a benchmark for evaluating retrieval systems in terms of fairness to different content providers in addition to classic notions of relevance. As part of the benchmark, we defined standardized fairness metrics with evaluation protocols and released a dataset for the fair ranking problem. The 2019 task focused on reranking academic paper abstracts given a query. The objective was to fairly represent relevant authors from several groups that were unknown at the system submission time. Thus, the track emphasized the development of systems which have robust performance across a variety of group definitions. Participants were provided with querylog data (queries, documents, and relevance) from Semantic Scholar. This paper presents an overview of the track, including the task definition, descriptions of the data and the annotation process, as well as a comparison of the performance of submitted systems.

1 Introduction

Modern information access systems influence both the *consumers* and the *producers* of the content that they mediate. Some systems are quite explicitly two-sided, such as online dating platforms [5]. Hiring platforms, where employers "consume" rankings of job-seekers, and micro-lending platforms where recommender algorithms help lenders identify people they wish to invest in [2], are also clearly two-sided. More traditional environments, such as music, video, and book recommendation systems where direct users consume rankings of works by different artists; community question answering forums where some members consume answers written by other members; and various e-commerce platforms where customers consume rankings of products offered by different sellers can also be considered two-sided because of the indirect matching that happens between the users and content creators. To date, however, evaluation methodologies in information retrieval and recommender systems have largely focused on quantifying the experience of the consumers through metrics focused on accuracy, diversity, and novelty.

Failing to adequately measure and address the effect of information access on content producers can have significant impact, both socially and to the information platform. If content producers feel that a platform does not enable them to connect to their audience effectively, they may move their content to a different platform (harming the platform's inventory) or leave the industry entirely. A system in which different groups of content producers experience disparate discoverability will harm the less-discoverable groups; if that division falls along lines of historical inequities, such as a music service where indigenous musicians have more difficulty being discovered, it may reinforce or even exacerbate biases and discrimination. Ineffective promotion of less popular musical genres may also distort the development of culture more broadly [6]. Similar problems were recognized by the early work on fair rankings [1, 3, 7, 8, 9, 10].

[*] Data and code are available at: https://fair-trec.github.io/2019/

The goals of the Fair Ranking track are to,

- develop metrics for fair exposure of individuals or groups in retrieval scenarios,
- design an experimentation protocol for fair ranking,
- release a data set for benchmarking fair ranking algorithms, and
- promoting the development of fair ranking algorithms.

For 2019, we adopted an **academic search task**, where we have a corpus of academic article abstracts and queries submitted to a production academic search engine. The central goal of the Fair Ranking track is to provide **fair exposure to different groups of authors** (a *group fairness* framing).

We recognize that there may be multiple group definitions (e.g. based on demographics, stature, topic) and hoped for the systems to be robust to these. As such, participants were expected to develop systems to optimize for fairness and relevance for **arbitrary group definitions**, and we did not reveal the exact group definitions until *after* the evaluation runs were submitted.

The track was set up as a **reranking** task. We provided participants with a sequence of queries, each accompanied by a varying-size set of documents for each query; the task was to rerank the documents to produce result lists that are fair and relevant.

2 Task Description

Prior work in the area of fair ranking shows that reasoning about fairness should be done over sequences of rankings rather than individual rankings [1, 7]. For the evaluation in this track, we thus provided participants with a sequence $\mathcal{Q}$ of queries accompanied by unordered sets of documents to rank. The document sets were of varying size. For each request (query q and set of documents $\mathcal{D}_q$), participants provided a ranked list of the documents from $\mathcal{D}_q$. The final system output was a sequence of rankings. Algorithm 1 presents a pseudocode of the evaluation protocol.

Participants were instructed to make their system optimize the ranking sequence for two goals: (1) be relevant to the consumers and (2) be fair to the producers.

Algorithm 1 Evaluation protocol

```
$\Pi \leftarrow \{\}$
for $q, \mathcal{D}_q \in \mathcal{Q}$ do
    $\pi \leftarrow \textsc{System}(q, \mathcal{D}_q)$
    $\Pi \leftarrow \Pi + [\pi]$
end for
return $\Pi$
```

3 Evaluation

Unlike previous TREC tracks, participants received multiple copies of the same query text–with varying query ids–and were allowed to submit different rankings for each instance of the query. At evaluation time, we measured *amortized performance* over rankings produced for each given query, as well as across all rankings and queries (macro- and micro- amortization, respectively.)

Given a sequence of queries $\mathcal{Q}$ and associated system rankings, we evaluate systems according to fair exposure of authors (Section 3.1.1) and relevance of documents (Section 3.2).

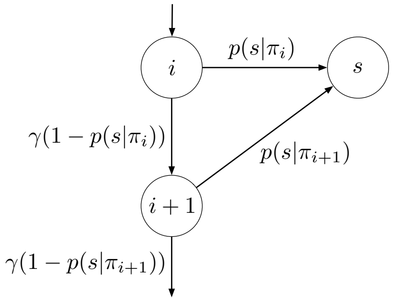


Figure 1: Attention model.

3.1 Measuring Fairness

3.1.1 Measuring Author Exposure for a Single Ranking

In order to measure exposure, we adopt the browsing model underlying the Expected Reciprocal Rank metric [4]. Given a ranking π, the exposure of author a is,

$$e_a^\pi = \sum_{i=1}^{n} \left[\gamma^{i-1} \prod_{j=1}^{i-1} (1 - p(s|\pi_j)) \right] I(\pi_i \in \mathcal{D}_a)$$

$$\begin{array}{rl} n & \text{number of documents in ranking } \pi \\ \mathcal{D}_a & \text{documents including } a \text{ as an author} \\ \pi_i & \text{document at position } i \\ \gamma & \text{continuation probability (fixed to 0 for the final position in the ranking)} \\ p(s|d) & \text{probability of stopping given user examined } d \end{array} \quad (1)$$

We present a graphical depiction of this model in Figure 1.

We fixed the value of the discounting factor γ, and assumed $p(s|d) = f(r_d)$, where r_d is the relevance of the document d and f is a monotonic transform of that relevance into a probability of being satisfied.

We compute the amortized exposure for a as,

$$e_a = \sum_{\pi \in \Pi} e_a^\pi \quad (2)$$

where Π is the sequence of all system rankings.

3.1.2 Measuring Author Relevance for a Single Ranking

The author relevance for a ranking π is defined as,

$$r_a^\pi = \sum_{d \in \mathcal{D}_a} p(s|d) \quad (3)$$

Notice that this metric is independent of the ranking but not the query. As with amortized exposure, we define amortized relevance as the sum over all rankings Π.

3.1.3 Measuring Group Fairness

Assume that each author is assigned to exactly one of $|\mathcal{G}|$ groups. Let $\mathcal{A}_g$ be the set of all authors in group g. The group exposure and relevance metrics are defined as,

$$\mathcal{E}_g = \frac{\sum_{a \in \mathcal{A}_g} e_a}{\sum_{g' \in \mathcal{G}} \sum_{a \in \mathcal{A}_{g'}} e_a} \tag{4}$$

$$\mathcal{R}_g = \frac{\sum_{a \in \mathcal{A}_g} r_a}{\sum_{g' \in \mathcal{G}} \sum_{a \in \mathcal{A}_{g'}} r_a} \tag{5}$$

We assume that groups should receive exposure proportional to relevance. We adopt the following measure to quantify the deviation from this ideal exposure,

$$\Delta_g = \mathcal{E}_g - \mathcal{R}_g \tag{6}$$

Given this relevance-normalized measure of exposure, we can compute the fair exposure using the ℓ_2 norm,

$$\Delta = \sqrt{\sum_{g \in \mathcal{G}} \Delta_g^2} \tag{7}$$

Because this metric does not capture some of the nuance of how over- and under-exposure is distributed, we will adopt secondary metrics in our final analysis of results.

3.2 Measuring Relevance

We measured the quality of a ranking for the searchers as the expected utility, assuming the same attention model as used for our fairness metric,

$$u^\pi = \sum_{i=1}^{n} \left[\gamma^{i-1} \prod_{j=1}^{i-1} (1 - p(s|\pi_j)) \right] p(s|\pi_i) \tag{8}$$

We average all utilities of rankings, $\mathrm{U} = \frac{1}{|\Pi|} \sum_{\pi \in \Pi} u^\pi$ as our final relevance metric.

3.3 Trading Off Fairness and Relevance

Although a system could, in theory, achieve the optimal relevance and fairness, in practice, relevance will degrade as fairness improves. We therefoe measured the trade-offs between fairness to producers and quality for consumers as an auxiliary metric.

4 Data

4.1 Input

There were three main inputs available to participants: the *corpus* of articles to rank, the *example group definition* file to help develop and test solutions, and the *queries*.

4.1.1 Corpus

The corpus for this track was the Semantic Scholar (S2) Open Corpus from the Allen Institute for Artificial Intelligence. It can be downloaded from `http://api.semanticscholar.org/corpus/`, and consists of 47 1GB data files. Each file is compressed JSON, where each line is a JSON object describing one paper.

The following data are available for most papers:

- S2 Paper ID
- DOI
- Title
- Abstract
- Authors (resolved to author IDs)
- Inbound and outbound citations (resolved to S2 paper IDs)

4.1.2 Queries

Query data. Query sequences were constructed based on a query log provided by Semantic Scholar[1]. This query log specified, for each query, a list of documents that were displayed at each ranking position (up to 6), and the number of clicks observed for each document. In case the list of documents displayed in the top positions changed over time for a query, we merged all the documents observed as a response to the query into the query reranking pool.

Because of the exhaustive annotation process that required annotating group memberships of all document authors, we sampled a smaller number of queries to construct evaluation sequences. We released 652 training and 635 evaluation queries. For both the training and evaluation data, these queries were selected first by random sampling, and then by a number of filtering steps. More specifically,

- We included only queries with at least three observed clicks, and where the clicks occur at at least two different ranking positions (this heuristic helped remove noise and filter out many of the known-item queries).
- We further manually cleaned the sample to remove any known-item queries, queries containing people's names, and queries with offensive and sensitive keywords.
- For evaluation queries, we additionally considered whether the documents in reranking pools appeared in the S2 corpus. We kept queries that had at least 5 resolvable documents, at least 1 of which has been clicked.

Data that was made publicly available contained, for each query:

- Query ID,
- Query string,
- Query normalized frequency,
- A list of documents to rerank (with relevance for training queries, and without relevance for evaluation queries).

Relevance was binary. We considered a document to be relevant to a query if there was at least one observed click on that document for the query, and not relevant otherwise.

Query sequences. Runs were submitted over *query sequences*: ordered sequences of queries that may contain duplicates. For training and development, we provided participants with training queries (with relevance and frequency data) and a script to generate query sequences. For evaluation, we provided query data (without relevance) not included in the training data, and generated five evaluation query sequences, each containing 25k queries. Sequences were generated by sampling queries from a distribution based on query frequencies.

[1] `https://www.semanticscholar.org`

Documents	5,620
Annotated Documents	2,866
Have Country Data	2,823
Advanced Econ Papers	2,372
Developing Econ Papers	308
Mixed Econ Papers	138
Advanced Econ Authors	7,018
Developing Econ Authors	1,184

Table 1: Annotation Outcome Summary

4.1.3 Annotations

NIST assessors annotated returned papers with the country in which each author was operating (based on their affiliation data in the paper manuscript). Not all papers were able to be annotated. These are the known reasons a paper may not have annotations:

- It has a large author list (> 25). We excluded such long papers because there were not very many of them, and large-team papers require special treatment in how we consider their author lists, particularly when authors may be from different groups.
- Due to a bug in the large-paper exclusion logic, a number of papers with a medium number of authors were also excluded.
- Some papers did not have an accessible source with sufficient affiliation information to provide annotations (e.g. no available PDF file, and a paper information page that either did not contain affiliation details or was not accessible from the annotation interface).
- Some papers may not provide sufficient information to determine an author's affiliation location.

Table 1 shows the coverage of annotations, and the number in each group, after merging and integrating data sources. For these statistics, to aggregate each paper's authors into a single economic designation for the paper, we considered a paper to be from an advanced or developing economy if all authors' locations had the same economic designation; otherwise, we list it as a 'mixed' economy paper.

4.1.4 Group Definitions

Group definition accompanying the training data. To help participants get started, we provided a file containing group membership definitions for authors in the S2 corpus. This definition was based on author i–10 indices (i–10 denotes the number of papers with at least 10 citations a person has coauthored). This definition was not used in the final evaluation, but was meant as a starting point for system development. For each author, the data consisted of:

- the author's S2 ID,
- the author's group identifier.

Authors were split into 7 groups, based on the value of their i–10 index.

Group definitions for evaluation. For evaluation, we used two different statistics to derive group definitions. The first definition was based on the NIST assessors' country annotations. We combined these annotations with economic development levels from the International Monetary Fund. With this definition, the fairness target is to ensure fair exposure for papers written in countries with more- and less-developed economies. The evaluation itself uses individual author-level annotations; the exposure a mixed-economy paper receives counts towards both developing and advanced economy exposure. Under this definition, authors are split into two groups.

The second definition is similar to the training group definition; but based on the h-index (h-index denotes the number h of papers with at least h citations a person has coauthored) of paper authors. Authors are split into four groups based on the value of their h-index: $h < 5$, $5 \leq h < 15$, $15 \leq h < 30$, $h \geq 30$.

4.2 Output

Each run output was submitted as a JSON file with the following contents:

- qnum: ⟨sequence id⟩.⟨query number in sequence⟩
- qid (to look up in the query file): ⟨query id⟩
- ranking: an ordered list of document IDs (of the documents to be reranked for the query).

5 Results

5.1 Evaluation parameters

The results in this notebook were computed using the following parameter values:

- Continuation probability (Eq. 1): $\gamma = 0.5$;
- Stopping probability given a document (Eq. 1 and 3): $p(s|d) = f(r_d) = 0.7 * r_d$.

Relevance scores r_d were binary, and computed from the click data as described in Sec. 4.1.2.

5.2 Overall results

Figure 2 presents the performance of all submitted runs in terms of unfairness (Eq. 7) and expected utility (Eq. 8) for two different group definitions (based on the economic levels of author affiliation countries and on h-indices of authors). Note that a run that randomly shuffles the reranked documents (fair_random) performs well in terms of amortized fairness, but proves to be the worst approach in terms of ranking quality. Based on short run descriptions provided by participants, other runs are based on a number of different approaches: learning to rank without explicit fairness modeling (fair_LambdaMART), BERT embeddings without explicit fairness modeling (first), an approach where the final ranking is a weighted merge of search results for different fields; weights are adjusted throughout the sequence (MacEwanBase), an approach that models fairness and unfairness distributions over groups (QUARTZ-*), various approaches based on result diversification (uognle*).

We generally observe that the relative ordering of the systems in terms of unfairness is not robust to varying group definitions.

6 Discussion

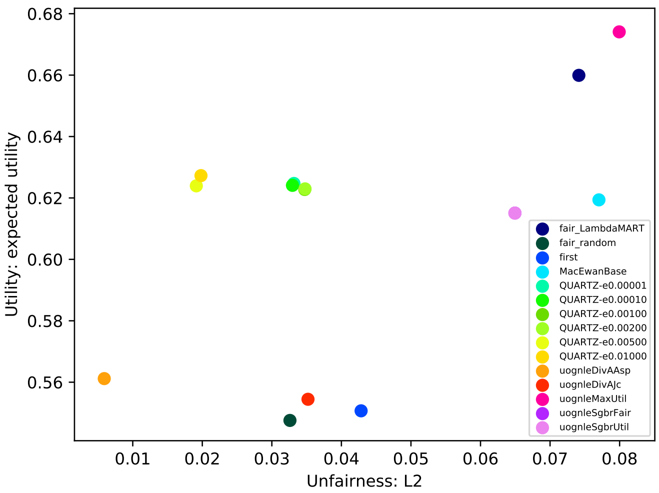
6.1 Limitations

There are a number of limitations to the data and evaluation in the current version of the track. We are actively looking to address several of them in the 2020 edition.

First, we derive a binary relevance construct from click logs. There are at least three problems with this:

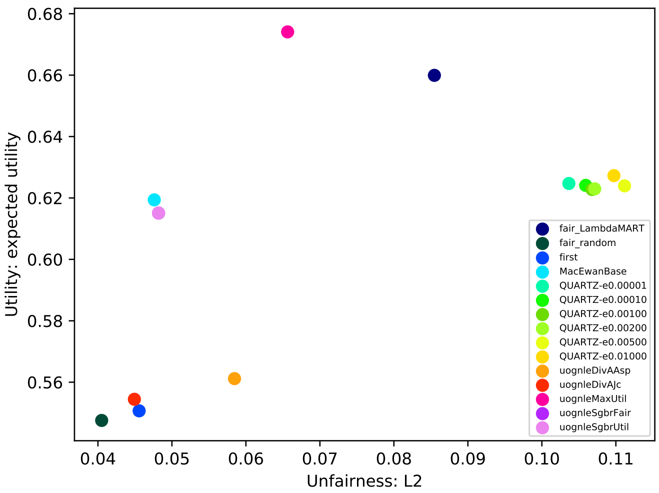
1. We do not know if the retrieved document was actually relevant to the user's query, or just looked relevant enough (or surprising enough) to receive a click.

run	utility	unfairness
uognleDivAAsp	0.5612	0.0059
QUARTZ-e0.00500	0.6239	0.0191
QUARTZ-e0.01000	0.6273	0.0198
fair_random	0.5476	0.0326
QUARTZ-e0.00010	0.6241	0.0330
QUARTZ-e0.00001	0.6247	0.0332
QUARTZ-e0.00100	0.6228	0.0347
QUARTZ-e0.00200	0.6230	0.0348
uognleDivAJc	0.5544	0.0352
first	0.5507	0.0428
uognleSgbrFair	0.6151	0.0649
uognleSgbrUtil	0.6151	0.0649
fair_LambdaMART	0.6599	0.0741
MacEwanBase	0.6194	0.0770
uognleMaxUtil	0.6741	0.0799



(a) Group definition: IMF level with 2 groups

run	utility	unfairness
fair_random	0.5476	0.0405
uognleDivAJc	0.5544	0.0449
first	0.5507	0.0456
MacEwanBase	0.6194	0.0476
uognleSgbrFair	0.6151	0.0482
uognleSgbrUtil	0.6151	0.0482
uognleDivAAsp	0.5612	0.0585
uognleMaxUtil	0.6741	0.0656
fair_LambdaMART	0.6599	0.0855
QUARTZ-e0.00001	0.6247	0.1036
QUARTZ-e0.00010	0.6241	0.1059
QUARTZ-e0.00100	0.6228	0.1068
QUARTZ-e0.00200	0.6230	0.1071
QUARTZ-e0.01000	0.6273	0.1097
QUARTZ-e0.00500	0.6239	0.1112



(b) Group definition: h-index with 4 groups

Figure 2: Performance of all submitted runs in terms of unfairness and expected utility for two different group definitions. Runs in tables ranked in decreasing order of fairness. Points toward the upper left area of the figures are preferred to those toward the lower right.

2. Clicks are going to be biased, possibly with the same biases that we are seeking to correct for. Our overall fairness goal in this track is to counteract prestige effects, but author recognition and prestige are likely to play a role in determining which article a user will click among articles with comparably relevant titles.

3. Binary relevance precludes approaches reliant on graded relevance.

Second, our group annotations are quite incomplete. We do not know the extent to which missingness correlates with protected characteristics, or the impact this has on final system performance.

Third, the rerank task is based on relatively small numbers of documents. If the source search engine is biased so that the top 6 positions are often given to dominant-group papers, re-ranking systems cannot correct for that bias and give exposure to relevant documents from protected groups that the original search engine ranked at lower positions.

6.2 TREC 2019 Fair Ranking Track: Lessons Learned

Running the Fair Ranking track at TREC 2019 taught us a number of valuable lessons that were not apparent in the prior work on fairness in information retrieval, primarily because published work to date had not had to engage with many of the practicalities of system evaluation. The primary challenges we faced were related to the fairness metric, label annotations, and data release.

- The fairness metric used in the 2019 task required exhaustive annotation of both group membership of document authors as well as document relevance. Because of this limitation, we decided that the benchmark needs to be a reranking task. It is quite possible that evaluation of this year's results will show that reranking of just a few documents is a trivial fairness problem; we are preparing to revisit the fairness metrics to account both for a recall task and the resource limitations of TREC assessment.

- Many datasets we considered could not be judged for relevance by non-expert annotators. The dataset of scholarly paper abstracts and queries from Semantic Scholar that the 2019 track was based on contained click information. We used this click information to derive paper relevance. However, our goal in the fairness task was to correct for the tendency of the searchers to select items authored by well-known researchers and institutions. The click information is likely to be biased in exactly this way: Searchers clicking more on abstracts of well-known authors and institutions. This bias can be only quantified after collecting the annotations necessary for producer group assignments.

- Identifying a sensitive attribute usable in the TREC setting was difficult. Many papers focus on gender, but we cannot use gender for a benchmark when gender identities are not available in existing data, because releasing gender annotations is problematic for ethical and privacy reasons. Prestige in the university setting is difficult to quantify; while the Carnegie Classification exists for U.S. universities, there is not a mapping other institutions.

References

[1] A. J. Biega, K. P. Gummadi, and G. Weikum. Equity of attention: Amortizing individual fairness in rankings. In *Proc. SIGIR '18*, 2018.

[2] R. Burke. Multisided fairness for recommendation. *CoRR*, 1707.00093, 2017. URL `http://arxiv.org/abs/1707.00093`.

[3] L. E. Celis, D. Straszak, and N. K. Vishnoi. Ranking with fairness constraints. *arXiv preprint arXiv:1704.06840*, 2017.

[4] O. Chapelle, D. Metzler, Y. Zhang, and P. Grinspan. Expected reciprocal rank for graded relevance. In *Proc. CIKM '09*, 2009.

[5] F. Diaz, D. Metzler, and S. Amer-Yahia. Relevance and ranking in online dating systems. In *Proc. SIGIR '10*, 2010.

[6] R. Mehrotra, J. McInerney, H. Bouchard, M. Lalmas, and F. Diaz. Towards a fair marketplace: Counterfactual evaluation of the trade-off between relevance, fairness & satisfaction in recommendation systems. In *Proc. CIKM '18*, 2018.

[7] A. Singh and T. Joachims. Fairness of exposure in rankings. In *Proc. SIGKDD '18*, 2018.

[8] Y. Wu, L. Zhang, and X. Wu. On discrimination discovery and removal in ranked data using causal graph. In *Proc. KDD '18*, 2018.

[9] K. Yang and J. Stoyanovich. Measuring fairness in ranked outputs. In *Proc. SSDBM '17*, 2017.

[10] M. Zehlike, F. Bonchi, C. Castillo, S. Hajian, M. Megahed, and R. Baeza-Yates. Fa*ir: A fair top-k ranking algorithm. In *Proc. CIKM '17*, 2017.