TREC 2022 Fair Ranking Track Participant Instructions

Michael D. Ekstrand
michaelekstrand@boisestate.edu

Graham McDonald
graham.mcdonald@glasgow.ac.uk

Amifa Raj
amifaraj@u.boisestate.edu

Isaac Johnson
isaac@wikimedia.org

June 14, 2022

These instructions may be updated; see `https://fair-trec.github.io` for the latest version.

1 Introduction

The TREC Fair Ranking Track aims to provide a platform for participants to develop and evaluate novel retrieval algorithms that can provide a fair exposure to a mixture of demographics or attributes, such as gender, that are represented by relevant documents in response to a search query. For example, particular demographics or attributes can be represented by the documents' topical content or authors.

The 2022 Fair Ranking Track adopts a resource allocation task. The task is focused on supporting Wikipedia editors who are looking to improve the encyclopedia's coverage of topics under the purview of a WikiProject.[1] WikiProject coordinators and/or Wikipedia editors search for Wikipedia documents that are in need of editing to improve the quality of the article.

The Fair Ranking track aims to ensure that documents that are about, or somehow represent, certain protected characteristics receive a fair exposure to the Wikipedia editors, so that they have an equal opportunity of being well represented in Wikipedia. The under-representation of particular protected characteristics in Wikipedia can result in systematic biases that can have a negative human, social, and economic impact, particularly for disadvantaged or protected societal groups [3, 4]. In particular, for TREC 2022, the Fair Ranking Track will focus on *intersectionality* fairness. There are many different sets, or groups, of protected characteristics that can hold a distinct point of view as to whether a ranking is fair. Moreover, a document could possibly represent many different protected characteristics. This leads to many interesting fairness related questions, such as "Does optimising a ranking to be fair to many protected characteristics result in a system that is unfair to particular protected characteristics?" and "Are similar protected characteristics treated similarly by particular fairness strategies?". Therefore, the TREC 2022 Fair Ranking Track will focus on evaluating submitted systems on how fairly they treat many different protected characteristics and the impacts for particular subsets of those protected characteristics.

2 Task Definition

The Fair Ranking Track uses an *ad hoc* retrieval protocol. Participants will be provided with a corpus of documents (a subset of the English language Wikipedia) and a set of queries. A query will be of the form of a short list of search terms that represent a WikiProject. Each document in the corpus is relevant to zero to many WikiProjects and associated with potentially many fairness categories.

[1] `https://en.wikipedia.org/wiki/WikiProject`

There are two tasks in the 2022 Fair Ranking Track. In each of the tasks, for a given query, participants are to produce document rankings that are:

1. Relevant to a particular WikiProject.
2. Provide a fair exposure to articles that are associated to particular protected attributes.

The tasks share a topic set, the corpus, the basic problem structure and the fairness objective. However, they differ in their target user persona, system output (static ranking vs. sequences of rankings) and evaluation metrics. The common problem setup is as follows:

- **Queries** are provided by the organizers and derived from the documents in particular WikiProjects.
- **Documents** are Wikipedia articles that may or may not be relevant to any particular WikiProject that is represented by a query.
- **Rankings** should be ranked lists of articles for editors to consider working on.
- **Fairness** of exposure should be achieved with respect to the protected attributes associated with the documents. Documents can be associated to many different fairness attributes. The official track evaluation will focus on intersectional fairness and, as such, will evaluate how fairly systems rank documents with respect to all of the fairness categories. However, individual teams can choose whether to optimise their systems with respect to all, a subset of, or individual fairness categories.

2.1 Task 1: WikiProject Coordinators

The first task is focused on WikiProject coordinators as users of the search system; their goal is to search for relevant articles and produce a ranked list of articles needing work that editors can then consult when looking for work to do.

Output: The output for this task is a **single ranking per query**, consisting of **500 articles**.

Evaluation will be a multi-objective assessment of rankings by the following two criteria:

- Relevance to a WikiProject topic. We will provide relevance assessments for the articles derived from existing Wikipedia data; Ranking relevance will be computed with nDCG, using binary relevance and logarithmic decay.
- Fairness with respect to the exposure of different fairness categories associated to the articles returned in response to a query.

We will use attention-weighted rank fairness [5] to measure the fairness of each ranked list. This compares cumulative exposure $\boldsymbol{\epsilon}$ across groups with a *population estimator* $\hat{\mathbf{p}}$ reflecting the target distribution; the system is more fair if the cumulative group exposure is close to the target distribution. This yields the following metric for a result list L:

$$\text{AWRF}(L) = \Delta(\boldsymbol{\epsilon}(L), \hat{\mathbf{p}}) \tag{1}$$

We will use the same logarithmic decay for both nDCG and AWRF (where k is the 1-based rank position):

$$\text{disc}(k) = \frac{1}{\log_2 \max(k, 2)} \tag{2}$$

$$\epsilon_G(L) \propto \sum_{k \in G} \text{disc}(k) \tag{3}$$

To compare the exposure distributions with a metric with the same range and similar interpretation as nDCG, we will use the one minus the Jenson-Shannon divergence:

$$\Delta(P_1, P_2) = 1 - \frac{1}{2}\left(D_{\mathrm{KL}}(P_1|M) + D_{\mathrm{KL}}(P_2|M)\right) \tag{4}$$

$$M = \frac{1}{2}(P_1 + P_2) \tag{5}$$

The final metric will be the product of AWRF and nDCG:

$$\mathrm{Metric}_1 = \mathrm{nDCG}(L) \cdot \mathrm{AWRF}(L) \tag{6}$$

This has the effect of requiring a system to do well on both relevance and fairness simultaneously in order to score well on the overall task.

2.2 Task 2: Wikipedia Editors

The second task is focused on individual Wikipedia editors looking for work associated with a project. The conceptual model is that rather than maintaining a fixed work list as in Task 1, a WikiProject coordinator would create a saved search, and when an editor looks for work they re-run the search. This means that different editors may receive different rankings for the same query, and differences in these rankings may be leveraged for providing fairness.

Output: The output of this task is **100 rankings per query**, each consisting of **20 articles**.

Evaluation will be a multi-objective assessment of rankings by the following three criteria:

- Relevance to a WikiProject topic. We will provide relevance assessments for articles derived from existing Wikipedia data. Ranking relevance will be computed with nDCG.
- Work needed on the article (articles needing more work preferred). We provide the output of an article quality assessment tool for each article in the corpus; for the purposes of this track, we assume lower-quality articles need more work.
- Fairness with respect to the exposure of different fairness categories associated to the articles returned in response to a query.

This task will use *expected exposure* to compare the exposure articles receive in result rankings to the *ideal* (or *target*) *exposure* they would receive based on their relevance and work-needed [1]. This addresses fundamental limits in the ability to provide fair exposure in a single ranking by examining the exposure over multiple rankings.

Given a query q, a ranking policy will provide a distribution π_q over rankings, the set of all (truncated) permutations of documents. We consider the 100 rankings to be samples from this distribution. Note that this is how we interpret the queries, but it does not mean that a stochastic policy is how the system must be implemented — other implementation designs are certainly possible. The objective is to provide comparable exposure to the fairness categories of documents with comparable relevance and work-needed; to operationalize this, we define an an ideal policy τ.

These policies are then used with a browsing model $\eta : \mathcal{L} \to \mathbb{R}^n$ to compute the per-query system exposure $\epsilon_q = \mathrm{E}_{\pi_q}[\eta]$ and target exposure $\epsilon^*_q = \mathrm{E}_{\tau_q}[\eta]$; for consistency with the Task 1 metric, we will use base-2 logarithmic weighting for the exposure function.

To address a weakness in previous years' metrics in which an algorithm could achieve good performance by exposing articles from the correct group, regardless of their relevance, we will do the following

1. Normalize system and target exposure to be distributions (each sums to 1).
2. Relate system exposure to target exposure on a *page-by-page* basis.
3. Only consider *underexposure*, where the system exposure is less than the target exposure; when system and target exposure are normalized to be distributions, under-exposure on one page will correspond to over-exposure on another.
4. Aggregate underexposure by group, to compute the *total underexposure* each group experiences.
5. Measure the L_2 norm of the groupwise underexposure. This metric is 0 only when there is no underexposure; between two systems with the same underexposure, the one that distributes that underexposure most equally between groups will have the lowest L_2 norm.

The resulting metric we call *equity of expected under-exposure*. We will also compute and report EE-L, EE-R, and EE-D from previous years, but equity of under-exposure will be the primary metric for comparing systems. Code to compute these metrics will be provided with the training queries.

3 Data

This section provides details of the format of the test collection, topics and ground truth.

3.1 Obtaining the Data

The data set is distributed via Globus, and can be obtained in two ways. First, it can be obtained via Globus, from our repository at `https://boi.st/3rIyniA`. From this site, you can log in using your institution's Globus account or your own Google account, and synchronize it to your local Globus install or download it with Globus Connect Personal[2]. This method has robust support for restarting downloads and dealing with intermittent connections. Second, it can be downloaded directly via HTTP from:
`https://data.boisestate.edu/library/Ekstrand/TRECFairRanking/`.

3.2 Corpus

The corpus, stored in `corpus`, consists of articles from English Wikipedia. We have removed all redirect articles, but have left the wikitext (markup Wikipedia uses to describe formatting) intact. This is available in three different formats, each provided as a JSON file, with one record per line, and compressed with gzip (e.g. `trec_corpus_20220301_plain.json.gz`).

Each record contains the following fields:

id: The unique numeric Wikipedia article identifier.

title: The article title.

url: The article URL, to comply with Wikipedia licensing attribution requirements.

The three available formats of the corpus are as follows:

text: The full article text, with Wiki markup (`text` file only)

plain: The full article text, without Wiki markup (`plain` file only)

html: The full article text, rendered into HTML (`html` file only)

[2] `https://www.globus.org/globus-connect-personal`

The contents of this corpus are prepared in accordance with, and licensed under, the CC BY-SA 3.0 license.[3] The raw Wikipedia dump files used to produce this corpus are available in the `source` directory; this is primarily for archival purposes, because Wikipedia does not publish dumps indefinitely.

Also included with the data under a `source` folder are various SQL dumps that contain additional extracted metadata related to the articles—e.g., categories, links. You can ignore these files but they are included for longevity to support alternative approaches for representing the Wikipedia articles.

3.3 Queries

The queries are in `2022`, in the file `train_topics_meta.jsonl`. Each of the queries map to a single Wikiproject. The queries are constructed from extracted keywords from articles that are relevant to a Wikiproject. The following fields are provided:

id A query identifier (int)

title The Wikiproject title (string)

keywords A collection of search keywords forming the query text (list of str). We cleaned and parsed the Wiki articles and then used KeyBert [2] to extract the most representative words of those articles. For each Wikiproject, we aggregated the extracted keywords from relevant articles and, after some manual filtering, used those as query texts for that particular Wikiproject.

homepage The URL for the Wikiproject. This is provided for attribution and not expected to be used by your system as it will not be present in the evaluation data (string)

rel_docs A list of the page IDs of relevant pages (list of int)

The keywords are the primary query text. The scope is there to provide some additional context and potentially support techniques for refining system queries.

In addition to query relevance, for Task 2: Wikipedia Editors (Section 2.2), participants will also be expected to return relevant documents that need more editing work done more highly than relevant documents that need less work done.

3.4 Fairness Categories

Fairness ground truth labels for the following fairness categories are also in the `2022` directory, in the `trec_2022_articles_discrete.json.gz` file. While we provide the raw values for each fairness category with the data, for most categories we also map the raw values to a reduced, fixed set of categories that will be used to judging systems.[4]

Geographic location (article topic) The geographical location associated with the article topic. Both the associated countries—e.g., United Kingdom—and sub-continental regions—e.g., Northern Europe—are provided but systems will be evaluated using sub-continental regions (and not countries). An article can have 0 to many regions associated with it.

Geographic location (article sources) The geographic location associated with the article based on the article's sources. Same categories as article geographic location above.

Gender (biographies only) The gender of the individual about which the biography pertains. Gender has been reduced to four distinct categories: Man, Woman, Non-binary, and Unknown (missing data or not a biography).

[3] `https://creativecommons.org/licenses/by-sa/3.0/`

[4] For more information, see: `https://public.paws.wmcloud.org/User:Isaac_(WMF)/TREC/TREC_2022_Data.ipynb`

Age of the topic How old the subject of the article is. For example, the birth date of a person in a biographical article, the date that an event occurred for articles that are about an event, or the creation date of a piece of art or music when the article is about the piece of art or music. The raw years are mapped to four distinct categories: Unknown, Pre-1900s, 20th century, and 21st century.

Occupation (biographies only) The occupation of the subject of an article. An article have 0 (unknown) to many occupations associated with it. There are 32 distinct occupation categories included in the data.

Alphabetical Editors often work through articles in alphabetical order and this can result in articles about subjects / topics that start with letters that appear earlier in the alphabet getting more exposure to the editors. Therefore, it is important that articles from later in the alphabet also get a fair exposure to the editors. The first letter is mapped to four discrete categories: a-d, e-k, l-r, and s-.

Age of the article The length of time the article has existed. The date is mapped to one of four discrete categories: 2001-2006, 2007-2011, 2012-2016, and 2017-2022.

Popularity (# pageviews) Number of times the page was viewed in February 2022. The number of pageviews are normalized and mapped to four discrete categories: Low, Medium-Low, Medium-High, and High.

Replication of articles in other languages The number of other language Wikipedias that the article is replicated in. This can range from English-only to all 300+ languages of Wikipedia but is mapped to three discrete categories: English only, 2-4 languages, and 5+ languages.

3.5 Metadata

We provide a simple Wikimedia quality score (a float between 0 and 1 where 0 is no content on the page and 1 is high quality) for optimizing for work-needed in Task 2. Work-needed can be operationalized as the reverse—i.e. 1 minus this quality score. The discretized quality scores will be used as work-needed for final system evaluation.

This data is provided together in a metadata file (`trec_metadata.json.gz`), in which each line is the metadata for one article represented as a JSON record with the following keys:

page_id Unique page identifier (int)

quality_score Continuous measure of article quality with 0 representing low quality and 1 representing high quality (float in range $[0, 1]$)

quality_score_disc Discrete quality score in which the quality score is mapped to six ordinal categories from low to high: Stub, Start, C, B, GA, FA (string)

Group Alignments The group alignments associated to an article as described in Section 3.4.

3.6 Output

For **Task 1**, participants should output results in rank order in a tab-separated file with two columns:

id The query ID for the topic

page_id ID for the recommended article

For **Task 2**, this file should have 3 columns, to account for repeated rankings per query:

id Query ID

rep_number Repeat Number (1-100)

page_id ID for the recommended article

4 Submission Instructions

TBA with eval data.

References

[1] F. Diaz, B. Mitra, M. D. Ekstrand, A. J. Biega, and B. Carterette. Evaluating stochastic rankings with expected exposure. In *Proc. CIKM '20*, 2020. URL `https://arxiv.org/abs/2004.13157`.

[2] M. Grootendorst. Keybert: Minimal keyword extraction with bert., 2020. URL `https://doi.org/10.5281/zenodo.4461265`.

[3] D. Pedreshi, S. Ruggieri, and F. Turini. Discrimination-aware data mining. In *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 560–568, 2008.

[4] M. Redi, M. Gerlach, I. Johnson, J. Morgan, and L. Zia. A taxonomy of knowledge gaps for wikimedia projects (first draft). *arXiv preprint arXiv:2008.12314*, 2020.

[5] P. Sapiezynski, W. Zeng, R. E Robertson, A. Mislove, and C. Wilson. Quantifying the impact of user attentionon fair group representation in ranked lists. In *Companion Proceedings of The 2019 World Wide Web Conference*, pages 553–562, 2019.