TREC 2021 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

Morten Warncke-Wang
mwang@wikimedia.org

May 5, 2021

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 ethnicity, 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 2021 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].

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 zero to many fairness categories.

There are two tasks in the 2021 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:

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

- **Queries** are provided by the organizers and derived from the topics of existing or hypothetical 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 **geographic location** of the articles (we will provide geographic location annotations). For the evaluation topics, in addition to geographic fairness, to the extent that biographical articles are relevant to the topic, the rankings should also be fair with respect to an undisclosed **demographic attribute** of the people that the biographies cover.

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 other editors can then consult when looking for work to do.

Output: The output for this task is a **single ranking per query**, consisting of **1000 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 articles for the training queries derived from existing Wikipedia data; evaluation query relevance will be assessed by NIST assessors. Ranking relevance will be computed with nDCG, using binary relevance and logarithmic decay.
- Fairness with respect to the exposure of different fairness categories in 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 ϵ 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:

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

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:

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

$$M = \frac{1}{2}(P_1 + P_2) \tag{3}$$

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

$$\mathrm{Metric}_1 = \mathrm{nDCG}(L) \cdot \mathrm{AWRF}(L) \tag{4}$$

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 **50 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 for the training queries derived from existing Wikipedia data; evaluation query relevance will be assessed by NIST assessors. 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 in the articles returned in response to a query.

This task will use *expected exposure* to compare the exposure article subjects receive in result rankings to the *ideal* (or *target*) *exposure* they would receive based on their relevance and work-needed [2]. 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 documents of comparable relevance and work-needed; to operationalize this, we define an an ideal policy τ.

These policies are then used with a browsing model $\eta : \mathcal{L} \to R^n$ to compute the per-query system exposure $\epsilon_q = \mathrm{E}_{\pi_q}[\eta]$ and target exposure $\epsilon_q^* = \mathrm{E}_{\tau_q}[\eta]$ (Diaz et al. [1] define two ways of computing these; we tentatively plan to use the ERR-based formula). The final metric is the *expected exposure loss*, the squared Euclidean distance between system and target expected exposure:

$$\mathrm{EEL}_q = \|\epsilon_q - \epsilon_q^*\|_2^2 \tag{5}$$

$$= \|\epsilon_q\|_2^2 - 2\epsilon_q^{\mathrm{T}}\epsilon_q^* + \|\epsilon_q^*\|_2^2 \tag{6}$$

This metric decomposes into two sub-metrics that can be used to assess fairness/relevance tradeoffs: the *expected exposure disparity* $EED = \|\epsilon\|_2^2$ and the *expected exposure relevance* $EER = 2\epsilon^{\mathrm{T}}\epsilon^*$. Our primary evaluation metric will be group-aggregated EEL, to measure *group fairness*; for comparison, we will also measure individual fairness for documents or document authors, and report EED and EER to characterize fairness/relevance tradeoffs. Since topical relevance and work-needed are integrated into expected exposure loss through the target policy, we will not be combining EEL with separate relevance metrics; systems will be ranked on their EEL.

3 Data

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

3.1 Corpus

The 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 provided as a JSON file, with one record per line, and compressed with gzip (`trec_corpus.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.

text The full article text.

The contents of this corpus are prepared in accordance with, and licensed under, the CC BY-SA 3.0 license[2].

3.2 Topics

Each of the track's training topics is based on a single Wikiproject. The topic is also GZIP-compressed JSON lines (file `trec_topics.json.gz`), with each record containing:

id A query identifier (int)

title The Wikiproject title (string)

keywords A collection of search keywords forming the query text (list of str)

scope A textual description of the project scope, from its project page (string)

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 topical 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.

3.3 Metadata and Fairness Categories

For training data, participants will be provided with a geographical fairness ground truth. For the evaluation data, submitted systems will be evaluated on how fair their rankings are to the geographical fairness category and an undisclosed personal demographic attribute.

We also 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)

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

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)

geographic_locations Continents that are associated with the article topic. Zero or many of: Africa, Antarctica, Asia, Europe, Latin America and the Caribbean, Northern America, Oceania (list of string)

3.4 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.

5 Limitations

The data and metrics in this task address a few specific types of unfairness, and do so partially. This is fundamentally true of any fairness intervention, and does not in any way diminish the value of the effort — it is impossible for any data set, task definition, or metric to fully capture fairness in a universal way, and all data and analyses have limitations.

We will provide a fuller accounting of known limitations in the notebook paper, but some include:

- For each Wikipedia article, we ascertain which, if any, continents are relevant to the content. This is determined by directly looking up several community-maintained (Wikidata) structured data statements about the article. These properties are checked for the presence of countries, which are then mapped to continents via the United Nation's geoscheme.[3] While this data must meet Wikidata's verifiability guidelines,[4] it does suffer from varying levels of incompleteness. For example, only 73% of people on Wikidata have a country of citizenship property.[5] Furthermore, structured data is itself limited—e.g., country of citizenship does not appropriately capture people who are considered stateless though these people may have many strong ties to a country. It is not easy to evaluate whether this data is missing at random or biased against certain regions of the world. Care should be taken when interpreting the absence of associated continents in the data, though evaluations for TREC will be based solely on the known values. Further details can be found in the code repository.[6]

[3] `https://en.wikipedia.org/wiki/United_Nations_geoscheme`
[4] `https://www.wikidata.org/wiki/Wikidata:Verifiability`
[5] `https://humaniki.wmcloud.org/gender-by-country`
[6] `https://github.com/geohci/wiki-region-groundtruth`

- Our proxy for work-needed is a coarse proxy. It is based on just a few simple features (page length, sections, images, and references) and does not reflect the nuances of the work needed to craft a top-quality Wikipedia article. A fully-fledged system for supporting Wikiprojects would also include a more nuanced approach to understanding the work needed for each article and how to appropriately allocate this work.
- The task is limited to topics for which Wikipedia already has articles. These tasks are not able to counteract biases in the processes by which articles come to exist (or are deleted) — recommending articles that should exist but don't is an interesting area for future study.

References

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

[2] 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`.

[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.