TREC Fair Ranking Track
Participant Instructions

June 2, 2019

For 2019, we will be adopting 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 hope for the systems to be robust to these. As such, participants are expected to develop systems to optimize for fairness and relevance for arbitrary group definitions. We will not reveal the exact group definitions until *after* the evaluation runs are submitted.

The track is set up as a *reranking* task. The track organizers will provide a sequence of queries, each accompanied by a varying-size set of documents; the task is to rerank the documents to produce result lists that are fair and relevant.

1 Protocol

For our fair ranking evaluation, we will be providing participants with a sequence $\mathcal{Q}$ of queries accompanied by unordered sets of documents to rank. The document sets are of varying size. For each request (query q and set of documents $\mathcal{D}_q$), participants should provide a ranked list of the documents from $\mathcal{D}_q$. The final system output is a sequence of rankings. Algorithm 1 presents a pseudocode of the evaluation protocol.

The rankings produced in response to queries in the sequence should balance 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 Π

2 Evaluation

Unlike previous TREC tracks, you will receive multiple copies of the same query text–although the query id will be different–and you may submit different rankings for each instance of the query. At evaluation time, we will measure *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 will be evaluating systems according to fair exposure of authors (Section 2.1.1) and relevance of documents (Section 2.2).

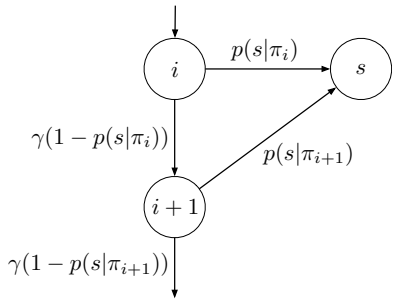

Figure 1: Attention model.

2.1 Measuring Fairness

2.1.1 Measuring Author Exposure for a Single Ranking

In order to measure exposure, we adopt the browsing model underlying the Expected Reciprocal Rank metric [1]. Given a ranking π, the exposure of author a is,

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

n	number of documents in ranking π
$\mathcal{D}_a$	documents including a as an author
π_i	document at position i
γ	continuation probability (fixed to 0 for the final position in the ranking)
$p(s\|d)$	probability of stopping given user examined d

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

We will provide the value of the discounting factor γ, and will assume $p(s|d) = f(r_d)$, 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,

$$e_a = \sum_{\pi \in \Pi} e_a^\pi \tag{1}$$

where Π is the sequence of all system rankings.

2.1.2 Measuring Author Relevance for a Single Ranking

The author relevance for a ranking π is defined as,

$$r_a^\pi = \sum_{d \in \mathcal{D}_a} p(s|d) \tag{2}$$

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.

2.1.3 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,

$$\mathcal{E}_g = \frac{\sum_{a \in \mathcal{A}_g} e_a}{\sum_{g' \in \mathcal{G}} \sum_{a \in \mathcal{A}_{g'}} e_a} \tag{3}$$

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

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

$$\Delta_g = |\mathcal{E}_g - \mathcal{R}_g| \tag{5}$$

Given this relevance-normalized measure of exposure, we can compute the fair exposure using the Gini coefficient,

$$\Delta = \frac{\sum_{g,g' \in \mathcal{G}} |\Delta_g - \Delta_{g'}|}{2|\mathcal{G}| \sum_{g \in \mathcal{G}} \Delta_g} \tag{6}$$

2.2 Measuring Relevance

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

$$u^\pi = \sum_{i=1}^{n} \left[\gamma^{i-1} \prod_{j=1}^{i-1} (1 - p(s|\pi_j)) \right] p(s|\pi_i) \tag{7}$$

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

2.3 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. As such, we will measure the trade-offs between fairness to producers and quality for consumers as an auxiliary metric.

3 Data

3.1 Input

There are three main inputs available to you: the *corpus* of articles to search, the *example group definition* file to help you develop and test your solution, and the *queries*.

3.1.1 Corpus

The corpus for this project is 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)

3.1.2 Example Group Definition Data

To help you get started, we have provided the file `fair-TREC-sample-author-groups.csv` containing group ids for authors in the S2 corpus. This group definition will not be our final group definition (we will evaluate using multiple various group definitions), but you can use it to start working on the task.

This CSV file contains two columns:

1. The `author` column has the S2 ID of the author.
2. The `gid` column has the author's group identifier.

3.1.3 Queries

The training queries are in `fair-TREC-training-sample.json`; this is a JSON file where each line is a JSON object (a dictionary) describing one query:

1. Query ID ('qid')
2. Query string ('query')
3. Query frequency ('frequency')
4. A list of documents ('documents') with relevance information; this is a list of dictionaries with two keys: 'doc_id' and 'relevance'

Runs will be submitted over *query sequences*: ordered sequences of queries that may contain duplicates. We will provide the set of query sequences for official runs 1 month prior to the deadline; for training and development, use the provided script (`query-sequence-generator.py`) to generate query sequences from the training queries distributions.

A query sequence is a CSV file with the following fields:

1. Query number (in sequence, 1–N)
2. Query ID (to look up in query file)

The evaluation query sequences will be accompanied by a query file that has no relevance information and has all query frequencies set to 0, to enable you to only look up the queries and document sets to rerank.

3.2 Output

For each query sequence, your submitted ranking will be a JSON file where each line is a JSON object (a dictionary) containing your ranking results:

- Query number (position in query sequence) ('q_num')
- Query ID (from the query lookup file) ('qid')
- An ordered list of document IDs (of the documents to be reranked for the query) ('ranking')

References

[1] O. Chapelle, D. Metzler, Y. Zhang, and P. Grinspan. Expected reciprocal rank for graded relevance. In *Proceedings of the 18th ACM conference on Information and knowledge management*, CIKM '09, pages 621–630, New York, NY, USA, 2009. ACM.