Towards Fairness-Aware Ranking by Defining Latent Groups Using Inferred Features

Yunhe Feng[1], Daniel Saelid[1], Ke Li[1], Ruoyuan Gao[2], Chirag Shah[1]
{yunhe,saeliddp,kel28,chirags}@uw.edu ruoyuan.gao@rutgers.edu

[1]University of Washington [2]Rutgers University

Abstract. Group fairness in search and recommendation is drawing increasing attention in recent years. This paper explores how to define latent groups, which cannot be determined by self-contained features but must be inferred from external data sources, for fairness-aware ranking. In particular, taking the Semantic Scholar dataset released in TREC 2020 Fairness Ranking Track as a case study, we infer and extract multiple fairness related dimensions of author identity including gender and location to construct groups. Furthermore, we propose a fairness-aware re-ranking algorithm incorporating both weighted relevance and diversity of returned items for given queries. Our experimental results demonstrate that different combinations of relative weights assigned to relevance, gender, and location groups perform as expected.

Keywords: Fair Ranking · Text Retrieval · Fair Exposure · Information Retrieval · Fairness · Ranking

1 Introduction

As one of the emerging topics in fairness-aware information systems, presenting relevant results to the users while ensuring fair exposure of the content suppliers have raised more and more attention. Fairer information retrieval and search systems not only provide relevant search results with higher diversity and transparency, but also offer reasonable discoverability for underrepresented groups. For example, a high-quality academic paper from small institutions, which have very limited media outlets and resources, should also be treated equally to get its deserved exposures in search systems, especially at the early stage of publication when such papers are more likely to suffer from cold-start problems.

This paper investigates fairness ranking within an academic search task context, where the goal was to provide fair exposure of different groups of authors while maintaining good relevance of the ranked papers regarding given queries. However, it is difficult to achieve such a goal due to the following challenges.

– **Openness and complexity of defining the author group.** Defining the author group is not a trivial task. This requires an in-depth understanding of what should be considered as important group attributes that not only separate different authors but also aggregate similar authors. The challenges

in this task include and are not limited to, how many groups should be identified, and how to identify and extract the features from authors and their publications for the group classification task.
- **Algorithm Robustness on different applications.** The definition of author groups may change from application to application. A good fairness ranking algorithm should be robust to a broad range of group definitions in various scenarios. In other words, fairness-aware ranking algorithms should demonstrate a high generalization capability when processing application-wise group definitions.
- **Trade-off between relevance and fairness.** The re-ranking algorithm based on a list of candidate items needs to optimize for both the relevance of the re-ranked results and the fairness of the exposed author groups, while carefully balancing between the two.

We aimed to design and implement fair ranking and retrieval algorithms to enhance the fairness for scholarly search. On the subset of the Semantic Scholar (S2) Open Corpus [1] provided by the Allen Institute for Artificial Intelligence, we defined multiple author groups, inferred demographic characteristics of authors, and developed fairness-aware algorithms to achieve a flexible trade-off between relevance and fairness by tuning principal component weights.

2 Data Description

The Semantic Scholar (S2) Open Corpus released by TREC 2020 Fairness Ranking Track consists of extracted fields of academic papers. For most papers, the available fields include the S2 paper ID, title, abstract, authors, inbound and outbound citations. In addition, another three auxiliary datasets are provided. The first dataset maps paper ids to a list of corresponding author positions with their corpus id. The second one contains paper information such as paper id, title, year of publication, venue, number of citations, and number of key citations. The last one contains author features including author's name, number of citations, h-index (and a dependent feature, h-class), i10-Index, and number of papers published.

3 Methodology

We first defined author groups based on general demographic characteristics including genders and countries. Then, we utilized Okapi BM25 [5] to estimate the relevance of papers for given search queries. Based on the group definition and BM25 relevance score, we proposed our fairness-aware re-ranking algorithm.

3.1 Group Definition

When defining author groups, we considered genders and countries of authors because the two demographic features are general enough for different applications. Re-ranking algorithms based on such group definitions are more likely to demonstrate strong robustness in various scenarios.

Gender Inference To predict the binary gender of a given author, we called the genderize.io API [2], which is powered by a large dataset that maps first names to binary genders. Given a name, genderize.io will return 'male' if there are more instances of the name associated with men, and it will return 'female' otherwise. If the dataset contains no instances of the given name, no gender prediction will be returned. For the authors in our sub-corpus, the returned gender predictions are shown in Table 1.

Table 1. The distribution of inferred genders by genderize.io

Gender	Count	Percentage
Male	18810	58.8%
Female	6235	19.5%
Unidentified	6930	21.7%
Total	31975	100%

Table 2. The economy distribution of inferred locations

Locations	Count	Percentage
Advanced	15106	47.2%
Developing	3926	12.3%
Unidentified	12933	40.5%
Total	31975	100%

Country Inference In contrast with gender prediction, we could not rely on a single API call for location prediction. To begin the process, we searched for the author by name in Google Scholar using the Scholarly API [3]. Since there are often many authors with a given full name on Google Scholar, we picked a single author by comparing our author citation data with Google Scholar's data. After choosing the closest match, we retrieved email extension and 'affiliation' data from Google Scholar. If we successfully retrieved this author data, we followed the below procedure, moving to each consecutive step if the prior was unsuccessful. As listed as the last step, if no author data was retrieved from Google Scholar, we tried finding the author's homepage and parsing its URL for country code.

1. Parse the email extension for a country code (e.g. .uk → United Kingdom).
2. Parse the affiliation for a university name, then return the country in which that university is located.[1]
3. Parse the affiliation for a city name, then return that city's country.[2]
4. Search author name, author affiliation on Google, scrape the first URL, then parse for country code.
5. Call Google Places API with affiliations, then return associated countries.
6. Search author name + 'homepage' on Google, scrape the first URL, then parse for country code.

Once all authors had been processed, we mapped each author's affiliated country to 'advanced economy' or 'developing economy' based on the IMF's October 2019 World Economic Outlook report [4]. The results are shown in Table 2. Here, 'unidentified' means that no country was predicted for that author.

[1] `https://www.4icu.org/reviews/index0001.htm`

[2] `https://en.wikipedia.org/wiki/List_of_towns_and_cities_with_100,000_or_more_inhabitants/cityname:_A`

3.2 Pure Relevance with BM25

We used Okapi BM25, a popular ranking algorithm adopted by many search engines, to estimate the relevance of a document based on a given query. Since complete paper contents are unavailable, we instead chose the paper's abstract and title to represent the corresponding document. The papers were written in 28 different languages including English, Arabian, German, Chinese, etc., while all queries were in English only. However, BM25 functions are incompatible with certain languages that cannot be tokenized by whitespace. Therefore, we decided to translate all needed documents into English first and stored the tokenized text in the database for further usage.

Then we started the BM25 process. We first translated and tokenized the queries since some of them contained Unicode. After that, for each query, we calculated the BM25 score as the base relevance score for each document, and then arranged the documents based on their scores in descending order. This sorted list was used as the pure ranking list for the given query.

3.3 Fairness-aware Re-ranking Algorithm

We proposed a fairness-aware re-ranking algorithm incorporating both relevance and diversity of documents. The main idea was to estimate the cost of adding a document to the rank list $\mathbf{R}$ from the perspective of relevance and fairness. For a document of d, we used $F(d, \mathcal{D}, q)$, the reversed normalized BM25 score of d in a corpus $\mathcal{D}$ given a query q, to represent its relevance cost, where 0 corresponds to most relevant, and 1 corresponds to least relevant.

For a given query q, we first retrieved the top relevant documents to build a candidate corpus $\mathcal{D}'$. To ensure ranking fairness, it is intuitive to make the probability of defined groups over the rank list R and the candidate corpus $\mathcal{D}'$ very similar. Specifically, let $p(v, \mathcal{D})$ be the probability distribution of a discrete group variable v over the the document corpus $\mathcal{D}$. Based on our group definitions, v could be either the group of `gender` g or `country` c, i.e., $v \in \{g, c\}$. Note that this is flexible to be extended to other group definitions. Then we use the Kullback-Leibler (KL) divergence of the group distribution probability between the updated current rank list $\mathbf{R}$ and the whole candidate corpus $\mathcal{D}'$ to measure their similarities. We also assigned weights $\mathbf{w}$ for relevance cost and fairness cost for each defined group. The cost function is expressed as:

$$C(d, \mathbf{w}, \mathbf{R}, \mathcal{D}', q) = w_r * F(d, \mathcal{D}', q) + \sum_{v \in \{g,c\}} w_v * KL(p(v, \mathbf{R}+\{d\}) \parallel p(v, \mathcal{D}')) \quad (1)$$

where $\mathbf{w} = \{w_r, w_g, w_c\}$ and $w_r + w_g + w_c = 1$; $F(d, \mathcal{D}', q)$ is the reversed normalized BM25 score of a document d such that 0 corresponds to most relevant, and 1 corresponds to least relevant; and $KL(p(v, \mathbf{R} + \{d\}) \parallel p(v, \mathcal{D}'))$ is the Kullback-Leibler divergence regarding group v between the updated $\mathbf{R}$ by appending document d and the overall candidate corpus $\mathcal{D}'$. Then, we built

our re-ranked list by repeatedly appending the document with the minimal cost $C(d, \mathbf{w}, \mathbf{R}, \mathcal{D}', q)$. The proposed fairness-aware re-ranking algorithm as illustrated in Algorithm 1.

Since many documents were missing group definitions for at least one author, we adopted a systematic way to address it. For every author missing a group definition, we assigned a group value based on the overall group distribution in the corpus. For instance, if 75% of the authors in the corpus were identified as male, we choose 'male' for an unidentified author with a probability of 75%.

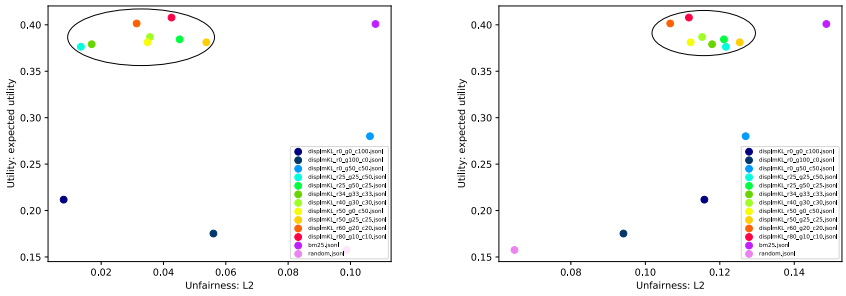
Algorithm 1: Fairness-aware Re-ranking Algorithm

Input: $\mathcal{D}$: document corpus; q: query of interest; l: length of expected ranked list ; $\mathbf{w}$: component weight vector
Output: $\mathbf{R}$: re-ranked list of relevant documents
$\mathbf{R} \leftarrow \emptyset$; // initialize the ranked list as empty
$\mathcal{D}', \mathcal{D}'' \leftarrow$ Retrieve relevant document candidates from $\mathcal{D}$ for query q ;
// document candidate corpus for q
for $i = 1 \rightarrow l$ **do**
  $c_{min} \leftarrow A\ Large\ Integer$; // initialize the minimal cost
  $d_{min} \leftarrow None$; // initialize the document with the minimal cost
  for $d \in \mathcal{D}''$ **do**
    Calculate the cost $C(d, \mathbf{w}, \mathbf{R}, \mathcal{D}', q)$ according to Equation 1 ;
    // calculate the cost of adding d into $\mathbf{R}$
    if $C(d, \mathbf{w}, \mathbf{R}, \mathcal{D}', q) < c_{min}$ **then**
      $d_{min} \leftarrow d$; // update the document with the minimal cost
      $c_{min} \leftarrow C(d, \mathbf{w}, \mathbf{R}, \mathcal{D}', q)$; // update the minimal cost
    end
  end
  append d_{min} to $\mathbf{R}$; // add the document with the minimal cost into the re-ranked list $\mathbf{R}$
  $\mathcal{D}'' \leftarrow \mathcal{D}'' - \{d_{min}\}$; // remove the added document d_{min} from $\mathcal{D}''$
end
return R

4 Results and Discussion

We evaluated the utility and unfairness with different combinations of w_r, w_g, w_c in Equation 1 from the perspective of relevance, the group of gender, and the group of country, as shown in Figure 1. In both gender and country groups, BM25 demonstrates a relatively high utility score but a low fairness score, implying that BM25 fails to take fairness into account when calculating the ranking. Another interesting finding is that the random ranking achieves lower fairness than most of our proposed methods on the country group but the highest fairness on the gender group. So, the fairness performance of random ranking methods is sensitive to the definition of groups. In other words, the definition of groups is not a trivial task as we claimed in Section 1. As we expected, our methods' utility drops greatly when BM25 scores are excluded ($w_r = 0$). When w_r is assigned a positive value, the performance of our methods with different combinations of

w_r, w_g, w_c are comparable on both country and gender groups (see the cluster on left top in Figure 1(a), and the cluster on the middle top in Figure 1(b)).



(a) Country (b) Gender

Fig. 1. Utility versus unfairness.

5 Conclusion

This paper presents how to define latent groups using inferred features for fair ranking. Specifically, we construct gender and location groups, which are generalized but not contained in the raw dataset, to promote search result fairness. We also propose a fairness-aware retrieval and re-ranking algorithm incorporating both relevance and fairness for Semantic Scholar data. Evaluation results with different weights of relevance, gender, and location information demonstrated that our algorithm was flexible and explainable.

Acknowledgements

A part of this work is supported by the US National Science Foundation (NSF) award number IIS-1910154.

References

1. Ammar, W., Groeneveld, D., Bhagavatula, C., Beltagy, I., Crawford, M., Downey, D., Dunkelberger, J., Elgohary, A., Feldman, S., Ha, V., Kinney, R., Kohlmeier, S., Lo, K., Murray, T., Ooi, H.H., Peters, M., Power, J., Skjonsberg, S., Wang, L.L., Wilhelm, C., Yuan, Z., van Zuylen, M., Etzioni, O.: Construction of the literature graph in semantic scholar. In: NAACL (2018)
2. ApS, D.: genderize.io (2020), `https://genderize.io/`
3. Cholewiak, S.A., Ipeirotis, P., Revision, V.S.: scholarly: Simple access to Google Scholar authors and citations (2020), `https://pypi.org/project/scholarly/`
4. Dept., I.M.F.R.: World economic outlook. World Economic Outlook, INTERNATIONAL MONETARY FUND (2019). https://doi.org/http://dx.doi.org/10.5089/9781513508214.081
5. Robertson, S.E., Walker, S., Jones, S., Hancock-Beaulieu, M., Gatford, M.: Okapi at TREC-3 (1994), `https://fair-trec.github.io`