Information Retrieval Research 1 (2025) 247-280 Submitted 05/25; Revised 11/25; Published 12/25

An Eigensystem for Topic Term Weighting yields Fair and Effective Document Rankings

Massimo Melucci MASSIMO.MELUCCI@UNIPD.IT
Department of Information Engineering
University of Padova, 35131 Padova, Italy, EU

Editor: Johanne Trippas

Abstract

By incorporating fairness into the ranking function at retrieval time through topic term weights, this paper suggests a way to at the same time achieve effective and fair rankings in document collections. The topic term weights are calculated using the eigensystem of a matrix that linearly combines effectiveness and fairness matrices in a single matrix, whose main eigenvector provides the weights. The latter can then be utilized as a new topic representation, thus providing a fair ranking while minimizing the impact on effectiveness. The experiments described in the paper demonstrate that the proposed approach works.

Keywords: group fairness, probability ranking principle, eigensystem

1 Introduction

Many information management systems strive to suit the user's information needs by providing access to enormous data repositories and delivering optimal outcomes by organizing output data in several formats. Ranking is a common form used by recommender and Information Retrieval (IR) systems. Because of the introduction of normative conceptions such as fairness or equality in the late 2010s, all of these systems have struggled to balance both user satisfaction and producer or author visibility. Visibility is crucial for ensuring that diverse voices and perspectives are represented in the information ecosystem.

In IR, one finds that the principal task has traditionally been articulated in terms of relevance: retrieve those documents most likely to be relevant to a user's query or pertinent to a topic.[1] Much effort has been devoted to the construction of models that estimate, more or less precisely, this elusive probability of relevance. Within this framework, the representation of both documents and queries as weight distributions over terms is critical. To be specific, a document d is at retrieval time assigned a usually non-negative score measuring the likelihood that the IR system will retrieve it to respond to a given topic q. Let

$$y(q,d) = \sum_{t \in q \cap d} w(t,d)\kappa(t,q) \tag{1}$$

be the retrieval score given to the document to answer a topic, where $w(t,d)$ is a function capturing the association between term t and document, such as term frequency, probabilis-

1. In this paper, we also use "topic" because the test collections utilized in the experiments include topics rather than real queries.

©2025 held by the author(s) License: CC-BY 4.0 WWW: https://irrj.org DOI: 10.54195/irrj.23480

tic relevance score, or contextual embedding similarity, and $\kappa(t, q)$ has analogous meaning yet it is usually set to 1. Next, we rank the retrieved documents based on their score. Topic term weights are often set to a constant value, e.g., 1, because experimenters have no clue, nor do users provide information as regards the importance of topic terms.

One particular concern is that retrieval systems may systematically under-rank documents associated with certain categories, especially when these are underrepresented in data or associated with less frequent terms. We consider the problem of fair ranking—that is, ranking documents most likely to be relevant to a user's query or pertinent to a topic and fairly representative of the categories to which they refer. The problem bears some resemblance to the approaches for fairness in classification: one adjusts the influence of particular inputs to ensure parity across outputs. In IR, however, the problem is compounded by the necessity of preserving relevance.

The difficulty also recalls the remarks made by Robertson (1977) on the trade-off between optimal ranking and clustering. According to Robertson, ranking documents by decreasing their relevance to information contradicts the cluster hypothesis, which states that the best document clusters should be ranked first. This is because relevant documents tend to cluster together more than non-relevant documents. This clustering effect can lead to a situation where the most relevant documents are not necessarily the top-ranked ones, as they may be grouped with other relevant documents in the same cluster. As a result, this trade-off requires careful consideration when designing retrieval systems to balance the benefits of both ranking and clustering effectively.

To obtain a fair yet effective ranking, a suite of options is available to the designers of a system. One option is to directly permute document rankings and optimize a certain function of utility that balances both effectiveness and fairness. Another option is to indirectly obtain a new ranking by changing the term weights. It is in the manipulation of the term weight distributions that we encounter opportunities not only to improve retrieval effectiveness but also to simultaneously address broader concerns, notably that of fairness. Suppose we identify that certain terms—associated with marginalized topics, communities, or discourses—are chronically underweighted in a baseline model. By explicitly increasing the weights of such terms during re-ranking, we can enhance the visibility of documents pertaining to these underrepresented areas. A possible procedure may start from the ranking function (1) which can be expanded as:

$$y(q, d) = \sum_{t \in q \cap d} w(t, d) x(t, q) \tag{2}$$

where $x(t, q)$ is the weight assigned to the term. The crux here is that by using x, we can raise or lower the profile of particular topic terms, thus reshaping the final ranking also to the aim of making the ranking fairer than the initial ranking. As a consequence, the variation of a topic term weight should not only award the terms that promote relevant documents and demote non-relevant documents, it should also award the terms that promote underrepresented category documents and demote overrepresented category documents. The challenge is to balance promotions and demotions.

However, a naive reweighting that elevates fairness at the cost of user satisfaction is unlikely to be acceptable. Thus, the art lies in achieving a balance wherein gains in fairness do not substantially degrade effectiveness, or where modest losses in effectiveness are justified

by substantial fairness improvements. Such intervention, to be clear, need not be ad hoc; on the contrary, a principled approach was adopted in this work.

Group fairness is concerned with the balanced exposure of groups. In the event of IR, group fairness is concerned with the balanced exposure of the groups of document author's organization to the end users which interact with a system to retrieved documents and meet information needs. The requirement of group fairness derives from the hypothesis that minority organization groups are less likely exposed to the end user by means of the retrieved documents.

We considered group fairness and addressed the problem of fair ranking as follows. An initial list of documents most likely to be relevant to a user's query or pertinent to a topic is ranked by some baseline method. By assigning weights to individual terms according to their topical relevance, one may modulate the influence of different query or topic components on the final ranking. This method may, under suitable conditions, yield improved effectiveness. But our claim here is more ambitious: that topic term weighting, suitably construed, can serve also as an instrument of fairness. The topic term weights thus serve a dual role: guiding the model toward content representation and simultaneously adjusting for societal disparities encoded in the data.

2 Previous Work on Fairness and Information Retrieval

The problem of fairness in IR is novel in the sense that explicit discussions of fairness have only relatively recently entered the vocabulary of IR researchers, in part due to a broader societal concern with bias, accountability, and transparency in algorithmic systems, as witnessed by Baeza-Yates (2018). But it is not novel in the sense that any systems that provide ranked outputs have always made distinctions, highlighted some items and omitted others, and thereby implicitly favored some items over others, as investigated in result diversification; see McDonald et al. (2022); Chapelle et al. (2011). The question, then, is not whether IR systems are fair, but rather: how do they treat the concept of fairness, and ought they be doing more?

Actually, the classical IR framework was never designed with fairness in mind, at least not as a first-class concern. The objective was to retrieve documents that are relevant to a given query or pertinent to a given topic; see Van Rijsbergen (1979); Salton (1968). Relevance was at the center of the whole process; see Robertson et al. (1982). The evaluation framework, exemplified by Cranfield-style test collections, built upon this centrality of relevance: precision, recall, average precision, and later, NDCG, all serve to measure relevance.

But of course, as the scope of retrieval systems has expanded—from digital libraries to web search, from enterprise to social media—the stakes have changed. In recommending job ads, loan offers, or news articles, the system is not only reflecting some notion of relevance, which itself is complex as well explained by Saracevic (1975, 2007a,b) and by Mizzaro (1997), but it is also participating in a social process, one in which the distribution of attention, opportunity, and information may have significant consequences. A discussion was provided by Balagopalan et al. (2023) which applied concepts from social sciences, information retrieval, and machine learning to empirically investigate if breaches of criteria

for combining effectiveness and fairness had an influence, and if so, how to avoid such concerns.

Let us take an example. Suppose a job search platform retrieves and ranks positions for a user based on their profile, queries, topics, and past interactions. The system may optimize for click-through rate or predicted user satisfaction. But what if this optimization consistently results in showing high-paying technical roles to male users and lower-paying service jobs to female users, even when their qualifications are similar? Is the system "fair"? In the traditional sense of relevance, perhaps it is—if those results align with inferred user preferences. But from a broader societal view, the answer is more troubling.

This disjunction—between individual relevance and societal fairness—is at the heart of the fairness-in-IR discussion. It invites a number of further questions. Can we define fairness in retrieval systems in a rigorous way? If so, how might we measure it? And—most crucially—how might it trade off against other objectives, notably relevance?

Several proposals have emerged in recent literature. One line of work attempts to define group fairness in ranked outputs: ensuring that members of different groups (say, based on gender or ethnicity) are represented equitably in the top-k results; see Binns (2020); Jaenich et al. (2024); Melucci (2024a,b, 2025a,b); Morik et al. (2020); Sakai et al. (2023). Another strand focuses on individual fairness, the idea that similar individuals (in some feature space) should receive similar outcomes; see Biega et al. (2018); Binns (2020); Draws et al. (2021); Ilvento et al. (2020); Kletti et al. (2022); Xiao et al. (2017). Both of these concepts, while intuitive, face considerable practical challenges. Group fairness may require knowledge of sensitive attributes (which may not be available or legally permissible to use); individual fairness depends critically on the definition of similarity (which is itself a value-laden and context-sensitive choice).

In retrieval settings, the notion of fairness becomes further complicated by the ranked nature of outputs. Not only must one consider which items are retrieved, but also where they appear in the ranking. Users typically examine only the top few results—so the fairness implications of a system are concentrated in those positions. A group may be underrepresented although the documents that belong to the group are ranked not far from the top-k hits. As a result, measures of fairness must be sensitive to rank position, much as traditional effectiveness measures are.

Here, one might consider analogues of well-known metrics. For instance, Normalized Discounted Cumulative Gain (NDCG) gives more weight to items at higher ranks—see Jarvëlin and Kekäläinen (2002). One might define Discounted Cumulative Fairness, where the representation of groups in top ranks is weighted more heavily. But this, too, introduces tension. A system that seeks to maximize NDCG may produce rankings that are highly effective but group-imbalanced; a system that enforces strict fairness constraints may compromise on user satisfaction. This tension is real, and it cannot be wished away.

Some researchers have proposed re-ranking methods to address this: first, produce a ranking based on relevance; then, adjust it to improve fairness. Adjusting a ranking for fairness may require demoting highly relevant items, which can degrade performance—we adopted this approach in our previous work; see Melucci (2024a,b, 2025a,b). Alternatively, one may attempt to build fairness into the scoring function itself—this, however, raises difficult questions of how to balance relevance and fairness at the model level. It is indeed this approach and the difficulties thereof we addressed in this work. Another approach was

reported by Bigdeli et al. (2021), where the authors empirically investigated the tradeoff between fairness and effectiveness and saw if it is possible to increase fairness while preserving equivalent retrieval utility. They investigated if this is feasible by rewriting the input query using a bias-aware pseudo-relevance feedback system.

Recent work by Jaenich et al. (2025) proposes generative query reformulation to improve fairness without compromising relevance—an approach closely aligned with the goals of this paper. Instead of modeling fair reranking as this paper proposes, Jaenich et al. (2025) found that Large Language Models (LLMs) may provide a significant improvement in some cases yet at the expenses of high computational costs.

3 Are Test Collections Fair?

The focus on relevance has driven the design of experiments in IR and, in particular, the preparation of test collections. The test documents are harvested from various sources and assembled in a repository according to some categories and objectives. At the time of document harvest, a set of test topics is selected. The experiments consist of document retrieval, i.e., an IR system retrieves documents and selects those likely relevant to a topic. At retrieval time, the system ranks the retrieved documents by likelihood of relevance to topics. A test collection also includes relevance assessments, i.e., some users judge document relevance to topics.

The preparation and the use of a test collection can affect IR fairness in different ways. At harvest time, designers and experimenters may affect the distribution of documents, topics, and relevance assessments because of data availability, source authoritativeness, technological reasons, language barriers, or cultural bias. At retrieval time, a system may affect the distribution of the retrieved documents because of stopword removal, stemming, or distribution of document and topic terms. When ranking documents, a system may affect the distribution of ranked documents because of fairness-unaware term weighting. The relevance assessments of a test collection may also affect the distribution of ranked documents by categories because of the assessors' bias towards unprotected categories and against protected categories.

Consider a topic q and a document subset X. In this paper, we define the score of X with respect to q as

$$y(q, X) = \sum_{d \in X} y(q, d) \ .$$

The score of X measures the degree to which a user will find relevant information from the documents in X if s/he examines all the documents. In other terms, the score of X is a measure provided by the system of the "quantity" of relevance found in X. Following the foundations of probability theory, the probability of X with respect to a topic is defined as

$$\Pr(X) = \frac{y(q, X)}{\sum_d y(q, d)} \ .$$

Let Y be another document subset. In this paper, we define

$$\Pr(X, Y) = \Pr(X \cap Y) \qquad \Pr(X|Y) = \frac{\Pr(X, Y)}{\Pr(Y)} \ .$$

Let $\Pr(C|X)$ be a measure of category exposure within X and can therefore be used to measure fairness. Let $\mathcal{C}$ be a set of categories.

The Gini index of mutability

$$G(\mathcal{C}|X) = 1 - \sum_{C \in \mathcal{C}} \Pr(C|X)^2$$

measures fairness for each distribution conditioned to X. Gini's measure of mutability was reported more than a century ago in (Gini, 1912, pages 142–144), but no English translation has been available until Ceriani and Verme (2024) published "an abridged and commented translation of this second part of" Gini's book. The translation is crucial since mutability is commonly misunderstood with variability and the former is often ignored. We believe it is beneficial to deliver a quick presentation.

When a category, such as gender, is observed in a group of people, an unordered list of category values is eventually produced. An unordered series cannot have extreme values, nor can median or average values be derived; this is why variance and quantiles cannot be calculated. Nonetheless, an unordered series can have its relative frequency estimated.

Let X be a set of individuals, and $\mathcal{C} = C_1, ..., C_m$ be a collection of categories such that $C_k \subseteq X$ and $C_\ell \cap C_k = \emptyset$. Let $f_k = |C_k|/|X|$, where $n = |X|$. When an individual belongs to C_k, one deviation $1 - f_k/n$ from the relative frequency and m deviations $f_\ell/n, \ell = 1, ..., k-1, k+1, ..., m$ from the other relative frequencies are observed. The sum of the variances caused by witnessing the k-th value in one individual is

$$1 - f_k/n + f_1/n + \cdots + f_{k-1}/n + f_{k+1}/n + \cdots + f_m/n = 2\frac{n - f_k}{n} \ .$$

Because the f_k individuals display the k-th value, the overall deviation will be $2 f_k \frac{n-f_k}{n}$. After seeing the attribute in each individual, the overall deviation is

$$\frac{2}{n} \sum_{k=1}^{m} f_k (n - f_k) \ .$$

Because each individual is considered twice, once for negative deviations and once for positive deviations, the average deviation can be written as

$$\begin{aligned} G(\mathcal{C}|X) &= \frac{1}{n^2} \sum_{k=1}^{m} f_k (n - f_k) \\ &= 1 - \sum_{k=1}^{m} \left(\frac{f_k}{n} \right)^2 \ . \end{aligned} \tag{3}$$

Equation (3) is Gini's index of mutability for the set X with respect to the collection of categories $\mathcal{C}$; see the original book by Gini (1912) as well as Ceriani and Verme (2024)'s article for a commented translation.

The test collections from the Text Retrieval Conference (TREC) Fair tracks in 2021 and 2022 were used to measure the extent to which fairness can be affected; see Ekstrand et al. (2022a, 2023) and Section 5. At TREC, evaluation takes place in two steps, i.e., training and

Track	Category	Group	$\Pr(C)$	Track	Category	Group	$\Pr(C)$
2021	gender	Female	0.056	2022	locations	S. Africa	0.003
2021	gender	Male	0.239	2022	locations	W. Africa	0.002
2021	gender	Third	0.000	2022	locations	C. America	0.005
2021	gender	Unknown	0.706	2022	locations	N. America	0.228
2021	locations	Africa	0.022	2022	locations	S. America	0.008
2021	locations	Antartica	0.002	2021	locations	Antartica	0.000
2021	locations	Asia	0.099	2022	locations	C. Asia	0.001
2021	locations	Europe	0.212	2022	locations	E. Asia	0.018
2021	locations	LAC	0.031	2022	locations	S. Asia	0.021
2021	locations	N. America	0.187	2022	locations	W. Asia	0.010
2021	locations	Oceania	0.026	2022	locations	S.-E. Asia	0.009
2021	locations	Unknown	0.422	2022	locations	Melanesia	0.000
2022	gender	Man	0.231	2022	locations	Micronesia	0.000
2022	gender	Non-binary	0.000	2022	locations	Caribbean	0.003
2022	gender	Woman	0.055	2022	locations	E. Europe	0.023
2022	gender	Unknown	0.714	2022	locations	N. Europe	0.114
2022	locations	ANZ	0.032	2022	locations	S. Europe	0.021
2022	locations	E. Africa	0.002	2022	locations	W. Europe	0.062
2022	locations	N. Africa	0.001	2022	locations	UNK	0.258
continue in the next column				2022	locations	Unknown	0.179

Table 1: This is the initial distribution of the test document across categories and groups. This distribution is the result of harvesting the data included in the test collection. It is worth noting that the proportions differ by one or even two orders of magnitude. LAC refers to Latin America and the Caribbean, and ANZ to Australia and New Zealand.

testing. At training time, the participants design and train their own system configurations by using the full, past data. At testing time, the participants submit the runs by using a new set of topics without knowledge of the relevance judgments, though.

The test collections show that *the distribution of documents appears as biased towards males, Europe, and Northern America (NA).* Let C be a subset of documents of a certain category. Document harvest creates the initial distribution $\Pr(C)$ reported in Table 1. Besides the fact that the test collections are biased, the order of magnitude of $\Pr(C)$ of males, NA, and European location documents influences the subsequent phases. Note that the unknown locations and genders take a quite significant portion of the collection. In principle, they should not be considered as a category because the unknown category documents cannot contribute to the calculation of the index of mutability. The values of the Gini index reported in this paper consider the proportion of unknonws—the purified values of the proportion of unknowns can be obtained with the following expression:

$$1 - \frac{\left(\sum_{C \in \mathcal{C} \setminus \{U\}} \Pr(C|X)^2\right)^2}{(1 - \Pr(C|U))^2} - \Pr(C|U)^2 \qquad \Pr(C|U) < 1$$

where U is the subset of items of the unknown group.

After document retrieval, *the distribution of retrieved documents remains as biased towards males, Europe, and NA*, thus implying that the state-of-the-art search algorithms reproduce the bias observed after collections are harvested. Let B be a subset of documents retrieved to match q. Retrieval decomposes $\Pr(C)$ into the $\Pr(C|B)$'s in Table 2. Retrieval creates document sets in which the distribution by category differs a little from the distribution $\Pr(C)$. In particular, the exposure of Africa and unknown groups has been made a little

Track	Phase	Category	Group	Pr($C\|B$)
2021	train	locations	Africa	0.017
2021	train	locations	Asia	0.087
2021	train	locations	Europe	0.09
2021	train	locations	LAC	0.023
2021	train	locations	N. America	0.192
2021	train	locations	Oceania	0.024
2021	train	locations	Unknown	0.567
2021	test	gender	Female	0.005
2021	test	gender	Male	0.017
2021	test	gender	Unknown	0.978
2021	test	locations	Africa	0.003
2021	test	locations	Asia	0.045
2021	test	locations	Europe	0.134
2021	test	locations	LAC	0.004
2021	test	locations	N. America	0.162
2021	test	locations	Oceania	0.024
2021	test	locations	Unknown	0.627
2022	train	gender	Man	0.065
2022	train	gender	Unknown	0.926
2022	train	gender	Woman	0.01
2022	train	locations	ANZ	0.028
2022	train	locations	Caribbean	0.003
2022	train	locations	C. America	0.002
2022	train	locations	C. Asia	0.001
2022	train	locations	E. Africa	0.0
2022	train	locations	E. Asia	0.022
2022	train	locations	E. Europe	0.02
2022	train	locations	Melanesia	0.0
2022	train	locations	Micronesia	0.0
2022	train	locations	M. Africa	0.0
2022	train	locations	N. Africa	0.0
2022	train	locations	N. America	0.281
2022	train	locations	N. Europe	0.12
2022	train	locations	Polynesia	0.0
continue in the next column				

Track	Phase	Category	Group	Pr($C\|B$)
2022	train	locations	S. America	0.004
2022	train	locations	S.-E. Asia	0.009
2022	train	locations	S. Africa	0.002
2022	train	locations	S. Asia	0.012
2022	train	locations	S. Europe	0.013
2022	train	locations	UNK	0.26
2022	train	locations	Unknown	0.155
2022	train	locations	W. Africa	0.0
2022	train	locations	W. Asia	0.009
2022	train	locations	W. Europe	0.058
2022	test	gender	Man	0.025
2022	test	gender	Unknown	0.97
2022	test	gender	Woman	0.004
2022	test	locations	ANZ	0.027
2022	test	locations	Caribbean	0.003
2022	test	locations	C. America	0.002
2022	test	locations	C. Asia	0.001
2022	test	locations	E. Africa	0.001
2022	test	locations	E. Asia	0.016
2022	test	locations	E. Europe	0.011
2022	test	locations	Melanesia	0.0
2022	test	locations	Micronesia	0.0
2022	test	locations	N. Africa	0.0
2022	test	locations	N. America	0.227
2022	test	locations	N. Europe	0.109
2022	test	locations	S. America	0.003
2022	test	locations	S.-E. Asia	0.01
2022	test	locations	S. Africa	0.002
2022	test	locations	S. Asia	0.008
2022	test	locations	S. Europe	0.01
2022	test	locations	UNK	0.22
2022	test	locations	Unknown	0.302
2022	test	locations	W. Africa	0.001
2022	test	locations	W. Asia	0.009
2022	test	locations	W. Europe	0.038

Table 2: This is the breakdown of retrieved documents by category and group. Each topic's top 1,000 ranked documents were retrieved. Because some topic sets lack another field to extract topic terms, the documents were retrieved to answer questions based solely on the topic title. The sum of document scores was calculated by category. Because the top 1,000 ranked documents produce the highest scores, the reported proportions can be regarded as a reliable estimate of the proportion obtained if the sum of all document scores were calculated. LAC refers to Latin America and the Caribbean, and ANZ to Australia and New Zealand.

larger than before retrieval against the exposure of the other categories. However, males, NA, and Europe still remain the majority because of the prior probability determined at collection harvest time. Note that Table 2 makes a distinction between "train" and "test" because retrieval took place at two different phases during the TREC Fair tracks. Table 3 summarizes Table 2 by using the Gini index.

After document ranking, *the distribution of retrieved documents remains as biased, and the Gini index of mutability does not increase or significantly decrease.* Let D be a subset of documents retrieved at certain ranks, such as the retrieved documents ranked between

Track	Phase	Category	$G(C\|B)$
2021	train	locations	0.729
2021	test	gender	0.065
2021	test	locations	0.653
2022	train	gender	0.207
2022	train	locations	0.845
2022	test	gender	0.088
2022	test	locations	0.832

Table 3: This table reports the Gini index of mutability with respect to each category measured for the retrieved documents from each subcollection.

the first and the tenth position of the first retrieved page. [2] Ranking allocates the retrieved documents into ranks at which the distributions $\Pr(C|B, D)$ by category and pages may differ from $\Pr(C|B)$. Figure 3 of Appendix A depicts the Gini indexes of the 100-document rankings segmented in 10-document pages such that page p includes the retrieved documents ranked from $10(p-1)+1$ to $10p$. Ten-document pages are common in many applications.

After relevance assessment, *the trade-off between fairness and effectiveness arises*. Let A be a subset of documents relevant to q. Assessment decomposes $\Pr(C|B, D)$ into the $\Pr(C|A, B, D)$'s. Figure 4 of Appendix A depicts the proportion of relevant documents that are ranked in each page, the latter being called "precision at page p" whereas Figure 3 of Appendix A depicts the Gini index. The heatmaps depicting the distribution of accuracy across the 10 document pages reveal that the top pages have the highest precision values and the lowest Gini index of mutability. However, the trade-off between accuracy and Gini index is less pronounced in the location category than in the gender category. Note that, the heatmap was necessary because of δ, which is absent from the recovery effectiveness measurement. Regarding the latter, a one-dimensional heatmap was used instead of other chart types because the x-axis represents the page, just like with equity heatmaps.

As a summary, a document page on average includes three to four NA documents, three European documents, one to two Asian documents, and the remaining documents are from the other locations. In parallel, five or six documents refer to males, whereas other genders are basically absent from the ranking pages. Ranking confirms the unbalanced distribution of documents by categories; for example, the top ten ranked documents on average are four North American documents, three European documents, and the remaining three are distributed over the other four locations. After assessment, the likelihood of relevance results as unbalanced over the locations. NA and Europe are the most likely locations of relevant documents in the top-ranking pages. The retrieval phase little reduces the imbalance against some unprotected categories.

4 Eigensystem for Topic Term Weighting

Given a document collection and the indices thereof, a way to achieve fair rankings is allowing the system to re-rank the retrieved documents. Re-ranking retrieved documents has the downside of reducing effectiveness, since the top-ranked relevant documents might be moved away from the top ranks. An alternative approach—which is adopted in this

2. In IR, Search Engine Retrieved Page (SERP) means the dynamic webpage listing the links to the retrieved documents.

paper—is to incorporate fairness in the ranking function at the retrieval time. The way is paved by the topic term weights.

The selection of the topic terms and the weights thereof can, in theory, impact ranking and, as a consequence, category exposure. The topic terms' impact on ranking by means of the number of terms, the document term weights, and the topic term weights. The number of terms and the document term weights depend on the test collection and on the weighting scheme adopted by a retrieval system—therefore, they can hardly be changed to accomplish the goal of fair ranking. A simple example, which might even seem trivial, arises when the documents of a privileged category are longer than those of a disadvantaged category. Since term weights increase with the length of the documents in which they appear, those from the privileged category will be shown before those from the disadvantaged category.

However, a system could in principle adjust topic term weights if some information about category exposure and distribution were available—in this paper, the topic term weights are automatically adjusted to promote the retrieved documents of unprotected groups by minimizing the impact on effectiveness. The retrieval document score can be utilized to this end in combination with category-document distribution, since document scores are sums of products between document term weights and the corresponding topic term weights. Accordingly, the document term weights remain unchanged, whereas the topic term weights can be adjusted to incorporate some information about category exposure, thus making protected categories exposed as unprotected ones.

Topic term weights can be adjusted to change the score of the documents matching the topic term. In particular, the weight of a topic term can be increased if the term occurs in documents of protected categories more frequently than in documents of unprotected categories. The scores of protected documents can thus be increased to take fairness into account. If topic term weights were adjusted to take fairness into account, the ranking would also be adjusted to increase the exposure of the documents of unprotected categories.

for all topic q **do**
  $e \leftarrow 1^k$ `-- unary representation of` q
  $y_e \leftarrow Be$ `-- rank documents matching` e
  $B \leftarrow$ term-document occurrence matrix
  $Q \leftarrow$ equation (4)
  $x \leftarrow$ main eigenvector of $B'B - B'QB$
  $y \leftarrow Bx$ `-- rank documents matching` x
  $C \leftarrow$ retrieved document-group matrix
  compute fairness and effectiveness measures
end for

Figure 1: The computational approach of Eigensystem for Topic Term Weighting (ET) is summarized in this pseudo code.

ET aims to maximize both fairness and effectiveness by automatically reweighting topic terms and re-ranking documents that are retrieved to answer the topic of which terms are equally weighted. Consider Eq. (2): ET finds the function called x in a principled way. To this end, a k-dimensional vector space over the real field represents k topic terms. The main eigenvectors of a symmetric matrix provide an alternative k-dimensional vector to

represent topics. Moreover, the main eigenvectors of the symmetric matrix maximize a function measuring effectiveness and fairness. The function takes a quadratic form whose matrix is the parameter. The computational approach is summarized in Figure 1.

In mathematical terms, consider $n \in \mathbb{N}$ retrieved documents, $k \in \mathbb{N}$ topic terms, $m \in \mathbb{N}, m > 1$ categories, and $r \in \mathbb{N}, r > 1$ relevance grades. Let $B \in \mathbb{R}^{n \times k}$ be the term-to-document matrix where a B's element is the weight of a term in a document, i.e., $w(t, d)$. Let $C \in \mathbb{R}^{m \times n}$ be the documents-to-categories matrix where an element is the degree of membership of a document to a category. Suppose the C's column is non-negative and sums to 1. Let CB be the terms-to-categories matrix, which is in $\mathbb{R}^{m \times k}$. Let

$$x \in \mathbb{R}^{k \times 1}$$

be the topic term weight vector such that

$$\|x\| = 1 \ .$$

Given a topic and a set of retrieved documents, let Bx be the retrieved document scores, which are in $\mathbb{R}^{n \times 1}$. Let

$$y = Bx$$

be the retrieved document normalized scores. Without loss of generality, y can be assumed as non-negative. Using L_1 normalization, let

$$y = Bx \,/\, 1'Bx$$

where $1 \in \{1\}^{n \times 1}$. Therefore,

$$y \geq 0 \qquad 1'y = 1 \ .$$

Using L_1 normalization, suppose the rows of C are normalized. The sum of scores of the retrieved documents of each category is given by

$$p = Cy$$

where $p \in [0, 1]^{m \times 1}$ and sums to 1, i.e., p is a probability distribution, since the C's columns are non-negative and sum to 1 as y does. The Gini index of mutability becomes

$$G = 1 - \sum_{j=1}^{m} p_j^2 \ .$$

The index ranges between 0 and $1 - 1/m$. Let's replace p with Cy, and let c_j be the j-row of C:

$$p_j^2 = (c_j y)(c_j y) = (c_j y)'(c_j y) = (y' c_j')(c_j y) = y' c_j' c_j y \ .$$

Therefore,

$$\sum_{j=1}^{m} p_j^2 = \sum_{j=1}^{m} y' c_j' c_j y = y' \left(\sum_{j=1}^{m} c_j' c_j \right) y = y' Q y$$

where

$$Q = \sum_{j=1}^{m} c_j' c_j \tag{4}$$

As $y = Bx$, the Gini index becomes a function of x defined as

$$G(x) = 1 - x'B'QBx \ .$$

The latter is a symmetric quadratic form since $B'QB$ is symmetric. Moreover, G is a continuous function in a closed interval; therefore, a maximal point does exist according to Weierstrass' theorem. The maximum of G is achieved when x is one of the main eigenvectors of $I - B'QB$. As a consequence, a ranking of n retrieved documents exhibits the highest fairness if the topic terms are weighted according to x and then is obtained by Bx. Let

$$F = y'y$$

be the sum of squared retrieved document scores. As $y = Bx$,

$$F = F(x) \qquad F(x) = x'B'Bx$$

can be viewed as a measure of the likelihood that an n-document ranking represents relevant information to the topic. The main eigenvector of $B'B$ determines the n-document ranking for which F is maximum. If y is a measure of probability of relevance of the n items, the maximal value of F is a necessary condition of the Probability Ranking Principle (PRP), which has been used, in one form or another, by various people since M. E. Maron and J. L. Kuhns. W. S. Cooper gave a formal statement of the principle as reported by S. E. Robertson; see Robertson (1977).

> If a reference retrieval system's response to each request is a ranking of the documents in the collections in order of decreasing probability of usefulness to the user who submitted the request, where the probabilities are estimated as accurately as possible on the basis of whatever data has been made available to the system for this purpose, then the overall effectiveness of the system to its users will be the best that is obtainable on the basis of that data.

Apart from the ranking by non-decreasing probability of usefulness or relevance, the optimal probability estimation is one key concept of the principle. The estimation of the probability that the i-th item is relevant or useful is provided by (1) or (2), depending on whether the topic term weights are subject to estimation or not, respectively.

Whereas $w(t, d)$ can be referred to as relevance and then to effectiveness, $x(t, q)$ can be leveraged to integrate fairness and effectiveness when ranking documents. The potential provided by x is then what the PRP requires to make a ranking optimal, since x can integrate w, thus providing a—hopefully—better estimation of the probability of usefulness to the user.

The objective of a fair ranking is to maximize both effectiveness and fairness; therefore, $F + G$ should be maximized, i.e., both F and G should be. As both F and G are functions of x, then $F + G$ is also a function of x, i.e.,

$$F + G = (F + G)(x) \ .$$

Therefore,

$$(F + G)(x) = F(x) + G(x) = x'B'Bx + 1 - x'B'QBx = x'(B'B + I - B'QB)x \ .$$

It follows that the maximal point of $F + G$ is one of the main eigenvectors of

$$B'B + I - B'QB \ .$$

Therefore, finding the maximum point of $(F + G)(x)$ means finding the maximum point of $x'(B'B + I - B'QB)x$ where I is the $k \times k$ identity matrix. The latter is a quadratic form in the operator $B'B + I - B'QB$ whose maximum point is one out of the main eigenvectors. For any z such that $z'z = 1$, we have that

$$z'(B'B + I - B'QB)z = z'(B'B - B'QB)z + 1 \ .$$

The latter reaches its maximum value if z is one out of the main eigenvectors called Eigensystem for Topic Term Weightings (ETs) of $B'B - B'QB$.

Actually, any vector of the subspace spanned by the ETs are maximal points of $F + G$. Indeed, a quadratic form may in general have more than one maximal point corresponding to the ETs, since multiple eigenvectors may share the same largest eigenvalue. In particular, there might be more than one ETs because $B'B - B'QB$ might not be primitive, and one cannot leverage Perron and Frobenius' theorem stating that a primitive matrix has one main eigenvector; see Seneta (2006).

The multiplicity of the ETs is of little importance, though, since one eigenvector out of them is sufficient to maximize $F + G$. Nevertheless, the ETs are still mutually orthogonal, and therefore they represent alternative topic term reweighting schemes. Finally, note that, as the scale of the eigenvalues of $B'B$ may largely differ from those of $B'QB$, the matrices $B'B$ and $B'QB$ should be divided by the respective main eigenvalue.

5 Experiments

5.1 Data and Measures

The test collections from the TREC Fair tracks in 2021 and 2022 were used to evaluate the algorithms. The experimental data were obtained from Wikimedia[3]; see Ekstrand et al. (2022a, 2023). The public availability of the experimental collections makes the utilization for an empirical study such as that reported in this section acceptable despite the unavoidable limitations of any dataset, as explained by Ekstrand et al. (2023). Apart from the experimental documents, the test collections also comprise a series of experimental inquiries, the metadata (information about category membership), and evaluations of the documents' pertinence. Table 4 reports on the test collections' main statistics.

	2021	2022
Number of documents	6,280,328	6,475,401
Average document size (bytes)	6,748	3,197
Number of train topics	47	46
Number of test topics	22	20
Average number of train topic words	5.38	61.3
Average number of test topic words	2.72	3.35

Table 4: The main statistics of the test collections

3. https://en.wikipedia.org/wiki/WikiProject

The 2021 and 2022 Fair Ranking Tracks were an ad hoc retrieval methodology where participants were given a collection of documents and a series of topics to rank documents based on relevance to a specific WikiProject.[4] The challenges shared a topic set, corpus, fundamental problem structure, and fairness aim. The topics were prepared from WikiProject subjects, documents being English Wikipedia articles, rankings being lists of articles that editors may consider relevant to the topics—NIST assessors annotated the retrieved documents with binary relevance scores for given topics—and articles being fairly exposed based on geographical location and gender.

As for the 2021 track, the categories were defined by gender and geographic locations, the latter being used for training. Within TREC, "training" refers to the experimental phase for tuning the participating systems and precedes "test," which is the actual measurement of the effectiveness of the systems for the reference year, test data, and gender, which was only used for the test data.

The 2022 Fair Ranking Track required the participants to deal with a larger number of categories other than geographic location and gender. As for the 2022 track, the categories were defined by geographic locations, gender, age of the topic, occupation, alphabetical order of name, age of the article, page popularity, and replication of articles in other languages as detailed by Ekstrand et al. (2022b). In order to compare the tracks, the experiments were only carried out with geographic locations and genders for the 2022 track as well.

The experiments were carried out with the following input:

- the experimental topics completed with the relevance assessments,
- the experimental document identifiers completed with category memberships, i.e.:
  - gender-based category with male, female, or non-binary, also known as "third";
  - geographical location-based with different continents or subcontinent regions;
- the experimental full-text documents.
- the document retrieval scores returned by an experimental search system to compute A on the basis of both: (Elasticsearch v 7.17.1 and the standard configuration thereof were utilized.)
  - either title-only topics or full topics, and
  - either title-only document or full document;

For each topic, the list of relevant documents was retrieved from the test collection. A term-based representation of the topic was built by concatenating the words found in the topic fields, such as title and keywords. If there were at least two topic terms, a list of retrieved documents was ranked by (1). The $n = 100$ documents retrieved for each topic were ranked by Fielded Best Match 25 (BM25F); as said, ten-document pages are common in many applications, and Figure 3 of Appendix A can still be readable if ten pages are observed. Each document was scored by

$$\sum_{t \in d \cap q} \text{BM25F}(t, d)$$

4. https://en.wikipedia.org/wiki/WikiProject

where

$$\mathrm{BM25F}_d = \sum_{\ell=1}^{L} \sum_{t=1}^{k} \mathrm{IDF}_t \mathrm{SAT}_{d,t}$$

and

$$\mathrm{IDF}_t = \log 1 + \frac{N - n_t + \frac{1}{2}}{n_t + \frac{1}{2}} \qquad \mathrm{SAT}_{d,t} = \frac{f_{d,t}(k_1 + 1)}{f_{d,t} + k_1 \left(1 - b + b \frac{s_{d,\ell}}{s_\ell}\right)}$$

and ℓ refers to a field such as title or abstract, $f_{d,t}$ is the frequency of t in d, n_t is the number of documents indexed by t, $s_{d,\ell}$ is the size of ℓ of d, $k_1 = 1.2$ and $b = 0.75$ are hyperparameters, IDF stands for Inverse Document Frequency, SAT stands for "saturation," and BM25F stands for Fielded Best Match 25. Note that the topic term weight is assumed to be 1, thus not rescaling BM25F.

In parallel, an efficient implementation of B and C was obtained by in-memory, sparse matrices. The ranking obtained for the topic was browsed, and the Compressed Sparse Row (CSR)-based representation of each matrix was allocated. Using the CSR-based representation, the ℓ-th matrix non-zero entry v of row i and column j was allocated into three array entries as follows:

$$\mathrm{value}[\ell] = v, \mathrm{row}[\ell] = i, \mathrm{column}[\ell] = j \ .$$

In this way, the CSR-based matrices for A, B, and C were allocated where $n = 100$, k topic terms, and m categories.

The eigensystem was then computed, thus obtaining an $\mathrm{ET}(q)$ of a given topic. To be specific, using the ad-hoc subroutines for the CSR-based representations, the products $R = B'B$ and $S = B'QB$ were computed. Both R and S were divided by the respective main eigenvalue to make their scale neutral. Each document was then rescored by

$$\sum_{t \in d \cap q} \mathrm{BM25F}(t, d) \mathrm{ET}(t, q) \ .$$

where $\mathrm{ET}(t, q)$ is the eigenvector component of a term of the topic that determined the ranking. Note that the topic term weight is $\mathrm{ET}(t, q)$, thus rescaling BM25F according to the ET.

For each $\delta \in \mathbb{R}$ selected from a predefined set of values that was decided at the experimentation time, the eigensystem of $Q = R - \delta S$ was computed. The Q's main eigenvector, x, was multiplied by B, thus obtaining the n-dimensional vector $y = Bx$ of document scores recomputed by using non-unitary topic term weights. In this way, this ranking was compared with the ranking obtained by Be, where $e \in 1^k$ is the n-dimensional vector y_e of document scores recomputed by using unitary topic term weights. Note that the entries of the main eigenvector of Q might be non-positive. As a consequence, the y might not be non-negative. To work around the issue of negativity of Q, G was obtained as follows: First, y was translated to $y + \min y$ and then L_1-normalized, thus obtaining a non-negative, normalized retrieval score vector. Then, Q was computed.

The following measures were computed for each topic, i.e., for each baseline ranking and the re-rankings thereof: [5]

5. A re-ranking is a ranking of documents rescored by an ET.

- Gini's index of mutability (G),
- Precision at 10-document page (P), and
- a Global Measure, i.e., $M = G \times P$.

The latter function is non-decreasing for both P and G—it aims at measuring the extent to which reranking documents may not decrease if not even increase effectiveness or fairness. The sum of G and P is not appropriate because they are derived from different measurement units and phenomena. Instead, there are some justifications for multiplying different quantities, such as those used in Macroeconomics. Other fairness measures are occasionally used, such as Attention-Weighted Rank Fairness (AWRF); see Raj and Ekstrand (2022). Other measures of effectiveness, such as NDCG and Average Precision (AP), are also available. The former was defined for non-binary relevance grades, which are not applicable in our case. The latter considers relevant document ranks—we used it in the previous work; see Melucci (2024a,b, 2025a,b). We preferred to investigate the same issues from a different perspective and then use different measures. AWRF assumes an attention model, but we wanted to avoid additional assumptions and take an assumption-free approach. A comparison, albeit indirect, can be made by observing the results in Melucci (2025b).

5.2 Main Results

Overall, ET can yield fairer and more effective rankings than the baseline depending on δ and on category. Figures 2 and 5(a) of Appendix A depict the baseline values of M and the values of M for each δ and page and both tracks as a whole. As a summary, moderate to high delta values (1.00–1.75) yield the most improvement in the global measure under gender bias correction. Moreover, peak gains are concentrated around the middle-ranking pages (5–7). Finally, too high or too low delta settings tend to be less effective, especially on the tails of the ranking. As for location, bias mitigation for location yields strong and consistent improvements, especially from delta 1.25 onward. In addition, the ranking system appears to benefit more clearly from fairness adjustments for location than for gender, likely due to higher initial disparity. Pages 5 and 6 are again the sweet spot for fairness or utility gains. In order to understand what may affect the global measure, the results have been reported by measure, i.e., G and P. Both measures increase as δ does, and in particular the mid-ranked pages show the largest values. Both measures cannot be increased at the top-ranked pages, where effectiveness and fairness remain relatively low. Figures 5(b) and 5(c) of Appendix A depict the values of G, P for each δ and page and both tracks as a whole.

Figures 6 and 7 of Appendix A show that *ET can improve fairness to varying degrees in most cases.* The variations of the Gini index of mutability clearly depend on the fairness, i.e., bias against some groups caused by the harvest and retrieval phases. This bias can lead to unequal outcomes, which the ET method seeks to address by redistributing weights more equitably. By analyzing the Gini index, we can better understand the impact of these adjustments on fairness across different demographic groups. For example, gender appears to be the most biased category, and as a result, the little variations of G achieved by ET should be considered as quite large. In contrast, categories with relatively high Gini index values can become more equitable than the baseline to varying degrees. This

track	phase	category	δ	training M	testing M
2021	test	gender	1.25	0.046	0.039
2021	test	locations	0.0	0.466	0.453
2021	train	locations	0.25	0.461	0.458
2022	test	gender	1.25	0.156	0.100
2022	test	locations	1.5	0.467	0.465
2022	train	gender	1.25	0.322	0.287
2022	train	locations	2.0	0.528	0.524

Table 5: The maximum δ for each track, phase, and category was determined for the training topic set and used to calculate $M = G \times P$ for the test set. Except in one case, using $\delta > 1$ enhances effectiveness and fairness.

shift towards greater fairness can lead to improved outcomes for underrepresented groups, fostering inclusivity and fairness in decision-making processes. By concentrating on these adjustments, a more equitable environment that serves the interests of all parties involved can be established.

As shown in Figures 6 and 8 of Appendix A, *ET may decrease efficacy for most cases*, even for only the top pages, whereas effectiveness can enhance for the mid or bottom pages. Heatmaps illustrate that precision at top pages decreases as δ grows. As δ grows, the decrease in precision slows or even becomes an increment when browsing from top to bottom pages. This shows that, while the top pages may lose precision, the middle and bottom pages may benefit from changes to the browsing parameters. As a result, this variety emphasizes the necessity for personalized solutions that improve user experience across various levels of content. This necessity for personalized solutions highlights the importance of adaptive algorithms that can cater to individual user preferences and behaviors.

Considering known groups only slightly changes the overall results. In Section 3, we discussed the significant proportion of papers whose group is known and hypothesized that this proportion may affect the metrics of fairness. The issue is similar to the lack of relevance judgments, making it difficult to develop effective relevance feedback algorithms in IR. Despite a significant number of unknown-group documents, we found that the overall efficacy of ET is barely affected. Figure 9 of Appendix A depicts heatmaps of the results obtained by omitting these documents. The consistent performance of ET suggests that the presence of unknown-group documents does not significantly compromise the fairness metrics we are investigating. Therefore, the conclusion should not change. This finding demonstrates the robustness of our approach in handling varying levels of group knowledge within the dataset.

Topic term weighting based on ET and $\delta > 0$ is a good decision in terms of global measure, to some extent depending on the track, phase, and category. Cross-validation has been used to determine the best *delta* for each track, phase, and category. To accomplish this, the topic set was divided into training and test sets using the $3/4, 1/4$ rule. The training set included 75% of the topics, while the test set included the remaining 25%. The training set was selected at random and repeated twenty times. The fairness and effectiveness measures, G and P, were calculated for each page and δ. The global measure was then computed and averaged across the pages to produce a global measure for ranking purposes. Table 5 shows that δ should be greater than 1 to achieve the best performance in terms of M, except for one case.

	Fairness: 2021 train locations										Effectiveness: 2021 train locations									
δ	1	2	3	4	5	6	7	8	9	10	1	2	3	4	5	6	7	8	9	10
.25	.02		.01			.04		.04	.03	.02	.0	.0							.01	.03
.50	.04		.04		.04				.02		.0									
.75					.04			.02	.04	.01	.0	.01							.03	
1.00	.02				.05				.0		.0		.04		.02			.04		
1.25									.01	.0	.0				.04					
1.50	.04		.03								.0	.01			.01				.02	
1.75	.02							.02	.0	.0	.0	.01			.0				.01	
2.00	.02		.04					.0	.0	.0	.0	.0			.01			.03	.01	

Table 6: The table reports the p-values for each δ and page for the year 2021, the training phase, and the category "geographical location". The blank cells indicate non-significance. If a cell is not blank then the number is the p-value observed–the lower the p-value the more significant the variation of either G or P caused by δ with respect to the baseline for the given track, phase, and category.

The darker areas of the heatmaps do *not always correspond to significant variations in fairness or effectiveness*. To provide a measure of the statistical significance of the results, the Wilcoxon signed-rank test was conducted, and only the cases where the p-value was less than or equal to 0.05 were considered significant. The Gini index and precision were observed for each track, phase, category, topic, and ten-document page for $\delta = 0$ and then for each $\delta > 0$ as mentioned previously. So, a page of ten documents obtained with $\delta = 0$ was compared to the corresponding one with a certain $\delta > 0$. See Table 6.

The original bias against a category may impact on the significance of the variations and as a consequence fairness can hardly be improved if the initial conditions as rather unfavourable. The following table shows that the significant variations are less frequently significant with gender than in the case of locations:

	Fairness: 2021 test gender									
δ	1	2	3	4	5	6	7	8	9	10
0.25			0.04		0.04					
0.50										
0.75			0.04		0.03					
1.00					0.02					
1.25					0.02					
1.50										
1.75										
2.00										

	Effectiveness: 2021 test gender									
δ	1	2	3	4	5	6	7	8	9	10
0.25										
0.5										
0.75										
1.00	0.0						0.02			
1.25	0.0					0.05	0.05		0.02	
1.5	0.0						0.05		0.02	
1.75	0.0						0.05		0.02	
2.0	0.0						0.05		0.02	

As for the 2021 test set and the location category and the 2022 sets, a very few differences were statistically significant as reported in Appendix B.

The experimental results presented in this section are expressed using two distinct measures of effectiveness and fairness, namely G and P, which are then summarized into an overall measure, $M = G \times P$. This is not the only possible suite of measures. In conclusion to this section, the results obtained with the measurement proposed recently by Sakai et al. (2023) are presented in order to provide a comprehensive overview and, hopefully, make

them comparable with other experiments. In the work by Sakai et al. (2023), a measure called the Global Fairness and Relevance (GFR) is introduced, defined as follows, and which is reproduced in this work with the same notation as the authors of this measure:

$$\text{GFR}(L) = w_0\text{RELEVANCE}(L) + w_1\text{GF}(L) \tag{5}$$

where RELEVANCE(L) is a measure of the utility of the ranked list L and Global Fairness (GF) measures the user-perceived difference of fairness between a "fair" ranking and the actual ranking. In particular,

$$\text{GF}(L) = \sum_{k=1}^{|L|} \text{DECAY}(L,k)\text{DISTRSIM}(L,k)$$

where DECAY measures the event a user accesses the k-th item in L and DISTRSIM how far the distribution of groups is from an ideal, uniform distribution. Expression (5) has been computed for each topic of every track, phase, category, and δ, and for each of the following methods implementing the three functions occurring in (5):

- DECAY: Expected Reciprocal Rank (ERR) and Ranked Biased Precision (RBP),
- DISTRSIM was defined as

  $$1 - \text{DIVERGENCE}(p||p^*)$$

  where p is the observed group distribution and p^* is the fair group distribution and DIVERGENCE was either Jensen-Shannon Divergence (), Root Normalised Order-aware Divergence (RNOD), or Normalised Match Distance (NMD).[6]

We were interested to the trend of GFR as δ increases to check whether GFR tends to increase, decrease or keep stable if the importance to the fairness operator in the calculation of the eigentopics increases. The average of the GFR values obtained for each topic was then calculated for each track, phase, category, and δ and classified by the aforementioned methods used to implement the three functions of (5). Table 7 summarizes the results where the average and the standard deviation values of GFR have been computed over all the possible combinations of the methods used for decay and distribution similarity.

Table 8 breaks the GFR values down the decay and distribution similarity methods. Overall, GFR makes an improvement of fairness *and* effectiveness visible for different decay and distribution similarity methods. The variations of GFR while δ increases depend on the variations of the measures of effectiveness and fairness. For example, as for the ERR-based decay method the variations of precision and fairness are reported below on the left and on the right, respectively:

δ	All	2021	2022	Gndr	Locat	Train	Test	All	2021	2022	Gndr	Locat	Train	Test
0.0	0.495	0.386	0.576	0.282	0.507	0.608	0.41	0.162	0.135	0.181	0.039	0.249	0.19	0.14
0.25	0.481	0.365	0.568	0.272	0.491	0.571	0.414	0.19	0.167	0.208	0.05	0.27	0.26	0.138
0.5	0.48	0.363	0.567	0.265	0.494	0.57	0.412	0.191	0.169	0.207	0.049	0.269	0.265	0.135
0.75	0.468	0.346	0.56	0.257	0.477	0.566	0.395	0.185	0.171	0.197	0.05	0.273	0.247	0.139
1.0	0.442	0.294	0.553	0.17	0.464	0.574	0.343	0.215	0.232	0.203	0.121	0.311	0.243	0.195
1.25	0.432	0.286	0.543	0.152	0.449	0.568	0.331	0.209	0.223	0.198	0.101	0.305	0.239	0.186
1.5	0.433	0.28	0.548	0.15	0.45	0.567	0.333	0.212	0.221	0.206	0.101	0.312	0.24	0.192
1.75	0.434	0.278	0.552	0.15	0.45	0.569	0.333	0.212	0.221	0.206	0.101	0.313	0.24	0.192
2.0	0.431	0.278	0.547	0.15	0.449	0.561	0.334	0.209	0.219	0.201	0.101	0.305	0.241	0.185

6. The details have been described by Sakai et al. (2023) in their paper.

δ	All	2021	2022	Gndr	Locat	Train	Test
0.0	0.33 ± 0.10	0.27 ± 0.08	0.38 ± 0.01	0.15 ± 0.08	0.39	0.41 ± 0.06	0.28 ± 0.10
0.25	0.34 ± 0.10	0.27 ± 0.07	0.39 ± 0.01	0.15 ± 0.08	0.39	0.42 ± 0.04	0.28 ± 0.10
0.5	0.34 ± 0.11	0.27 ± 0.07	0.39 ± 0.01	0.15 ± 0.09	0.39	0.42 ± 0.05	0.28 ± 0.10
0.75	0.33 ± 0.11	0.26 ± 0.08	0.38 ± 0.01	0.15 ± 0.09	0.38	0.41 ± 0.05	0.27 ± 0.10
1.0	0.33 ± 0.12	0.26 ± 0.09	0.38 ± 0.02	0.13 ± 0.10	0.39	0.41 ± 0.06	0.26 ± 0.11
1.25	0.32 ± 0.13	0.25 ± 0.08	0.37 ± 0.01	0.11 ± 0.10	0.38	0.41 ± 0.05	0.25 ± 0.11
1.5	0.32 ± 0.12	0.25 ± 0.08	0.38 ± 0.01	0.11 ± 0.10	0.38	0.41 ± 0.05	0.26 ± 0.12
1.75	0.32 ± 0.12	0.24 ± 0.08	0.38 ± 0.01	0.11 ± 0.10	0.38	0.41 ± 0.06	0.26 ± 0.12
2.0	0.32 ± 0.12	0.24 ± 0.08	0.38 ± 0.01	0.11 ± 0.10	0.38	0.40 ± 0.06	0.26 ± 0.12

Table 7: The average GFR $\pm$ the standard deviations thereof have been computed over the mean GFRs obtained for each decay or distribution similarity method and classified by track, phase, and category. These averages measures the scale for each track, phase, and category. The standard deviations suggest that there is variability from a method to another, being the methods used for gender showing less variations between methods than the other criteria.

The variations of GFR then depend on the weights assigned to precision and GF in Expression (5). The need of deciding the weights can be considered the sympton of a GFR's weakness. First, the sum of two different measures can often be problematic if these quantities measure different phenomena because of different measure units, for instance. Second, the indications provided by GFR strongly depend on weights, thus asking a question left to the future work as Sakai et al. (2023) admit.

6 Final Remarks

Group fairness is considered in this paper, but there is the question of whose fairness we are talking about. In other circumstances, fairness may be defined in terms of what is displayed (e.g., job ads or products). In such a case, search result diversity would be an acceptable phrase if the items were not affiliated with real organizations who profit from exposure. To be precise, fairness should be defined in terms of the producers or authors, particularly in systems that serve multiple stakeholders. There is a third possibility as well: that fairness pertains to subjects of content, such as people mentioned in news articles or search results, the latter being best termed as a privacy issue. Finally, topics might be pertinent to fairness: some topics might not be acceptable in some contexts, whereas they are in other contexts. In sum, the multi-stakeholder nature of IR further complicates fairness definitions.

So where does this leave us? First, we must recognize, at the risk of looking generic, that fairness is not a property of an algorithm alone, but of a system in context. Despite the trend to consider the distinct phases of search, a fair system is a product of design choices, data distributions, user behaviors, and societal norms. No amount of technical tinkering will yield a "perfectly fair" system if the underlying assumptions are flawed or the inputs are biased. Second, fairness cannot be divorced from trade-offs. We must make explicit our priorities—between relevance and (fair) representation, between user satisfaction and (fair) societal impact. Third, evaluation becomes a central challenge. Unlike relevance,

(a) ERR

δ	All	2021	2022	Gndr	Locat	Train	Test
0.0	0.328	0.261	0.379	0.161	0.378	0.399	0.275
0.25	0.336	0.266	0.388	0.161	0.380	0.416	0.276
0.5	0.335	0.266	0.387	0.157	0.381	0.417	0.273
0.75	0.327	0.259	0.378	0.154	0.375	0.407	0.267
1.0	0.329	0.263	0.378	0.145	0.388	0.409	0.269
1.25	0.321	0.254	0.370	0.126	0.377	0.403	0.258
1.5	0.323	0.250	0.377	0.125	0.381	0.404	0.262
1.75	0.323	0.250	0.378	0.126	0.382	0.404	0.262
2.0	0.320	0.248	0.374	0.125	0.377	0.401	0.259

(b) RBP

All	2021	2022	Gndr	Locat	Train	Test
0.34	0.278	0.387	0.145	0.404	0.416	0.283
0.344	0.273	0.397	0.146	0.399	0.424	0.284
0.342	0.273	0.394	0.141	0.401	0.423	0.281
0.333	0.261	0.387	0.138	0.39	0.413	0.273
0.323	0.247	0.381	0.111	0.391	0.413	0.257
0.32	0.246	0.376	0.103	0.384	0.413	0.251
0.32	0.241	0.379	0.102	0.384	0.409	0.253
0.32	0.239	0.381	0.102	0.384	0.41	0.253
0.318	0.239	0.377	0.102	0.381	0.407	0.251

(c) JSD

δ	All	2021	2022	Gndr	Locat	Train	Test
0.0	0.327	0.261	0.377	0.153	0.379	0.403	0.271
0.25	0.334	0.264	0.387	0.155	0.38	0.411	0.277
0.5	0.329	0.26	0.38	0.151	0.375	0.401	0.274
0.75	0.322	0.256	0.371	0.147	0.369	0.395	0.267
1.0	0.317	0.244	0.371	0.127	0.375	0.402	0.253
1.25	0.309	0.242	0.36	0.115	0.363	0.396	0.245
1.5	0.311	0.24	0.365	0.114	0.366	0.398	0.247
1.75	0.313	0.24	0.367	0.114	0.368	0.4	0.247
2.0	0.311	0.239	0.364	0.114	0.365	0.397	0.246

(d) RNOD

All	2021	2022	Gndr	Locat	Train	Test
0.338	0.285	0.377	0.154	0.396	0.409	0.284
0.343	0.284	0.388	0.154	0.394	0.427	0.28
0.344	0.286	0.388	0.149	0.399	0.434	0.277
0.334	0.272	0.381	0.146	0.389	0.42	0.27
0.332	0.274	0.376	0.132	0.398	0.418	0.269
0.328	0.265	0.375	0.118	0.391	0.417	0.261
0.329	0.262	0.379	0.117	0.394	0.415	0.265
0.329	0.261	0.38	0.117	0.393	0.415	0.264
0.325	0.26	0.375	0.117	0.388	0.411	0.261

(e) NMD

δ	All	2021	2022	Gndr	Locat	Train	Test
0.0	0.338	0.262	0.395	0.152	0.398	0.41	0.283
0.25	0.342	0.261	0.403	0.152	0.395	0.422	0.282
0.5	0.342	0.263	0.402	0.147	0.399	0.426	0.28
0.75	0.333	0.251	0.395	0.145	0.389	0.414	0.273
1.0	0.329	0.247	0.39	0.125	0.395	0.412	0.266
1.25	0.324	0.242	0.385	0.112	0.387	0.412	0.258
1.5	0.324	0.235	0.391	0.111	0.388	0.406	0.262
1.75	0.324	0.234	0.392	0.111	0.388	0.407	0.262
2.0	0.321	0.232	0.387	0.111	0.384	0.404	0.259

Table 8: In these tables, GFR has been detailed by decay and distribution similarity methods. There are no big differences except for scale.

where judgments can be solicited from assessors (albeit imperfectly) as noted by (Voorhees and Harman, 2005, page 24) for example, fairness is often a normative question as noted by (Balagopalan et al., 2023, page 2658), Fang et al. (2024), Zehlike et al. (2022a,b) for example. What counts as a fair outcome may vary across contexts, cultures, and communities. Thus, fairness evaluation may need to rely not only on quantitative metrics but also on qualitative assessments, stakeholder consultation, and ethical reflection.

Regarding evaluation, two questions naturally arise. First of all, can document collections be fair in general? Since document collections are built from broader sources, such as catalogs, portals, or subsets of the web, they will necessarily reflect the content present in those same sources. Therefore, if the content of the sources contains biases, the collections will also contain them. For example, in the Fair Track collections, there is a clear bias in favor of males, Europe, and North America. Does this bias represent the actual

content? Unfortunately, yes. The statistics presented in Section 3 speak for themselves. These statistics also suggest that bias might not be eliminated by an automated system that retrieves and ranks documents in response to a topic, nor by a user who evaluates the retrieved documents as relevant or not, since users, like the authors of the documents, are affected by the same prejudices.

Nevertheless, the approaches to balance effectiveness and group fairness, like the one illustrated in this paper, ignore the context from which groups and categories are normed in the sense that only the frequency distribution of items over groups counts. Such a view may inevitably be seen as specific or even restrictive to the eyes of social experts; however, the balance is also a matter of statistics and computation. Computation plays a crucial role in quantifying disparities, yet it often fails to capture the nuanced realities of social dynamics. Therefore, integrating qualitative insights alongside quantitative methods is essential for a more comprehensive understanding of fairness and effectiveness in decision-making processes.

We are aware that the conclusions drawn from this paper may be conditioned by the context and the objectives of the experimental data. This specific test collections are a niche use case. The definition of bias would likely change in the case of general-purpose search in which not only Wikipedia articles are retrieved and typical web queries are sent to a search engine.

Fairness in information retrieval requires us to broaden our view of what retrieval systems are for and whom they serve. It asks us to reflect not only on what our systems retrieve but on what they leave out—-and whom they may leave behind. In doing so, we may discover that fairness, like relevance before it, is not a static concept but a moving target—one that will evolve as our systems and our societies do.

Acknowledgments and Disclosure of Funding

I thank the editor and the anonymous reviewers for the comments. No external funding was received in support of this work.

References

Ricardo Baeza-Yates. Bias on the web. *Communication of the ACM*, 61(6):54–61, 2018.

Aparna Balagopalan, Abigail Z. Jacobs, and Asia J. Biega. The role of relevance in fair ranking. In *Proceedings of SIGIR*, pages 2650–2660, 2023. doi: 10.1145/3539618.3591933.

Asia J. Biega, Krishna P. Gummadi, and Gerhard Weikum. Equity of attention: Amortizing individual fairness in rankings. In *Proceedings of SIGIR*, pages 405–414. ACM, 2018. doi: 10.1145/3209978.3210063.

Amin Bigdeli, Negar Arabzadeh, Shirin Seyedsalehi, Morteza Zihayat, and Ebrahim Bagheri. On the orthogonality of bias and utility in ad hoc retrieval. In *Proceedings of SIGIR*, SIGIR '21, pages 1748–1752, New York, NY, USA, 2021. Association for Computing Machinery. ISBN 9781450380379. doi: 10.1145/3404835.3463110.

Reuben Binns. On the apparent conflict between individual and group fairness. In *Proceedings of FAccT*, 2020.

Lidia Ceriani and Paolo Verme. Gini on mutability. *METRON*, 82(3):269–292, 2024. doi: 10.1007/s40300-024-00279-2.

Olivier Chapelle, Shihao Ji, Ciya Liao, Emre Velipasaoglu, Larry Lai, and Su-Lin Wu. Intent-based diversification of web search results: metrics and algorithms. *Information Retrieval*, 14(6):572–592, 2011.

Tim Draws, Nava Tintarev, Ujwal Gadiraju, Alessandro Bozzon, and Benjamin Timmermans. This is not what we ordered: Exploring why biased search result rankings affect user attitudes on debated topics. In *Proceedings of SIGIR*, pages 295–305, 2021. doi: 10.1145/3404835.3462851.

Michael D. Ekstrand, Graham McDonald, Amifa Raj, and Isaac Johnson. Overview of the trec 2021 fair ranking track. In *The Thirtieth Text REtrieval Conference (TREC 2021) Proceedings*, 2022a.

Michael D. Ekstrand, Graham McDonald, Amifa Raj, and Isaac Johnson. TREC 2022 Fair Ranking Track Participant Instructions. `https://fair-trec.github.io/docs/Fair_Ranking_2022_Participant_Instructions.pdf`, 2022b.

Michael D. Ekstrand, Graham McDonald, Amifa Raj, and Isaac Johnson. Overview of the trec 2022 fair ranking track. In *The Thirtieth Text REtrieval Conference (TREC 2021) Proceedings*, 2023.

Yi Fang, Ashudeep Singh, and Zhiqiang Tao. Fairness in search systems. *Foundations and Trends in Information Retrieval*, 18(3):262–416, 2024. doi: 10.1561/1500000101.

Corrado Gini. *Variabilità e mutabilità*. Tipografia di Paolo Cuppin, 1912.

Christina Ilvento, Meena Jagadeesan, and Shuchi Chawla. Multi-category fairness in sponsored search auctions. In *Proceedings of FAccT*, FAT* '20, pages 348–358. ACM, 2020. ISBN 9781450369367. doi: 10.1145/3351095.3372848.

Thomas Jaenich, Graham McDonald, and Iadh Ounis. Fairness-aware exposure allocation via adaptive reranking. In *Proceedings of SIGIR*, SIGIR '24, pages 1504–1513, New York, NY, USA, 2024. Association for Computing Machinery. ISBN 9798400704314. doi: 10.1145/3626772.3657794.

Thomas Jaenich, Graham McDonald, and Iadh Ounis. Fair exposure allocation using generative query expansion. In Claudia Hauff, Craig Macdonald, Dietmar Jannach, Gabriella Kazai, Franco Maria Nardini, Fabio Pinelli, Fabrizio Silvestri, and Nicola Tonellotto, editors, *Advances in Information Retrieval*, pages 267–281, Cham, 2025. Springer Nature Switzerland. ISBN 978-3-031-88717-8.

K. Jarvëlin and J. Kekäläinen. Cumulated gain-based evaluation of IR techniques. *ACM Transactions on Information Systems*, 20(4):422–446, 2002.

Till Kletti, Jean-Michel Renders, and Patrick Loiseau. Introducing the expohedron for efficient pareto-optimal fairness-utility amortizations in repeated rankings. In *Proceedings of WWW*, pages 498–507, 2022. doi: 10.1145/3488560.3498490.

Graham McDonald, Craig Macdonald, and Iadh Ounis. Search results diversification for effective fair ranking in academic search. *Information Retrieval*, 25(1):1–26, March 2022. doi: 10.1007/s10791-021-09399-z.

Massimo Melucci. A model of the relationship between the variations of effectiveness and fairness in information retrieval. *Discover Computing*, 27(3), 2024a.

Massimo Melucci. On the trade-off between ranking effectiveness and fairness. *Expert Systems with Applications*, 241, 2024b.

Massimo Melucci. Preference eigensystems for fair ranking. *Expert Systems with Applications*, 269:126324, 2025a. doi: 10.1016/j.eswa.2024.126324. https://www.sciencedirect.com/science/article/pii/S0957417424031919.

Massimo Melucci. Three methods for fair ranking of multiple protected items. *Scientific Reports*, to appear, 2025b.

Stefano Mizzaro. Relevance: The whole history. *Journal of the American Society for Information Science*, 48(9):810–832, 1997.

Marco Morik, Ashudeep Singh, Jessica Hong, and Thorsten Joachims. Controlling fairness and bias in dynamic learning-to-rank. In *Proceedings of SIGIR*, pages 429–438. ACM, 2020. doi: 10.1145/3397271.3401100.

Amifa Raj and Michael D. Ekstrand. Comparing fair ranking metrics, 2022. Arxiv. https://arxiv.org/abs/2009.01311.

S.E. Robertson, M.E. Maron, and W.S. Cooper. Probability of relevance: A unification of two competing models for document retrieval. *Information Technology: Research and Development*, 1:1–21, 1982.

Stephen E. Robertson. The probability ranking principle in information retrieval. *Journal of Documentation*, 33(4):294–304, 1977.

Tetsuya Sakai, Jin Young Kim, and Inho Kang. A versatile framework for evaluating ranked lists in terms of group fairness and relevance. *ACM Transactions on Information Systems*, 42(1), 2023. doi: 10.1145/3589763.

G. Salton. *Automatic Information Organization and Retrieval*. Mc Graw Hill, New York, NY, 1968.

T. Saracevic. Relevance: a review of and a framework for thinking on the notion in information science. *Journal of the American Society for Information Science*, 26:321–343, 1975.

T. Saracevic. Relevance: a review of the literature and a framework for thinking on the notion in information science. Part II: nature and manifestations of relevance. *Journal of the American Society for Information Science and Technology*, 58(3):1915–1933, 2007a.

T. Saracevic. Relevance: a review of the literature and a framework for thinking on the notion in information science. Part III: Behavior and effects of relevance. *Journal of the American Society for Information Science and Technology*, 58(13):2126–2144, 2007b.

E. Seneta. *Non-negative Matrices and Markov Chains*. Springer, 2006.

Cornelis Joost Van Rijsbergen. *Information Retrieval*. Butterworths, London, 1979.

E.M. Voorhees and D.K. Harman, editors. *TREC: Experiment and Evaluation in Information Retrieval*. The MIT Press, Cambridge, MA, 2005.

Lin Xiao, Zhang Min, Zhang Yongfeng, Gu Zhaoquan, Liu Yiqun, and Ma Shaoping. Fairness-aware group recommendation with Pareto-efficiency. In *Proceedings of RecSys*, pages 107–115, 2017.

Meike Zehlike, Ke Yang, and Julia Stoyanovich. Fairness in ranking, part i: Score-based ranking. *ACM Computing Surveys*, 55(6), December 2022a. ISSN 0360-0300. doi: 10.1145/3533379.

Meike Zehlike, Ke Yang, and Julia Stoyanovich. Fairness in ranking, part ii: Learning-to-rank and recommender systems. *ACM Computing Surveys*, 55(6), December 2022b. ISSN 0360-0300. doi: 10.1145/3533380.

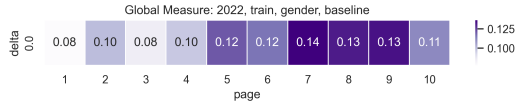
Appendix A. Heatmaps

(a) $M = G \times P$ as for baseline, 2021, training set, geographic location

(b) $M = G \times P$ as for baseline, 2021, test set, geographic location

(c) $M = G \times P$ as for baseline, 2021, test set, gender

(d) $M = G \times P$ as for baseline, 2022, training set, geographic location

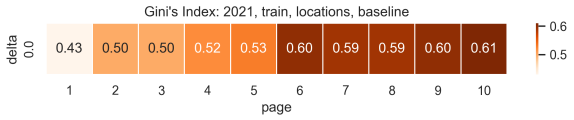


(e) $M = G \times P$ as for baseline, 2022, training set, gender

(f) $M = G \times P$ as for baseline, 2022, test set, geographic location

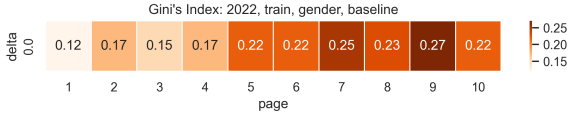
(g) $M = G \times P$ as for baseline, 2022, test set, gender

Figure 2: The heatmaps of this figure depicts the *baseline* global measure $M = G \times P$ by track (2021, 2022), phase (train, test), category (location, gender). It is the baseline $M = G \times P$ because ET was not activated in this case, i.e. $\delta = 0$. These heatmaps should be read in conjunction with Figures 3 and 4. In particular, Figure 4 not surprisingly shows that the baseline precision is higher at the top pages than at the bottom pages for all tracks, phases, and categories. Therefore, the variations of the baseline global measures mostly depend on the variations of Gini's Index as depicted by Figure 3. In sum, the rankings tend to be biased with respect gender to a larger extent than location.

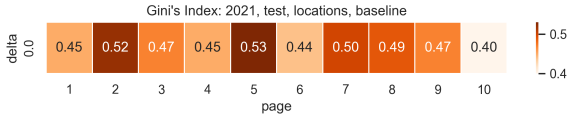


(a) G as for baseline, 2021, training set, geographic location

(b) G as for baseline, 2021, test set, geographic location

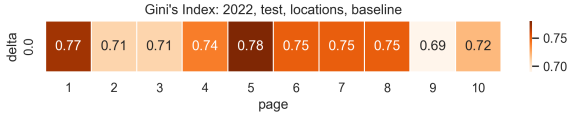


(c) G as for baseline, 2021, test set, gender

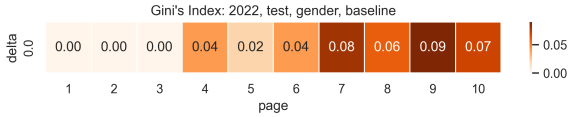


(d) G as for baseline, 2022, training set, geographic location

(e) G as for baseline, 2022, training set, gender

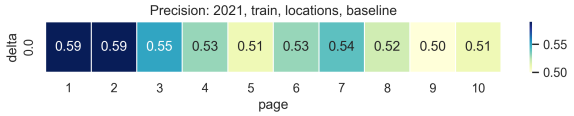


(f) G as for baseline, 2022, test set, geographic location



(g) G as for baseline, 2022, test set, gender

Figure 3: The rankings provided by a retrieval system are generally biased against certain groups, particularly non-males and non-Western regions, with the exception of location-based groups in the 2022 track.

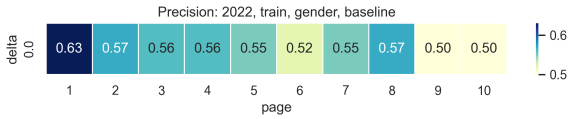


(a) P as for baseline, 2021, training set, geographic location

(b) P as for baseline, 2021, test set, geographic location

(c) P as for baseline, 2021, test set, gender

(d) P as for baseline, 2022, training set, geographic location

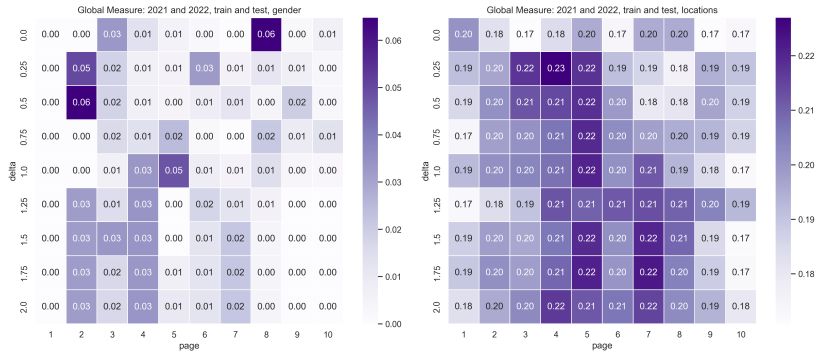


(e) P as for baseline, 2022, training set, gender

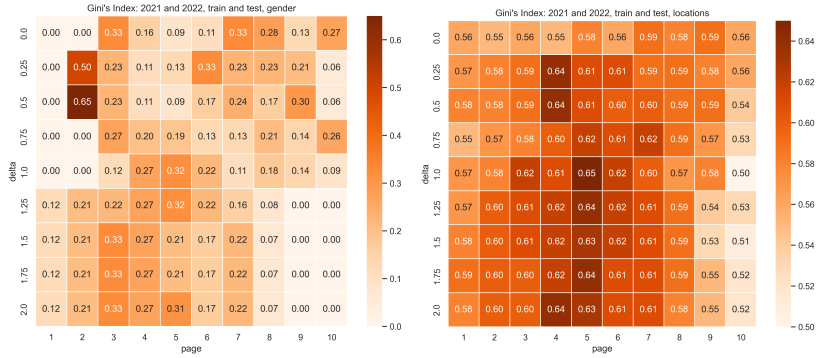
(f) P as for baseline, 2022, test set, geographic location

(g) P as for baseline, 2022, test set, gender

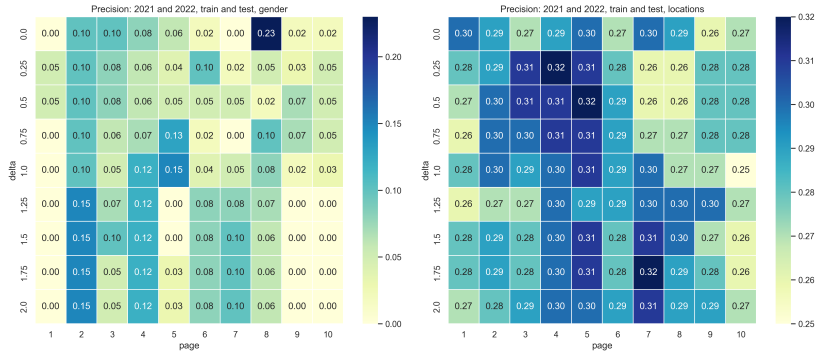
Figure 4: Retrieval effectiveness is relatively high, with minor variations across the 10-document pages. However, the top two pages are the most dense of relevant documents.



(a) $M = G \times P$ as for gender and geographic location



(b) G as for gender and geographic location

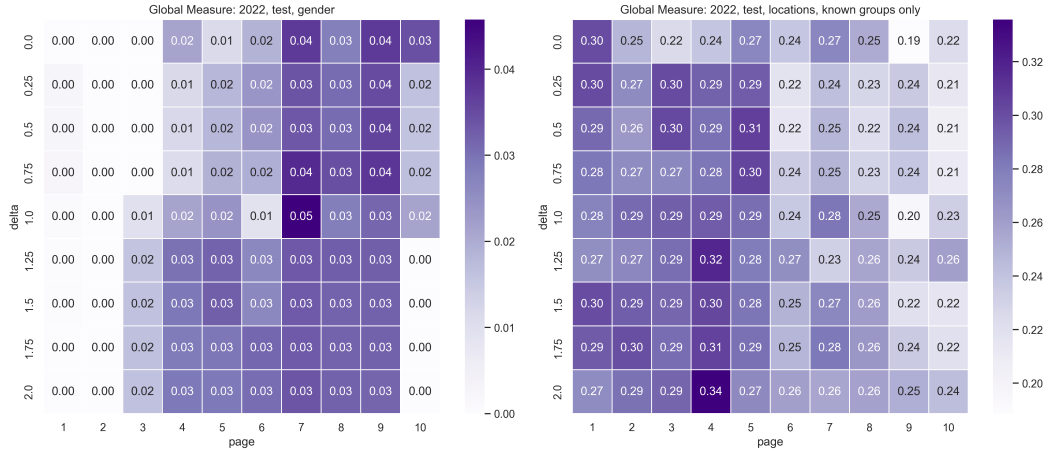


(c) P as for gender and geographic location

Figure 5: The figure shows heatmaps of global measures $M = G \times P$ for each 10-document page, ranging from 0 to 2. The heatmaps show that precision is no longer the highest at the top pages, indicating that the global measures observed may depend on both variations of Gini's Index and precision. Improving $M = G \times P$ requires increasing δ, indicating that ET is effective in improving fair, effective retrieval and ranking, as well as controlling variations of fairness and effectiveness. Some improvement can also be seen in gender, which is the most difficult bias due to the high prevalence of males in the test collection.

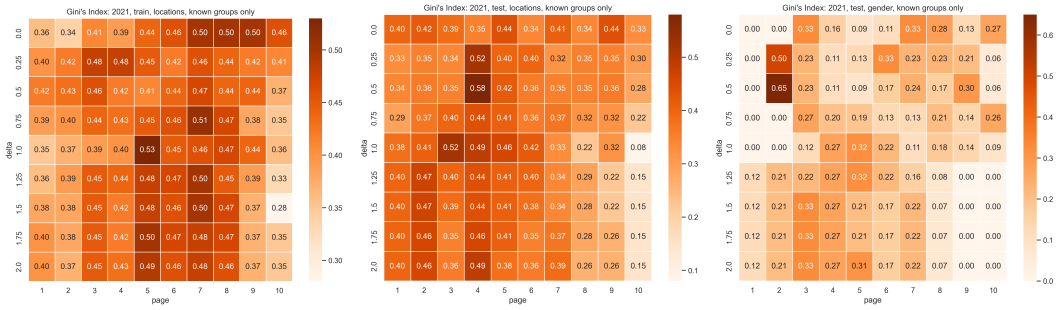
(a) $M = G \times P$ as for 2021, train, geographic location (b) $M = G \times P$ as for 2021, test set, geographic location (c) $M = G \times P$ as for 2021, test set, gender

(d) $M = G \times P$ as for 2022, training set, gender (e) $M = G \times P$ as for 2022, training set, geographic location



(f) $M = G \times P$ as for 2022, test set, gender (g) $M = G \times P$ as for 2022, test set, geographic locations

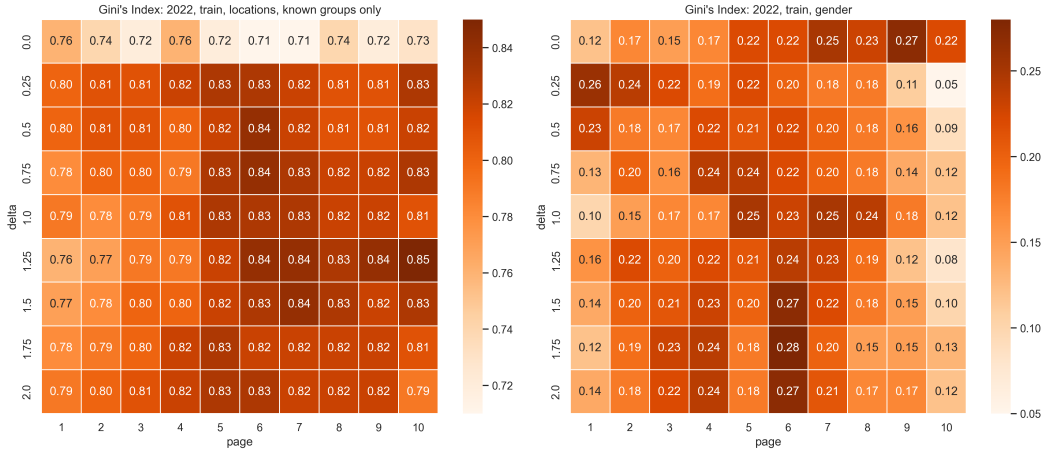
Figure 6: In contrast to Figure 5(a), these heatmaps show $M = G \times P$ values by track, phase, and category.



(a) G as for 2021, training set, geographic location

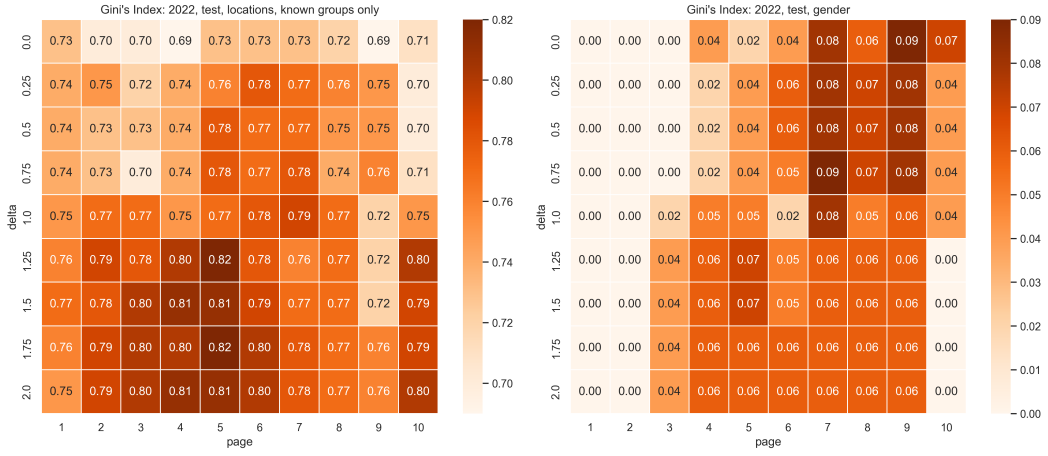
(b) G as for 2021, test set, geographic location

(c) G as for 2021, test set, gender



(d) G as for 2022, training set, geographic location

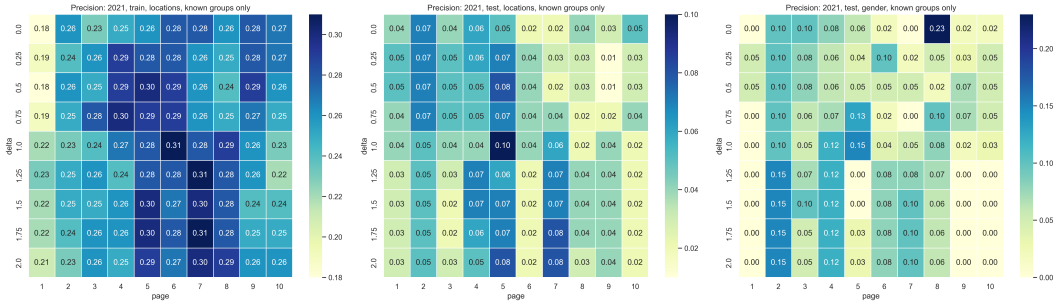
(e) G as for 2022, training set, gender



(f) G as for 2022, test set, geographic location

(g) G as for 2022, test set, gender

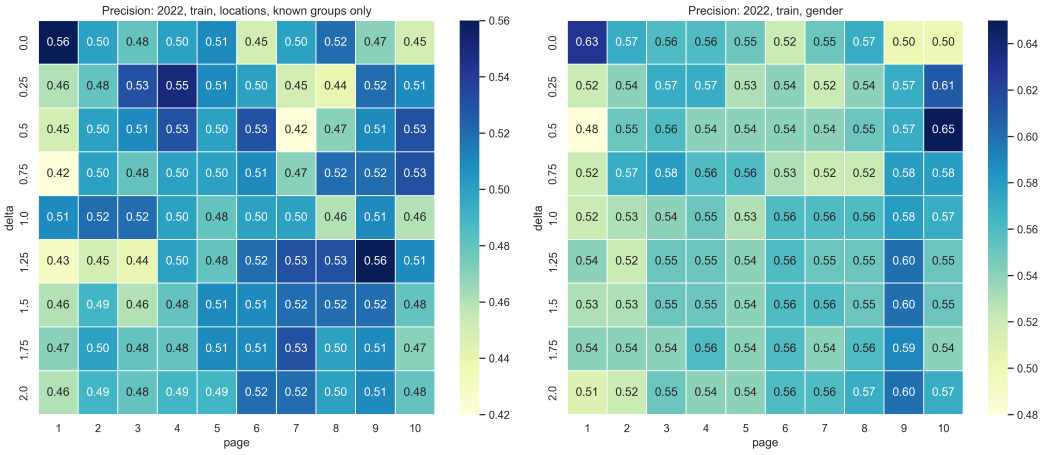
Figure 7: Unlike Figure 5(b), these heatmaps show G values by track, phase, and category.



(a) P as for 2021, training set, geographic location

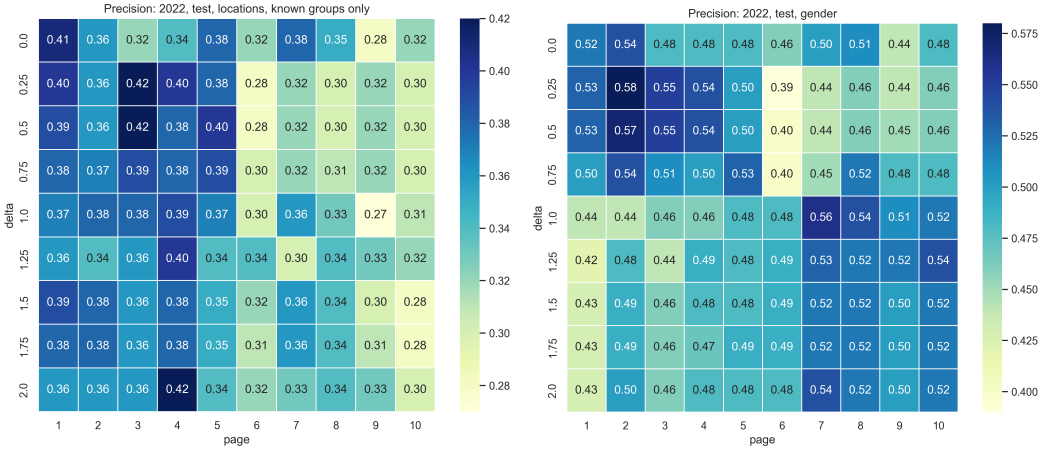
(b) P as for 2021, test set, geographic location

(c) P as for 2021, test set, gender



(d) P as for 2022, training set, geographic location

(e) P as for 2022, training set, gender

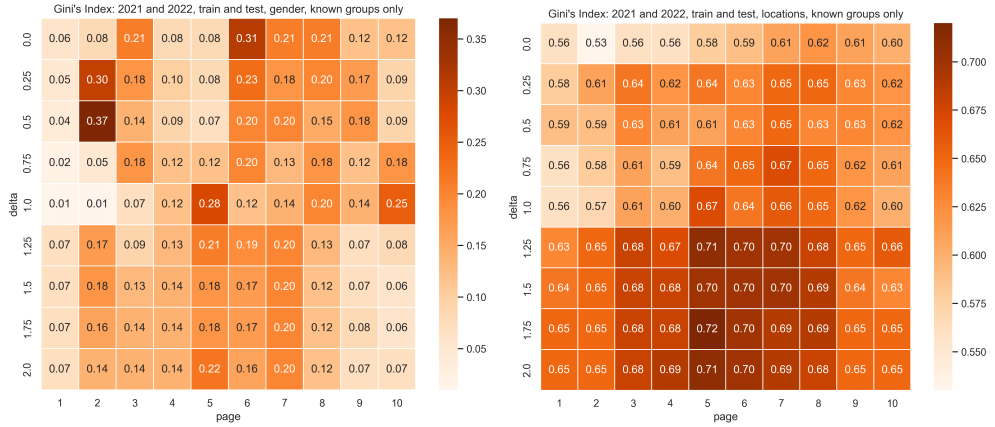


(f) P as for 2022, test set, geographic location

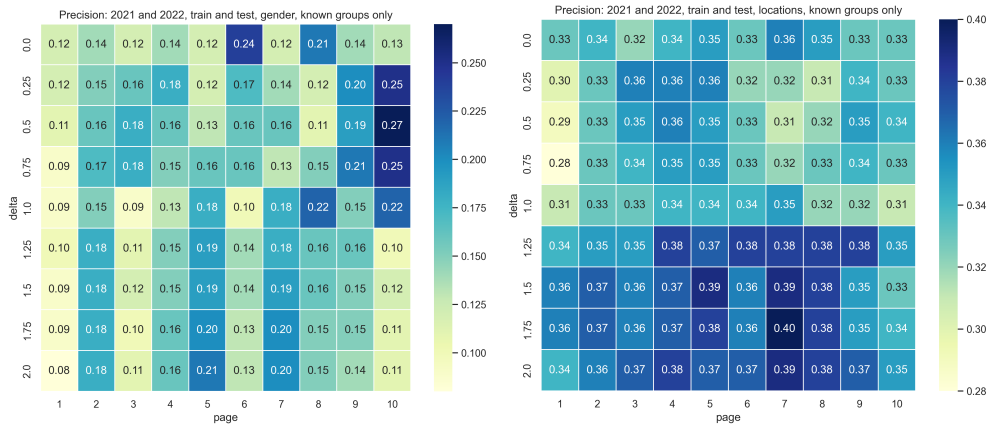
(g) P as for 2022, test set, gender

Figure 8: These heatmaps show P values by track, phase, and category, rather than combining tracks and phases as in Figure 5(c).

(a) $M = G \times P$ as for gender and geographic location



(b) G as for gender and geographic location



(c) P as for gender and geographic location

Figure 9: Unknown group membership appear to have a minor impact. Compare heatmaps to Figures 5(a), 5(b), and 5(c).

Appendix B. Statistical Significance Testing

Fairness: 2021 test locations										
δ	1	2	3	4	5	6	7	8	9	10
0.25										
0.50										
0.75	.02									
1.00										
1.25										
1.50										
1.75										
2.00										

Effectiness: 2021 test locations										
δ	1	2	3	4	5	6	7	8	9	10
0.25										
0.50										
0.75										
1.00	.0						.01			
1.25	.0						.05			
1.50	.0								.04	
1.75	.0								.04	
2.00	.0								.02	

Fairness: 2022 train gender										
δ	1	2	3	4	5	6	7	8	9	10
0.25										
0.50										
0.75										
1.00										
1.25										
1.50										
1.75										
2.00										

Effectiness: 2022 train gender										
δ	1	2	3	4	5	6	7	8	9	10
0.25	.02									.01
0.50	.0								.03	.0
0.75	.01								.05	
1.00	.01								.02	
1.25	.03								.0	
1.50	.02								.01	
1.75	.03								.01	
2.00	.01								.01	

Fairness: 2022 train locations										
δ	1	2	3	4	5	6	7	8	9	10
0.25										
0.50										
0.75	.02	.01								
1.00		.02								
1.25										
1.50										
1.75	.01									
2.00		.01								

Effectiness: 2022 train locations										
δ	1	2	3	4	5	6	7	8	9	10
0.25	.01									.0
0.50	.02					.04				.0
0.75	.0									.01
1.00										
1.25	.0	.04							.0	.02
1.50	.0									
1.75	.01									
2.00	.01									

Fairness: 2022 test gender										
δ	1	2	3	4	5	6	7	8	9	10
0.25										
0.50										
0.75										
1.00										
1.25										
1.50										
1.75										
2.00										

Effectiness: 2022 test gender										
δ	1	2	3	4	5	6	7	8	9	10
0.25										
0.50										
0.75										
1.00		.04								
1.25										
1.50										
1.75										
2.00										

Fairness: 2022 test locations										
δ	1	2	3	4	5	6	7	8	9	10
0.25				.03		.02			.03	
0.50				.02					.04	
0.75									.04	
1.00		.0	.0	.0		.01				
1.25	.05	.0	.0	.0	.01	.0			.05	
1.50	.01	.0	.0	.0	.02	.01			.04	
1.75	.04	.0	.0	.01	.01	.0	.04			
2.00		.0	.0	.0	.03	.0			.05	

Effectiness: 2022 test locations										
δ	1	2	3	4	5	6	7	8	9	10
0.25										
0.50			.04							
0.75										
1.00										
1.25										
1.50										
1.75										
2.00										