Information Retrieval Journal manuscript No.
(will be inserted by the editor)

Search Results Diversification for Effective Fair Ranking in Academic Search

Graham McDonald · Craig Macdonald · Iadh Ounis

Received: date / Accepted: date

Abstract Providing users with relevant search results has been the primary focus of information retrieval research. However, focusing on relevance alone can lead to undesirable side effects. For example, small differences between the relevance scores of documents that are ranked by relevance alone can result in large differences in the exposure that the authors of relevant documents receive, i.e., the likelihood that the documents will be seen by searchers. Therefore, developing fair ranking techniques to try to ensure that search results are not dominated, for example, by certain information sources is of growing interest, to mitigate against such biases. In this work, we argue that generating fair rankings can be cast as a search results diversification problem across a number of assumed fairness groups, where groups can represent the demographics or other characteristics of information sources. In the context of academic search, as in the TREC Fair Ranking Track, which aims to be fair to *unknown* groups of authors, we evaluate three well-known search results diversification approaches from the literature to generate rankings that are fair to multiple assumed fairness groups, e.g. early-career researchers vs. highly-experienced authors. Our experiments on the 2019 and 2020 TREC datasets show that explicit search results diversification is a viable approach for generating effective rankings that are fair to information sources. In particular, we show that building on xQuAD diversification as a fairness component can result in a significant ($p < 0.05$) increase (up to 50% in our experiments) in the fairness of exposure that authors from unknown protected groups receive.

Keywords Fair Ranking · Search Results Diversification · Expected Exposure · Academic Search

G. McDonald
University of Glasgow, Scotland, UK
0000-0002-1266-5996
E-mail: Graham.McDonald@Glasgow.ac.uk

C. Macdonald
University of Glasgow, Scotland, UK
0000-0003-3143-279X
E-mail: Craig.Macdonald@Glasgow.ac.uk

I. Ounis
University of Glasgow, Scotland, UK
0000-0003-4701-3223
E-mail: Iadh.Ounis@Glasgow.ac.uk

1 Introduction

The objective of an information retrieval (IR) system has traditionally been seen as being to maximise the fraction of results presented to the user that are relevant to the user's query or to address the user's information need as close to the top rank position as possible. However, many studies have shown that focusing on relevance alone as a measure of search success can lead to undesirable side effects that can have negative societal impacts (Baeza-Yates, 2018; Epstein et al., 2017; Kay et al., 2015; Mehrotra et al., 2018). Interventions in computational decision-support systems, such as IR systems, cannot solve such societal issues (Abebe et al., 2020). However, developing systematic interventions can support broader attempts at understanding and addressing social problems. Therefore, in recent years, there has been an increased interest in the societal implications of how IR systems select or present documents to users and the potential for IR systems to systematically discriminate against particular groups of people (Pleiss et al., 2017).

IR systems and machine learned models can encode and perpetuate any biases that exist in the test collections that they use (Zehlike et al., 2017; Bender et al., 2021). Moreover, search engines that are used to find, for example, jobs or news can have a significant negative impact on information sources that produce relevant content but are often unfairly under-represented in the search results. Therefore, it is imperative that the IR community focuses on minimising the potentially negative human, social, and economic impact of such biases in search systems (Culpepper et al., 2018), particularly for disadvantaged or protected groups of society (Pedreschi et al., 2008). One way to mitigate against such biases that is receiving increasing attention in the IR community is to develop *fair* ranking strategies to try to ensure that certain users or information sources are not discriminated against (Culpepper et al., 2018; Ekstrand et al., 2019; Olteanu et al., 2019b). The increasing importance this topic is exemplified by the Text REtrieval Conference (TREC) Fair Ranking Track (Biega et al., 2020).

When defining fairness for ranking strategies, we agree with Singh and Joachims (2018) that there is not one definitive definition, and a judgement of fairness is context specific. In this work, we take the view that for a search engine's ranking to be considered as being fair, relevant information sources should be given a fair *exposure* to the search engine's users. Following Singh and Joachims (2018), we consider a fair exposure to mean that the exposure that a document receives should be proportional to the relevance of the document with respect to a user's query. In the context of an IR system that presents documents to a user that are ranked in decreasing order of their *estimated* relevance to the user's query, documents that are placed lower in the ranking will receive less exposure than higher ranked documents. The reduction (or drop-off) in exposure that a document at position j (Pos_j) in a ranking gets can be estimated as $\text{Exposure}(Pos_j) = \frac{1}{\log(1+Pos_j)}$. This is the user model that encapsulates position bias that is commonly used in the Discounted Cumulative Gain (DCG) (Järvelin and Kekäläinen, 2002) measure.

Moreover, the amount of exposure that a document receives accumulates over repeated instances of a query (we refer to this as a sequence of rankings).

Therefore, in this work, as our definitions of fairness we adopt the *Disparate Treatment* and *Disparate Impact* fairness constraints of Singh and Joachims (2018). Disparate Treatment enforces the exposure of two fairness groups, G_0 and G_1, (e.g., authors from a protected societal group) to be proportional to the average relevance of the group's documents, and is defined as:

$$\frac{\text{Exposure}(G_0 \mid \mathbf{P})}{\text{U}(G_0 \mid q)} = \frac{\text{Exposure}(G_1 \mid \mathbf{P})}{\text{U}(G_1 \mid q)} \tag{1}$$

where $\mathbf{P}$ is a doubly stochastic matrix where the cell $\mathbf{P}_{i,j}$ is the probability that a ranking r places document d_i at rank j, and the average utility (relevance) of a group, $\text{U}(G_k \mid q)$ is calculated as $\text{U}(G_k \mid q) = \frac{1}{|G_k|} \sum_{d_i \in G_k} \mathbf{u}_i$, where $\mathbf{u}$ is the individual utility scores of each of the documents in the group G_k.

The Disparate Impact constraint builds on Disparate Treatment with the additional constraint that the clickthrough rates for the groups, as determined by their exposure and relevance, are proportional to their average utility. Disparate Impact is defined as:

$$\frac{\text{CTR}(G_0 \mid \mathbf{P})}{\text{U}(G_0 \mid q)} = \frac{\text{CTR}(G_1 \mid \mathbf{P})}{\text{U}(G_1 \mid q)} \tag{2}$$

where the average clickthrough rate of a group, $\text{CTR}(G_0 \mid \mathbf{P})$, is defined as:

$$\text{CTR}(G_k \mid \mathbf{P}) = \frac{1}{|G_k|} \sum_{i \in G_k} \sum_{j=1}^{N} \mathbf{P}_{i,j} \mathbf{u}_i \mathbf{v}_j \tag{3}$$

for N documents with utility, $\mathbf{u}$, and attention (i.e., exposure drop-off), $\mathbf{v}$. The probability of a document being clicked is calculated using the click model of Richardson et al. (2007) as follows:

$$\begin{aligned} \text{P}(\text{click on document } i) &= \text{P}(\text{examining } i) \times \text{P}(i \text{ is relevant}) \\ &= \text{Exposure}(d_i \mid \mathbf{P}) \times \text{P}(i \text{ is relevant}) \\ &= \left(\sum_{j=1}^{N} \mathbf{P}_{i,j} \mathbf{v}_j \right) \times \mathbf{u}_i \end{aligned} \tag{4}$$

In our experiments, we view the Disparate Treatment and the Disparate Impact constraints as the *target* exposures for each of the fairness groups, G_0 and G_1, under two different fairness constraints, and measure how much a sequence of rankings violates each of the constraints.

Many approaches in recent years have tried to ensure that items, e.g., documents, that represent particular societal groups, e.g., gender or ethnicity, receive a fair exposure within a single ranking. However, queries are often searched repeatedly (either by the same user over a period of time or by multiple users) and if the same static ranking is produced for each instance of the

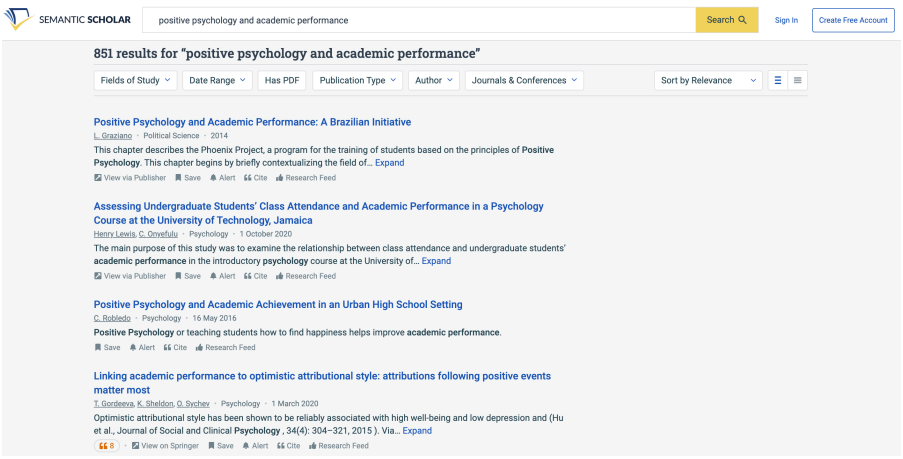

Fig. 1 An example academic search engine UI. The figure illustrates the *10-blue-links* UX of the Semantic Scholar search engine.

query then inequalities of exposure can emerge over time. For example, consider a fairness policy that ensures that, in a single ranking, 50% of relevant documents are by authors from a protected societal group. If, for a particular query, the highest ranked 5% of the documents are not by authors from the protected group, then although the single ranking could be considered to be fair, over time the documents in the top-ranked positions would cumulatively receive more exposure than the authors from the protected group that have also produced relevant documents. However, there is also the potential to compensate for any under-exposure of the documents in previous rankings if the search engine introduces a fair ranking policy (Biega et al., 2018). This scenario, where authors receive exposure to users over repeated queries, is addressed in the context of academic search by the TREC Fair Ranking Track (Biega et al., 2020).

Academic search addresses the scenario in which a search engine indexes scientific articles, such as research papers, books or theses, that have been published in academic journals or the proceedings of scientific conferences. Examples of such search engines include Semantic Scholar,[1] Google Scholar[2] and the Cornell University search engine arXiv.[3] Typically, academic search engines adhere to the *ten-blue-links* UX presentation strategy for presenting a user with the results of their query, Figure 1 provides an illustration of the user interface of the Semantic Scholar search engine.

In an academic search scenario, a fair ranking should not be dominated by a single author or institution if there are equally relevant papers from other authors or institutions. Moreover, an author or institution should not receive a disproportionately high (or low) exposure over time, compared to other authors or institutions that produce relevant papers. For example, in response to the query *information retrieval*, the results at the top rank positions should not all be from the same research group, since there will be multiple groups

[1] `https://semanticscholar.org` [2] `https://scholar.google.com` [3] `https://arxiv.org`

that have produced relevant papers. Moreover, as a query is repeated over time, different relevant information sources should be given the opportunity to appear in higher ranked positions.

We argue that generating a fair ranking towards information sources can be cast as a search results diversification (Santos et al., 2015) task. In search results diversification, an IR system aims to maximise the number of sub-topics, or *aspects*, of an ambiguous query that are represented within the results list. We postulate that fair rankings can be generated by viewing the characteristics of groups that we aim to be fair to as latent aspects of relevance and maximising the number of such groups that are represented within the search results.

In this work, we evaluate three well-known search results diversification approaches from the literature as fair ranking strategies. Our experiments on the 2019 and 2020 TREC Fair Ranking Track datasets show that explicit search results diversification is particularly effective for generating rankings that provide a fair exposure for authors from protected societal groups, when the definition of the protected groups are *unknown*. In particular, we show that leveraging xQuAD (Santos et al., 2010) search result diversification as a fair ranking strategy can result in a significant ($p < 0.05$) increase (up to 50% in our experiments) in the fairness of exposure that authors from unknown protected groups receive, when exposure is evaluated for multiple instances of a repeated query. Indeed, the tailored xQuAD search result diversification model with our proposed *assumed* fairness groups was the best performing system submitted to the TREC 2019 Fair Ranking Track in terms of fairness, for both of the official TREC evaluation groupings. Moreover, in this work, we show that diversifying over assumed fairness groups that model the topical contents of documents is a particularly promising approach and can result in significantly increased group fairness, relative to the relevance of the documents from the group, compared to when the ranking is optimised for relevance only.

The remainder of this paper is as follows: In Section 2 we discuss prior work on fairness in information access systems. We introduce the fair ranking task in Section 3 before defining our proposed assumed fairness groups in Section 4. We present how we propose to cast fair ranking as a search result diversification task in Section 5 before presenting our experimental setup in Section 6, then our results in Section 7. Concluding remarks follow in Section 8.

2 Related Work

In this section, we firstly discuss work related to fairness in classification systems and search engines, before presenting prior work on search results diversification.

Fairness: Most of the previous work on measuring or enforcing fairness in information access systems has focused on fairness in machine learning classifiers. Such classifiers might be deployed in decision-making tasks, such as loan or parole applications (Chouldechova, 2017; Hardt et al., 2016; Kleinberg

et al., 2016; Woodworth et al., 2017; Zafar et al., 2017), where discrimination amongst individuals or sections of society can have serious implications for those who are discriminated against. Many approaches for developing fair classifiers have focused on removing bias from the data that the classifier is trained on, e.g. (Hajian and Domingo-Ferrer, 2013; Kamiran and Calders, 2009; Zemel et al., 2013), or from external resources, such as word embeddings (Bolukbasi et al., 2016). However, the majority of the literature on fair classification focuses on enforcing fairness constraints in the classifier's predictions, for example (Dwork et al., 2012; Calders and Verwer, 2010; Pleiss et al., 2017; Woodworth et al., 2017; Zafar et al., 2017).

There are two main notions of fairness that classifiers typically try to integrate. Firstly, in *individual fairness* (Dwork et al., 2012), a classifier's probability (or confidence) score should be comparable for all individuals irrespective of the classification group that they truly belong to. For example, in the case of loan repayment, for all subjects (i.e. individuals) a classification score of 0.8 should represent the same likelihood of repayment irrespective of the actual class that the subject belongs to.

The second notion of fairness that is commonly enforced in classification systems is *group fairness*. Examples of group fairness include *statistical parity* and *Balancing for the positive/negative class*. For statistical parity, e.g. as in (Calders and Verwer, 2010; Kamiran and Calders, 2009; Kamishima et al., 2011), equal percentages of each of the protected groups should be classified as belonging to the positive class. Differently, in balancing for the positive/negative class (Kleinberg et al., 2016), the average prediction score for the positive/negative class should be the same for each of the protected groups.

The ethical implications, and the potential effects on society, that arise from search engines have been recognised for many years, e.g. see (Belkin and Robertson, 1976). Variations in queries issued by different demographic groups can result in, for example, differences in satisfaction levels between older and younger users (Mehrotra et al., 2017). However, there has been much less work on encoding or measuring fairness in search engines compared to classification systems (Castillo, 2018). Lately, integrating fairness into ranking algorithms has received more attention in the literature, for example at the inaugural FACTS-IR workshop (Olteanu et al., 2019a,b) and the TREC Fair Ranking Track (Biega et al., 2020). Indeed, at the recent Strategic Workshop on Information Retrieval in Lorne (Culpepper et al., 2018), fairness in search systems was identified as one of the most important emerging topics for IR research.

There are four main differences in the ways in which fairness is implemented and evaluated in search systems compared to classification systems (Ekstrand et al., 2019). Firstly, evaluating a search system requires a *user model*. For example, the usefulness of a search result can depend on the other results that a user has previously looked at. Moreover, the probability that a user will view any particular document will also vary depending on the position that the document appears in the ranking and how far down the ranked list the

user is prepared to look for relevant documents. Secondly, search queries can be repeated within the same, or over multiple, search sessions. This provides an opportunity to compensate for any unfairness in the results of previous instances of the query. Thirdly, the desired outcome of a search system, i.e. relevance or utility, is a subjective notion that can be confounded by other intentions of the system, such as the personalisation of results. Fourthly, search engines have multiple sets of stakeholders that each have their own fairness concerns. For example, information consumers (i.e. users) want fairness in access to information, whereas information producers want a fair opportunity to be discovered by users (Mehrotra et al., 2018).

Most of the work on fairness in ranking systems has investigated latent biases in search engines, e.g. (White, 2013; Baeza-Yates, 2018; De-Arteaga et al., 2019) or correcting for biases in learning to rank scenarios, e.g., (Singh and Joachims, 2019; Yadav et al., 2019; Morik et al., 2020). However, recently there has also been an interest in developing ranking algorithms that aim to enforce group fairness through fairness constraints that require the ranker to assign a certain portion of the top rank positions to members of protected or minority classes, e.g. (Zehlike et al., 2017; Celis et al., 2018; Singh and Joachims, 2018). Differently from constraint-based approaches, which rely on the protected groups being known a priori, in this work we propose to cast the fair ranking task, where the protected groups are *unknown* a priori, as a search results diversification task. Indeed, Gao and Shah (2020) recently showed that diversity and relevance are highly correlated with statistical parity fairness.

Identifying the most appropriate method of evaluating fairness in systems that output ranked results is a developing area of research (Diaz et al., 2020). Early work on developing fair ranking metrics focused on directly applying fairness approaches from classification, such as statistical parity (Yang and Stoyanovich, 2017) and group fairness (Sapiezynski et al., 2019). Biega et al. (2018) proposed a fairness evaluation metric akin to evaluating individual fairness (Dwork et al., 2012). Their approach evaluated position bias and was modelled on the premises that (1) the attention of searchers should be distributed fairly and (2) information producers should receive attention from users in proportion to their relevance to a given search task. To account for the fact that no single ranking can achieve individual fairness, the authors introduced *amortized fairness*, where attention is accumulated over a series of rankings. In this work, we are interested evaluating the fairness of the exposure that authors from protected groups are likely to receive in a ranking. Singh and Joachims (2018) introduced the Disparate Treatment Ratio (DTR) and Disparate Impact Ratio (DIR) metrics to evaluate such a scenario. Therefore, we select to use these metrics to evaluate our proposed approaches. We present full details of DTR and DIR in Section 6.

Search Results Diversification: Queries submitted to a Web search engine are often short and ambiguous (Spärck-Jones et al., 2007). Therefore, it is often desirable to diversify the search results to include relevant documents for mul-

tiple *senses*, or *aspects*, of the query. There are two main families of search result diversification approaches, namely *implicit* and *explicit* diversification.

Implicit diversification approaches, for example (Carbonell and Goldstein, 1998; Chen and Karger, 2006; Wang and Zhu, 2009; Radlinski et al., 2008), assume that documents that are similar in content will cover the same aspects of a query. Such approaches increase the coverage of aspects in a ranking by demoting to lower ranked documents that are similar to higher-ranked documents. Maximal Marginal Relevance (MMR) (Carbonell and Goldstein, 1998) was one of the first implicit diversification approaches. MMR aims to increase the amount of novel information in a ranked list by selecting documents that are dissimilar to the documents that have already been selected for the results list. In this work, we evaluate MMR as a fair ranking strategy based on implicit diversification. However, differently from when MMR is deployed for search result diversification, e.g. as in (Carbonell and Goldstein, 1998), we instead evaluate the effectiveness of selecting documents from information sources that have dissimilar characteristics. We provide details of how we build on MMR in Section 5.1.

Explicit diversification approaches, e.g. (Santos et al., 2010; Radlinski and Dumais, 2006; Agrawal et al., 2009; Dang and Croft, 2012), directly model the query aspects with an aim to maximise the coverage of aspects that are represented in the search results. For example, eXplicit Query Aspect Diversification(Santos et al., 2010) (xQuAD) uses query reformulations to represent possible information needs of an ambiguous query and iteratively generates a ranking by selecting documents that maximise the novelty and diversity of the search results. Dang and Croft (Dang and Croft, 2012) proposed an explicit approach, called PM-2, that was based on *proportional representation*. The intuition of PM-2 is that, for each aspect of an ambiguous query, the number of documents relating to the aspect that are included in the search results should be proportional to the number of documents relating to the aspect in a larger ranked list of documents that the search results are sampled from. For example, for the query *java*, if 10% of the documents in the list of ranked documents that the search results are sampled from are about *java the island* then 10% of the search results should also be about *java the island*. In this work, we build on the xQuAD (Santos et al., 2010) and PM-2 (Dang and Croft, 2012) explicit diversification approaches to generate fair rankings. However, differently from the work of (Santos et al., 2010) and (Dang and Croft, 2012), we explicitly diversify over the characteristics of *assumed* groups that we wish to be fair to. We provide details of how we leverage xQuAD and PM-2 in Section 5.2.

Castillo (2018) argued that search results diversification differs from fair ranking in that the former focuses on utility for the searcher while the latter focuses on the utility of sources of relevant information. We agree that intuitively the tasks are different. However, in this work, we postulate that fair rankings that also provide utility for the search engine users can be generated by diversifying over a set of assumed groups that we aim to be fair to, to maximise the representation of such groups in the search results.

roman chamomile oil
drones
bystander effect
positive psychology and academic performance
cost management theory

Table 1 Examples of informational queries in the context of academic search. The queries are topics from the TREC 2019 Fair Ranking Track and are real queries from the Semantic Scholar academic search engine.

3 Fair Ranking Task

In this section, we introduce the fair ranking task and provide a formal definition of the problem, as it is set out by the TREC Fair Ranking Track (Biega et al., 2020, 2021). As previously stated in Section 1, the task is set within the context of academic search, i.e., in response to an informational query (Broder, 2002) (i.e., there may be multiple relevant documents for each query), a search system should return a ranked list of relevant published (or potentially pre-published) research papers. Table 1 provides examples of such informational queries in the context of academic search. The queries presented in Table 1 are queries from the TREC 2019 Fair Ranking Track test collection (Biega et al., 2020). They are real users' queries from the query logs of the Semantic Scholar academic search engine. For a search system to be fair to the information producers, for such generic topic-based queries, the generated rankings should not be dominated by a single author or institution. Moreover, an author or institution should not receive a disproportionately high (or low) amount of exposure to users over a period of time, compared to relevant work from other authors or institutions. For example, adapting the simplified example from Singh and Joachims (2018), we consider a ranking of six documents that are all judged to be relevant by the search engine's users. The authors of 3 of the documents are from *institution A* and the authors of the other 3 documents are from *institution B*. The search engine's *estimated* relevance scores for the documents from institution A are 0.80, 0.79 and 0.78 while the estimated relevance scores for the documents from institution B are 0.77, 0.76 and 0.75. Following Singh and Joachims (2018), to allow meaningful argument on their relative difference, we assume that the estimated relevance scores are probabilities. If the documents are ranked in accordance with the Probability Ranking Principle (Robertson, 1977), according to position bias user model of the DCG measure, the exposure drop-off of a document at position j in the ranking is $\text{Exposure}(Pos_j) = \frac{1}{\log(1+Pos_j)}$. In other words, we expect users to start at the top of the ranking and consider the relevance of each document in-turn. The further down the ranking a relevant document appears, the more likely it is that a user will have stopped before reaching the relevant document. In our example ranking, the documents from institution B will receive 30% less exposure than the documents from institution A. However, the difference in average estimated relevance between the documents from institution A and institution B is just 0.03. In other words, a small difference in estimated rel-

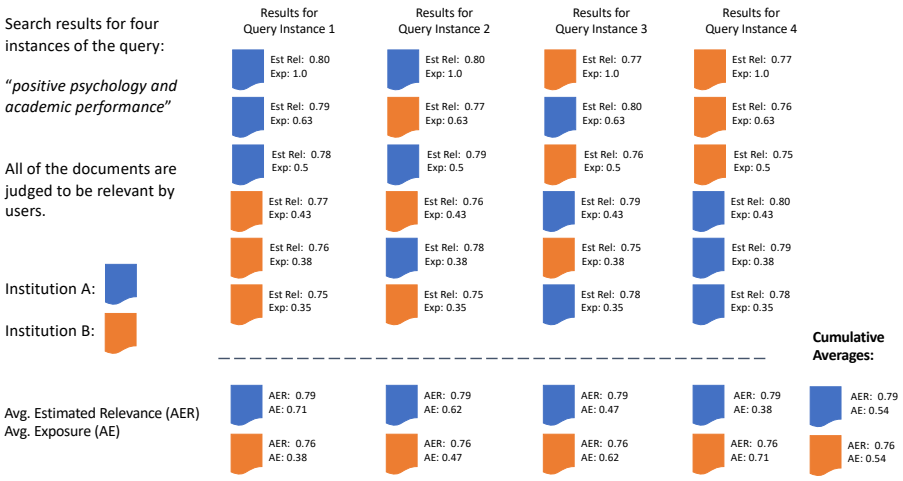

Fig. 2 Illustration of exposure (Exp) vs estimated relevance (Est Rel) in the context of the fair ranking task, where the fairness of the search results are evaluated over multiple instances of a repeated query.

evance can lead to a large difference in the exposure (or opportunity to be seen by the users) that the documents receive. Moreover, if the same ranking is displayed to users each time the query is issued to the search engine, this disparity in exposure will increase over time.

With this in mind, fairness in ranking systems should be evaluated over sequences of queries (Biega et al., 2018) to enable a system to address any potential unfairness that might have been present in the results of a previous instance of a repeated query (Ekstrand et al., 2019), such that the rankings are both (1) relevant to the users and (2) fair to the information produces. The fair ranking task addresses such a scenario. In response to repeated instances of a query, a fair IR system should therefore output a sequence of ranked results that balances the trade-off between maximising the relevance (or *utility*) of the results for the user and minimising any *unfairness* in the exposure that the information producers get over the sequence of queries. In practice, there are three main approaches that IR systems typically take for balancing this trade-off, namely (1) optimising for relevance while enforcing fairness constraints (Zehlike et al., 2017; Celis et al., 2018), (2) deploying fairness-focused regularisation (Mehrotra et al., 2018) or (3) jointly optimising for relevance and fairness (Mehrotra et al., 2018).

Figure 2 illustrates how the unfair exposure that the documents in our example receive can be addressed within the context of the fair ranking task. As illustrated in Figure 2, the first instance of the query returns the documents ranked as we previously described. However, in subsequent instances of the query, the system tries to compensate for any unfairness of exposure by re-ordering the presentation of the documents to the user(s). In our illustration, after four instances of the query have been submitted to the search engine the average cumulative exposure for institutions A and B is equal. Hence, the

exposure that the two institutions receive are equal to their relative relevances (as judged by the user(s)).

It is important to note that in this simplified example the groups that we wish to be fair to are known. However, in the fair ranking task the protected groups are *unknown*. The fair ranking task is, therefore, defined as follows: For a query, q_i, with associated set of candidate documents, r_i, generate a sequence of rankings, s_i, for repeated instances of the query q_i, where $r_{i,j}$ is the generated permutation of r_i for the j'th instance of q_i, such that items that are relevant for q_i get a fair exposure to users over the sequence of rankings s_i. The initial set of candidate documents r_i for the query q_i is the same for every instance of q_i. However, the number of candidate documents associated to different queries can vary.

4 Defining Fairness Groups for Academic Search Fairness

In this section, we provide details of the assumed fairness groups that we propose to use for generating fair rankings. In the context of academic search, the information sources that we aim to be fair to are the authors of papers. In particular, we aim to be fair to members of sub-groups of *unknown* groups of authors, where each group is defined by a certain demographic or characteristic. For example, a gender grouping would include sub-groups for *male* authors and *female* authors (and potentially other additional sub-groups depending on the definition of the grouping). In this example, male authors and female authors should both receive a fair exposure in the ranking.

There are many characteristics of authors that it may be desirable to provide a fair exposure to. For example, we may wish to be fair to the characteristic *experience* and ensure that the search results provide a fair exposure to the sub-groups *early career researcher* and *highly experienced / professorial*. In fact, the available options for selecting characteristics, or groups, that we would like to be fair to are potentially infinite. Moreover, there are characteristics of a document that can potentially be used to uncover latent author characteristics. For example, the specific topics that a document is about could be a good indicator of an author's main research interests. Therefore, our approach to fairness is based on defining fairness groups by identifying attributes of a document, i.e., a document's authors, publication venue or topics, that have characteristics that, we argue, are intuitively desirable to be fair to in the context of academic search. We note that, in practice, our assumed fairness group definitions may or may not match the *unknown* protected groups. Indeed, this is akin to the official vs. system generated sub-topics in search result diversification (Santos et al., 2015).

To generate fair rankings with respect to an assumed fairness group, for each paper that is to be ranked, we need a list of scores that represent the amount that a specific instance of a document attribute (i.e., an author, a publication venue or a topic) represents a characteristic that we wish to be fair to (i.e., *experience* in the case of authors, *popularity / exposure* in the

case of publication venues, or *aboutness* in the case of topics). A high (or low) score is not intended as a measure of how good (or bad) a paper is but is an indicator of how *representative* the paper is of a particular characteristic of an assumed fairness group. For example, when considering topics, each score represents the probability that a document is about a particular topic. Each of our proposed approaches that we present in this section outputs a list of such scores that can be directly used as an input for each of the diversification approaches that we present later in Section 5.

In the remainder of this section, we provide details of, and define, our three proposed approaches for generating assumed fairness groupings for evaluating the effectiveness of search results diversification for generating fair rankings in the context of academic search.

4.1 Author Experience

The first assumed fairness group that we propose aims to provide a fair exposure to authors that are at different stages of their careers: e.g. early career researchers vs. highly experienced researchers. Intuitively this approach aims to reduce the preponderance of individual authors at the top of the ranking for a given query. Hence, prolific or highly experienced authors should not overwhelm other authors in the ranking and relevant work that is produced by early career researchers should receive a fair exposure to the users.

For defining this fairness group, we need a score to represent the amount of *experience* that an author has. There are many possible signals that could be used as proxies for estimating the amount of experience that an author has. For example, the number and/or dates of the author's publications, or the dates and/or trends of their citations. In this work, to estimate an author's experience we use the total number of citations that the author has in the collection, calculated as follows:

$$\text{Experience}(a) = \sum_{d \in D_a} \text{Citations}(d) \tag{5}$$

where d is a document authored by a in the set of all documents, D_a, that a is an author of and Citations(d) is the number of documents that cite d. A given document, d, is then represented as a list of author experience scores, one for each of the authors of d. This document representation can then be used as input to the search results diversification approaches that we evaluate. Approaches that use our *Author Experience* assumed fairness group for diversification are denoted with the subscript A in Section 7.

In this approach, the group characteristic that we are aiming to be fair to is the authors' experience and the sub-groups of this assumed fairness group, $g_i \in G_A$, are *early career researchers* and *highly experienced / professorial* researchers. Authors with a high Experience(a) score are likely to be more senior researchers that have accumulated more citations over time.

4.2 Journal Exposure

The second assumed fairness group that we propose aims to provide a fair exposure to papers from different publication venues in the corpus. Intuitively, this approach aims to surface relevant search results from publication venues that may usually be underrepresented by search systems. For example, within the IR research field, the search results for a particular query may be unfairly dominated by papers from conference proceedings, e.g. SIGIR, and journals that output a lot of material, while other sources, e.g. smaller journals or TREC notebooks, may be underrepresented with respect to their relevance or utility to the user. Moreover, it is also the case that papers that are published in more well-known venues, such as the ACM Digital Library,[4] are likely to unfairly benefit from rich-get-richer dynamics, compared to lesser-known venues. Therefore, our Journal Exposure strategy aims to provide a fair exposure to papers from venues that are underrepresented in the search results.

In this approach, the group characteristic that we are aiming to be fair to is the paper's exposure through publication venues. For a given document, d, and the publication venues (e.g. journals), V, that d is published in, the journal coverage score, Coverage(v), for a venue, $v \in V$, is the total number of documents (in the collection) that are published in v, i.e, the coverage of the publication venue. For our proposed Journal Exposure assumed fairness group, d is then represented as a list of Coverage(v) scores, one for each of the venues that d is published in, and the list of scores are input into the diversification approaches that we evaluate.[5] The sub-groups of this assumed fairness group, $gi \in G_A$, that we aim to be fair to are *low coverage* and *high coverage* publication venues. Approaches that use our *Journal Exposure* assumed fairness group for diversification are denoted with the subscript J in Section 7.

4.3 Topical Grouping

The third, and final, assumed fairness group that we propose aims to provide a fair exposure to different authors that publish papers on the same research topics. Our intuition for this grouping is that a query's results may be dominated by, for example, an author that primarily publishes work on a particularly popular sub-topic of the query. This may be problematic for broadly defined queries that can have multiple relevant sub-topics, as per the examples that were presented in Table 1. For example, for the query *interactive information retrieval* (IIR), it may be the case that the retrieved results are dominated by documents that discuss IIR *user studies*. Moreover, these results may be

[4] https://dl.acm.org/ [5] In the TREC Fair Ranking Track test collection, each paper is only published in a single venue. Therefore, in this work, there is only one non-zero value in the list for each document (the size of the list is equal to the number of publications venues in the collection). However, in practice, it is often the case that a paper can be available through multiple venues (for example through the ACM digital library and through arXiv). Our proposed approach handles such a case without any adaptation.

further dominated by an author that is particularly well-known for IIR user studies. However, other sub-topics of IIR are also likely to be relevant, for example *user modelling* or *interaction simulation*, and the authors that publish in these fields may be under-exposed. We expect that diversifying the rankings over the topics that are discussed in the retrieved documents will likely give more exposure to such under-exposed authors. To account for this potential disparity in exposure for different sub-topics within the relevant search results, we build on a topic modelling approach to generate topical groupings based on the textual content of the documents. We note that our topical grouping may not be the most appropriate approach for very specific or narrowly defined queries, such as when a searcher is looking for papers to cite. However, as previously discussed in Section 3, the majority of the queries in the TREC Fair Ranking task, which are from the query logs of the Semantic Scholar academic search engine, are broadly-defined informational queries.

For generating our topical groupings, we use a topic modelling approach to identify the main latent topics that are discusses in the documents' text. The topical grouping is defined as the probability that a document, d, discusses a latent topic, $z_i \in Z$, where Z is the set of latent topics that are discussed in all of the documents in a collection. A topic, z_i, can then be seen as a group characteristic that we wish to be fair to.

In this approach, the group characteristic that we are aiming to be fair to is the topics that the paper is about. The sub-groups of this assumed fairness group, $gi \in G_A$, are, therefore, the topics that are discussed by the papers in the collection. With this in mind, when diversifying over topics, a document is represented by its top k topics, where the document's score for a top k topic, z_i, is $p(z_i \mid d)$. The document's score for all other topics that are discussed in the collection is 0. In other words, each document is associated with a fixed number of topics, k, and the probability that a document discusses each of the k topics varies. Approaches that use our *Topical* assumed fairness group for diversification are denoted with the subscript T in Section 7.

5 Casting Fair Ranking as Search Results Diversification

As previously discussed in Section 1, we argue that generating a fair ranking towards information sources can be cast as a search results diversification task, by viewing the characteristics of groups that we aim to be fair to as latent aspects of relevance and maximising the number of groups that are represented within the top rank positions of the search results (as is the objective of search results diversification). In this section, we present the three search results diversification approaches from the literature that we build on and how we propose to adapt and tailor each of them to generate fair rankings. Sections 5.1 presents the *implicit* diversification approach that we evaluate, while 5.2 presents the two *explicit* diversification approaches that we evaluate.

5.1 Implicit Diversification for Fairness

As our implicit diversification approach to fairness, we leverage the well-known Maximal Marginal Relevance (Carbonell and Goldstein, 1998) (MMR) diversification approach. At each iteration of the algorithm, MMR selects from the remaining documents the one with the maximal marginal relevance, calculated as follows:

$$\text{MMR} \stackrel{\text{def}}{=} \text{Arg} \max_{d_i \in r \setminus s} \left[\lambda \left(\text{Sim}_1 (d_i, q) - (1 - \lambda) \max_{d_j \in s} \text{Sim}_2 (d_i, d_j) \right) \right] \tag{6}$$

where q is a query, r is a ranking of the subset of documents in the collection that are candidate documents with respect to their relevance to q, s is the subset of documents in r that have already been selected and $r \setminus s$ is the set of documents in r that have not been selected. Sim_1 is a metric that measures the relevance of a document w.r.t. a query and Sim_2 is a metric to measure the similarity of a document and each of the previously selected documents.

To leverage MMR as a fairness component, we define the dissimilarity function, FairSim, for two document representations, $\mathbf{d_i}$ and $\mathbf{d_j}$, that are output from one of our proposed approaches presented in Section 4, as follows:

$$\text{FairSim}(\mathbf{d_i}, \mathbf{d_j}) = \frac{\sum_{a \in A_i \cap A_j} \mid a_i - a_j \mid}{\mid A_i \cap A_j \mid} \tag{7}$$

where $a \in A_i \cap A_j$ is the set of attributes that are common to both $\mathbf{d_i}$ and $\mathbf{d_j}$, $\mid a_i - a_j \mid$ is the absolute difference in the scores for a particular attribute, and $\mid A_i \cap A_j \mid$ is the number of attributes that are common to $\mathbf{d_i}$ and $\mathbf{d_j}$. If d_i and d_j do not have any common attributes then $\text{FairSim}(\mathbf{d_i}, \mathbf{d_j}) = 0$. FairSim is designed to identify to what extent two documents represent the same (or similar) fairness sub-group of an assumed fairness group ($g \in G_A$). In practice, an attribute is an index of $\mathbf{d}$ that has a non-zero value. However, when deploying $\text{FairSim}(\mathbf{d_i}, \mathbf{d_j})$ for our Author Experience grouping, we assume that any index that is non-zero in either $\mathbf{d_i}$ or $\mathbf{d_j}$ is non-zero in both $\mathbf{d_i}$ and $\mathbf{d_j}$.

For example, when deploying our topical groupings presented in Section 4.3, a document can discuss many topics, e.g. *user studies* and *user modelling*. The score that represents how much the document is about a particular topic, z, is the probability of observing the topic given the document d_i, $p(z_i \mid d)$. If $\mathbf{d_i}$ and $\mathbf{d_j}$ have three common topics and the topic scores for each are $\mathbf{d_i} = \{0.3, 0.3, 0.3\}$ and $\mathbf{d_j} = \{0.2, 0.2, 0.2\}$, then $\text{FairSim}(\mathbf{d_i}, \mathbf{d_j}) = \frac{(0.3-0.2)+(0.3-0.2)+(0.3-0.2)}{3} = 0.1$. However, if the scores are $\mathbf{d_i} = \{0.3, 0.3, 0.3\}$ and $\mathbf{d_j} = \{0.1, 0.1, 0.1\}$, then $\text{FairSim}(\mathbf{d_i}, \mathbf{d_j}) = \frac{(0.3-0.1)+(0.3-0.1)+(0.3-0.1)}{3} = 0.2$. In other words, the documents in the second example are less similar than the documents in the first example because there is a greater difference in how strongly they are related to their common topics.

In practice, this is could be viewed as a hybrid approach (as apposed to a purely implicit diversification approach) since instead of calculating the similarity over the entire text of a document, as in the original MMR formulation,

MMR calculates the documents' similarities based on the characteristic scores that are output from our proposed assumed fairness groups approaches. MMR then generates fair rankings by selecting the documents that are most dissimilar from the previously selected documents, with respect to the characteristics of assumed fairness groups.

5.2 Explicit Diversification for Fairness

The first explicit diversification approach that we build on is xQuAD (Santos et al., 2010). xQuAD is a probabilistic framework for explicit search result diversification that guides the diversification process of an ambiguous query through a set of sub-queries, each having a possible interpretation for the original query. For a given query, q, and an initial ranking, r, xQuAD builds a new ranking, s, by iteratively selecting the highest scored document from r with the following probability mixture model:

$$(1 - \lambda)\mathrm{P}(d \mid q) + \lambda\mathrm{P}(d, \bar{s} \mid q) \tag{8}$$

where $\mathrm{P}(d \mid q)$ is the estimated relevance of a document, d, with respect to the initial query q, and $\mathrm{P}(d, \bar{s} \mid q)$ is the *diversity* of d with respect to s, i.e., how relevant d is to the subtopic queries that are least represented in s. xQuAD's objective is to cover as many of the interpretations of the queries in the search results, while also ensuring novelty. To generate fair rankings using xQuAD, we leverage the fact that, for a given sub-query, q_i, $\mathrm{P}\left(d, \bar{s} \mid q_i\right)$ is calculated as:

$$\mathrm{P}\left(d, \bar{s} \mid q_i\right) = \mathrm{P}\left(d \mid q_i\right) \mathrm{P}\left(\bar{s} \mid q_i\right) \tag{9}$$

where $\mathrm{P}\left(d \mid q_i\right)$ is the probability of document d being relevant to the sub-query q_i and $\mathrm{P}\left(\bar{s} \mid q_i\right)$ provides a measure of *novelty*, i.e. the probability of q_i not being satisfied by any of the documents already selected in s. We view the documents' attributes, i.e., authors, publication venues or topics, as sub-queries and a document attribute's characteristic score as a measure of the attribute's fairness sub-group coverage. We then calculate the relevance and novelty of a document, d, with respect to a fairness sub-group, g_i, as follows:

$$\mathrm{P}\left(d, \bar{s} \mid g_i\right) = \mathrm{P}\left(d \mid g_i\right) \mathrm{P}\left(\bar{s} \mid g_i\right) \tag{10}$$

where $\mathrm{P}\left(d \mid g_i\right)$ is the probability of document d being associated to the group g_i and $\mathrm{P}\left(\bar{s} \mid g_i\right)$ is the probability of g_i not being associated to any of the documents already selected in s. $\mathrm{P}\left(\bar{s} \mid g_i\right)$ is obtained using $1 - P(s \mid g_i)$, while $P(s \mid g_i)$ is directly observable.

In other words, xQuAD iteratively adds documents to s by prioritising (1) documents that belong to assumed fairness sub-groups that have relatively few documents belonging to them, $\mathrm{P}\left(d \mid g_i\right)$, and (2) documents that belong to fairness sub-groups that do not have many documents belonging to them in the partially constructed ranking s, $\mathrm{P}\left(\bar{s} \mid g_i\right)$. For example, when deploying our topical assumed fairness groups, if there are relatively few documents that

belong to the latent topic *interaction simulation* then these documents will be prioritised for selection, unless relatively many of the documents that have previously been selected for s also belong to this assumed fairness group. In that case, documents that belong to another, less rare but underrepresented, group will be prioritised. In doing so, the coverage of the assumed fairness groups is maximised in the top rank positions by promoting documents that belong to underrepresented groups.

We now move on to discuss the second explicit diversification approach that we build on for generating fair rankings, namely PM-2 (Dang and Croft, 2012). PM-2 is a proportional representation approach that aims to generate a useful and diversified ranked list of search results (of any given size) by sampling documents from a larger list of documents that have been ranked with respect to their relevance to a user's query. The aim of PM-2 is to generate a list of search results in which the number of documents relating to a query aspect, $a_i \in A$, that are included in the search results are proportional to the number of documents relating to the aspect in the larger list of documents that the search results are sampled from. In other words, for the query *java*, if 90% of the documents in the larger ranked list of documents are about *java the island* then 90% of the search results should also be about the island.

PM-2 selects documents to add to the ranking, s, as follows:

$$d \leftarrow \lambda \times qa\left[i^*\right] \times \mathrm{P}\left(d_j \mid a_{i^*}\right) + (1-\lambda) \sum_{i \neq i^*} qa[i] \times \mathrm{P}\left(d_j \mid a_i\right) \qquad (11)$$

where $qa\left[i^*\right]$ is $\frac{v_i}{2s_i+1}$, v_i is the number of documents that discuss aspect a_i, s_i is the number of rank positions that are assigned to a_i (proportional to the popularity of a_i in the larger list of documents that the search results are sampled from), $\mathrm{P}\left(d_j \mid a_i\right)$ is a document's fairness characteristic score for aspect, a_i, and a_{i^*} is an aspect that has already been selected for s.

When generating fair rankings with PM-2, we view the documents' attributes as the query aspects, $a_i \in A$. We view the proportionality of an aspect, a_i, as the fraction of documents in the whole document collection, D, that also contain the aspect and replace v_i in Equation (11) with the probability, $\mathrm{p}(a_i \mid D)$, of the aspect, a_i, in the collection D – i.e. the fraction of documents in the collection that contain a_i.

In other words, for each of the attributes in an assumed fairness group in turn, documents that have a relatively large characteristic score for that attribute, but also have characteristic scores for many attributes, are prioritised for selection in the ranking until the allocated portion of the ranking, s (proportional to the frequency of the attribute in the entire collection), is filled by documents that contain a_i. As a consequence, this ensures the promotion of documents that contain group fairness attributes, which are underrepresented with respect to their proportionality in the collection.

Table 2 Per-query statistics for relevant and non-relevant candidate documents for the 2019 and 2020 TREC Fair Ranking Track evaluation queries.

	2019 Collection				2020 Collection			
	Max	Min	Mean	Std.	Max	Min	Mean	Std.
Relevant	20	1	3.35	1.50	14	2	3.48	2.43
Non-relevant	25	0	2.48	3.48	299	5	20.58	22.66
All Candidate Documents	32	5	6.83	2.73	312	10	24.07	23.66

6 Experimental Setup

In this section, we present our experimental setup for evaluating the effectiveness of leveraging search results diversification to generate fair rankings of search results. We aim to answer the following research questions:

- RQ1: Is leveraging search results diversification as a fairness component effective for generating fair rankings?
- RQ2: Which family of search results diversification, i.e. explicit vs. implicit, is most effective as a fairness component?
- RQ3: Does diversifying over multiple assumed fairness groupings results in increased fairness?

We evaluate our research questions on the test collections of the 2019 and 2020 TREC Fair Ranking Tracks. As previously stated in Section 3, it is appropriate to evaluate the fairness of an IR system over a sequence of possibly repeating queries to allow the system to correct for any potential unfairness in the results of previous query instances. The TREC Fair Ranking Track is designed to evaluate such a scenario within the context of an academic search application.

The 2019 and 2020 Fair Ranking Track test collections both consist of documents (academic paper abstracts) sampled from the Semantic Scholar (S2) Open corpus (Ammar et al., 2018) from the Allen Institute for Artificial Intelligence,[6] along with training and evaluation queries. Both of the collections are constructed from the same 7903 document abstracts. However, each of the collections have a different set of queries. The approaches that we evaluate in this work are all unsupervised approaches. Therefore, we use the evaluation queries from each of the collections. The task is setup as a re-ranking task, where each of the queries has an associated set of candidate document with relevance judgements and fairness group ground truth labels. The number of candidate documents that are to be re-ranked varies per-query, ranging from 5 to 312. Table 2 provides statistics about the (per-query) candidate documents for the evaluation collections. There are 4040 documents that have relevance judgements for the 635 evaluation queries of the 2019 collection and 4693 documents have relevance judgements for the 200 evaluation queries of the 2020 collection. In our experiments, we evaluate our approaches over 100 instances of each of the queries.

The collections include relevance assessments for two *unknown* evaluation fairness groups, i.e. the groups were not known by the track participants and

[6] https://allenai.org/

did not influence our proposed fairness approaches. Both of the evaluation fairness groups define 2 sub-groups that a system should be fair to. The first evaluation group is the H-index of a paper's authors. This evaluation group evaluates if a system gives a fair exposure to papers that have authors from a *low* H-index sub-group (H-index < 15) and a *high* H-index sub-group (H-index ≥ 15). The second evaluation group is the International Monetary Fund[7] (IMF) economic development level of the countries of the authors' affiliations. The sub-groups that this group evaluates if a system gives a fair exposure to are papers that have authors from *less* developed countries and *more* developed countries.[8] For the H-index and IMF evaluation groups, a paper can have authors from both of the defined sub-groups. In this case, in our experiments, we assign the paper to the *low* H-index sub-group or the *less* developed country sub-group and not to the *high* H-index sub-group or the *more* developed country sub-group. For example, in the case of the H-index fairness sub-group, if a paper has three authors and the H-indices of the three authors are 5, 7, and 20, then the paper is assigned to the *low* H-index fairness sub-group in the ground truth since at least one of the authors has an H-index of < 15.

As noted by Biega et al. (2020), identifying fairness groups for generating a ground truth evaluation is a difficult task, since attributes of the authors such as gender or prestige are not readily available. Nevertheless, the TREC Fair Ranking Track collections, despite their limitations, are currently the only public IR test collections that enable us to evaluate approaches for this emerging and important (Culpepper et al., 2018) topic in IR.

To index the corpus, we use the Terrier.org Information Retrieval (IR) platform v5.2 (Macdonald et al., 2012; Ounis et al., 2006) and apply standard stopword removal and Porter stemming. We deploy the DPH (He et al., 2008) parameter free document weighting model from the Divergence from Randomness (DFR) framework as a relevance-oriented baseline (i.e. there is no explicit fairness component deployed in this approach), denoted as DPH in Section 7. Moreover, we use the relevance scores from the DPH baseline approach as the relevance component for each of the diversification approaches that we evaluate.

As our metrics, we report the mean Disparate Treatment Ratio (denoted as DTR) and mean Disparate Impact Ratio (denoted as DIR) that were proposed by Singh and Joachims (2018). DTR and DIR measure how much a sequence of rankings violates the Disparate Treatment and Disparate Impact constraints, that we introduced in Section 1, respectively. For two groups, G_0 and G_1, DTR measures the extent that the groups' exposures are proportional to their utility. For a given query, q, and a doubly stochastic matrix $\mathbf{P}$ that estimates the probability of each candidate document being ranked at each rank position over a distribution of rankings that have maximal utility (see Singh and Joachims (2018) for full details of how $\mathbf{P}$ is computed), DTR is defined as:

[7] https://www.imf.org [8] The threshold IMF economic development level that was used to separate countries into less or more developed has not been disclosed by the TREC Fair Ranking Track organisers.

$$\mathrm{DTR}(G_0, G_1 \mid \boldsymbol{P}, q) = \frac{\mathrm{Exposure}(G_0 \mid \mathbf{P})/\mathrm{U}(G_0 \mid q)}{\mathrm{Exposure}(G_1 \mid \mathbf{P})/\mathrm{U}(G_1 \mid q)} \tag{12}$$

where, the utility, U, of a group G_k, is calculated as the sum of the binary relevances, u, of each of the documents, d_i, in G_k, and is defined as:

$$\mathrm{U}\left(G_k \mid q\right) = \frac{1}{|G_k|} \sum_{d_i \in G_k} u_i \tag{13}$$

Following Singh and Joachims (2018), we estimate the exposure drop-off of a document at position j (Pos_j) in a ranking using the position bias user model of DCG (Järvelin and Kekäläinen, 2002), i.e., $\mathrm{Exposure}(Pos_j) = \frac{1}{\log(1+Pos_j)}$.

DIR measures the contribution of each of a group's members (i.e., documents) to the overall utility of the group, defined as:

$$\mathrm{DIR}(G_0, G_1 \mid \mathbf{P}, q) = \frac{\mathrm{CTR}(G_0 \mid \mathbf{P})/\mathrm{U}(G_0 \mid q)}{\mathrm{CTR}(G_1 \mid \mathbf{P})/\mathrm{U}(G_1 \mid q)} \tag{14}$$

where CTR is the sum of the expected click-through rates of the documents in group G_k, and the click-through rate of a document, d_i, is estimated as $\mathrm{Exposure}(d_i \mid \mathbf{P})(d_i \text{ is relevant})$.

For DTR and DIR, a value of 1 shows that both of the groups have a proportionate *exposure* and *impact*, respectively, within the generated rankings. Values less than or greater than 1 show the amount that one of the groups is being disadvantaged by the rankings, with respect to the utility (i.e, relevance) of the documents in the group. We note again here that the number of candidate documents that are associated to a query varies on a per-query basis. Therefore, the size of the ranking and the depth to which DTR and DIR is calculated also varies per-query.

To test for statistical significance, we use the paired t-test over all of the query instances. We select $p < 0.05$ as our significance threshold and apply Bonferroni correction (Dunn, 1961) to adjust for multiple comparisons. Approaches that perform significantly better than the next best performing system with the same assumed fairness groups configuration for an individual metric (e.g., DTR) are denoted with †. For example, for the systems that diversify over the *Author Experience* (A) and *Topical* (T) assumed fairness groups together (denoted by subscript AT), a system is compared with the next best performing system in a pairwise manner (e.g., $PM2_{AT}$ vs. MMR_{AT}) w.r.t. the specific metric (e.g., DTR). If there is a significant difference in the systems' performance then the best performing system is denoted by †. Approaches that perform significantly better than the DPH relevance-only approach for an individual metric are denoted with ‡.

7 Results

In this section, we report the results of our experiments. When evaluating the effectiveness of our proposed approaches, we are primarily concerned with

Table 3 Mean Disparate Treatment Ratio (DTR) and mean Disparate Impact Ratio (DIR) for each of the approaches w.r.t. the H-Index and IMF evaluation groups of the 2019 and 2020 TREC Fair Ranking Track test collections. The table show the results for each of our assumed fairness groups: author experience (A), journal exposure (J) and topical (T) groups. Also included are our DPH relevance-only baseline and a random permutation of the result set r, where r is the candidate set of documents that have been returned by the relevance only retrieval model, denoted as *Random*. For each metric, approaches that are significantly better than the next best performing system with the same assumed fairness groups configuration, e.g., $PM2_{AT}$ vs. MMR_{AT}, are denoted as ‡, while approaches that perform significantly better than the relevance-only DPH approach are denoted as †.

	2019				2020			
	H-Index		IMF		H-Index		IMF	
	DTR	DIR	DTR	DIR	DTR	DIR	DTR	DIR
Single Groupings								
MMR_A	5.63†	**1.99**†	1.61	0.28	4.90†	**1.91**†	3.76	0.44
$xQuAD_A$	5.24†‡	2.06†	0.61	0.32	4.77†	1.99†	**1.10**†‡	0.48†
$PM2_A$	5.51†	**1.99**†	0.88†‡	0.28	4.95†	**1.91**†	1.43†‡	0.44
MMR_J	5.38†	2.04†	0.91†	0.29	4.68†‡	1.99†	1.43†	0.47†
$xQuAD_J$	5.61†	2.03†	1.01†	0.29	5.14†	1.95†	1.42†	0.45
$PM2_J$	5.60†	2.05†	0.84	0.30	5.25†	1.96†	1.51†	0.47†
MMR_T	5.72†	2.02†	0.91†	0.29	5.00†	1.95†	1.54†	0.45
$xQuAD_T$	5.37†‡	2.04†	**1.00**	0.29	**4.48**†‡	1.96†	1.54†	0.45
$PM2_T$	5.66†	2.02†	0.92	0.29	5.00†	1.95†	1.51†	0.45
Paired Groupings								
MMR_{AJ}	5.34†	2.04†	0.77	0.29	4.65†	1.97†	1.19†	0.47†
$xQuAD_{AJ}$	**5.20**†	2.07†	0.73	**0.32**	4.73†	2.00†	1.20†	0.48†
$PM2_{AJ}$	5.51†	2.05†	0.73	0.30	5.13†‡	1.96†	1.31†	0.47†
MMR_{AT}	5.43†	2.02†	0.80	0.29	4.81†	1.94†	1.39†	0.45
$xQuAD_{AT}$	5.24†	2.06†	0.61	**0.32**	4.77†	1.99†	**1.10**†‡	**0.49**†
$PM2_{AT}$	5.49†	2.01†	0.82	0.28	4.62†	1.93†	1.31†	0.45
MMR_{JT}	5.31†	2.05†	0.79	0.29	4.61†	1.99†	1.16†	0.47†
$xQuAD_{JT}$	5.22†	2.05†	0.68	0.30	4.73†‡	1.96†	1.18†	0.46
$PM2_{JT}$	5.52†	2.05†	0.74	0.30	5.13†	1.96†	1.29†	0.47†
All Groupings								
MMR_{AJT}	5.33†	2.05†	0.76	0.29	4.63†‡	1.98†	1.19†	0.47†
$xQuAD_{AJT}$	5.17†	2.07†	0.75	0.31	4.60†	2.00†	1.13†	0.47†
$PM2_{AJT}$	5.51†	2.05†	0.73	0.30	5.13†	1.96†	1.31†	0.47†
No Fairness Component								
DPH	7.99	4.05	1.19	0.11	6.83	3.92	2.19	0.18
Random	8.14	3.03	1.39	0.21	7.37	2.91	1.97	0.33

the suitability of search results diversification for generating rankings that provide a *fair* exposure to *unknown* protected groups. In other words, a protected group should receive an exposure that is proportional to the average relevance of the group, with respect to a user's query. With this in mind, the metrics that we report in this section, namely Disparate Treatment Ratio (DTR) and Disparate Impact Ratio (DIR), consider the exposure that the (authors of) papers from the sub-groups of a protected group receive in proportion to the relevance (utility) of the papers. It is important to note that for both of the metrics, DTR and DIR, a protected group has two sub-groups (e.g., *low* H-index and *high* H-index) and a score of 1.0 denotes that both of the sub-groups receive an exposure that is proportional to the utility / relevance of the documents in the sub-group. Values less than or greater than 1 show that one of

the sub-groups is disadvantaged by the rankings, with respect to the utility (i.e, relevance) of the documents in the sub-group.

Table 3 presents the performance of our diversification approaches for fair rankings, namely MMR, xQuAD and PM-2, with each of our assumed fairness groups: Author Experience (denoted as A), Journal Exposure (denoted as J) and Topics (denoted as T) individually and combined. The table presents each of the proposed approaches' performance in terms of DTR and DIR w.r.t. each of the official TREC evaluation groups, namely H-Index and IMF Economic Level, denoted as IMF, for the 2019 and 2020 TREC Fair Ranking Track collections. In addition, the table shows the results of our DPH relevance-only baseline and a random permutation of the result set, r, returned by the relevance-only model,[9] denoted as Random.

Firstly, addressing RQ1, we are interested in whether leveraging search results diversification as a fairness component is effective for generating fair rankings. We note from Table 3 that all of our proposed fair diversification approaches result in a fairer exposure for the authors of relevant documents than both the relevance-only DPH approach and the Random permutation approach in terms of DTR and DIR, for both of the evaluation fairness groups (H-index and IMF) on both the 2019 and the 2020 collections. Moreover, twelve of the twenty one approaches that we evaluate result in significantly fairer levels of exposure for the H-index evaluation fairness grouping on the 2019 collection and both of the evaluation groupings on the 2020 collection, in terms of DTR and DIR ($p < 0.05$, denoted as † in Table3).

All of the diversification approaches for fairness that we evaluate in this work use our DPH approach for estimating relevance. This shows that, for the diversification approaches that we evaluate, leveraging diversification to integrate a fairness component into the rankings strategy does indeed lead to protected groups receiving a fairer exposure that is more in-line with their utility, or relevance. Therefore, in response to RQ1, we conclude that diversifying over *assumed* fairness groupings can indeed result in fairer rankings when the actual protected groups are not known.

Moving to RQ2, which addresses which of the families of search results diversification, i.e. implicit or explicit, is the most effective for deploying as a fairness component. In terms of DTR, xQuAD explicit diversification consistently results in rankings that are the fairest in terms of ensuring that a protected group receives an exposure proportional to the overall utility of the documents from the group. This observation is true when xQuAD is deployed on either of the 2019 or 2020 collections and evaluated for either of the evaluation fairness groupings (H-index or IMF). On the 2019 collection, $xQuAD_J$ achieves 5.20 DTR for the H-index evaluation grouping and $xQuAD_T$ achieves perfect 1.00 DTR for the IMF evaluation grouping. On the 2020 collection, $xQuAD_T$ achieves 4.48 DTR for the H-index evaluation grouping, while $xQuAD_A$ and $xQuAD_{AT}$ both achieve 1.10 DTR for the IMF evaluation grouping. We note that the 1.10 DTR achieved by $xQuAD_A$ and $xQuAD_{AT}$ for the IMF fairness

[9] Note that the size of r is equal to the number of candidate documents that are associated to a query and varies on a per-query basis.

grouping on the 2020 collection is a 49.7% increase in fairness of exposure compared to the 2.19 DTR of the DPH relevance-only approach.

When deployed on the 2020 Fair Ranking collection, the xQuAD approaches perform significantly better in terms of DTR than the next best performing diversification approach deployed with the same assumed fairness groupings (denoted as ‡ in Table 3). For the H-index evaluation group, $xQuAD_T$ achieves 4.48 DTR while $PM2_T$ and MMR_T only achieve 5.00 DTR. For the IMF evaluation group, $xQuAD_{AT}$ achieves 1.10 DTR while $PM2_{AT}$ only achieves 1.31 DTR. Moreover, $xQuAD_A$ achieves 1.10 DTR while $PM2_A$ only achieves 1.43 DTR. However, we note that, in terms of DTR, the differences between diversification approaches that are deployed with the same assumed fairness groupings are not significant when the approaches are deployed on the 2019 TREC Fair Ranking Track collection.

Turning our attention to how well explicit and implicit diversification approaches perform in terms of DIR, we note from Table 3 that there is no approach that consistently performs best for both of the evaluation groups or on both of the TREC Fair Ranking collections. For the H-index evaluation grouping, MMR_A and $PM2_A$ achieve the best DIR score for on the 2019 and 2020 collections, achieving 1.99 DIR and 1.91 DIR respectively. However, the approaches are not significantly better than the 2.06 DIR (2019) and 1.99 (2020) DIR that is achieved by $xQuAD_A$.

For the IMF evaluation grouping, on the 2019 collection $xQuAD_{AJ}$ and $xQuAD_{AT}$ are the best performing approaches and achieve 0.32 DIR. However, they do not perform significantly better than the MMR or PM2 approaches in terms of DIR. Moreover, notably, none of the fair diversification approaches actually perform significantly better than the DPH relevance-only approach or the Random approach for the IMF evaluation grouping on the 2019 collection. When deployed on the 2020 collection, $xQuAD_{AT}$ performs best in terms of DIR for the IMF evaluation grouping (0.49 DIR). However, $xQuAD_{AT}$ is not significantly better than MMR_{AT} or $PM2_A$ in terms of DIR for the IMF evaluation grouping on the 2020 collection.

These findings provide some evidence that, from the approaches that we evaluate, explicit search results diversification is potentially the more viable diversification approach for developing fair ranking strategies within an academic search context. This finding is supported by the observation that xQuAD explicit search results diversification is consistently the best performing approach in terms of DTR for both of the evaluation fairness groupings when deployed on either of the TREC Fair Ranking Track collections.

Therefore, in response to RQ2, we conclude that explicit search results diversification appears to be the most effective approach, within an academic search context, for ensuring that protected groups receive a fair exposure that is proportional to their utility (relevance) to the users (as is measured by DTR). However, more work needs to be done to identify what is the most effective diversification approach for ensuring that each of the members of a protected group contribute a proportionate amount of gain to the protected

group's overall exposure, i.e., the individual exposure of each of the members of the protected group (as is measured by DIR).

Lastly, addressing RQ3, we conclude that, on our experiments, diversifying over multiple assumed fairness groups does not lead to increased fairness. As can be seen from the numbers in bold in Table 3, three out of the four best performing approaches in terms of DTR diversify over a single assumed fairness group, namely $xQuAD_T$ for the IMF evaluation grouping on the 2019 collection and for the H-index evaluation grouping on the 2020 collection, and $xQuAD_A$ for the IMF grouping on the 2020 collection. Moreover, in terms of DIR, MMR_A and $PM2_A$ achieve the best scores for the H-index evaluation grouping on both the 2019 and 2020 collection. The remaining best performing approaches, namely $xQuAD_{AJ}$ and $xQuAD_{AT}$ both diversify over two assumed fairness groups. However, none of the approaches perform best for any of the evaluation fairness groupings of TREC collections when they diversify over all three assumed fairness groups. This suggests that further work is needed to adequately integrate multiple assumed fairness groups in a diversification approach. We expect that explicitly diversifying across multiple dimensions (Yigit-Sert et al., 2021) of the groups will improve this. However, we leave this interesting area of research to future work.

Finally, we note that in our experiments, diversifying over the documents' authors' experience seems to be a particularly promising approach for generating fair ranking strategies in academic search. Five of the six best performing diversification approaches in terms of DTR and DIR diversify over this assumed fairness group, either as a single group or in combination with one other assumed fairness group, i.e., MMR_A, $xQuAD_A$, $PM2_A$, $xQuAD_{AJ}$ and $xQuAD_{AT}$. We note however, that our approach for calculating an author's experience is only a first reasonable attempt to model this assumed fairness grouping and there remains room for improvement. For example, it is possible that when a lesser known researcher is an author on a very highly cited paper, such as a resource paper, this will potentially skew the system's view of the author's experience. Moreover, we note that our experiments only investigate the exposure that the groups receive for a single browsing model. In practice, variations in users' browsing behaviour will potentially lead to varying exposures for individual papers and authors, and for the overall group that the paper and/or author belong to. Furthermore, the diversification approaches that we evaluate in this work are deterministic processes. A next logical step in developing diversification approaches for fairness would seem to be to introduce a non-deterministic element to proactively compensate for the under, or over, exposure of protected groups. However, we leave the investigation of these interesting questions to future work.

8 Conclusions

In this work, we proposed to cast the task of generating rankings that provide a fair exposure to *unknown* protected groups of authors as a search results di-

versification task. We leveraged three well-known search results diversification models from the literature as fair ranking strategies. Moreover, we proposed to adapt search results diversification to diversify the search results with respect to multiple *assumed* fairness group definitions, such as early-career researchers vs. highly-experienced authors. Our experiments on the 2019 and 2020 TREC Fair Ranking Track datasets showed that leveraging adequately tailored search results diversification can be an effective approach for generating fair rankings within the context of academic search. Moreover, we found that explicit search results diversification performed better than implicit diversification for providing a fair exposure for protected author groups, while ensuring that the group's exposure is in-line with the utility, or relevance, of the groups' papers. In terms of Disparate Treatment Ratio (DTR), xQuAD explicit search results diversification was the most effective approach for generating fair rankings w.r.t. both of the TREC Fair Ranking Track evaluation groupings (H-index and IMF) when the approach was deployed on either of the 2019 or 2020 collections.

This work has provided an in-depth analysis of how search results diversification can be effective as an approach for addressing the important topic of ensuring fairness of exposure in the results of search systems. The search results diversification literature is very broad ranging and, although diversification is not the same task as fairness of exposure, there are potentially many other interesting and useful approaches that can build on the similarities between the tasks to improve the exposure of disadvantaged, or under-represented, societal groups within the results of search engines. In summary, this work has provided a foundation on which future work on integrating fairness into IR systems, and in-particular diversification-based approaches, can build on as this emerging field continues to develop.

9 Acknowledgements

We wish to thank the Associate Editor and the three peer reviewers for their thorough comments and helpful suggestions.

References

Abebe R, Barocas S, Kleinberg JM, Levy K, Raghavan M, Robinson DG (2020) Roles for computing in social change. In: Proceedings of the FAT* Conference on Fairness, Accountability, and Transparency, Barcelona, Spain, January 27-30, 2020, ACM, pp 252–260, URL https://doi.org/10.1145/3351095.3372871

Agrawal R, Gollapudi S, Halverson A, Ieong S (2009) Diversifying search results. In: Proceedings of the Second International Conference on Web Search and Web Data Mining, WSDM 2009, Barcelona, Spain, February 9-11, 2009, ACM, pp 5–14, URL https://doi.org/10.1145/1498759.1498766

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, Peters ME, Power J, Skjonsberg S, Wang LL, Wilhelm C, Yuan Z, van Zuylen M, Etzioni O (2018) Construction of the literature graph in semantic scholar. In: Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2018, New Orleans, Louisiana, USA, June 1-6, 2018, Volume 3 (Industry Papers), Association for Computational Linguistics, pp 84–91, URL https://doi.org/10.18653/v1/n18-3011

Baeza-Yates R (2018) Bias on the web. Commun ACM 61(6):54–61, URL https://doi.org/10.1145/3209581

Belkin NJ, Robertson SE (1976) Some ethical implications of theoretical research in information science. In: The ASIS Annual Meeting

Bender EM, Gebru T, McMillan-Major A, Shmitchell S (2021) On the dangers of stochastic parrots: Can language models be too big? In: Proceedings of the FAccT '21: 2021 ACM Conference on Fairness, Accountability, and Transparency, Virtual Event / Toronto, Canada, March 3-10, 2021, ACM, pp 610–623, URL https://doi.org/10.1145/3442188.3445922

Biega AJ, Gummadi KP, Weikum G (2018) Equity of attention: Amortizing individual fairness in rankings. In: Proceedings of the 41st International ACM SIGIR Conference on Research & Development in Information Retrieval, SIGIR 2018, Ann Arbor, MI, USA, July 08-12, 2018, ACM, pp 405–414, URL https://doi.org/10.1145/3209978.3210063

Biega AJ, Diaz F, Ekstrand MD, Kohlmeier S (2020) Overview of the TREC 2019 fair ranking track. CoRR abs/2003.11650, URL https://arxiv.org/abs/2003.11650, 2003.11650

Biega AJ, Diaz F, Ekstrand MD, Feldman S, Kohlmeier S (2021) Overview of the TREC 2020 fair ranking track. CoRR abs/2108.05135, URL https://arxiv.org/abs/2108.05135, 2108.05135

Bolukbasi T, Chang K, Zou JY, Saligrama V, Kalai AT (2016) Man is to computer programmer as woman is to homemaker? debiasing word embeddings. In: Proceedings of the Advances in Neural Information Processing Systems 29: Annual Conference on Neural Information Processing Systems 2016, December 5-10, 2016, Barcelona, Spain, pp 4349–4357

Broder AZ (2002) A taxonomy of web search. SIGIR Forum 36(2):3–10, URL https://doi.org/10.1145/792550.792552

Calders T, Verwer S (2010) Three naive bayes approaches for discrimination-free classification. Data Min Knowl Discov 21(2):277–292, URL https://doi.org/10.1007/s10618-010-0190-x

Carbonell JG, Goldstein J (1998) The use of MMR, diversity-based reranking for reordering documents and producing summaries. In: Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, August 24-28 1998, Melbourne, Australia, ACM, pp 335–336, URL https://doi.org/10.1145/290941.291025

Castillo C (2018) Fairness and transparency in ranking. SIGIR Forum 52(2):64–71, URL https://doi.org/10.1145/3308774.3308783

Celis LE, Straszak D, Vishnoi NK (2018) Ranking with fairness constraints. In: Proceedings of the 45th International Colloquium on Automata, Languages, and Programming, ICALP 2018, July 9-13, 2018, Prague, Czech Republic, LIPIcs, vol 107, pp 28:1–28:15, URL https://doi.org/10.4230/LIPIcs.ICALP.2018.28

Chen H, Karger DR (2006) Less is more: probabilistic models for retrieving fewer relevant documents. In: Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Seattle, Washington, USA, August 6-11, 2006, ACM, pp 429–436, URL https://doi.org/10.1145/1148170.1148245

Chouldechova A (2017) Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. Big Data 5(2):153–163, URL https://doi.org/10.1089/big.2016.0047

Culpepper JS, Diaz F, Smucker MD (2018) Research frontiers in information retrieval: Report from the third strategic workshop on information retrieval in Lorne (SWIRL 2018). SIGIR Forum 52(1):34–90, URL https://doi.org/10.1145/3274784.3274788

Dang V, Croft WB (2012) Diversity by proportionality: an election-based approach to search result diversification. In: Proceedings of the 35th International ACM SIGIR conference on research and development in Information Retrieval, SIGIR '12, Portland, OR, USA, August 12-16, 2012, ACM, pp 65–74, URL https://doi.org/10.1145/2348283.2348296

De-Arteaga M, Romanov A, Wallach HM, Chayes JT, Borgs C, Chouldechova A, Geyik SC, Kenthapadi K, Kalai AT (2019) Bias in bios: A case study of semantic representation bias in a high-stakes setting. In: Proceedings of the Conference on Fairness, Accountability, and Transparency, FAT* 2019, Atlanta, GA, USA, January 29-31, 2019, ACM, pp 120–128, URL https://doi.org/10.1145/3287560.3287572

Diaz F, Mitra B, Ekstrand MD, Biega AJ, Carterette B (2020) Evaluating stochastic rankings with expected exposure. In: Proceedings of the 29th ACM International Conference on Information and Knowledge Management, Virtual Event, Ireland, October 19-23, 2020, ACM, pp 275–284, URL https://doi.org/10.1145/3340531.3411962

Dunn OJ (1961) Multiple comparisons among means. Journal of the American statistical association 56(293):52–64

Dwork C, Hardt M, Pitassi T, Reingold O, Zemel RS (2012) Fairness through awareness. In: Proceedings of the Innovations in Theoretical Computer Science Conference, Cambridge, MA, USA, January 8-10, 2012, ACM, pp 214–226, URL https://doi.org/10.1145/2090236.2090255

Ekstrand MD, Burke R, Diaz F (2019) Fairness and discrimination in retrieval and recommendation. In: Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2019, Paris, France, July 21-25, 2019, ACM, pp 1403–1404, URL https://doi.org/10.1145/3331184.3331380

Epstein R, Robertson RE, Lazer D, Wilson C (2017) Suppressing the search engine manipulation effect (SEME). ACM Hum Comput Interact 1(CSCW):42:1–42:22, URL https://doi.org/10.1145/3134677

Gao R, Shah C (2020) Toward creating a fairer ranking in search engine results. Inf Process Manag 57(1), URL https://doi.org/10.1016/j.ipm.2019.102138

Hajian S, Domingo-Ferrer J (2013) A methodology for direct and indirect discrimination prevention in data mining. IEEE Trans Knowl Data Eng 25(7):1445–1459, URL https://doi.org/10.1109/TKDE.2012.72

Hardt M, Price E, Srebro N (2016) Equality of opportunity in supervised learning. In: Proceedings of the Neural Information Processing Systems Annual Conference, December 5-10, 2016, Barcelona, Spain, pp 3315–3323

He B, Macdonald C, Ounis I, Peng J, Santos RLT (2008) University of Glasgow at TREC 2008: Experiments in blog, enterprise, and relevance feedback tracks with Terrier. In: Proceedings of the Seventeenth Text REtrieval Conference, TREC 2008, Gaithersburg, Maryland, USA, November 18-21, 2008, NIST Special Publication, vol 500-277, URL http://trec.nist.gov/pubs/trec17/papers/uglasgow.blog.ent.rf.rev.pdf

Järvelin K, Kekäläinen J (2002) Cumulated gain-based evaluation of IR techniques. ACM Trans Inf Syst 20(4):422–446, URL http://doi.acm.org/10.1145/582415.582418

Kamiran F, Calders T (2009) Classifying without discriminating. In: Proceedings of the 2nd International Conference on Computer, Control and Communication, IEEE, pp 1–6, DOI 10.1109/IC4.2009.4909197

Kamishima T, Akaho S, Sakuma J (2011) Fairness-aware learning through regularization approach. In: Proceedings of the 11th International Conference on Data Mining Workshops, Vancouver, BC, Canada, December 11, 2011, IEEE Computer Society, pp 643–650, URL https://doi.org/10.1109/ICDMW.2011.83

Kay M, Matuszek C, Munson SA (2015) Unequal representation and gender stereotypes in image search results for occupations. In: Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems, CHI 2015, Seoul, Republic of Korea, April 18-23, 2015, ACM, pp 3819–3828, URL https://doi.org/10.1145/2702123.2702520

Kleinberg JM, Mullainathan S, Raghavan M (2016) Inherent trade-offs in the fair determination of risk scores. CoRR abs/1609.05807, URL http://arxiv.org/abs/1609.05807

Macdonald C, McCreadie R, Santos RLT, Ounis I (2012) From puppy to maturity: Experiences in developing Terrier. In: Proceedings of the SIGIR 2012 Workshop on Open Source Information Retrieval, OSIR@SIGIR 2012, Portland, Oregon, USA, 16th August 2012, University of Otago, Dunedin, New Zealand, pp 60–63

Mehrotra R, Anderson A, Diaz F, Sharma A, Wallach HM, Yilmaz E (2017) Auditing search engines for differential satisfaction across demographics. In: Proceedings of the 26th International Conference on World Wide Web Companion, Perth, Australia, April 3-7, 2017, ACM, pp 626–633, URL

https://doi.org/10.1145/3041021.3054197

Mehrotra R, McInerney J, Bouchard H, Lalmas M, Diaz F (2018) Towards a fair marketplace: Counterfactual evaluation of the trade-off between relevance, fairness & satisfaction in recommendation systems. In: Proceedings of the 27th ACM International Conference on Information and Knowledge Management, CIKM 2018, Torino, Italy, October 22-26, 2018, ACM, pp 2243–2251, URL https://doi.org/10.1145/3269206.3272027

Morik M, Singh A, Hong J, Joachims T (2020) Controlling fairness and bias in dynamic learning-to-rank. In: Proceedings of the 43rd International ACM SIGIR conference on research and development in Information Retrieval, SIGIR 2020, Virtual Event, China, July 25-30, 2020, ACM, pp 429–438, DOI 10.1145/3397271.3401100, URL https://doi.org/10.1145/3397271.3401100

Olteanu A, Garcia-Gathright J, de Rijke M, Ekstrand MD (2019a) Proceedings of FACTS-IR. CoRR abs/1907.05755, URL http://arxiv.org/abs/1907.05755

Olteanu A, Garcia-Gathright J, de Rijke M, Ekstrand MD, Roegiest A, Lipani A, Beutel A, Lucic A, Stoica A, Das A, Biega A, Voorn B, Hauff C, Spina D, Lewis DD, Oard DW, Yilmaz E, Hasibi F, Kazai G, McDonald G, Haned H, Ounis I, van der Linden I, Baan J, Lau KN, Balog K, Sayed MF, Panteli M, Sanderson M, Lease M, Lahoti P, Kamishima T (2019b) FACTS-IR: fairness, accountability, confidentiality, transparency, and safety in information retrieval. SIGIR Forum 53(2):20–43, URL https://doi.org/10.1145/3458553.3458556

Ounis I, Amati G, Plachouras V, He B, Macdonald C, Lioma C (2006) Terrier: A high performance and scalable information retrieval platform. In: Proceedings of the SIGIR 2006 Workshop on Open Source Information Retrieval, OSIR@SIGIR 2006, Seattle, WA, USA, August 2006, pp 18–25

Pedreschi D, Ruggieri S, Turini F (2008) Discrimination-aware data mining. In: Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Las Vegas, Nevada, USA, August 24-27, 2008, ACM, pp 560–568, URL https://doi.org/10.1145/1401890.1401959

Pleiss G, Raghavan M, Wu F, Kleinberg JM, Weinberger KQ (2017) On fairness and calibration. In: Proceedings of the Advances in Neural Information Processing Systems Conference, December 4-9, 2017, Long Beach, CA, USA, pp 5680–5689

Radlinski F, Dumais ST (2006) Improving personalized web search using result diversification. In: Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Seattle, Washington, USA, August 6-11, 2006, ACM, pp 691–692, URL https://doi.org/10.1145/1148170.1148320

Radlinski F, Kleinberg R, Joachims T (2008) Learning diverse rankings with multi-armed bandits. In: Proceedings of the Twenty-Fifth International Conference on Machine Learning, Helsinki, Finland, June 5-9, 2008, ACM, ACM International Conference Proceeding Series, vol 307, pp 784–791, URL

https://doi.org/10.1145/1390156.1390255
Richardson M, Dominowska E, Ragno R (2007) Predicting clicks: estimating the click-through rate for new ads. In: Proceedings of the 16th International Conference on World Wide Web, WWW 2007, Banff, Alberta, Canada, May 8-12, 2007, ACM, pp 521–530, DOI 10.1145/1242572.1242643, URL https://doi.org/10.1145/1242572.1242643
Robertson SE (1977) The probability ranking principle in IR. Journal of documentation
Santos RLT, Peng J, Macdonald C, Ounis I (2010) Explicit search result diversification through sub-queries. In: Proceedings of the 32nd European Conference on Information Retrieval, Milton Keynes, UK, March 28-31, 2010. Proceedings, Springer, Lecture Notes in Computer Science, vol 5993, pp 87–99, URL https://doi.org/10.1007/978-3-642-12275-0_11
Santos RLT, MacDonald C, Ounis I (2015) Search result diversification. Found Trends Inf Retr 9(1):1–90, URL https://doi.org/10.1561/1500000040
Sapiezynski P, Zeng W, Robertson RE, Mislove A, Wilson C (2019) Quantifying the impact of user attentionon fair group representation in ranked lists. In: Companion of The 2019 World Wide Web Conference, WWW 2019, San Francisco, CA, USA, May 13-17, 2019, ACM, pp 553–562, URL https://doi.org/10.1145/3308560.3317595
Singh A, Joachims T (2018) Fairness of exposure in rankings. In: Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, KDD 2018, London, UK, August 19-23, 2018, ACM, pp 2219–2228, URL https://doi.org/10.1145/3219819.3220088
Singh A, Joachims T (2019) Policy learning for fairness in ranking. CoRR abs/1902.04056, URL http://arxiv.org/abs/1902.04056, 1902.04056
Spärck-Jones K, Robertson SE, Sanderson M (2007) Ambiguous requests: Implications for retrieval tests, systems and theories. In: ACM SIGIR Forum, vol 41, pp 8–17, URL https://doi.org/10.1145/1328964.1328965
Wang J, Zhu J (2009) Portfolio theory of information retrieval. In: Proceedings of the 32nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2009, Boston, MA, USA, July 19-23, 2009, ACM, pp 115–122, DOI 10.1145/1571941.1571963, URL https://doi.org/10.1145/1571941.1571963
White R (2013) Beliefs and biases in web search. In: Proceedings of the 36th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR '13, Dublin, Ireland - July 28 - August 01, 2013, ACM, pp 3–12, URL https://doi.org/10.1145/2484028.2484053
Woodworth BE, Gunasekar S, Ohannessian MI, Srebro N (2017) Learning non-discriminatory predictors. CoRR abs/1702.06081, URL http://arxiv.org/abs/1702.06081
Yadav H, Du Z, Joachims T (2019) Fair learning-to-rank from implicit feedback. CoRR abs/1911.08054, URL http://arxiv.org/abs/1911.08054, 1911.08054
Yang K, Stoyanovich J (2017) Measuring fairness in ranked outputs. In: Proceedings of the 29th International Conference on Scientific and Statistical

Database Management, Chicago, IL, USA, June 27-29, 2017, ACM, pp 22:1–22:6, URL https://doi.org/10.1145/3085504.3085526

Yigit-Sert S, Altingovde IS, Macdonald C, Ounis I, Ulusoy Ö (2021) Explicit diversification of search results across multiple dimensions for educational search. J Assoc Inf Sci Technol 72(3):315–330, URL https://doi.org/10.1002/asi.24403

Zafar MB, Valera I, Gomez-Rodriguez M, Gummadi KP (2017) Fairness beyond disparate treatment & disparate impact: Learning classification without disparate mistreatment. In: Proceedings of the 26th International Conference on World Wide Web, WWW 2017, Perth, Australia, April 3-7, 2017, ACM, pp 1171–1180, URL https://doi.org/10.1145/3038912.3052660

Zehlike M, Bonchi F, Castillo C, Hajian S, Megahed M, Baeza-Yates R (2017) FA*IR: A fair top-k ranking algorithm. In: Proceedings of the 2017 ACM Conference on Information and Knowledge Management, CIKM 2017, Singapore, November 06 - 10, 2017, ACM, pp 1569–1578, URL https://doi.org/10.1145/3132847.3132938

Zemel RS, Wu Y, Swersky K, Pitassi T, Dwork C (2013) Learning fair representations. In: Proceedings of the 30th International Conference on Machine Learning, ICML 2013, Atlanta, GA, USA, 16-21 June 2013, JMLR.org, JMLR Workshop and Conference Proceedings, vol 28, pp 325–333, URL http://proceedings.mlr.press/v28/zemel13.html

10 Declarations

10.1 Funding

Not Applicable

10.2 Conflicts of Interest/Competing Interests

None

10.3 Availability of Data and Material

The datasets are available from the TREC fair Ranking Track, NIST USA. Our assumed fairness groups can be directly calculated from the datasets.

10.4 Code Availability

All of the diversification approaches that we leverage are in the public domain.