# Diversification and Fairness in Search: Two Sides of the Same Coin?

SACHIN PATHIYAN CHERUMANAL, RMIT University, Australia
FALK SCHOLER, RMIT University, Australia
DAMIANO SPINA, RMIT University, Australia

Information retrieval systems aim to return relevant and useful content to users and are often biased towards popular items. This implies that an under-represented group or attribute will not receive a fair share of a user's attention in search results. For example, while a ranked results list for a query such as "physicists" might be fair according to a particular attribute such as gender, nationality, or social group, it might not be fair for all of them. Ideally, while providing relevant answers, a results list should also provide fair exposure across a broad range of attributes. We demonstrate that while a system can be fair towards multiple attributes, they are not necessarily diverse (i.e., redundancy/minimal novelty). To this end, we include an additional dimension to the study, i.e., diversity, and explore the relationship between fairness and diversity measures by exploring popular search result diversification techniques using the test collections from TREC 2021 Fair Ranking Track, TREC 2022 Fair Ranking Track and NTCIR-17 FairWeb-1. Furthermore, we study the impact of such diversification techniques along both nominal and ordinal attributes, as well as for intersectional fairness. Our results indicate that explicit search results diversification techniques showed improved results when the attributes were nominal but failed to provide fairer and more diverse results when the attributes were ordinal in nature. Additionally, in terms of intersectional fairness explicit search results diversification also performed significantly better than baseline retrieval runs.

CCS Concepts: • **Information systems** → **Retrieval effectiveness**; **Test collections**.

Additional Key Words and Phrases: fairness; group fairness; search results diversification; information retrieval

**ACM Reference Format:**
Sachin Pathiyan Cherumanal, Falk Scholer, and Damiano Spina. 2018. Diversification and Fairness in Search: Two Sides of the Same Coin?. In *Woodstock '18: ACM Symposium on Neural Gaze Detection, June 03–05, 2018, Woodstock, NY*. ACM, New York, NY, USA, 30 pages. https://doi.org/10.1145/1122445.1122456

## 1 INTRODUCTION

Information retrieval systems play a crucial role in our modern society, shaping how we access and consume information. However, recent studies show that biases in search results can be problematic and can create new unwanted biases or reinforce existing societal biases. The presence of such biases in information access systems has profound implications. It can lead to preferential treatment of certain groups, limiting exposure for marginalised communities. Information access was primarily focused on providing relevant results. However, with the above-mentioned implications in mind, researchers have shifted focus to making information access more fair. In this effort, researchers have proposed various fairness-aware algorithms and metrics. Some of the recent efforts have been along the lines of shared tasks like the Fair Ranking Tracks of TREC 2021 (TRECFair21) [30], TREC 2022 (TRECFair22) [29], and NTCIR FairWeb-1 (NTCIRFair23) [81], which provides a platform for researchers to experiment with systems in a controlled setting. A search result needs to be relevant to a user's query and should be fair along one or more aspects. Consider the example of a user query "Physicists"


Manuscript submitted to ACM

with the intent to find a list of famous physicists. In this search, we should ideally observe a diverse representation of fairness aspects. However, in Figure 1, it can be observed that the search results mainly present male physicists, thereby providing less exposure to other gender categories. Moreover, the majority of the top results are associated with North American and European regions. This poses a new challenge where we require the system to be fair towards multiple fairness attributes (in this particular example, gender and geographic locations). Recent work on fairness in IR focuses on binomial attribute fairness, multinomial attribute fairness, and multi-protected group fairness, but not much work has considered multi-attribute fairness. Furthermore, the recent shared tasks-TRECFair21, TRECFair22 and NTCIRFair23, are evidence that multi-attribute fairness is still an open problem. In this work, we consider the notion of group fairness, which requires that the demographics of those receiving a certain treatment are proportional to the demographics in the overall population – also known as statistical parity [27]. Considering this definition of fairness, we explore re-ranking techniques to ensure fair exposure to non-topical aspects. From an IR perspective, multi-attribute fairness optimisation can be seen as the task of balancing aspects in a ranked list, while maintaining relevance, which is similar to search results diversification (SRD) [2, 75].

The primary motivation behind SRD is to minimise redundancy and maximise novelty. For instance, let us consider the popular single-term query "apple" [12]. This could either mean the user was looking for the fruit or the company "Apple Inc." The search results for this query should include top-ranking results for both topics, rather than focusing solely on one, to ensure a balanced presentation of perspectives. However, as noted by Gao and Shah [32], search engines do not always achieve this balance and may privilege certain viewpoints over others. This limitation becomes particularly salient in queries with societal implications. For instance, in the case of the query "coffee health", the search engine may predominantly present the benefits of drinking coffee while neglecting to highlight potential and possibly lesser-known risks. This can lead to a biased view of the subject, as the search results position the benefits of coffee in the top rankings. Recent diversity-based re-ranking techniques try to achieve maximum coverage, and minimal redundancy of multiple aspects. However, fairness-aware re-ranking optimises for certain fairness criteria and relevance. In this work, we start by analysing the systems submitted to TRECFair21, TRECFair22, and NTCIRFair23. Motivated by these examples, we conducted empirical studies to answer the following research questions:

**RQ1:** Is a system fairer towards multiple fairness attributes necessarily more diverse (i.e., maximise novelty and minimise redundancy)? Using participant submissions from TRECFair21, TRECFair22, and NTCIRFair23, we analyse the systems submitted to these tracks. By comparing the rankings of systems from track-provided fairness metrics with diversity measures and examining their correlation, we empirically demonstrate that the most fair system is not necessarily the most diverse (see Section 5).

**RQ2:** Can SRD along multiple fairness attributes achieve fairer search results? Considering the example in Figure 1, it can be observed that physicists with a lower $h$-index are not represented at the top of the rankings. Consequently, physicists with a lower $h$-index would not receive due exposure. Unlike certain attributes (gender or geographic locations), attributes can also be ordinal, as is the case with the $h$-index. This leads us to our third research question:

**RQ3:** Can SRD achieve fairer results when both nominal and ordinal fairness attributes are involved? In light of the research questions, we highlight the contributions of this paper as follows: (i) We leverage our participations at TRECFair21 [60], TRECFair22 [57], NTCIRFair23 [58], and conduct a comprehensive analysis of the systems submitted to the tracks, while unifying notations and frameworks used in the aforementioned shared tasks. (ii) We empirically and mathematically demonstrate the significance of constraints on the weights used in the Group Fairness and Relevance

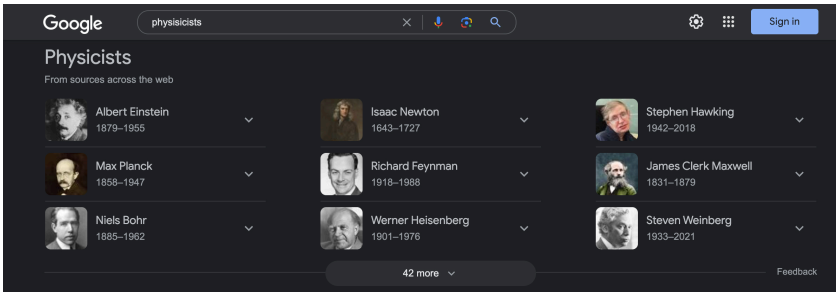



Fig. 1. Search result for the query "physicists".

framework (*GFR*) [71][1] and propose a new constraint for the same, namely $GFR^{\psi}$, which is later used to study the relationship between fairness and diversity. (iii) Leveraging the runs submitted to three shared tasks, our meta-evaluation demonstrates that a fair system is not necessarily diverse, and confirms that diversity needs to be considered as a third dimension for IR evaluation. Moreover, this corroborates preliminary exploration of Pathiyan Cherumanal et al. [59] and Sakai et al. [71]. The former utilized the CLEF 2020 Touché Lab [10] collection to explore the relationship between relevance, fairness, and diversity treating *Arguments* as documents and their associated *Stance* ( i.e., PRO/Supporting or CON/Attacking) as aspects. The latter [71] also utilised the same collection to examine the relationship between fairness and diversity measures, demonstrating that while these measures remain highly correlated in a hard group membership binary attribute setting, they assess different dimensions. To this end, we propose using harmonic mean (H-Score) as a compound metric that evaluates along the three dimensions, while highlighting the limitations of the *dot product*, which was used in TRECFair21 and TRECFair22 for aggregation. (iv) Using the proposed H-Score as well as the measures used in three shared tasks, we evaluate our proposed approach (i.e., SRD + ranking fusion) and conclude that explicit SRD technique can significantly improve fairness along different characteristics of the fairness attributes (Section 5). The source code associated with this work is made available at: https://github.com/sachinpc1993/re-ranking-lightweight.

The remainder of this paper is structured as follows, Section 2 discusses the different notions of fairness (Section 2.1), metrics used to measure fairness particularly in ranked systems (Section 2.2), the shared tasks and the participant runs that were leveraged to address the research questions (Section 2.3), the significance of diversity in information access systems and metrics for the same (Section 2.4), SRD in a fairness problem (Section 2.5). Section 3 discusses our approach (i.e., SRD + ranking fusion), and Section 4 discusses experimental setup, including the test collections (Section 4.1), metrics for fairness and diversity (4.2) along with the meta-evaluation strategy (Section 4.3).

## 2 RELATED WORK

### 2.1 Notions of Fairness

A recent survey by Lerman and Hogg [48] discusses the various stages at which biases can occur in a data-driven system, including data (e.g., under-represented categories for a classification task), algorithmic biases (e.g., ranking biases), and user biases (cognitive or behavioural biases).

Of the three types, our work focuses primarily on algorithmic biases. The use of data-driven algorithms on a large scale may have negative impacts on society [3, 47], thus highlighting the need for more fair algorithms and evaluation

[1]GFR is discussed in detail in Section 4.2.2.

techniques. While we used the term "algorithm" in the broad sense, Gao and Shah [32] differentiates how the notions of fairness vary depending on the downstream task. For instance, the notion of fairness and its evaluation and mitigation strategies for a supervised machine learning setting [35] would be different from that in a ranked system (e.g., search engines, recommender systems), mainly due to position bias [48]. Even the slightest changes in ranking can significantly impact the user's attention [62]. Therefore, it is essential to carefully place items in a ranked list to ensure that under-represented items in a ranked list get fair exposure. For instance, a biography website about a public figure from an under-represented or protected group would not receive fair exposure if it were ranked lower on the Search Engine Results Page (SERP). Note that an under-represented and protected group, in the literature, often refers to the "protected group"[2] defined by the United States anti-discrimination law [25]. According to the literature, there are two notions of fairness. *Individual fairness* where similar individuals are treated similarly, and *group fairness* where demographics of those receiving a certain treatment are proportional to the demographics in the overall population - also known as statistical parity [27, 59]. In this work, we focus on group fairness, specifically in ranked systems [17, 31, 67].

Our attention is directed towards *provider*-side group fairness, treating providers as analogous to documents, aiming to reduce imbalanced exposure based on document attributes (e.g., gender of the author, geography associated with the document, time of creation, and so on). In the upcoming Section 2.2, we discuss previous work that measures group fairness in ranked systems and highlight the measures we use in our study.

### 2.2 Measuring Fairness in Ranked Systems

Algorithm fairness has been studied across multiple disciplines and has attracted attention from researchers in IR [22]. Given the widespread use of ranking systems in our daily lives and the vast socio-economic impact of their applications (e.g., college admissions, product listings, candidate short-listing for jobs), a lot of emphasis has been placed on developing metrics to evaluate fairness in such ranked systems. While fairness measures have been proposed in the past for classification-based systems, they are quite distinct from those required for ranked systems since the position of a particular document in the results needs to be accounted for, unlike in classification-based systems. Additionally, such ranked systems primarily involve two stakeholders (*consumer* and *provider*) [28]. The consumer-side fairness is concerned with how an information access system impacts users and sub-groups of users, and whether those effects are fair or result in unjust harms. The provider-side fairness is concerned with the opportunity an item in a search result has to be interacted with by a user. Our work focuses on *provider-side group fairness*, where the providers are analogous to the documents themselves, and the objective of fairness is to reduce imbalanced exposure to the documents, and the attributes would be the features associated with the document (e.g., gender of author, geography associated with the document, time of creation and so on). For the remainder of this paper, when we refer to fairness, we mean provider-side group fairness.

Several measures have been introduced over the recent years that facilitate evaluating fairness, such as rKL [90], Skew and NDKL [33], Expected Cumulative Exposure (ECE) [76], Attention Bias Ratio (ABR) [34], NDJS [26], and Attention-Weighted Rank Fairness (AWRF) [30]. Moreover, a modified version of AWRF was used in TRECFair21 [30] and TRECFair22 [29] to evaluate intersectional fairness [31] i.e., combining the different groups of multiple attributes. For instance, if we had an attribute $A_1$ = Gender, and the groups in $A$, $G_A = \{\text{Male}, \text{Female}, \text{Other}\}$ and attribute $B$ = Age, and the groups in $B$, $G_B = \{> 50, \leq 50\}$, the intersectional groups ($IG$) would be a Cartesian product of $G_A$ and $G_B$ i.e., $IG = \{(\text{Male}, > 50), (\text{Female}, > 50) \ldots (Other, \leq 50)\}$. We use AWRF [30] in our study to evaluate how our approach

[2] A class of people who share a trait upon which a recommendation or classification should not be discriminatory. This includes discrimination based on race, gender, religion, and similar traits.

works for intersectional fairness.[3] One of the limitations of a Cartesian product approach is that it ignores the ordinal nature of the attribute $B$. Recent work by Sakai et al. [71] distinguishes between nominal and ordinal attributes and calls for a separate evaluation along each attribute so that the ordinal and nominal nature can be accounted for. Following the recent NTCIRFair23 [81], we account for these characteristics in our study, by using Root Normalised Order-aware Divergence (RNOD) [70].

Overall, to evaluate ranked systems we use, Jensen Shannon Divergence (JSD) [26] for nominal attributes, RNOD for ordinal attributes, and AWRF [30] for intersectional group fairness, as outlined in Section 4.2.2. Details of the ranked systems used in the study are provided in Section 2.3, and the test collections used are described in Section 4.1 as a part of our experimental setup.

### 2.3 Shared Tasks

The primary objective of shared tasks in information retrieval is for researchers to collectively address a problem using standardised test collections and evaluation metrics. However, a major challenge is to consolidate knowledge from multiple shared tasks, which may each aim to address different issues, using different test collections, and evaluation metrics that measure different dimensions (sometimes even in cases where the primary focus problem of the tasks is the same e.g., fairness in ranked systems). Since the TREC 2019 Fair Ranking Track [8], shared tasks on fairness in ranked systems have evolved, with subsequent initiatives like the TREC Fair Ranking Tracks of 2020, 2021, 2022, and NTCIR-17 FairWeb-1 introducing new challenges and expanding the scope to address various aspects of fairness in IR. The TREC 2019 Fair Ranking Track focused on an academic search task, where the organisers provided a corpus of academic article abstracts and queries submitted to a production academic search engine. The main objective of this track was to ensure fair representation of various groups of authors. The TREC 2019 Fair Ranking Track was established as a reranking task, while the TREC 2020 Fair Ranking Track [7] expanded to include both retrieval and reranking approaches. Both the TREC 2019 and 2020 tracks focused on evaluating fairness across a sequence of rankings rather than individual rankings [9, 77]. Our work focuses on individual rankings; therefore, we leverage test collections and evaluation metrics from the TREC 2021 Fair Ranking Track onwards i.e., TREC 2021 Fair Ranking Track (TRECFair21) [30], TREC 2022 Fair Ranking Track (TRECFair22) [29], NTCIR-17 FairWeb-1 (NTCIRFair23) [81]. Details about the test collections, including the corpus and fairness attributes associated with these shared tasks, are summarised in Table 2, and the evaluation measures are discussed in detail in Section 4.2.

The TREC 2021 (TRECFair21) and 2022 (TRECFair22) versions of the Fair Ranking Track aim to provide equal representation of Wikimedia articles from different marginalised groups based on different attributes. Note that the aforementioned tracks involved evaluating fair exposure to nominal-type fairness attributes. Consequently, to enable evaluation of ordinal-type fairness attributes, we use the test collection from NTCIRFair23. We examine the similarities and differences between these tasks, highlighting how they inform our approach and reviewing the systems submitted to each shared task.

#### *2.3.1 TREC 2021 Fair Ranking Track (TRECFair21).*

***Description.*** The TRECFair21 adopts a resource allocation task that supports Wikipedia coordinators who search for documents that need to be improved. During such a task, it is imperative that the documents about, or that somehow

[3]Building on prior work by Sakai et al. [71], we use a simplified version of "intersectional fairness" represented using a Cartesian product. However, we acknowledge that this simplification may have obscured the complexities of the social theory of intersectionality, and operationalising such theories requires further exploration [38].

represent, certain attributes receive a fair exposure. For instance, one of the attributes provided by the track organisers was the geographic location. However, the organisers evaluate the fairness of the systems across an undisclosed protected attribute as well.

***Participation.*** In total, for Task 1 of the track, 13 runs were submitted to TRECFair21 by four groups. Most of the submissions take the approach of starting from an unfair ranking that maximises for relevance (ad hoc retrieval) i.e., a system that does not take into account the fairness attributes involved, and then uses a re-ranking step. For the initial unfair ranking, most teams used retrieval systems like BM25 [66] from the Pyserini implementation[53], Divergence from Randomness (DFR) from the Terrier platform [55], and LambdaMART [11] from the LightGBM [44] implementation. For the re-ranking stage, the submissions varied across techniques like Pairwise Swapping (where items were iteratively swapped to result in the largest marginal gain for the objective), iterative re-ranking using BM25 score and individual fairness scores, proportionally allocating protected group position in a new ranking based on the relevance score, and the implicit SRD technique called Maximal Marginal Relevance (MMR) [13]. However, note that while most of the techniques explicitly utilised the fairness attributes for re-ranking, MMR utilised the text of the document to re-rank the initial retrieved list. Out of the 13 systems that were evaluated, it can be seen that UoGTrDExpDisT1 [40], a run based on exposure disparity, performed the best overall.

#### *2.3.2 TREC 2022 Fair Ranking Track (TRECFair22).*

***Description.*** The goal of Task 1 at TRECFair22 is to support WikiProject coordinators in producing a ranked list of articles that editors can then consult when looking for work to do. Given a query (i.e., keywords in a WikiProject title), the system needs to produce a single ranking consisting of 500 articles retrieved from a collection of Wikimedia articles (a dump of the English Wikipedia).

***Participation.*** In total, for Task 1 of the track, 27 runs were submitted to TRECFair22 by five teams. In this overview of systems from the track, we can discuss only the systems from three of the five teams, as only these teams have been added to the proceedings. Similar to TRECFair21, most teams took a two-step approach of retrieval and then re-ranking. For the retrieval stage, most teams used BM25. Conversely, certain runs employed dense retrieval methods in the first-stage retrieval, specifically utilising ColBERT [46]. Note that UoGTrT1ColPRF and UoGRelvOnlyT1 by Jaenich et al. [41] performed equally across metrics and hence are treated equally throughout this manuscript, where UoGTrT1ColPRF serves as a relevance-only baseline, implemented using ColBERT-PRF [87].

#### *2.3.3 NTCIR-17 FairWeb-1 Task (NTCIRFair23).*

***Description.*** The overall goal of the NTCIRFair23 task is to select a diverse group of IR researchers for roles such as conference organisers. This requires consideration of various dimensions, including career stages (e.g., junior researchers), gender, and nationality. However, existing search engines often yield biased results, predominantly highlighting well-known researchers (e.g., high $h$-index) and may fail to ensure adequate gender and demographic diversity. While the goal may seem similar to TRECFair21 and TRECFair22, the significant difference lies in the attribute characteristics and the evaluation [81]. For instance, some attributes may not just be categorical but also ordinal, as in the case of a researcher's $h$-index when a user searches for a specific researcher's profile [81]. The differences between the evaluation metrics used at the TRECFair22 and NTCIRFair23 have been discussed in detail by Sakai et al. [71]. However, for quick reference, Table 2 compares and contrasts NTCIRFair23 with TRECFair21 and TRECFair22 along the following task

features: aim, corpus, topics in test collection, sample queries, fairness attributes, the characteristics of the attributes, and the evaluation metrics.

***Participation.*** In total, 28 runs were submitted to NTCIRFair23 from five participant teams (i.e., THUIR [82], UDinfo-Lab [16], RSLFW [50], and rmit_ir [58]),[4] including six runs from the organisers. Similar to TRECFair21 and TRECFair22, most runs used a two-step approach of ad-hoc retrieval and re-ranking. For the retrieval stage, most teams used the baseline runs provided by the organisers, which mainly consisted of BM25 and query-likelihood with Dirichlet smoothing and Jelinek-Mercer smoothing. These runs were created based on the query and description fields separately and provided to the participants. For the re-ranking stage, the techniques spanned from leveraging demographic-specific embeddings for re-ranking baselines to query augmentation, fusing sparse and dense relevance features, prompt-based feature scoring, learning-to-rank (LTR) using LightGBM [44], and SRD techniques (e.g., xQuAD [74], PM-1 [24], and PM-2 [24]). Among the 28 evaluated systems, the top-performing ones leverage Learning-to-Rank (LTR) and SRD-based approaches [82].

In summary, we examined recent initiatives to improve fairness in ranked retrieval systems through shared tasks and analysed the characteristics of the systems submitted for evaluation. Section 2.4 discusses the importance of incorporating diversity as a complementary dimension to fairness and relevance in retrieval evaluation and reviews prior works on diversity evaluation. Section 2.5 articulates the motivation for exploring SRD techniques in a fairness problem.

### 2.4 Diversity in Information Access

Diversity has gained attention across various information access applications. For instance, one such application is in news recommendations where diversity could be defined as the "heterogeneity of media content in terms of one or more specified characteristics" [5, 83]. According to Bernstein et al. [5] and Helberger [36], diversity of perspectives in news is crucial for informed citizens in democratic societies. Moreover, a lot of attention has shifted to studying and improving diversity in information access systems recently [18, 37, 64, 86].

Recent work by Raza et al. [64] highlights the significance of exposing users to a variety of perspectives/information in recommender systems and hence proposes a novel two-tower architecture that aims to achieve a balance between accuracy and diversity. The authors used a combination of nDCG and F1-Score to measure accuracy and used Gini Index (GINI) [80] to evaluate distribution of popularity, where a higher Gini indicated uneven distribution and bias. From an IR perspective, in a two-tiered recommender system, the first stage involves the retrieval of a subset of related items from a large corpus in response to a user's query, and then a ranking model ranks the retrieved items based on users' actions (i.e., clicks or ratings) [64]. Additionally, a recent survey by Zhao et al. [92] indicates that fairness and diversity have been exhaustively investigated independently and calls for studying the intrinsic connection between the two. Recent studies highlight the importance of examining fairness alongside diversity, establishing it as a significant research problem [6, 89]. Given the distinction between fairness and diversity, it is essential to balance both within a ranked list while preserving relevance. This necessitates a three-dimensional evaluation framework encompassing relevance, fairness, and diversity.

[4]One participating team was declared as "undisclosed" by the shared task organisers and was therefore excluded from this study. See [81].

### 2.5 Search Results Diversification

Information retrieval systems often provide a long list of results in response to a user's query. This list may include redundant results, which can fail to satisfy user information needs due to ambiguous queries, impacting user satisfaction [1]. SRD is a technique that aims to minimise this redundancy and maximise novelty in the list of documents presented to the user, implying that a system must strike a balance between the relevance of information presented and its diversity i.e., the trade-off between relevance and diversity aims to find an optimal ranked list which is both relevant and diverse. The overall framework of SRD comprises two main attributes: the relevance (i.e., the similarity between the input query and the document), which is often estimated using a retrieval component (e.g., BM25 [65]) to generate a candidate list, and a diversification component which re-ranks whole/subset of the candidate list.

SRD is generally classified into two categories: (i) implicit SRD, and (ii) explicit SRD. The implicit SRD re-ranks search results to enhance diversity without explicit knowledge of user intent or attributes. A notable technique is Maximal Marginal Relevance (MMR) [13], which balances document relevance and diversity by adjusting their weights via a tunable parameter $\lambda$. While MMR is a heuristic implicit SRD method, several supervised learning-based implicit SRD techniques have been developed. One such method is Passage-Aware SRD, which leverages multiple passages to better represent documents, as different passages may reflect varying aspects (or subtopics) [79].

On the other hand, explicit SRD techniques use explicitly defined aspects of the document to achieve diversity.[5] For instance, Santos et al. [74] proposed xQuAD that calculates a document's probability across various aspects to promote diverse search results. Another explicit SRD technique, PM-2, proposed by Dang and Croft [24], selects documents to ensure the final result list reflects the proportional popularity of different aspects. Recent explicit SRD techniques like HxQuAD and HPM-2 (based on xQuAD and PM-2, respectively), proposed by Hu et al. [39], use hierarchical representations of aspects instead of earlier methods that use a flat list of aspects (or subtopics).

Revisiting the earlier example of an ambiguous user query "apple", the dimensions in this example are *company* and $fruit$. So in this case, explicit SRD would aim to reorder the documents associated with these dimensions. This would prevent any single aspect (i.e., *company* or $fruit$) from dominating the list and receiving the majority of user attention. This aim of preventing domination of an aspect is similar to the notion of fairness that was discussed earlier in Section 2.1. However, the objective of diversification and fairness still vary [14, 59, 71], where fairness-aware rankings aim to provide equity in opportunity for a document to be viewed [32]. For example, a document that presents under-represented political or socio-economic perspectives and is ranked lower on the list may not be seen by users. This neglect could potentially lead to radicalisation and reinforce existing biases or stereotypes. Fairness-aware rankings seek to address these challenges by encouraging a balanced representation and exposure [32, 56]. Given these similarities and differences between fairness-aware ranking and diversified ranking, we raise the question: given a retrieved list, can SRD techniques provide fairer results?

To address this question, we examine one implicit and one explicit SRD technique: MMR, and PM-2, respectively. To the best of our knowledge, the closest relatable study was done by McDonald et al. [54], where the authors examined the effectiveness of SRD techniques in promoting fairness over a sequence of rankings [9, 77] using the test collections from TREC 2019 and TREC 2020 Fair Ranking Track. The results indicated that explicit SRD techniques like xQuAD and PM-2 showed fairer results, in a sequence of ranking with assumed attributes and groups. In contrast, our work focuses on group fairness achieved by SRD techniques within a single ranking. As mentioned earlier, the study [54]

[5] An *aspect* in this context refers to different dimensions related to a search query that a user might be interested in.

Table 1. Notations used.

| Notation | Description |
|---|---|
| $q$ | Query. |
| $R$ | Candidate list. |
| $S$ | Re-ranked list |
| $C_d$ | Candidate documents. |
| $d$ | An instance of a document, such that $d \subset D$. |
| $k$ | Ranking cutoff or depth. |
| $pos$ | Position in ranking. |
| $\mathcal{F}$ | Fairness attribute e.g., Gender, geographic location. |
| $f_i$ | Fairness aspect (or group), such that $f_i \subset \mathcal{F}$. |
| $m_d$ | Membership distribution. |
| $T_d$ | Target distribution. |
| $Dis$ | Distance function. e.g., Jensen Shannon, KL Divergence and so on. |

used assumed fairness attributes and groups.[6] This could mean that the top-performing system in the study may have excelled due to the approximations made regarding these attributes.

## 3 APPROACH

In this section, we present our 3-stage approach for studying the impact of SRD on fairness: (i) Ad-hoc retrieval, (ii) SRD-based re-ranking by fairness attribute, and (iii) Ranking fusion. Each stage is discussed in detail, with the overall architecture illustrated in Figure 2. Additionally, the notations used throughout this paper are summarised in Table 1.

### 3.1 Stage 1: Ad-hoc Retrieval

The first stage involves retrieving Wikipedia articles based on a WikiProject topic by leveraging the information in the "keywords" field, specifically the parsed title provided by the organisers. For indexing purposes, we use the "plain" collection,[7] representing the full article text, without any Wiki markup (plain file only) [29]. The retrieval is performed using the Pyserini implementation of the Okapi BM25 model [53] with default parameters ($k1 = 0.9, b = 0.4$).

The BM25 model [65] adopts the TF-IDF signal to measure term weights and calculate the relevance score between a query $q$, and a document $d$, to finally generate an ad-hoc retrieval list $R$, containing a set of candidate documents $C_d$. Given a query $q$ comprising of keywords $\{t_1, t_2, t_3, \ldots t_n\}$, the relevance score of the document, $d$, is calculated as shown in Equation 1.

$$\text{BM25}(d, q) = \sum_{t_i} \frac{\text{IDF}(t_i) \times \text{TF}(t_i, d) \times (k1 + 1)}{\text{TF}(t_i, d) + k1 \times (1 - b + b \times \frac{\text{len(d)}}{\text{avgdl}})} \tag{1}$$

where, $\text{TF}(t_i, d)$ denotes the frequency of query term $t_i$ in the document $d$ (i.e., term frequency) and len(d) denotes the document length. avgdl is the average document length in the corpus. $\text{IDF}(t_i)$ is the inverse document weight of the query term $t_i$ is calculated using Equation 2 as follows:

[6] Note that the TREC 2019 Fair Ranking Track [8] hid certain fairness attributes from participants during the development phase to evaluate the robustness of systems across various group definitions.
[7] https://data.boisestate.edu/library/Ekstrand/TRECFairRanking/corpus/trec_corpus_20220301_plain.json.gz

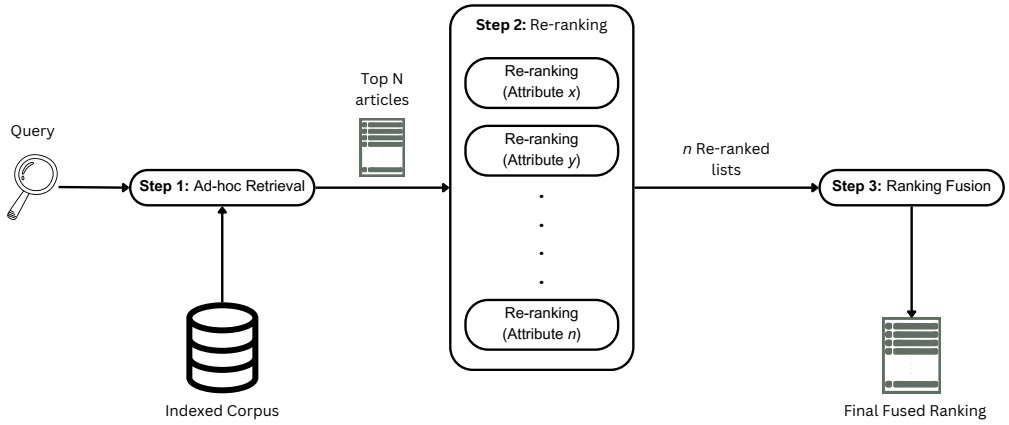



Fig. 2. Architecture diagram depicting the different components used in our approach.

$$\text{IDF}(t_i) = \log \frac{N - n(t_i) + 0.5}{n(t_i) + 0.5} \tag{2}$$

where $n(t_i)$ denotes the number of documents containing the query term $t_i$, and $N$ denotes the number of documents.

### 3.2 Stage 2: Re-Ranking using Search Results Diversification

As discussed in Section 2, our objective is to determine to what extent SRD methods can diversify a ranking based on a fairness attribute ($\mathcal{F}$). Given a fairness attribute, values for that particular attribute are treated as *aspects* or sub-topics $f_i$, aligned with the literature on SRD [23, 24, 51, 85]. In the re-ranking stage we employ the following SRD techniques (i) MMR [13] and (ii) PM-2 [24]. The rationale for selecting these SRD techniques is detailed in Section 2.5.

*3.2.1 Implicit Search Results Diversification.* For implicit SRD we use the popular MMR. According to Carbonell and Goldstein [13], high marginal relevance is achieved when a document is both relevant to the query and contains maximal dissimilarity to previously selected documents.

$$\text{MMR} = \underset{D \epsilon R \backslash S}{\arg\max} \left[ \lambda(\text{Sim}_1(d_i, q) - (1 - \lambda) \max_{d_j \epsilon S} \text{Sim}_2(d_i, d_j)) \right] \tag{3}$$

where $q$ is the query, $R$ is the retrieved list of documents, $S$ is the list of documents already selected and $R \backslash S$ is the set difference i.e., the set of as yet unselected documents in $R$. $\text{Sim}_1$ is the similarity between a document and the query. $\text{Sim}_2$ is the similarity between the document from the retrieved list and the documents already selected for the re-ranked list $S$. Although $\text{Sim}_1$ and $\text{Sim}_2$ can use different similarity measures, in our experiments, both are computed using cosine similarity[8] which can be formulated as shown in Equation 4.

$$\text{Sim}(\overrightarrow{q}, \overrightarrow{d}) = \frac{\overrightarrow{q} \cdot \overrightarrow{d}}{|\overrightarrow{q}||\overrightarrow{d}|} \tag{4}$$

[8]The cosine similarity is defined as the cosine of the angle between the vector representations of a query $q$ and a document $d$ [73].

The $\lambda$ parameter used in Equation 3 refers to the degree of diversification where $\lambda$ can take any value between 0 and 1. When $\lambda = 0$, MMR calculates a maximal diverse ranking, and a standard relevance-ranked list when $\lambda = 1$.

*3.2.2 Explicit Search Results Diversification.* We use the proportionality-aware diversification algorithm called PM-2. PM-2 iteratively picks the best aspect which maintains overall proportionality and then selects the best document for the aspect for each position [24]. Additionally, it is worth noting that PM-2 uses a probabilistic approach, which considers that each document can be associated with multiple aspects. The technique assumes that all documents $d_j \epsilon D$ are relevant to all aspects $t_j \epsilon T$ and can be formulated as shown in Equation 5.

$$d^* \leftarrow \arg\max_{d_j \epsilon R} \left[\lambda \times \text{quotient}[i^*] \times P(d_j|t_{i^*}) + (1-\lambda)\Sigma_{i \neq i^*} \text{quotient}[i] \times P(d_j|t_i)\right] \tag{5}$$

where $\text{quotient}[i] = \frac{v_i}{2s_i+1}$, and $i^* = \arg\max_i \text{quotient}[i]$. Here, $R$ is the candidate list, and $v_i$ is the number of aspects in $R$, $s_i$ is the number of aspects already allocated a position in the new ranked list, $S$.

### 3.3 Stage 3: Ranking Fusion

Ranking fusion is a data fusion technique that combines multiple ranked lists from different IR systems into a single ranking [88]. Data fusion can follow two main approaches: (i) score-based fusion and (ii) ranking fusion. Leveraging the effectiveness of Reciprocal Ranking Fusion (RRF) [21], a ranking fusion technique, over score-based fusion, and the impracticality of exploring the vast design space for fusion exhaustively, we use RRF[9] to combine the diversified rankings from Stage 2 of our approach, as illustrated in Figure 2.[10] RRF is defined in Equation 6:

$$\text{RRF}(d) = \sum_{i=1}^{n} \frac{1}{\beta + pos_i} \tag{6}$$

where $\beta$[11] is a constant set to a default value of 60 following Cormack et al. [21], $pos_i$ is the rank of a document in the i-th ranked list where $n$ ranked lists are fused. The RRF score is calculated by summing the reciprocal of the rank and the constant $\beta$ for each document across $n$ lists and the document with the highest RRF score is considered the most relevant.

## 4 EXPERIMENTAL SETUP

### 4.1 Test Collections

We use three collections in our study. (1) the TREC 2021 Fair Ranking Track collection (TRECFair21) [30] which contains 49 queries and articles labeled with binary relevance. This track focuses on providing fair exposure to articles from different groups specifically gender and geographic locations [30] (2) TREC 2022 Fair Ranking Track collection (TRECFair22) [29] which focuses on fair exposure to articles from 11 fairness attributes[12] (3) The NTCIR-17 FairWeb-1 Task (NTCIRFair23) [81] is based on the Chuweb21D-60 corpus [19] and the test collection contains 45 test queries which can be categorized

[9]Implemented via polyfuse https://github.com/rmit-ir/polyfuse.
[10]Note that ranking fusion (i.e., Stage 3 of our approach) is applicable only when explicit SRD techniques are used, as implicit SRD relies solely on document text and does not consider explicitly defined attributes (e.g., gender).
[11]The constant $\beta$ mitigates the impact of high rankings by outlier systems [21].
[12]The TRECFair21 and TRECFair22 shared tasks had two tracks (1) Single Rankings (2) Sequenced Rankings. Our study focuses on the single ranking settings to investigate the effects of SRD on fairness.

into three types based on the entity being sought – 15 about Researchers (R-topics), 15 about Movies (M-topics), and 15 about YouTube videos (Y-topics).[13]

Table 2. Comparison of the three test collections – TRECFair21, TRECFair22, and NTCIRFair23 – based on their aims, corpora, topics, sample queries, fairness attributes, attribute types, and evaluation metrics.

| **Features** | **TRECFair21 [30]** | **TRECFair22 [29]** | **NTCIRFair23 [81]** |
|---|---|---|---|
| Aim | Documents representing nominal fairness attributes receive a fair exposure to the Wikipedia editors | Documents representing nominal fairness attributes receive a fair exposure to the Wikipedia editors | Tackling fairness for nominal and ordinal attributes. |
| Corpus | Wikimedia [14] | Wikimedia[15] | Chuweb21D-60 [19] |
| Topics | 49 | 47 | 45 |
| Sample Queries | Keyword Queries: ["government", "cabinet", "administration", "regime member"] …["computer engineer", "computer scientist", "programmer", "developer", "coder", "hacker", "software engineer"] | Keyword Queries: "Baseball", "Athletics" …"Chemistry" | Query Phrases: "Daniel Craig 007 movies", "information retrieval researchers" …"Coldplay covers on YouTube" |
| Attributes | geographic locations, gender | topic_countries, topic_regions, sources_countries, sources_regions, gender, topic_age, occupations, alphabetical, creation_date, pageviews and languages | gender, $h$-index, origin, rating, subscriber |
| No.of fairness attributes | 2 | 11 | 5 |
| Attribute Features | Nominal (e.g., Gender) | Nominal | 3 Ordinal (e.g., $h$-index) and 2 Nominal |
| Metrics used | Score = dot(nDCG, AWRF) | Score = dot(nDCG, AWRF) | GFR = GF(RNOD) + R(iRBU) |

### 4.2 Evaluation Measures

As discussed earlier in Sections 2.2 and 2.4, our study uses the following metrics for evaluation across multiple dimensions (i) For relevance, we use Normalized Discounted Cumulative Gain (nDCG) [43] and intentwise Rank-biased Utility (iRBU) [72] - an adaptation for the ad-hoc retrieval scenario of the Rank-Biased Utility (RBU) proposed by Amigó et al. [4].[16]; (ii) for diversity, we use $\alpha$-nDCG (discussed in Section 4.2.1); and (iii) for group fairness, we use Attention-Weighted Rank Fairness (AWRF) – intersectional fairness, JSD – nominal attributes, and RNOD – ordinal attributes (discussed in Section 4.2.2).

[13] In the NTCIRFair23 task, relevant entities meet the conditions in the topic description. As the topics are constructed around three distinct entity types (i.e., R-Researcher, M-Movie, Y-YouTube), they also function as topic categories within the evaluation framework. More details about the topic types can be found in [81].
[14] See http://boi.st/TREC2021Globus.
[15] See https://boi.st/TREC2022Globus.
[16] We use iRBU in this work, following Tao et al. [81] at NTCIRFair23.

*4.2.1 Diversity.* For evaluating diversity, we use $\alpha$-nDCG proposed by Clarke et al. [20] which is an aspect based variation of nDCG [43]. The main idea of this measure is to compute the gain of each document in terms of information nuggets, which are ultimately interpreted as sub-topics, or, as in our case, aspects ($\mathcal{T}$). Given a ranking $R$, a cutoff $k$, a set of aspects $\mathcal{T}$, and $\text{Rel}(d_i, t)$ indicating the relevance of document $d_i \in C_d$ for aspect value $t \in \mathcal{T}$:

$$\alpha\text{-DCG@}k(R) = \sum_{i=1}^{k} \frac{\sum_{t \in \mathcal{T}} \text{Rel}(d_i, t)(1-\alpha)^{c(i,t)}}{\log(i+1)} \tag{7}$$

where $c(i, t)$ represents the amount of documents previously observed that capture the aspect $t \in \mathcal{T}$. The parameter $\alpha \in [0, 1]$ controls the reward of diversity and novelty in the ranking. When $\alpha = 0$, $\alpha$-nDCG is equivalent to nDCG, and the higher the $\alpha$ value, the more reward is obtained by diversifying the ranking—at the expense of reward obtained by relevance. Similar to the original nDCG - an ad-hoc retrieval effectiveness measure, $\alpha$-DCG@k is normalized with an ideal gain vector ($\alpha$-DCG'@k) to compute $\alpha$-nDCG, where $\alpha$-DCG'@k represents the ideal ordering that maximizes cumulative gain at all levels, determined using the greedy approach by Clarke et al. [20].

*4.2.2 Fairness Measures.* In this work, we use two fairness measures: (i) Attention-Weighted Rank Fairness (AWRF) [76] and (ii) Group Fairness and Relevance (*GFR*) [71].

*Attention-Weighted Rank Fairness (AWRF):.* We use AWRF to measure intersectional group fairness. AWRF compares cumulative exposure $\epsilon$, of a ranking R, with a population estimator $P$, reflecting the target distribution, $T_d$.

$$\text{AWRF}(R) = \Delta(\epsilon(R), T_d) \tag{8}$$

A result list $R$ is more fair when $\epsilon$ is closer to the target distribution. The $\Delta$ is calculated as follows with Dis function using Jensen-Shannon Divergence (JSD) as shown in Equation 9.

$$\Delta(P_1, P_2) = 1 - \frac{1}{2}(\text{Dis}(P_1, T_d) + \text{Dis}(P_2, T_d)) \tag{9}$$

$$\text{where } T_d = \frac{1}{2}(P_1 + P_2) \tag{10}$$

*Group Fairness and Relevance (GFR).* The *GFR* [71] framework evaluates a ranking $R$, at a cutoff $k$, by using a weighted linear combination of utility (relevance in this context) and the distance between a document's group membership and the gold distribution (Shown in Equation 11).

$$\text{GFR}(R)@k = \sum_{pos=1}^{k} \text{Decay}(R, pos)\left(w_0 \text{Utility}(R, pos) + \sum_{f=1}^{|\mathcal{F}|} w_f \text{DistrSim}(R, pos)\right) \tag{11}$$

Let us say that we have a ranking $R$, for a topic $T$, at $pos = 1$, and we are trying to evaluate the *GFR* score along a set of two attributes $\mathcal{F} = \{f_1, f_2\}$. For convenience, we simplify some of the function names above where $\gamma = \text{Decay}(R, pos)$, $U = \text{Utility}(R, pos)$, and $\Delta = \text{DistrSim}(R, pos)$. A simplified version of Equation 11 becomes as shown in Equation 12. We follow Tao et al. [81], and set the weights to be all equal i.e., $w_U = w_{f_1} = w_{f_2} = \frac{1}{3}$. Equation 12 becomes as shown in Equation 13.

$$\text{GFR}(R) = \gamma\left(w_U.U + w_{f_1}.\Delta_{f_1} + w_{f_2}.\Delta_{f_2}\right) \tag{12}$$

$$\text{GFR}(R) = \frac{1}{3} \cdot \gamma \left( U + \Delta_{f_1} + \Delta_{f_2} \right) \tag{13}$$

Note that Equation 13 was derived from Equation 11 for a scenario with two fairness attributes ($|\mathcal{F}| = 2$), $k = 1$, and $w_U = w_{f_1} = w_{f_2} = \frac{1}{3}$, following Tao et al. [81] at NTCIRFair23. Equation 13 reveals that the current *GFR* evaluation gives equal weights to $U$ (i.e., Relevance), $\Delta_{f_1}$(i.e., fairness for attribute $f_1$), and $\Delta_{f_2}$ (i.e., fairness for attribute $f_2$). This would imply that, as the number of fairness attributes in the evaluation increases (when $w_{f_1} = w_{f_2} = \ldots = w_{f_n}$), the weight assigned to the fairness components of *GFR* (i.e., $\Delta_{f_1}, \Delta_{f_2}, \ldots, \Delta_{f_n}$) increases while the weight assigned to the utility component reduces. Note that this issue does not arise when $|\mathcal{F}| = 1$.

To address this imbalance, we impose a constraint $\psi$ on *GFR* (Equation 11), redefining it as Equation 14 where $\psi$ is the constraint such that $\boldsymbol{\psi = w_1 + w_2 + \cdots + w_n = (1 - w_0)}$, making $GFR^{\psi}$ a linear combination without compromising the weights to Utility as was in the original version (Equation 11).

$$\text{GFR}^{\psi}(R)@k = \sum_{pos=1}^{k} \text{Decay}(R, pos) \left( w_0 \text{Utility}(R, pos) + \sum_{f=1}^{|\mathcal{F}|} w_f \text{DistrSim}_f(R, pos) \right) \quad : \boldsymbol{\psi = w_1 + w_2 + \cdots + w_n = (1 - w_0)} \tag{14}$$

Additionally, we illustrate the distinction between *GFR* and $GFR^{\psi}$, empirically, by examining the variation in the ranking of the systems submitted to NTCIRFair23 using both *GFR* and $GFR^{\psi}$ (see Tables 3 4 5, and 6). Although the overall system ranks remain unchanged (Table 3), variations emerge across individual topic types. From Table 4, along $GFR^{\psi}$ for M-Topic, RSLFW-Q-RR-5 and UDinfo-Q-RR-2 improved ranking by 3 positions, and THUIR-QD-RR-3 went up by 1 position. While THUIR-QD-RR-4, run.qljm-depThre3-query and UDinfo-Q-RR-4 maintained their positions for both *GFR* and $GFR^{\psi}$, THUIR-QD-RG-2, went down by 1 position. Moreover, THUIR-D-RR-5 and RSLFW-Q-RR-4 fell 3 positions in $GFR^{\psi}$. From Table 5, along with $GFR^{\psi}$ for R-Topic, THUIR-QD-RR-3 went up in rankings by 2 positions, meanwhile THUIR-QD-RR-4, run.bm25-depThre3-query, and UDinfo-Q-RR-2 went up in ranking only by 1 position. Moreover, UDinfo-D-RR-3 and UDinfo-Q-RR-4 saw a drop in ranking by 2 positions and run.qld-depThre3-query by 1 position. Note that the top three systems - THUIR-QD-RG-2, THUIR-QD-RG-1, run.qld-depThre3-description maintained their positions and so did THUIR-D-RR-5. Conversely, no variation is observed between *GFR* and $GFR^{\psi}$ for Y-Topic (see Table 6), as it involves only a single fairness attribute. Overall, with $GFR^{\psi}$, the weight for relevance component (iRBU) increased from 1/3 (i.e., in *GFR*) to 1/2, giving equal importance to relevance compared to *GFR*. This rebalancing caused previously lower-ranked systems under *GFR* to improve with $GFR^{\psi}$, highlighting the importance of the $GFR^{\psi}$ constraint.

In summary, our empirical comparison of $GFR^{\psi}$ and *GFR* (see Tables 4 and 5) reveals that $GFR^{\psi}$ produces a different ranking for the top systems, particularly when multiple fairness attributes are involved. Furthermore, we mathematically distinguish $GFR^{\psi}$ from *GFR* (Equations 12 to 14), justifying our use of $GFR^{\psi}$ throughout the remainder of this paper. The components of $GFR^{\psi}$ are defined as follows:

$$\text{DistrSim}(R, pos) = \text{DistrSim}(P_{R,pos} || T_d) = 1 - \text{Dis}(P_{R,pos} || T_d) \tag{15}$$

where $P_{R,pos}$ is the achieved distribution at rank *pos*, of a ranking $R$, and $T_d$ is the target distribution. Dis is JSD for $\mathcal{F}$ containing nominal groups (e.g., gender, geographic locations, and so on), and RNOD (Root Normalised Order-aware Divergence) [69, 70] for $\mathcal{F}$ containing ordinal groups (e.g., $h$-index, ratings and so on) following Tao et al. [81]. The Decay function in Equation 14 is based on Expected Reciprocal Rank (ERR) [15] and is defined as follows:

Table 3. Ranking of all systems submitted to NTCIRFair23 measured along $GFR$ and $GFR^{\psi}$ across all topics ($n = 45$). Along both $GFR$ and $GFR^{\psi}$, the ranking of the systems is strongly correlated ($\tau = 0.8400, \text{p} \approx 0.0$).

| Rank | Run | $GFR$ | $GFR^{\psi}$ |
|---|---|---|---|
| 1 | THUIR-QD-RG-2 | 0.5164 | 0.5283 |
| 2 | THUIR-QD-RR-4 | 0.5079 | 0.5219 |
| 3 | THUIR-QD-RR-3 | 0.4977 | 0.5120 |
| 4 | THUIR-D-RR-5 | 0.4590 | 0.4625 |
| 5 | UDinfo-Q-RR-2 | 0.4483 | 0.4577 |
| 6 | RSLFW-Q-RR-5 | 0.4456 | 0.4556 |
| 7 | UDinfo-Q-RR-4 | 0.4469 | 0.4544 |
| 8 | THUIR-QD-RG-1 | 0.4452 | 0.4526 |
| 9 | RSLFW-Q-RR-4 | 0.4411 | 0.4463 |
| 10 | rmit_ir-Q-RR-5 | 0.4297 | 0.4389 |
| 11 | UDinfo-D-RR-3 | 0.4088 | 0.4144 |
| 12 | run.qld-depThre3-query | 0.4069 | 0.4129 |
| 13 | rmit_ir-D-RR-1 | 0.4017 | 0.4076 |
| 14 | UDinfo-D-RR-1 | 0.4003 | 0.4066 |
| 15 | UDinfo-D-RR-5 | 0.3943 | 0.4000 |
| 16 | run.qljm-depThre3-query | 0.3885 | 0.3999 |
| 17 | run.bm25-depThre3-query | 0.3889 | 0.3971 |
| 18 | rmit_ir-D-RR-4 | 0.3838 | 0.3912 |
| 19 | rmit_ir-D-RR-3 | 0.3816 | 0.3860 |
| 20 | run.qljm-depThre3-description | 0.3645 | 0.3702 |
| 21 | run.qld-depThre3-description | 0.3630 | 0.3686 |
| 22 | rmit_ir-D-RR-2 | 0.3563 | 0.3608 |
| 23 | run.bm25-depThre3-description | 0.3357 | 0.3424 |
| 24 | RSLFW-Q-MN-1 | 0.3195 | 0.3258 |
| 25 | RSLFW-Q-MN-2 | 0.3055 | 0.3100 |
| 26 | RSLFW-Q-MN-3 | 0.2046 | 0.2087 |
| | **Mean** | 0.4016 | 0.4089 |
| | **Median** | 0.4010 | 0.4071 |
| | **Std Dev** | 0.0671 | 0.0690 |

$$\text{Decay}(R, pos) = P^{sat}_{R,pos} \prod_{j=1}^{k-1} (1 - P^{sat}_{R,j}) \quad (k > 1) \tag{16}$$

where $P^{sat}_{R,pos}$ is the satisfaction probability at rank $pos$ [81]. Meanwhile, the Utility function of $GFR^{\psi}$ is defined as follows:

$$\text{Utility}(R, pos) = \text{iRBU}(R, pos) = \phi^{pos} \quad \text{where}, \phi = 0.99 \tag{17}$$

As for the Dis function, Equation 14 uses JSD [52] when attribute $\mathcal{F}$ is nominal and defined as follows:

$$\text{JSD}(P, P^*) = \frac{\text{KLD}(P||P^M) + \text{KLD}(P^*||P^M)}{2} \qquad P^M_i = \frac{P_i + P^*_i}{2} \tag{18}$$

where $P^*$ denotes the entire target/gold distribution probability mass function, and KLD denotes Kullback-Leibler Divergence, which is defined as follows:

Table 4. Ranking of all systems submitted to NTCIRFair23 measured along $GFR$ and $GFR^{\psi}$ over the M-Topic ($n = 15$) of the track. The horizontal line in the table is used to distinguish where exactly in the table the rankings start varying. The ranking of the systems along $GFR$ and $GFR^{\psi}$ shows strong, significant positive correlation using Kendall's $\tau$ ($\tau = 0.91$, p $\approx 0.0$). However, the top 9 systems along both $GFR$ and $GFR^{\psi}$ are only mildly correlated ($\tau = 0.5$, p $= 0.075$). The $\uparrow$ *and* $\downarrow$ depict the change in position in the ranking compared to the ranking based on $GFR$.

| Rank | Run | iRBU | GF-RNOD | GF-JSD | $GFR$ | Rank | Run | $GFR^{\Psi}$ |
|---|---|---|---|---|---|---|---|---|
| 1 | THUIR-QD-RG-2 | 0.6923 | 0.5788 | 0.5684 | 0.6132 | 1 | THUIR-QD-RR-3 $\uparrow^{1}$ | 0.6383 |
| 2 | THUIR-QD-RR-3 | 0.7230 | 0.5683 | 0.5391 | 0.6101 | 2 | THUIR-QD-RG-2 $\downarrow^{1}$ | 0.6330 |
| 3 | THUIR-QD-RR-4 | 0.6859 | 0.5435 | 0.5332 | 0.5875 | 3 | THUIR-QD-RR-4 | 0.6121 |
| 4 | run.qljm-depThre3-query | 0.6026 | 0.4871 | 0.4716 | 0.5205 | 4 | run.qljm-depThre3-query | 0.5410 |
| 5 | THUIR-D-RR-5 | 0.5316 | 0.4983 | 0.4900 | 0.5066 | 5 | UDinfo-Q-RR-2 $\uparrow^{3}$ | 0.5134 |
| 6 | RSLFW-Q-RR-4 | 0.5463 | 0.4758 | 0.4768 | 0.4996 | 6 | RSLFW-Q-RR-5 $\uparrow^{3}$ | 0.5130 |
| 7 | UDinfo-Q-RR-4 | 0.5582 | 0.4750 | 0.4601 | 0.4978 | 7 | UDinfo-Q-RR-4 | 0.5129 |
| 8 | UDinfo-Q-RR-2 | 0.5668 | 0.4706 | 0.4493 | 0.4956 | 8 | THUIR-D-RR-5 $\downarrow^{3}$ | 0.5129 |
| 9 | RSLFW-Q-RR-5 | 0.5674 | 0.4693 | 0.4479 | 0.4949 | 9 | RSLFW-Q-RR-4 $\downarrow^{3}$ | 0.5113 |
| 10 | rmit_ir-Q-RR-5 | 0.5819 | 0.4480 | 0.4177 | 0.4825 | 10 | rmit_ir-Q-RR-5 | 0.5074 |
| 11 | UDinfo-D-RR-5 | 0.5136 | 0.4488 | 0.4543 | 0.4722 | 11 | UDinfo-D-RR-5 | 0.4826 |
| 12 | rmit_ir-D-RR-4 | 0.5239 | 0.4281 | 0.4211 | 0.4577 | 12 | rmit_ir-D-RR-4 | 0.4742 |
| 13 | run.qld-depThre3-query | 0.4958 | 0.4351 | 0.4275 | 0.4528 | 13 | run.qld-depThre3-query | 0.4635 |
| 14 | run.bm25-depThre3-query | 0.5035 | 0.4283 | 0.4135 | 0.4484 | 14 | run.bm25-depThre3-query | 0.4622 |
| 15 | run.qljm-depThre3-description | 0.4883 | 0.4211 | 0.4273 | 0.4455 | 15 | run.qljm-depThre3-description | 0.4562 |
| 16 | rmit_ir-D-RR-3 | 0.4741 | 0.4234 | 0.3989 | 0.4321 | 16 | rmit_ir-D-RR-3 | 0.4426 |
| 17 | rmit_ir-D-RR-1 | 0.4862 | 0.4062 | 0.3842 | 0.4255 | 17 | rmit_ir-D-RR-1 | 0.4407 |
| 18 | rmit_ir-D-RR-2 | 0.4579 | 0.4035 | 0.3772 | 0.4129 | 18 | rmit_ir-D-RR-2 | 0.4241 |
| 19 | UDinfo-D-RR-1 | 0.4585 | 0.3913 | 0.3672 | 0.4057 | 19 | UDinfo-D-RR-1 | 0.4189 |
| 20 | THUIR-QD-RG-1 | 0.4400 | 0.3684 | 0.3395 | 0.3826 | 20 | THUIR-QD-RG-1 | 0.3970 |
| 21 | run.bm25-depThre3-description | 0.4337 | 0.3630 | 0.3401 | 0.3789 | 21 | run.bm25-depThre3-description | 0.3926 |
| 22 | RSLFW-Q-MN-2 | 0.4250 | 0.3503 | 0.3514 | 0.3756 | 22 | RSLFW-Q-MN-2 | 0.3879 |
| 23 | RSLFW-Q-MN-3 | 0.4250 | 0.3503 | 0.3514 | 0.3756 | 23 | RSLFW-Q-MN-3 | 0.3879 |
| 24 | UDinfo-D-RR-3 | 0.4069 | 0.3569 | 0.3476 | 0.3705 | 24 | UDinfo-D-RR-3 | 0.3796 |
| 25 | RSLFW-Q-MN-1 | 0.4268 | 0.3296 | 0.3179 | 0.3581 | 25 | RSLFW-Q-MN-1 | 0.3753 |
| 26 | run.qld-depThre3-description | 0.3728 | 0.3208 | 0.3122 | 0.3353 | 26 | run.qld-depThre3-description | 0.3447 |

$$\mathrm{KLD}(P||P^*) = \sum_{i \in \mathcal{F} \ s.t. \ P_i > 0} P_i \log_2 \frac{P_i}{P_i^*} \tag{19}$$

When the attribute $\mathcal{F}$ is ordinal in nature, the Dis function employs RNOD [69, 70], which is defined as follows:

$$\mathrm{RNOD}(P||P^*) = \sqrt{\frac{\mathrm{OD}(P||P^*)}{|C| - 1}} \tag{20}$$

Let $C^* = i \in C | P_i^* > 0$ where $C^*(C)$ is the set of classes with a non-zero gold probability. Thereby, Order-Aware Divergence (OD) can be defined as shown in Equation 21.

$$\mathrm{OD}(P||P^*) = \frac{1}{|C^*|} \sum_{i \in C^*} DW_i \tag{21}$$

where $DW_i$ is the Distance-weighted sum of squares of Class i and can be defined as shown in Equation 22.

$$DW_i = \sum_{j \in C} \delta_{ij} (P_j - {P_j}^*)^2 \quad \delta_{ij} = |i - j| \tag{22}$$

Table 5. Ranking of all systems submitted to NTCIRFair23 measured along $GFR$ and $GFR^{\psi}$ over the R-Topic ($n = 15$) of the track. The horizontal lines in the table are used to distinguish where exactly in the table the rankings start varying. The ranking of the systems along $GFR$ and $GFR^{\psi}$ has strong significant positive correlation using Kendall's $\tau$ ($\tau = 0.7830$, p $\approx 0.0$). However, the top 11 systems along both $GFR$ and $GFR^{\psi}$ are only mildly significantly correlated ($\tau = 0.5272$, p $\approx 0.0263$). The $\uparrow$ *and* $\downarrow$ depict the change in position in the ranking compared to the ranking based on $GFR$.

| Rank | Run | iRBU | GF-RNOD | GF-JSD | $GFR$ | Rank | Run | $GFR^{\Psi}$ |
|---|---|---|---|---|---|---|---|---|
| 1 | THUIR-QD-RG-2 | 0.6560 | 0.5352 | 0.5831 | 0.5914 | 1 | THUIR-QD-RG-2 | 0.6075 |
| 2 | THUIR-QD-RG-1 | 0.6013 | 0.5257 | 0.5823 | 0.5698 | 2 | THUIR-QD-RG-1 | 0.5776 |
| 3 | run.qld-depThre3-description | 0.5695 | 0.4975 | 0.5497 | 0.5389 | 3 | run.qld-depThre3-description | 0.5466 |
| 4 | UDinfo-D-RR-3 | 0.5582 | 0.4866 | 0.5374 | 0.5274 | 4 | THUIR-QD-RR-4 $\uparrow^1$ | 0.5430 |
| 5 | THUIR-QD-RR-4 | 0.5957 | 0.4720 | 0.5086 | 0.5254 | 5 | THUIR-QD-RR-3 $\uparrow^2$ | 0.5367 |
| 6 | run.qld-depThre3-query | 0.5518 | 0.4807 | 0.5356 | 0.5227 | 6 | UDinfo-D-RR-3 $\downarrow^2$ | 0.5351 |
| 7 | THUIR-QD-RR-3 | 0.5804 | 0.4875 | 0.4987 | 0.5222 | 7 | run.qld-depThre3-query $\downarrow^1$ | 0.5300 |
| 8 | THUIR-D-RR-5 | 0.5351 | 0.4841 | 0.5351 | 0.5181 | 8 | THUIR-D-RR-5 | 0.5224 |
| 9 | UDinfo-Q-RR-4 | 0.5367 | 0.4650 | 0.5190 | 0.5069 | 9 | run.bm25-depThre3-query $\uparrow^1$ | 0.5170 |
| 10 | run.bm25-depThre3-query | 0.5489 | 0.4605 | 0.5096 | 0.5064 | 10 | UDinfo-Q-RR-2 $\uparrow^1$ | 0.5170 |
| 11 | UDinfo-Q-RR-2 | 0.5489 | 0.4605 | 0.5096 | 0.5064 | 11 | UDinfo-Q-RR-4 $\downarrow^2$ | 0.5144 |
| 12 | RSLFW-Q-RR-5 | 0.5488 | 0.4556 | 0.4986 | 0.5010 | 12 | RSLFW-Q-RR-5 | 0.5130 |
| 13 | UDinfo-D-RR-1 | 0.5055 | 0.4434 | 0.4985 | 0.4825 | 13 | UDinfo-D-RR-1 | 0.4882 |
| 14 | RSLFW-Q-RR-4 | 0.4974 | 0.4562 | 0.4886 | 0.4807 | 14 | RSLFW-Q-RR-4 | 0.4849 |
| 15 | rmit_ir-Q-RR-5 | 0.4880 | 0.4530 | 0.4927 | 0.4779 | 15 | rmit_ir-Q-RR-5 | 0.4804 |
| 16 | rmit_ir-D-RR-1 | 0.4816 | 0.4509 | 0.4819 | 0.4715 | 16 | rmit_ir-D-RR-1 | 0.4740 |
| 17 | run.bm25-depThre3-description | 0.4801 | 0.4155 | 0.4694 | 0.4550 | 17 | run.bm25-depThre3-description | 0.4612 |
| 18 | run.qljm-depThre3-query | 0.4971 | 0.3999 | 0.4315 | 0.4428 | 18 | run.qljm-depThre3-query | 0.4564 |
| 19 | run.qljm-depThre3-description | 0.4361 | 0.3824 | 0.4120 | 0.4101 | 19 | run.qljm-depThre3-description | 0.4166 |
| 20 | rmit_ir-D-RR-3 | 0.4133 | 0.3815 | 0.4125 | 0.4025 | 20 | rmit_ir-D-RR-3 | 0.4052 |
| 21 | rmit_ir-D-RR-4 | 0.4063 | 0.3613 | 0.3861 | 0.3846 | 21 | rmit_ir-D-RR-4 | 0.3900 |
| 22 | UDinfo-D-RR-5 | 0.4100 | 0.3554 | 0.3829 | 0.3828 | 22 | UDinfo-D-RR-5 | 0.3896 |
| 23 | RSLFW-Q-MN-1 | 0.3700 | 0.3395 | 0.3771 | 0.3622 | 23 | RSLFW-Q-MN-1 | 0.3642 |
| 24 | rmit_ir-D-RR-2 | 0.3551 | 0.3255 | 0.3572 | 0.3459 | 24 | rmit_ir-D-RR-2 | 0.3482 |
| 25 | RSLFW-Q-MN-2 | 0.3084 | 0.2916 | 0.3080 | 0.3027 | 25 | RSLFW-Q-MN-2 | 0.3041 |
| 26 | RSLFW-Q-MN-3 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 27 | RSLFW-Q-MN-3 | 0.0000 |

### 4.3 Meta-Evaluation

In order to examine how the ranking of systems in TRECFair21 and TRECFair22 has changed, we used Kendall's $\tau$ correlation to compare evaluation measures. The $\tau$ values in this correlation range from $[-1, 1]$. Note that Kendall's $\tau$ considers only the order of values being studied, not their magnitude. This makes it an appropriate test to analyse changes in the ranking of submissions. In both collections, the measures are averaged across topics and compared to determine system rankings.

We used a randomised version of Tukey HSD with paired observations (RTHSD) [68] for statistical significance. This is a distribution-free parametric, single-step multiple comparison procedure. It is free from assumptions of normality[17] and homoscedasticity[18] which are required for a significance test to be applicable between two or more systems [68]. Furthermore, it is not subject to the studentized range distribution (see [78]), which can limit certain tests. In this work, we report statistical significance using RTHSD with $B = 5,000$ trials and significance criterion $\alpha = 0.05$ using the

[17]The assumption that two or more systems shares a normal distribution for a significance test to be applicable.
[18]The assumption that two or more systems share common variance, $\sigma^2$.

Table 6. Ranking of all systems submitted to NTCIRFair23 measured along $GFR$ and $GFR^{\psi}$ over the Y-Topic ($n = 15$) of the track. The ranking of the systems along $GFR$ and $GFR^{\psi}$ does not change as indicated by Kendall's $\tau$ ($\tau = 1.0$, p $\approx 0.0$).

| Rank | Run | $GFR$ | Rank | Run | $GFR^{\Psi}$ |
|---|---|---|---|---|---|
| 1 | THUIR-QD-RR-4 | 0.4107 | 1 | THUIR-QD-RR-4 | 0.4107 |
| 2 | THUIR-QD-RG-1 | 0.3832 | 2 | THUIR-QD-RG-1 | 0.3832 |
| 3 | THUIR-QD-RR-3 | 0.3608 | 3 | THUIR-QD-RR-3 | 0.3608 |
| 4 | THUIR-D-RR-5 | 0.3523 | 4 | THUIR-D-RR-5 | 0.3523 |
| 5 | THUIR-QD-RG-2 | 0.3445 | 5 | THUIR-QD-RG-2 | 0.3445 |
| 6 | UDinfo-Q-RR-2 | 0.3428 | 6 | UDinfo-Q-RR-2 | 0.3428 |
| 7 | RSLFW-Q-RR-4 | 0.3428 | 7 | RSLFW-Q-RR-4 | 0.3428 |
| 8 | RSLFW-Q-RR-5 | 0.3408 | 8 | RSLFW-Q-RR-5 | 0.3408 |
| 9 | UDinfo-Q-RR-4 | 0.3361 | 9 | UDinfo-Q-RR-4 | 0.3361 |
| 10 | rmit_ir-Q-RR-5 | 0.3288 | 10 | rmit_ir-Q-RR-5 | 0.3288 |
| 11 | UDinfo-D-RR-3 | 0.3285 | 11 | UDinfo-D-RR-3 | 0.3285 |
| 12 | UDinfo-D-RR-5 | 0.3280 | 12 | UDinfo-D-RR-5 | 0.3280 |
| 13 | UDinfo-D-RR-1 | 0.3128 | 13 | UDinfo-D-RR-1 | 0.3128 |
| 14 | rmit_ir-D-RR-2 | 0.3101 | 14 | rmit_ir-D-RR-2 | 0.3101 |
| 15 | rmit_ir-D-RR-3 | 0.3101 | 15 | rmit_ir-D-RR-3 | 0.3101 |
| 16 | rmit_ir-D-RR-4 | 0.3092 | 16 | rmit_ir-D-RR-4 | 0.3092 |
| 17 | rmit_ir-D-RR-1 | 0.3081 | 17 | rmit_ir-D-RR-1 | 0.3081 |
| 18 | run.qld-depThre3-query | 0.2453 | 18 | run.qld-depThre3-query | 0.2453 |
| 19 | RSLFW-Q-MN-3 | 0.2381 | 19 | RSLFW-Q-MN-3 | 0.2381 |
| 20 | RSLFW-Q-MN-2 | 0.2381 | 20 | RSLFW-Q-MN-2 | 0.2381 |
| 21 | RSLFW-Q-MN-1 | 0.2381 | 21 | RSLFW-Q-MN-1 | 0.2381 |
| 22 | run.qljm-depThre3-description | 0.2377 | 22 | run.qljm-depThre3-description | 0.2377 |
| 23 | run.qld-depThre3-description | 0.2147 | 23 | run.qld-depThre3-description | 0.2147 |
| 24 | run.bm25-depThre3-query | 0.2121 | 24 | run.bm25-depThre3-query | 0.2121 |
| 25 | run.qljm-depThre3-query | 0.2024 | 25 | run.qljm-depThre3-query | 0.2024 |
| 26 | run.bm25-depThre3-description | 0.1733 | 26 | run.bm25-depThre3-description | 0.1733 |

Discpower toolkit.[19] There are $2^n$ different ways pairwise permutations can be made; however, as indicated by Sakai [68] and later followed by Tao et al. [81], we choose $B = 5{,}000$ ($\ll 2^n$) to save computation costs.[20]

## 5 RESULTS AND ANALYSIS

In this section, we report the results of our experiments.

**Diversity and Fairness**. We begin by evaluating the systems submitted to TRECFair21 and TRECFair22 across three dimensions: relevance (nDCG), fairness (AWRF), and diversity ($\alpha$-nDCG). According to Table 7, the most fair system among all the submitted runs is UoGTrDExpDisT1 (AWRF = 0.8299), whereas the least fair system is UoGTrRelT1 (AWRF = 0.6559), as indicated by their ranks 1 and 10, respectively.

The top group (>4-10) comprising 3 systems significantly outperforms the remaining 7 systems along AWRF. However, when the 10 systems were sorted based on their diversity ($\alpha$-nDCG) and relevance (nDCG) scores, we identified a change in the overall ranks. We observe that the most fair system, UoGTrDExpDisT1 ($\alpha$-nDCG = 0.4796), is no longer the best performing along $\alpha$-nDCG and at the second position with UoGTrRelDiT1 ($\alpha$-nDCG = 0.6248) being the best-performing system and significantly outperforms all the remaining systems ranked from 2 to 10. It is worth mentioning that

[19] See http://research.nii.ac.jp/ntcir/tools/discpower-en.html.
[20] A detailed explanation of the RTHSD significance test and its implementation is provided by Sakai [68].

Table 7. Ranking of top 10 systems submitted to TRECFair21 measured along AWRF (Fairness), $\alpha$-nDCG (Diversity), and nDCG (Relevance) over the 49 topics of the track. ">" indicates statistically significant improvement according to the RTHSD test with $B$ = 5,000 trials and significance level $\alpha = 0.05$ [68]. For example, if Group = (> 4–10) for Run $A$, it indicates that Run $A$ statistically significantly outperforms the systems ranked fourth through tenth. The horizontal lines in the table are used to distinguish different groups whose members do not significantly outperform each other. This does not imply they are equivalent; further equivalence tests may be required.

| Rank | Run | AWRF | Group |
|---|---|---|---|
| 1 | UoGTrDExpDisT1 | 0.8299 | (>4–10) |
| 2 | UoGTrDExpDisLT1 | 0.8197 | (>4–10) |
| 3 | UoGTrDRelDiT1 | 0.8072 | (>4–10) |
| 4 | UoGTrDivPropT1 | 0.7112 | – |
| 5 | 2step_pair | 0.6943 | – |
| 6 | 1step_pair | 0.6940 | – |
| 7 | 2step_pair_list | 0.6912 | – |
| 8 | 1step_pair_list | 0.6908 | – |
| 9 | RUN1 | 0.6627 | – |
| 10 | UoGTrRelT1 | 0.6559 | – |

| Rank | Run | $\alpha$-nDCG | Group |
|---|---|---|---|
| 1 | UoGTrDRelDiT1 | 0.6248 | (>2–10) |
| 2 | UoGTrDExpDisT1 | 0.4796 | (>3–10) |
| 3 | UoGTrDExpDisLT1 | 0.3647 | (>5–10) |
| 4 | UoGTrDivPropT1 | 0.2619 | (>9–10) |
| 5 | UoGTrRelT1 | 0.2307 | – |
| 6 | RUN1 | 0.1808 | – |
| 7 | 1step_pair | 0.1313 | – |
| 8 | 1step_pair_list | 0.1297 | – |
| 9 | 2step_pair | 0.1184 | – |
| 10 | 2step_pair_list | 0.1123 | – |

| Rank | Run | nDCG | Group |
|---|---|---|---|
| 1 | RUN1 | 0.2169 | (>7–10) |
| 2 | UoGTrDivPropT1 | 0.2157 | (>7–10) |
| 3 | UoGTrRelT1 | 0.2120 | (>7–10) |
| 4 | UoGTrDExpDisT1 | 0.2071 | (>7–10) |
| 5 | UoGTrDRelDiT1 | 0.2001 | (>7–10) |
| 6 | UoGTrDExpDisLT1 | 0.1776 | (>7–10) |
| 7 | 1step_pair | 0.0838 | – |
| 8 | 2step_pair | 0.0824 | – |
| 9 | 1step_pair_list | 0.0820 | – |
| 10 | 2step_pair_list | 0.0786 | – |

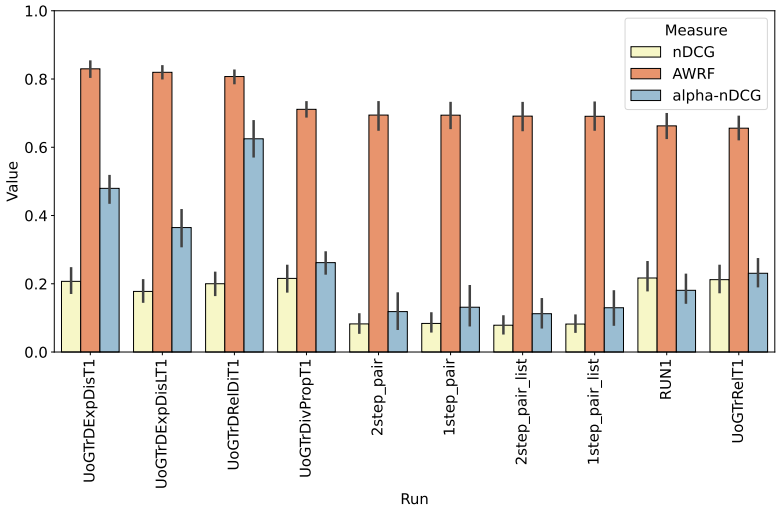
UoGTrDExpDisLT1 ($\alpha$-nDCG = 0.3647), which previously ranked second, along AWRF, is now pushed to the third position along $\alpha$-nDCG. Furthermore, it statistically significantly outperforms only systems ranked at 5 to 10. Moving onto the third dimension i.e., relevance measured using nDCG, we observe that the top group – comprising six systems – significantly outperforms the remaining, where the most relevant system is RUN1 (nDCG = 0.2169) and the least relevant system is 2step_pair_list (nDCG = 0.0786). In summary, results from TRECFair21 reveal that relevance does not necessarily imply fairness, and fair systems do not consistently achieve diversity.

Similarly, Table 8 indicates that the most fair system at TRECFair22 is UoGRelvOnlyT1 (AWRF = 0.5246). In contrast, FRT_attention (AWRF = 0.4484) exhibits the lowest fairness score. Notably, the top-performing system exhibits statistically significant improvements only over systems ranked 9 through 18. Consistent with the trends in TRECFair21, we observed a shift in the system ranks, wherein the most diverse system, ans_bm25 ($\alpha$-nDCG = 0.3503), is ranked in the lowest fairness group, at position 15. Conversely, the least diverse system UoGTrExpE2 ($\alpha$-nDCG = 0.1184), ranks among the top three systems in terms of fairness (AWRF = 0.5246). These observations highlight a trade-off between fairness and diversity, as the top-performing group – comprising four systems – significantly outperforms the top three fair systems in terms of diversity. Additionally, we observe that the most relevant system demonstrates statistically significant improvements over all systems ranked from position 3 onwards. In contrast, the second-best system exhibits statistically significant gains only over the least relevant system – UoGTrExpE1 (nDCG = 0.5176).

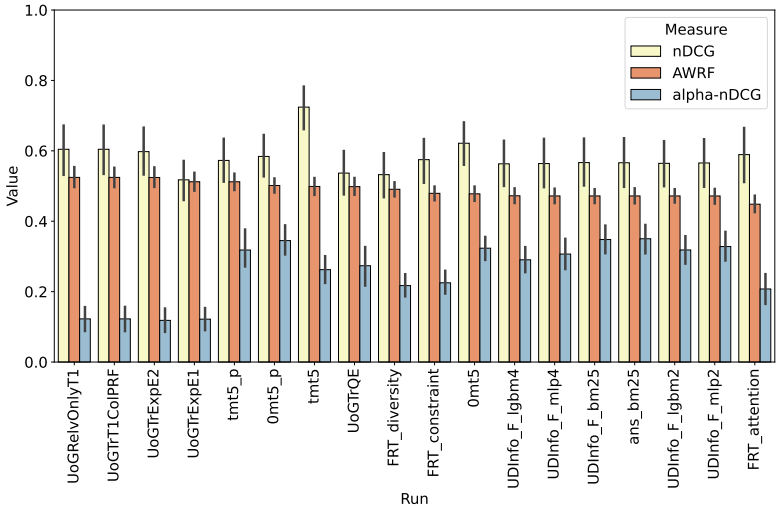
We corroborate our findings from the previous discussion of the TRECFair22, that fair systems are not necessarily diverse and vice-versa. For instance, we almost see an inversion of the ranking of systems in Table 8, where the most fair are least diverse and so on. Moreover, Figures 3a and 3b demonstrate that across TRECFair21 and TRECFair22, no consistent trends emerge among relevance (nDCG), fairness (AWRF), and diversity ($\alpha$-nDCG) indicating a non-monotonous relationship among these dimensions.[21] To investigate this relationship in greater detail, we perform a Kendall's $\tau$ correlation analysis [45] across multiple cutoff values $k = \{1, 3, 5, 10, 20\}$, as presented in Table 9.

In TRECFair21, for $k = 20$, we observe that AWRF and $\alpha$-nDCG are positively correlated, although not statistically significant. However, for $k = \{1, 3, 5, 10\}$ AWRF negatively correlates with $\alpha$-nDCG, although the correlations are low, and

[21] A non-monotonous relationship between three variables means that as one variable increases or decreases, the other two do not consistently follow a single trend. This suggests a complex relationship requiring trade-offs and a careful balance to optimise all three dimensions simultaneously.




(a) Top 10 systems at TRECFair21.




(b) Top 18 systems at TRECFair22.

Fig. 3. Comparison of nDCG, AWRF and $\alpha$-nDCG for systems submitted to TRECFair21 and TRECFair22, highlighting the non-monotonous relationship between the three dimensions. In both (a) and (b), the systems are shown in the decreasing order of fairness i.e., AWRF, and for all measures, higher value is better.

statistical significance was observed only for $k = 1$. This corroborates the relationship between fairness and diversity metrics explored by Pathiyan Cherumanal et al. [59] and Sakai et al. [71].

A similar pattern is observed in TRECFair22, where AWRF and $\alpha$-nDCG exhibit a positive correlation at $K = 20$, but negative correlations at $K = \{1, 3, 5, 10\}$. For nDCG, positive correlations are observed at $K = \{3, 5, 10\}$, with a slight negative correlation at $K = 1$. However, none of these correlations are statistically significant.

Table 8. Ranking of top 18 systems submitted to TRECFair22 measured along AWRF (Fairness), $\alpha$-nDCG (Diversity), and nDCG (Relevance) over the 47 topics. ">" indicates statistically significant improvement according to the RTHSD test with $B$ = 5,000 trials and significance level $\alpha = 0.05$ [68]. For example, if Group = (> 9–18) for Run $A$, it indicates that Run $A$ statistically significantly outperforms the systems ranked ninth through eighteenth. The horizontal lines in the table are used to distinguish different groups whose members do not significantly outperform each other. This does not imply they are equivalent; further equivalence tests may be required.

| Rank | Run | AWRF | Group |
|---|---|---|---|
| 1 | UoGRelvOnlyT1 | 0.5246 | (>9-18) |
| 2 | UoGTrT1ColPRF | 0.5246 | (>9–18) |
| 3 | UoGTrExpE2 | 0.5243 | (>9-18) |
| 4 | UoGTrExpE1 | 0.5122 | (>10–18) |
| 5 | tmt5_p | 0.5121 | (>10–18) |
| 6 | 0mt5_p | 0.5015 | (>12–18) |
| 7 | tmt5 | 0.4988 | (>18) |
| 8 | UoGTrQE | 0.4983 | (>18) |
| 9 | FRT_diversity | 0.4908 | (>18) |
| 10 | FRT_constraint | 0.4793 | (>18) |
| 11 | 0mt5 | 0.4778 | (>18) |
| 12 | UDInfo_F_lgbm4 | 0.4723 | – |
| 13 | UDInfo_F_mlp4 | 0.4719 | – |
| 14 | UDInfo_F_bm25 | 0.4719 | – |
| 15 | ans_bm25 | 0.4719 | – |
| 16 | UDInfo_F_lgbm2 | 0.4719 | – |
| 17 | UDInfo_F_mlp2 | 0.4718 | – |
| 18 | FRT_attention | 0.4484 | – |

| Rank | Run | $\alpha$-nDCG | Group |
|---|---|---|---|
| 1 | ans_bm25 | 0.3503 | (>12–18) |
| 2 | UDInfo_F_bm25 | 0.3480 | (>12–18) |
| 3 | 0mt5_p | 0.3450 | (>12–18) |
| 4 | UDInfo_F_mlp2 | 0.3281 | (>12–18) |
| 5 | 0mt5 | 0.3234 | (>13–18) |
| 6 | UDInfo_F_lgbm2 | 0.3184 | (>13–18) |
| 7 | tmt5_p | 0.3182 | (>13–18) |
| 8 | UDInfo_F_mlp4 | 0.3068 | (>14–18) |
| 9 | UDInfo_F_lgbm4 | 0.2904 | (>15–18) |
| 10 | UoGTrQE | 0.2736 | (>15–18) |
| 11 | tmt5 | 0.2625 | (>15–18) |
| 12 | FRT_constraint | 0.2251 | (>17–18) |
| 13 | FRT_diversity | 0.2171 | – |
| 14 | FRT_attention | 0.2076 | – |
| 15 | UoGTrT1ColPRF | 0.1226 | – |
| 16 | UoGRelvOnlyT1 | 0.1226 | – |
| 17 | UoGTrExpE1 | 0.1218 | – |
| 18 | UoGTrExpE2 | 0.1184 | – |

| Rank | Run | nDCG | Group |
|---|---|---|---|
| 1 | tmt5 | 0.7242 | (>3–18) |
| 2 | 0mt5 | 0.6216 | (>18) |
| 3 | UoGRelvOnlyT1 | 0.6044 | – |
| 4 | UoGTrT1ColPRF | 0.6044 | – |
| 5 | UoGTrExpE2 | 0.5977 | – |
| 6 | FRT_attention | 0.5893 | – |
| 7 | 0mt5_p | 0.5841 | – |
| 8 | FRT_constraint | 0.5749 | – |
| 9 | tmt5_p | 0.5728 | – |
| 10 | UDInfo_F_bm25 | 0.5666 | – |
| 11 | ans_bm25 | 0.5661 | – |
| 12 | UDInfo_F_mlp2 | 0.5655 | – |
| 13 | UDInfo_F_lgbm2 | 0.5645 | – |
| 14 | UDInfo_F_mlp4 | 0.5639 | – |
| 15 | UDInfo_F_lgbm4 | 0.5631 | – |
| 16 | UoGTrQE | 0.5368 | – |
| 17 | FRT_diversity | 0.5323 | – |
| 18 | UoGTrExpE1 | 0.5176 | – |

Table 9. Kendall's $\tau$ correlation (with 95% CIs) between AWRF, nDCG, and $\alpha$-nDCG for the TRECFair21 and TRECFair22 runs. The * indicates statistical significance ($p < 0.05$).

| Cutoff K | TRECFair21 | | TRECFair22 | |
|---|---|---|---|---|
| | nDCG | $\alpha$-nDCG | nDCG | $\alpha$-nDCG |
| 1 | −0.7333 | −0.3778* | −0.0882 | −0.1765 |
| 3 | −0.4222 | −0.2000 | 0.2647 | −0.2941 |
| 5 | −0.0667 | −0.3333 | 0.2059 | −0.1324 |
| 10 | 0.1556 | −0.1111 | 0.1765 | −0.3382 |
| 20 | 0.1556 | 0.2000 | 0.1029 | 0.1471 |

In summary, Table 9 reports the correlation results for systems submitted to TRECFair21 and TRECFair22, demonstrating that the fairness metric (AWRF) and the diversity metric ($\alpha$-nDCG) reflect distinct and complementary dimensions of system behaviour, as defined by the target distribution used in the shared tasks. To answer our *RQ1: Is a system fairer towards multiple fairness attributes necessarily more diverse (i.e., maximise novelty and minimise redundancy)*?, we can say that systems fairer towards multiple fairness attributes are not necessarily diverse i.e., may contain redundant fairness attributes and lack novelty. Therefore, to promote diversity alongside fairness and relevance, we propose using the harmonic mean (H-Score) as an aggregation function over AWRF, $\alpha$-nDCG, and nDCG, as defined in Equation 24. In prior works such as TRECFair21 and TRECFair22, system performance was aggregated using Score, computed as dot(AWRF, nDCG). Extending this approach to incorporate diversity would require computing dot(nDCG, AWRF, $\alpha$-nDCG). However, this extension may not adequately capture the overall performance of the system. For instance, the topic-wise measure for the best system from TRECFair21 (UoGTrDExpDisT1) would be computed as Score(nDCG = 0.207, AWRF = 0.8299, $\alpha$-nDCG = 0.4796) = 0.08. Despite strong fairness and moderate diversity, the overall score is substantially reduced due

Table 10. SRD on three shared tasks – TRECFair21, TRECFair22, and NTCIRFair23. For relevance, we measure nDCG and iRBU, for fairness, two measures are indicated – AWRF and GF. $\alpha$-nDCG is reported for the diversity measure. Aggregated scores from three shared tasks presented i.e., Score(nDCG,AWRF), $GFR^{\psi}$(iRBU,GF), and H-Score represent the harmonic mean between nDCG, AWRF, and $\alpha$-nDCG. $^*$ indicates statistically significant improvement over the initial ranking, BM25 ($p < 0.05$). "Best run" refers to the top-performing system as identified by the shared task organisers for each track. $^D$ denotes systems that use explicit SRD techniques.

| Configuration | | Relevance | | Fairness | | Diversity | Overall | | |
|---|---|---|---|---|---|---|---|---|---|
| **Shared Task** | **System** | **nDCG** | **iRBU** | **AWRF** | **GF** | **$\alpha$-nDCG** | **Score** | **$GFR^{\psi}$** | **H-Score(nDCG,AWRF,$\alpha$-nDCG)** |
| **TRECFair21** | BM25 | **0.2075** | 0.7895 | 0.6413 | 0.3140 | 0.2156 | 0.1331 | 0.4733 | 0.2723 |
| | BM25 + MMR | 0.1768 | 0.8023 | 0.6582 | 0.3216 | 0.2163 | 0.1164 | 0.4816 | 0.2543 |
| | BM25 + PM-2$^D$ | 0.1943 | 0.8919 | 0.7300$^*$ | **0.3473** | 0.4264$^*$ | 0.1418 | 0.5328 | 0.3385$^*$ |
| | UoGTrDExpDisT1 (Best run) | 0.2070 | **0.9394** | **0.8299**$^*$ | 0.2928 | **0.4796**$^*$ | **0.1718** | **0.5429** | **0.6079**$^*$ |
| **TRECFair22** | BM25 | 0.5438 | **0.8968** | 0.4416 | 0.6111 | 0.3264 | 0.2401 | 0.3046 | 0.3754 |
| | BM25 + MMR | 0.5244 | 0.8358 | 0.4498 | 0.6306 | 0.3283 | 0.2359 | 0.4225 | 0.3796 |
| | BM25 + PM-2$^D$ | 0.5385 | 0.8897 | **0.5017** | **0.6813** | **0.3540**$^*$ | 0.2702 | **0.5130**$^*$ | **0.4151**$^*$ |
| | tmt5 (Best run) | **0.7242**$^*$ | 0.8925 | 0.4988 | 0.6732 | 0.3182 | **0.3626** | 0.4630 | 0.3775 |
| **NTCIRFair23** | BM25 | 0.1686 | 0.4242 | 0.4616 | 0.3700 | 0.3026 | 0.0778 | 0.3971 | 0.3655 |
| | BM25 + MMR | 0.1313 | 0.3803 | 0.3952 | 0.3587 | 0.2640 | 0.0519 | 0.2857 | 0.3170 |
| | BM25 + PM-2$^D$ | 0.1593 | 0.4669$^*$ | 0.4057 | 0.3991$^*$ | 0.3184 | 0.0646 | 0.4330$^*$ | 0.3568 |
| | THUIR-QD-RG-2 (Best run) | **0.2013**$^*$ | **0.5744** | **0.5332**$^*$ | **0.4823** | **0.4013**$^*$ | **0.1073** | **0.5283** | **0.4580**$^*$ |

to relatively low relevance. This occurs because multiplicative aggregation aggressively penalises underperformance in any single dimension, driving the overall score toward zero more rapidly. Moreover, the impact of this limitation becomes more pronounced as additional evaluation dimensions are introduced. Therefore, in this work, we refrain from extending Score with an additional dimension. Similar to Van Rijsbergen [84], we compute the harmonic mean aggregating relevance, fairness, and diversity, as it equally prioritises each dimension and is particularly sensitive to low values.[22]

$$\text{Weighted H-Score} = \frac{w_1 + w_2 + w_3}{\left(\frac{w_1}{Relevance} + \frac{w_2}{Fairness} + \frac{w_3}{Diversity}\right)} \tag{23}$$

$$\text{H-Score} = \frac{3}{\left(\frac{1}{Relevance} + \frac{1}{Fairness} + \frac{1}{Diversity}\right)} \tag{24}$$

We investigate the effectiveness of two types of SRD techniques: an implicit approach based on SRD-MMR and an explicit approach using SRD-PM-2. Their performance is compared with retrieval baselines and the "Best run" from each shared task, as defined by official evaluation metrics. Note that TRECFair21 and TRECFair22 adopted Score, which aggregates nDCG and AWRF, while NTCIRFair23 employed GFR, which aggregates iRBU and GF. A detailed comparison of these shared tasks and their respective characteristics is presented in Table 2.

**Intersectional Fairness**. We use AWRF as the intersectional fairness measure (Section 2.2) for the systems reported in Table 10. For nominal attributes, results on TRECFair21 show that the explicit SRD method (PM-2) statistically significantly outperforms the retrieval baseline in terms of intersectional fairness but not the best system (UoGTrDExpDisT1), while the implicit SRD method (MMR) yields only marginal improvements over the retrieval baseline. A similar trend is observed in TRECFair22, where PM-2 achieves higher fairness than both the retrieval baseline as well as the best system from the

[22]Figures 11 and 12 present the intermediate computations involved in deriving the H-Score.

shared task, while MMR yields only marginal gains. For ordinal attributes in NTCIRFair23, both SRD-based techniques underperform compared to the top system from the shared task and also result in performance degradation relative to the retrieval baseline. Note that the top systems from NTCIRFair23 also utilise PM-2, and additionally utilise query augmentation/generation. *This implies that explicit* SRD *(*PM-2*) can provide relevant, fair and diverse rankings*; however, the impact of query augmentation/generation needs further investigation. Notably, Jaenich et al. [42] demonstrated that using generated queries consistently improves fairness while maintaining relevance in TRECFair22. Conversely, their influence on diversity remains uncertain and is an open research challenge beyond the scope of this study.

**Fairness**. As discussed in Section 4.2.2, we use two fairness measures in our analysis: AWRF, which captures intersectional fairness by considering combinations of attributes, and GF, which represents group fairness computed separately for each attribute (e.g., gender, popularity). Specifically, GF is computed as the mean of JSD and RNOD when ordinal attributes are involved, and forms the group fairness component of $GFR^{\psi}$. From Table 10 results based on GF show that PM-2 outperforms both the retrieval baseline and the implicit SRD method (MMR) in TRECFair21, achieving AWRF = 0.7300 and GF = 0.3473. Similarly, in TRECFair22, PM-2 improves over the retrieval baseline with AWRF = 0.5017 and GF = 0.6813, and additionally improved performance compared to the best system along GF for both TRECFair21 and TRECFair22. Accordingly, we answer **RQ2**: *Can diversification along multiple fairness attributes achieve fairer search results?* Explicit SRD method (PM-2) yields fairer search results than the retrieval baseline across multiple test collections and fairness measures, in a multi-attribute fairness setting. On the other hand, in the NTCIRFair23, while the explicit SRD (PM-2) performed better than the retrieval baseline, it was not consistent across AWRF and GF. We see PM-2 performing better than the retrieval baseline and implicit SRD (MMR) along GF but not AWRF. We see the best system from this track showed statistically significant improvement over all the systems with AWRF = 0.5332 and GF = 0.4823. We note that NTCIRFair23 evaluates fairness across both nominal and ordinal attributes, unlike TRECFair21 and TRECFair22, which involve only nominal attributes. Accordingly, we answer **RQ3**: *Can diversification achieve fairer results for nominal and ordinal fairness attributes?* The results demonstrate that both implicit and explicit SRD do not achieve fairer results when ordinal attributes are involved. Although results showed a slight improvement over the retrieval baseline, the improvements were not statistically significant.

**Diversity**. In TRECFair21, the explicit SRD method (PM-2) consistently outperforms both the retrieval baseline and MMR, achieving $\alpha$-nDCG = 0.4264. Similarly, in TRECFair22, PM-2 shows a statistically significant improvement over the retrieval baseline, MMR, and the best system from the track – tmt5. Conversely, in NTCIRFair23, the best system (THUIR-QD-RG-2) outperforms our approach, achieving $\alpha$-nDCG = 0.4013. In summary, PM-2 consistently improves diversity over the retrieval baseline in TRECFair21 and TRECFair22, but achieves only marginal gains in NTCIRFair23, which involves ordinal attributes.

**Overall**. As demonstrated earlier, a fair system does not necessarily mean a diverse system. Therefore, we use H-Score (see Equation 24) to measure relevance, fairness, and diversity. In both TRECFair21 and TRECFair22, which involve nominal attributes, the explicit SRD method (PM-2) consistently outperforms the retrieval baselines. Although similar trends are observed for the implicit SRD method (MMR), the gains are marginal. Conversely, in NTCIRFair23, which involves ordinal attributes, PM-2 underperforms compared to the retrieval baseline, likely due to its lack of native support for ordinal attributes (see Equation 5). Compared to the best systems from the shared tasks, PM-2 does not outperform the top system in TRECFair21, but does so in TRECFair22. A similar pattern is observed in NTCIRFair23, where PM-2 fails to outperform the best system, consistent with its limitations in handling ordinal attributes discussed earlier. In summary,

Table 11. H-Score(AWRF, $\alpha$-nDCG) and H-Score(nDCG, AWRF, $\alpha$-nDCG) measures for each submitted run over the 49 topics of TRECFair21. ">" indicates statistically significant improvement according to the RTHSD test with $B = 5{,}000$ trials and significance level $\alpha = 0.05$ [68]. The horizontal lines in the table are used to distinguish different groups whose members do not significantly outperform each other. This does not imply they are equivalent; further equivalence tests may be required.

| Rank | Run | H-Score(AWRF, $\alpha$-nDCG) | Group |
|---|---|---|---|
| 1 | UoGTrDRelDiT1 | 0.6861 | (>3–10) |
| 2 | UoGTrDExpDisT1 | 0.5947 | (>4–10) |
| 3 | UoGTrDExpDisLT1 | 0.4773 | (>6–10) |
| 4 | UoGTrDivPropT1 | 0.3709 | (>7–10) |
| 5 | UoGTrRelT1 | 0.3218 | (>9–10) |
| 6 | RUN1 | 0.2601 | – |
| 7 | 1step_pair_list | 0.1799 | – |
| 8 | 1step_pair | 0.1755 | – |
| 9 | 2step_pair | 0.1609 | – |
| 10 | 2step_pair_list | 0.1606 | – |

| Rank | Run | H-Score(nDCG,AWRF,$\alpha$-nDCG) | Group |
|---|---|---|---|
| 1 | UoGTrDRelDiT1 | 0.3512 | (>4-10) |
| 2 | UoGTrDExpDisT1 | 0.3334 | (>5–10) |
| 3 | UoGTrDExpDisLT1 | 0.2745 | (>7–10) |
| 4 | UoGTrDivPropT1 | 0.2614 | (>7–10) |
| 5 | UoGTrRelT1 | 0.2255 | (>7–10) |
| 6 | RUN1 | 0.2003 | – |
| 7 | 1step_pair | 0.0986 | – |
| 8 | 1step_pair_list | 0.0983 | – |
| 9 | 2step_pair | 0.0940 | – |
| 10 | 2step_pair_list | 0.0928 | – |

Table 12. H-Score(AWRF, $\alpha$-nDCG) and H-Score(nDCG,AWRF,$\alpha$-nDCG) measures for each submitted run over the 47 topics of TRECFair22. ">" indicates statistically significant improvement according to the RTHSD test with $B = 5{,}000$ trials and significance level $\alpha = 0.05$ [68]. The horizontal lines in the table are used to distinguish different groups whose members do not significantly outperform each other. This does not imply they are equivalent; further equivalence tests may be required.

| Rank | Run | H-Score(AWRF, $\alpha$-nDCG) | Group |
|---|---|---|---|
| 1 | 0mt5_p | 0.3934 | (>12–18) |
| 2 | ans_bm25 | 0.3842 | (>12–18) |
| 3 | UDInfo_F_bm25 | 0.3824 | (>12–18) |
| 4 | 0mt5 | 0.3713 | (>12–18) |
| 5 | UDInfo_F_mlp2 | 0.3674 | (>14–18) |
| 6 | tmt5_p | 0.3630 | (>14–18) |
| 7 | UDInfo_F_lgbm2 | 0.3617 | (>14–18) |
| 8 | UDInfo_F_mlp4 | 0.3505 | (>14–18) |
| 9 | UDInfo_F_lgbm4 | 0.3415 | (>14–18) |
| 10 | tmt5 | 0.3209 | (>15–18) |
| 11 | UoGTrQE | 0.3157 | (>15–18) |
| 12 | FRT_constraint | 0.2899 | (>15–18) |
| 13 | FRT_diversity | 0.2854 | (>15–18) |
| 14 | FRT_attention | 0.2495 | – |
| 15 | UoGTrT1ColPRF | 0.1727 | – |
| 16 | UoGRelvOnlyT1 | 0.1727 | – |
| 17 | UoGTrExpE1 | 0.1722 | – |
| 18 | UoGTrExpE2 | 0.1676 | – |

| Rank | Run | H-Score(nDCG,AWRF,$\alpha$-nDCG) | Group |
|---|---|---|---|
| 1 | 0mt5_p | 0.4239 | (>13–18) |
| 2 | ans_bm25 | 0.4152 | (>14–18) |
| 3 | UDInfo_F_bm25 | 0.4139 | (>14–18) |
| 4 | 0mt5 | 0.4121 | (>14–18) |
| 5 | UDInfo_F_mlp2 | 0.4019 | (>14–18) |
| 6 | tmt5_p | 0.3985 | (>14–18) |
| 7 | UDInfo_F_lgbm2 | 0.3957 | (>14–18) |
| 8 | UDInfo_F_mlp4 | 0.3896 | (>14–18) |
| 9 | tmt5 | 0.3775 | (>15–18) |
| 10 | UDInfo_F_lgbm4 | 0.3772 | (>15–18) |
| 11 | UoGTrQE | 0.3409 | (>15–18) |
| 12 | FRT_constraint | 0.3356 | (>15–18) |
| 13 | FRT_diversity | 0.3274 | (>15–18) |
| 14 | FRT_attention | 0.2913 | – |
| 15 | UoGTrT1ColPRF | 0.2122 | – |
| 16 | UoGRelvOnlyT1 | 0.2122 | – |
| 17 | UoGTrExpE2 | 0.2058 | – |
| 18 | UoGTrExpE1 | 0.2034 | – |

the explicit SRD method PM-2 consistently improves over the retrieval baseline across multi-attribute fairness and diversity when nominal attributes are involved.

## 6 CONCLUSION AND FUTURE WORK

Information access systems have been at the forefront and helping users make decisions in their day-to-day activities. However, certain biases in such systems can introduce new biases or reinforce existing biases in users, impacting society negatively. Recent research has focused on efforts to help mitigate such biases by attempting to achieve fairness in ranked systems. Research in this direction has focused on looking at relevance and fairness (as can be seen from the TREC and NTCIR fairness evaluation campaigns). However, our work shows that balancing fairness and relevance

affects diversity in rankings, and also demonstrates that explicit SRD combined with fusion techniques can enhance fairness and diversity in ranking systems.

**Limitations and Future Work**. While we provide a comprehensive analysis of the relationship between fairness and diversity, our work is not without limitations. For instance, we rely on relevance judgments from shared tasks, which may themselves be biased due to the lack of transparency on the annotators' backgrounds and preferences. Moreover, we study the relationship between fairness and diversity using systems submitted by the participants. To this end, we acknowledge that these systems were primarily developed to address fairness in search results, in alignment with the respective shared task objectives and evaluation criteria. These systems were not explicitly optimised for diversity. As such, comparatively lower performance on diversity measures should not be construed as a shortcoming of the system. While our study focuses on the effects of diversification as a re-ranking algorithm, the impact of other components in our approach (e.g., retrieval stage, fusion stage) requires further probing. The aforementioned problems could be addressed in the future by using a carefully curated synthetic test collection as demonstrated in a preliminary study earlier [59]. Regarding evaluation measures, we use H-Score as an aggregation technique, which is different from the linear combination style aggregation used in *GFR* family [71]. In the future, we plan to study the relation between these aggregation techniques as well as the impact of different user models in such evaluations. For instance, can we add another dimension (i.e., diversity) to the *GFR* framework? How does the number of fairness attributes impact the effectiveness of the aggregation technique? While our work proposes treating relevance, fairness, and diversity as key dimensions for IR evaluations, are there additional dimensions to consider in addressing biased information consumption (e.g., transparency, accountability, or safety)? Moreover, can different definitions of relevance, fairness, and diversity [72] impact evaluations? A recent study also highlights the challenges of combining fairness and relevance, offering suggestions from a recommender systems perspective [63]. This warrants further investigation in future work.

**Implications**. The emergence and rapid advancement of Large Language Models (LLMs) have transformed the field of information retrieval and how users consume information. One notable application of LLMs is the Retrieval-Augmented Generation (RAG) based search system where LLMs generate responses based on retrieved documents or passages. Recent research has empirically demonstrated that the quality of responses generated by the RAG system is influenced by the effectiveness of the retrieved list [61]. Consequently, this may indicate that a biased ranking can additionally lead to a biased generated response thereby highlighting the importance of ensuring fairness and diversity in the retrieval process. By fostering a deeper understanding of fairness and diversity within ranked systems, we can strive towards achieving fair and diverse responses in RAG-based search systems. Currently, we are unsure about the interplay between fairness and diversity in a RAG-based system. Exploring this may open up new challenges and research questions (both in information retrieval and the natural language community) such as (i) *"What is the impact of diversity and fairness of ranking in a RAG-based system?"*, (ii) *"How can these dimensions be measured in the final generated response of a RAG-based system?"*, and so on.

**Positionality Statement and Ethical Considerations**. The work described in this paper is strongly influenced by the perspectives of the authors as researchers working in the field of information retrieval. The notions of fairness and diversity have attracted significant attention from the research community in recent years (evident from the recent fairness-aware shared tasks like TRECFair21, TRECFair22, and NTCIRFair23) and the authors aim to further this cause by better understanding the relationship between various dimensions (i.e., relevance, fairness, and diversity) in the field of information retrieval. Consequently, the definition and interpretation of these dimensions are strongly influenced

by the aforementioned factors. For instance, a system we considered fair in our study may not align with the generic definition of fairness that is prevalent in society or our day-to-day lives. The study focused solely on the systems and did not involve any real users. However, the authors ensured that there was no bias towards specific attributes (such as gender, political stances, and cognitive impairment) associated with themselves at any stage of the decision-making process. The authors would like to acknowledge that this work utilises fairness attributes that have been defined and annotated by humans (not Large Language Models) as part of the shared tasks. Some of these attributes include socially defined characteristics such as gender, age, nationality, and so on. It is worth noting that the characteristics of the humans involved in the annotation processes, such as relevance judgments, may have influenced the results of the shared tasks. This may even warrant further studies to understand the effects of cognitive biases in relevance judgments. Furthermore, the authors recognise that there may be certain gaps or minority groups (attributes stating whether a document refers to First Nations peoples [49, 91]) whose voices may not be well represented in the test collections (nor the annotation and analysis processes) used in our work. Additionally, the authors strongly believe that the study should be replicated across other sensitive attributes (e.g., disability status) and more inclusive annotation processes before being exposed to users as a public-facing platform.

## ACKNOWLEDGMENTS

This research has been conducted in the unceded lands of the Wurundjeri and Boon Wurrung peoples of the eastern Kulin Nation. We pay our respects to Ancestors and Elders, past and present. This research is partially supported by the Australian Research Council (DE200100064, CE200100005). We sincerely thank the anonymous reviewers for their insightful comments and suggestions, which have greatly contributed to enhancing the quality of our paper. We express our gratitude to the participants and organizers of TREC 2021 and 2022 Fair Ranking Track, as well as the NTCIR-17 FairWeb-1 Task.

## REFERENCES

[1] Adnan Abid, Naveed Hussain, Kamran Abid, Farooq Ahmad, Muhammad Shoaib Farooq, Uzma Farooq, Sher Afzal Khan, Yaser Daanial Khan, Muhammad Azhar Naeem, and Nabeel Sabir. 2016. A Survey on Search Results Diversification Techniques. *Neural Computing and Applications* 27 (2016), 1207–1229. https://doi.org/10.1007/s00521-015-1945-5

[2] Rakesh Agrawal, Sreenivas Gollapudi, Alan Halverson, and Samuel Ieong. 2009. Diversifying Search Results. In *Proceedings of the Second ACM International Conference on Web Search and Data Mining* (Barcelona, Spain) *(WSDM '09)*. Association for Computing Machinery, New York, NY, USA, 5–14. https://doi.org/10.1145/1498759.1498766

[3] Shahriar Akter, Grace McCarthy, Shahriar Sajib, Katina Michael, Yogesh K. Dwivedi, John D'Ambra, and K.N. Shen. 2021. Algorithmic Bias in Data-Driven Innovation in the Age of AI. *International Journal of Information Management* 60 (2021), 102387. https://doi.org/10.1016/j.ijinfomgt.2021.102387

[4] Enrique Amigó, Damiano Spina, and Jorge Carrillo-de Albornoz. 2018. An Axiomatic Analysis of Diversity Evaluation Metrics: Introducing the Rank-Biased Utility Metric. In *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval* (Ann Arbor, MI, USA) *(SIGIR '18)*. Association for Computing Machinery, New York, NY, USA, 625–634. https://doi.org/10.1145/3209978.3210024

[5] Abraham Bernstein, Claes de Vreese, Natali Helberger, Wolfgang Schulz, Katharina Zweig, Christian Baden, Michael A. Beam, Marc P. Hauer, Lucien Heitz, Pascal Jürgens, Christian Katzenbach, Benjamin Kille, Beate Klimkiewicz, Wiebke Loosen, Judith Moeller, Goran Radanovic, Guy Shani, Nava Tintarev, Suzanne Tolmeijer, Wouter van Atteveldt, Sanne Vrijenhoek, and Theresa Zueger. 2021. Diversity in News Recommendation (Dagstuhl Perspectives Workshop 19482). *Dagstuhl Manifestos* 9, 1 (2021), 43–61. https://doi.org/10.4230/DagMan.9.1.43

[6] Abraham Bernstein, Claes De Vreese, Natali Helberger, Wolfgang Schulz, and Katharina A. Zweig. 2020. Diversity, Fairness, and Data-Driven Personalization in (News) Recommender System (Dagstuhl Perspectives Workshop 19482). *Dagstuhl Reports* 9, 11 (2020), 117–124. https://doi.org/10.4230/DagRep.9.11.117

[7] Asia J. Biega, Fernando Diaz, Michael D. Ekstrand, Sergey Feldman, and Sebastian Kohlmeier. 2021. Overview of the TREC 2020 Fair Ranking Track. In *The Twenty-Ninth Text REtrieval Conference (TREC 2020) Proceedings*. National Institute of Standards and Technology (NIST). https://trec.nist.gov/pubs/trec29/papers/OVERVIEW.FR.pdf

[8] Asia J. Biega, Fernando Diaz, Michael D. Ekstrand, and Sebastian Kohlmeier. 2020. Overview of the TREC 2019 Fair Ranking Track. In *The Twenty-Eighth Text REtrieval Conference (TREC 2019) Proceedings*. National Institute of Standards and Technology (NIST). https://trec.nist.gov/pubs/

trec28/papers/OVERVIEW.FR.pdf

[9] Asia J. Biega, Krishna P. Gummadi, and Gerhard Weikum. 2018. Equity of Attention: Amortizing Individual Fairness in Rankings. In *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval* (Ann Arbor, MI, USA) *(SIGIR '18)*. Association for Computing Machinery, New York, NY, USA, 405–414. https://doi.org/10.1145/3209978.3210063

[10] Alexander Bondarenko, Maik Fröbe, Meriem Beloucif, Lukas Gienapp, Yamen Ajjour, Alexander Panchenko, Chris Biemann, Benno Stein, Henning Wachsmuth, Martin Potthast, and Matthias Hagen. 2020. Overview of Touché 2020: Argument Retrieval: Extended Abstract. In *Experimental IR Meets Multilinguality, Multimodality, and Interaction: 11th International Conference of the CLEF Association, CLEF 2020, Thessaloniki, Greece, September 22–25, 2020, Proceedings* (Thessaloniki, Greece). Springer-Verlag, Berlin, Heidelberg, 384–395. https://doi.org/10.1007/978-3-030-58219-7_26

[11] Christopher JC Burges. 2010. From RankNet to LambdaRank to LambdaMART: An Overview. *Learning* 11, 23-581 (2010), 81.

[12] Gabriele Capannini, Franco Maria Nardini, Raffaele Perego, and Fabrizio Silvestri. 2011. Efficient Diversification of Web Search Results. *Proc. VLDB Endow.* 4, 7 (April 2011), 451–459. https://doi.org/10.14778/1988776.1988781

[13] Jaime Carbonell and Jade Goldstein. 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* (Melbourne, Australia) *(SIGIR '98)*. Association for Computing Machinery, New York, NY, USA, 335–336. https://doi.org/10.1145/290941.291025

[14] Carlos Castillo. 2019. Fairness and Transparency in Ranking. *SIGIR Forum* 52, 2 (jan 2019), 64–71. https://doi.org/10.1145/3308774.3308783

[15] Olivier Chapelle, Donald Metlzer, Ya Zhang, and Pierre Grinspan. 2009. Expected Reciprocal Rank for Graded Relevance. In *Proceedings of the 18th ACM Conference on Information and Knowledge Management* (Hong Kong, China) *(CIKM '09)*. Association for Computing Machinery, New York, NY, USA, 621–630. https://doi.org/10.1145/1645953.1646033

[16] Fumian Chen and Hui Fang. 2023. UDInfoLab at the NTCIR-17 FairWeb-1 Task. *Proceedings of NTCIR-17* (2023). https://doi.org/10.20736/0002001300

[17] Zhenpeng Chen, Jie M. Zhang, Federica Sarro, and Mark Harman. 2024. Fairness Improvement with Multiple Protected Attributes: How Far Are We?. In *Proceedings of the IEEE/ACM 46th International Conference on Software Engineering* (Lisbon, Portugal) *(ICSE '24)*. Association for Computing Machinery, New York, NY, USA, Article 160, 13 pages. https://doi.org/10.1145/3597503.3639083

[18] Seonghwan Choi, Hyeondey Kim, and Manjun Gim. 2022. Do Not Read the Same News! Enhancing Diversity and Personalization of News Recommendation. In *Companion Proceedings of the Web Conference 2022* (Virtual Event, Lyon, France) *(WWW '22)*. Association for Computing Machinery, New York, NY, USA, 1211–1215. https://doi.org/10.1145/3487553.3524936

[19] Zhumin Chu, Tetsuya Sakai, Qingyao Ai, and Yiqun Liu. 2023. Chuweb21D: A Deduped English Document Collection for Web Search Tasks. In *Proceedings of the Annual International ACM SIGIR Conference on Research and Development in Information Retrieval in the Asia Pacific Region* (Beijing, China) *(SIGIR-AP '23)*. Association for Computing Machinery, New York, NY, USA, 63–72. https://doi.org/10.1145/3624918.3625317

[20] Charles L.A. Clarke, Maheedhar Kolla, Gordon V. Cormack, Olga Vechtomova, Azin Ashkan, Stefan Büttcher, and Ian MacKinnon. 2008. Novelty and Diversity in Information Retrieval Evaluation. In *Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (Singapore, Singapore) *(SIGIR '08)*. Association for Computing Machinery, New York, NY, USA, 659–666. https://doi.org/10.1145/1390334.1390446

[21] Gordon V. Cormack, Charles L.A. Clarke, and Stefan Buettcher. 2009. Reciprocal Rank Fusion Outperforms Condorcet and Individual Rank Learning Methods. In *Proceedings of the 32nd International ACM SIGIR Conference on Research and Development in Information Retrieval* (Boston, MA, USA) *(SIGIR '09)*. Association for Computing Machinery, New York, NY, USA, 758–759. https://doi.org/10.1145/1571941.1572114

[22] J. Shane Culpepper, Fernando Diaz, and Mark D. Smucker. 2018. Research Frontiers in Information Retrieval: Report from the Third Strategic Workshop on Information Retrieval in Lorne (SWIRL 2018). *SIGIR Forum* 52, 1 (Aug. 2018), 34–90. https://doi.org/10.1145/3274784.3274788

[23] Van Dang and Bruce W. Croft. 2013. Term Level Search Result Diversification. In *Proceedings of the 36th International ACM SIGIR Conference on Research and Development in Information Retrieval* (Dublin, Ireland) *(SIGIR '13)*. Association for Computing Machinery, New York, NY, USA, 603–612. https://doi.org/10.1145/2484028.2484095

[24] Van Dang and W. Bruce Croft. 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* (Portland, Oregon, USA) *(SIGIR '12)*. Association for Computing Machinery, New York, NY, USA, 65–74. https://doi.org/10.1145/2348283.2348296

[25] John J. Donohue. 2016. *Anti-Discrimination Law.* Palgrave Macmillan UK, London, 1–10. https://doi.org/10.1057/978-1-349-95121-5_2146-1

[26] Tim Draws, Nava Tintarev, and Ujwal Gadiraju. 2021. Assessing Viewpoint Diversity in Search Results Using Ranking Fairness Metrics. *SIGKDD Explor. Newsl.* 23, 1 (may 2021), 50–58. https://doi.org/10.1145/3468507.3468515

[27] Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel. 2012. Fairness Through Awareness. In *Proceedings of the 3rd Innovations in Theoretical Computer Science Conference* (Cambridge, Massachusetts) *(ITCS '12)*. Association for Computing Machinery, New York, NY, USA, 214–226. https://doi.org/10.1145/2090236.2090255

[28] Michael D. Ekstrand, Anubrata Das, Robin Burke, and Fernando Diaz. 2022. Fairness in Information Access Systems. *Foundations and Trends® in Information Retrieval* 16, 1-2 (2022), 1–177. https://doi.org/10.1561/1500000079

[29] Michael D. Ekstrand, Graham McDonald, Amifa Raj, and Isaac Johnson. 2022. Overview of the TREC 2022 Fair Ranking Track. In *Proceedings of the Thirty-First Text REtrieval Conference, TREC 2022, online, November 15-19, 2022 (NIST Special Publication, Vol. 500-338)*, Ian Soboroff and Angela Ellis (Eds.). National Institute of Standards and Technology (NIST). https://trec.nist.gov/pubs/trec31/papers/Overview_fair.pdf

[30] Michael D. Ekstrand, Amifa Raj, Graham McDonald, and Isaac Johnson. 2021. Overview of the TREC 2021 Fair Ranking Track. In *Proceedings of the Thirtieth Text REtrieval Conference, TREC 2021, online, November 15-19, 2021 (NIST Special Publication, Vol. 500-335)*, Ian Soboroff and Angela Ellis

(Eds.). National Institute of Standards and Technology (NIST). https://trec.nist.gov/pubs/trec30/papers/Overview-F.pdf

[31] James R. Foulds, Rashidul Islam, Kamrun Naher Keya, and Shimei Pan. 2020. An Intersectional Definition of Fairness. In *2020 IEEE 36th International Conference on Data Engineering (ICDE)*. 1918–1921. https://doi.org/10.1109/ICDE48307.2020.00203

[32] Ruoyuan Gao and Chirag Shah. 2020. Toward Creating a Fairer Ranking in Search Engine Results. *Information Processing & Management* 57, 1 (2020), 102138. https://doi.org/10.1016/j.ipm.2019.102138

[33] Sahin Cem Geyik, Stuart Ambler, and Krishnaram Kenthapadi. 2019. Fairness-Aware Ranking in Search & Recommendation Systems with Application to LinkedIn Talent Search. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining* (Anchorage, AK, USA) *(KDD '19)*. Association for Computing Machinery, New York, NY, USA, 2221–2231. https://doi.org/10.1145/3292500.3330691

[34] Avijit Ghosh, Ritam Dutt, and Christo Wilson. 2021. When Fair Ranking Meets Uncertain Inference. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval* (Virtual Event, Canada) *(SIGIR '21)*. Association for Computing Machinery, New York, NY, USA, 1033–1043. https://doi.org/10.1145/3404835.3462850

[35] Moritz Hardt, Eric Price, and Nathan Srebro. 2016. Equality of Opportunity in Supervised Learning. In *Proceedings of the 30th International Conference on Neural Information Processing Systems* (Barcelona, Spain) *(NIPS'16)*. Curran Associates Inc., Red Hook, NY, USA, 3323–3331.

[36] Natali Helberger. 2019-09-14. On the Democratic Role of News Recommenders. *Digital journalism.* 7, 8 (2019-09-14). https://doi.org/10.1080/21670811.2019.1623700

[37] Jonathan Hendrickx, Annelien Smets, and Pieter Ballon. 2021. News Recommender Systems and News Diversity, Two of a Kind? A Case Study from a Small Media Market. *Journalism and Media* 2, 3 (2021), 515–528. https://doi.org/10.3390/journalmedia2030031

[38] Anna Lauren Hoffmann. 2019. Where Fairness Fails: Data, Algorithms, and the Limits of Antidiscrimination Discourse. *Information, Communication & Society* 22, 7 (2019), 900–915. https://doi.org/10.1080/1369118X.2019.1573912

[39] Sha Hu, Zhicheng Dou, Xiaojie Wang, Tetsuya Sakai, and Ji-Rong Wen. 2015. Search Result Diversification Based on Hierarchical Intents. In *Proceedings of the 24th ACM International on Conference on Information and Knowledge Management* (Melbourne, Australia) *(CIKM '15)*. Association for Computing Machinery, New York, NY, USA, 63–72. https://doi.org/10.1145/2806416.2806455

[40] Thomas Jaenich, Graham McDonald, and Iadh Ounis. 2021. University of Glasgow Terrier Team at the TREC 2021 Fair Ranking Track.. In *The Thirtieth Text REtrieval Conference (TREC 2021) Proceedings*. National Institute of Standards and Technology (NIST). https://trec.nist.gov/pubs/trec30/papers/uogTr-F.pdf

[41] Thomas Jaenich, Graham McDonald, and Iadh Ounis. 2023. University of Glasgow Terrier Team at the TREC 2022 Fair Ranking Track. (2023). https://trec.nist.gov/pubs/trec31/papers/UoGTr.F.pdf

[42] Thomas Jaenich, Graham McDonald, and Iadh Ounis. 2024. Improving Exposure Allocation in Rankings by Query Generation. In *Advances in Information Retrieval*. Springer Nature Switzerland, Cham, 121–129. https://doi.org/10.1007/978-3-031-56069-9_9

[43] Kalervo Järvelin and Jaana Kekäläinen. 2002. Cumulated Gain-Based Evaluation of IR Techniques. *ACM Trans. Inf. Syst.* 20, 4, 422–446. https://doi.org/10.1145/582415.582418

[44] Guolin Ke, Qi Meng, Thomas Finley, Taifeng Wang, Wei Chen, Weidong Ma, Qiwei Ye, and Tie-Yan Liu. 2017. LightGBM: A Highly Efficient Gradient Boosting Decision Tree. In *Advances in Neural Information Processing Systems*, I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (Eds.), Vol. 30. Curran Associates, Inc. https://proceedings.neurips.cc/paper_files/paper/2017/file/6449f44a102fde848669bdd9eb6b76fa-Paper.pdf

[45] Maurice G Kendall. 1938. A New Measure of Rank Correlation. *Biometrika* 30, 1/2 (1938), 81–93. https://www.jstor.org/stable/2332226

[46] Omar Khattab and Matei Zaharia. 2020. ColBERT: Efficient and Effective Passage Search via Contextualized Late Interaction over BERT. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval* (Virtual Event, China) *(SIGIR '20)*. Association for Computing Machinery, New York, NY, USA, 39–48. https://doi.org/10.1145/3397271.3401075

[47] Bruno Lepri, Jacopo Staiano, David Sangokoya, Emmanuel Letouzé, and Nuria Oliver. 2017. *The Tyranny of Data? The Bright and Dark Sides of Data-Driven Decision-Making for Social Good*. Springer International Publishing, Cham, 3–24. https://doi.org/10.1007/978-3-319-54024-5_1

[48] Kristina Lerman and Tad Hogg. 2014. Leveraging Position Bias to Improve Peer Recommendation. *PloS one* 9, 6 (2014), e98914. https://doi.org/10.1371/journal.pone.0098914

[49] Jason Edward Lewis, Angie Abdilla, Noelani Arista, Kaipulaumakaniolono Baker, Scott Benesiinaabandan, Michelle Brown, Melanie Cheung, Meredith Coleman, Ashley Cordes, Joel Davison, Kūpono Duncan, Sergio Garzon, D. Fox Harrell, Peter-Lucas Jones, Kekuhi Kealiikanakaoleohaililani, Megan Kelleher, Suzanne Kite, Olin Lagon, Jason Leigh, Maroussia Levesque, Keoni Mahelona, Caleb Moses, Isaac ('Ika'aka) Nahuewai, Kari Noe, Danielle Olson, 'Ōiwi Parker Jones, Caroline Running Wolf, Michael Running Wolf, Marlee Silva, Skawennati Fragnito, and Hēmi Whaanga. 2020. Indigenous Protocol and Artificial Intelligence Position Paper. (2020). https://doi.org/10.11573/SPECTRUM.LIBRARY.CONCORDIA.CA.00986506

[50] Fan Li, Kaize Shi, Nuo Chen Kenta Inaba, Sijie Tao, and Tetsuya Sakai. 2023. RSLFW at the NTCIR-17 FairWeb-1 Task. *Proceedings of NTCIR-17*. https://doi.org/10.20736/0002001303

[51] Shangsong Liang, Emine Yilmaz, Hong Shen, Maarten De Rijke, and W. Bruce Croft. 2017. Search Result Diversification in Short Text Streams. *ACM Trans. Inf. Syst.* 36, 1, Article 8 (jul 2017), 35 pages. https://doi.org/10.1145/3057282

[52] Jianhua Lin. 1991. Divergence Measures Based on the Shannon Entropy. *IEEE Transactions on Information Theory* 37, 1, 145–151. https://doi.org/10.1109/18.61115

[53] Jimmy Lin, Xueguang Ma, Sheng-Chieh Lin, Jheng-Hong Yang, Ronak Pradeep, and Rodrigo Nogueira. 2021. Pyserini: A Python Toolkit for Reproducible Information Retrieval Research with Sparse and Dense Representations. In *Proceedings of the 44th International ACM SIGIR Conference*

*on Research and Development in Information Retrieval* (Virtual Event, Canada) *(SIGIR '21)*. Association for Computing Machinery, New York, NY, USA, 2356–2362. https://doi.org/10.1145/3404835.3463238

[54] Graham McDonald, Craig Macdonald, and Iadh Ounis. 2022. Search Results Diversification for Effective Fair Ranking in Academic Search. *Information Retrieval Journal* 25, 1 (2022), 1–26. https://doi.org/10.1007/s10791-021-09399-z

[55] Iadh Ounis, Gianni Amati, Vassilis Plachouras, Ben He, Craig Macdonald, and Douglas Johnson. 2005. Terrier Information Retrieval Platform. In *Advances in Information Retrieval*. Springer Berlin Heidelberg, Berlin, Heidelberg, 517–519. https://doi.org/10.1007/978-3-540-31865-1_37

[56] Sachin Pathiyan Cherumanal. 2022. Fairness-Aware Question Answering for Intelligent Assistants. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval* (Madrid, Spain) *(SIGIR '22)*. Association for Computing Machinery, New York, NY, USA, 3492. https://doi.org/10.1145/3477495.3531682

[57] Sachin Pathiyan Cherumanal, Marwah Alaofi, Reham Abdullah Altalhi, Elham Naghizade, Falk Scholer, and Damiano Spina. 2022. RMIT CIDDA IR at the TREC 2022 Fair Ranking Track. In *The Thirty-First Text REtrieval Conference (TREC 2022) Proceedings*. National Institute of Standards and Technology (NIST). https://trec.nist.gov/pubs/trec31/papers/rmit_cidda_ir.F.pdf

[58] Sachin Pathiyan Cherumanal, Kaixin Ji, Danula Hettiachchi, Johanne R Trippas, Falk Scholer, and Damiano Spina. 2023. RMIT_IR at the NTCIR-17 FairWeb-1 Task. *Proceedings of NTCIR-17*. https://repository.nii.ac.jp/record/2001315/files/04-NTCIR17-FAIRWEB-CherumanalS.pdf

[59] Sachin Pathiyan Cherumanal, Damiano Spina, Falk Scholer, and W. Bruce Croft. 2021. Evaluating Fairness in Argument Retrieval. In *Proceedings of the 30th ACM International Conference on Information & Knowledge Management* (Virtual Event, Queensland, Australia) *(CIKM '21)*. Association for Computing Machinery, New York, NY, USA, 3363–3367. https://doi.org/10.1145/3459637.3482099

[60] Sachin Pathiyan Cherumanal, Damiano Spina, Falk Scholer, and W. Bruce Croft. 2022. RMIT at TREC 2021 Fair Ranking Track. In *The Thirtieth Text REtrieval Conference (TREC 2021) Proceedings*. National Institute of Standards and Technology (NIST). https://trec.nist.gov/pubs/trec30/papers/RMIT-IR-F.pdf

[61] Sachin Pathiyan Cherumanal, Lin Tian, Futoon M. Abushaqra, Angel Felipe Magnossão de Paula, Kaixin Ji, Halil Ali, Danula Hettiachchi, Johanne R. Trippas, Falk Scholer, and Damiano Spina. 2024. Walert: Putting Conversational Information Seeking Knowledge into Action by Building and Evaluating a Large Language Model-Powered Chatbot. In *Proceedings of the 2024 Conference on Human Information Interaction and Retrieval* (Sheffield, United Kingdom) *(CHIIR '24)*. Association for Computing Machinery, New York, NY, USA, 401–405. https://doi.org/10.1145/3627508.3638309

[62] Amifa Raj and Michael D. Ekstrand. 2022. Measuring Fairness in Ranked Results: An Analytical and Empirical Comparison. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval* (Madrid, Spain) *(SIGIR '22)*. Association for Computing Machinery, New York, NY, USA, 726–736. https://doi.org/10.1145/3477495.3532018

[63] Theresia Veronika Rampisela, Tuukka Ruotsalo, Maria Maistro, and Christina Lioma. 2024. Can We Trust Recommender System Fairness Evaluation? The Role of Fairness and Relevance. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval* (Washington DC, USA) *(SIGIR '24)*. Association for Computing Machinery, New York, NY, USA, 271–281. https://doi.org/10.1145/3626772.3657832

[64] Shaina Raza, Syed Raza Bashir, and Usman Naseem. 2022. Accuracy Meets Diversity in a News Recommender System. In *Proceedings of the 29th International Conference on Computational Linguistics*. International Committee on Computational Linguistics, Gyeongju, Republic of Korea, 3778–3787. https://aclanthology.org/2022.coling-1.332/

[65] Stephen Robertson and Hugo Zaragoza. 2009. *The Probabilistic Relevance Framework: BM25 and Beyond*. Vol. 3. Now Publishers, Inc. 333–389 pages. https://doi.org/10.1561/1500000019

[66] Stephen E Robertson, Steve Walker, Susan Jones, Micheline M Hancock-Beaulieu, Mike Gatford, et al. 1995. Okapi at TREC-3. *Nist Special Publication Sp* 109 (1995), 109.

[67] Arjun Roy, Vasileios Iosifidis, and Eirini Ntoutsi. 2022. Multi-fairness Under Class-Imbalance. In *Discovery Science*, Poncelet Pascal and Dino Ienco (Eds.). Springer Nature Switzerland, Cham, 286–301. https://doi.org/10.1007/978-3-031-18840-4_21

[68] Tetsuya Sakai. 2018. *Laboratory Experiments in Information Retrieval*. Laboratory Experiments in Information Retrieval, Vol. 40. Springer Singapore, Singapore. https://doi.org/10.1007/978-981-13-1199-4

[69] Tetsuya Sakai. 2021. A Closer Look at Evaluation Measures for Ordinal Quantification. *CEUR Workshop Proceedings* 3052. https://cir.nii.ac.jp/crid/1010857593591514249

[70] Tetsuya Sakai. 2021. Evaluating Evaluation Measures for Ordinal Classification and Ordinal Quantification. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, Chengqing Zong, Fei Xia, Wenjie Li, and Roberto Navigli (Eds.). Association for Computational Linguistics, Online, 2759–2769. https://doi.org/10.18653/v1/2021.acl-long.214

[71] Tetsuya Sakai, Jin Young Kim, and Inho Kang. 2023. A Versatile Framework for Evaluating Ranked Lists in Terms of Group Fairness and Relevance. *ACM Trans. Inf. Syst.* 42, 1, Article 11 (aug 2023), 36 pages. https://doi.org/10.1145/3589763

[72] Tetsuya Sakai and Zhaohao Zeng. 2019. Which Diversity Evaluation Measures Are "Good"?. In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval* (Paris, France) *(SIGIR'19)*. Association for Computing Machinery, New York, NY, USA, 595–604. https://doi.org/10.1145/3331184.3331215

[73] Gerard Salton and Christopher Buckley. 1988. Term-Weighting Approaches in Automatic Text Retrieval. *Information Processing & Management* 24, 5 (1988), 513–523. https://doi.org/10.1016/0306-4573(88)90021-0

[74] Rodrygo L.T. Santos, Craig Macdonald, and Iadh Ounis. 2010. Exploiting Query Reformulations for Web Search Result Diversification. In *Proceedings of the 19th International Conference on World Wide Web* (Raleigh, North Carolina, USA) *(WWW '10)*. Association for Computing Machinery, New

York, NY, USA, 881–890. https://doi.org/10.1145/1772690.1772780

[75] Rodrygo L. T. Santos, Craig Macdonald, and Iadh Ounis. 2015. Search Result Diversification. *Foundations and Trends® in Information Retrieval* 9, 1 (2015), 1–90. https://doi.org/10.1561/1500000040

[76] Piotr Sapiezynski, Wesley Zeng, Ronald E Robertson, Alan Mislove, and Christo Wilson. 2019. Quantifying the Impact of User Attentionon Fair Group Representation in Ranked Lists. In *Companion Proceedings of The 2019 World Wide Web Conference* (San Francisco, USA) *(WWW '19)*. Association for Computing Machinery, New York, NY, USA, 553–562. https://doi.org/10.1145/3308560.3317595

[77] Ashudeep Singh and Thorsten Joachims. 2018. Fairness of Exposure in Rankings. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining* (London, United Kingdom) *(KDD '18)*. Association for Computing Machinery, New York, NY, USA, 2219–2228. https://doi.org/10.1145/3219819.3220088

[78] Student. 1908. The Probable Error of a Mean. *Biometrika* 6, 1 (1908), 1–25. http://www.jstor.org/stable/2331554

[79] Zhan Su, Zhicheng Dou, Yutao Zhu, and Ji-Rong Wen. 2024. Passage-aware Search Result Diversification. *ACM Trans. Inf. Syst.* 42, 5, Article 136, 29 pages. https://doi.org/10.1145/3653672

[80] Wenlong Sun, Sami Khenissi, Olfa Nasraoui, and Patrick Shafto. 2019. Debiasing the Human-Recommender System Feedback Loop in Collaborative Filtering. In *Companion Proceedings of The 2019 World Wide Web Conference* (San Francisco, USA) *(WWW '19)*. Association for Computing Machinery, New York, NY, USA, 645–651. https://doi.org/10.1145/3308560.3317303

[81] Sijie Tao, Nuo Chen, Tetsuya Sakai, Zhumin Chu, Hiromi Arai, Ian Soboroff, Nicola Ferro, and Maria Maistro. 2023. Overview of the NTCIR-17 FairWeb-1 Task. In *Proceedings of the 17th NTCIR Conference on Evaluation of Information Access Technologies*. https://doi.org/10.20736/0002001318

[82] Yiteng Tu, Haitao Li, Zhumin Chu, Qingyao Ai, and Yiqun Liu. 2023. THUIR at the NTCIR-17 FairWeb-1 Task: An Initial Exploration of the Relationship Between Relevance and Fairness. *Proceedings of NTCIR-17* (2023). https://doi.org/10.20736/0002001317

[83] Jan Van Cuilenburg. 1999. On Competition, Access and Diversity in Media, Old and New: Some Remarks for Communications Policy in the Information Age. *New media & society* 1, 2 (1999), 183–207. https://doi.org/10.1177/14614449922225555

[84] C. J. Van Rijsbergen. 1979. *Information retrieval* (2d ed.). Butterworths, London.

[85] Lakshmi Vikraman, W. Bruce Croft, and Brendan O'Connor. 2018. Exploring Diversification In Non-Factoid Question Answering. In *Proceedings of the 2018 ACM SIGIR International Conference on Theory of Information Retrieval* (Tianjin, China) *(ICTIR '18)*. Association for Computing Machinery, New York, NY, USA, 223–226. https://doi.org/10.1145/3234944.3234973

[86] Sanne Vrijenhoek, Mesut Kaya, Nadia Metoui, Judith Möller, Daan Odijk, and Natali Helberger. 2021. Recommenders with a Mission: Assessing Diversity in News Recommendations. In *Proceedings of the 2021 Conference on Human Information Interaction and Retrieval* (Canberra ACT, Australia) *(CHIIR '21)*. Association for Computing Machinery, New York, NY, USA, 173–183. https://doi.org/10.1145/3406522.3446019

[87] Xiao Wang, Craig MacDonald, Nicola Tonellotto, and Iadh Ounis. 2023. ColBERT-PRF: Semantic Pseudo-Relevance Feedback for Dense Passage and Document Retrieval. *ACM Trans. Web* 17, 1, Article 3 (jan 2023), 39 pages. https://doi.org/10.1145/3572405

[88] Shengli Wu, Chunlan Huang, Liang Li, and Fabio Crestani. 2019. Fusion-Based Methods for Result Diversification in Web Search. *Information Fusion* 45 (2019), 16–26. https://doi.org/10.1016/j.inffus.2018.01.006

[89] Yao Wu, Jian Cao, and Guandong Xu. 2023. Fairness in Recommender Systems: Evaluation Approaches and Assurance Strategies. *ACM Trans. Knowl. Discov. Data* 18, 1, Article 10 (Aug. 2023), 37 pages. https://doi.org/10.1145/3604558

[90] Ke Yang and Julia Stoyanovich. 2017. Measuring Fairness in Ranked Outputs. In *Proceedings of the 29th International Conference on Scientific and Statistical Database Management* (Chicago, IL, USA) *(SSDBM '17)*. Association for Computing Machinery, New York, NY, USA, Article 22, 6 pages. https://doi.org/10.1145/3085504.3085526

[91] Tyson Yunkaporta. 2023. *Right Story, Wrong Story: Adventures in Indigenous Thinking*. Text Publishing Company.

[92] Yuying Zhao, Yu Wang, Yunchao Liu, Xueqi Cheng, Charu C. Aggarwal, and Tyler Derr. 2024. Fairness and Diversity in Recommender Systems: A Survey. *ACM Trans. Intell. Syst. Technol.* (May 2024). https://doi.org/10.1145/3664928