A Versatile Framework for Evaluating Ranked Lists in Terms of Group Fairness and Relevance

TETSUYA SAKAI, Waseda University/Naver Corporation, Japan
JIN YOUNG KIM and INHO KANG, Naver Corporation, Korea

We present a simple and versatile framework for evaluating ranked lists in terms of Group Fairness and Relevance, in which the groups (i.e., possible attribute values) can be either nominal or ordinal in nature. First, we demonstrate that when our framework is applied to a binary hard group membership setting, our Group Fairness and Relevance (GFR) measures can easily quantify the overall polarity of each ranked list. Second, by utilising an existing diversified search test collection and treating each intent as an attribute value, we demonstrate that our framework can also handle soft group membership and that the GFR measures are highly correlated with a diversified information retrieval (IR) measure in this context as well. Third, using real data from a Japanese local search service, we demonstrate how our framework enables researchers to study intersectional group fairness based on multiple attribute sets. We also show that the similarity function for comparing the achieved and target distributions over the attribute values should be chosen carefully when the attribute values are ordinal. For such situations, our recommendation is to use multiple similarity functions with our framework: for example, one based on Jensen-Shannon Divergence (which disregards the ordinal nature of the groups) and another based on Root Normalised Order-aware Divergence (which has been designed specifically for handling ordinal groups). In addition, we highlight the fundamental differences between our framework and Attention-Weighted Rank Fairness (AWRF), a group fairness measure used at the TREC Fair Ranking Track.

CCS Concepts: • Information systems → Test collections; Relevance assessment; Retrieval effectiveness;

Additional Key Words and Phrases: Evaluation, evaluation measures, fairness, group fairness, ordinal quantification

ACM Reference format:
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 (August 2023), 36 pages. https://doi.org/10.1145/3589763

1 INTRODUCTION

Bias can breed bias. If a ranked list of items presented to the user is “unfair” or “biased” from the viewpoint of ranked entities (e.g., people, opinions, products, shops) or their stakeholders, the
biased views may influence the users. Moreover, based on user feedback (e.g., clicks, views) on such ranked lists, the underlying search engine may tune itself and further intensify the bias. Hence, for example, those that are already enjoying much exposure or attention [9, 48] may further dominate the ranking, giving little or no room to those that have never been exposed to the users. It is our view that providers of search and ranking services should strive to prevent and eliminate such vicious circles. This is our motivation for addressing the problems of fairness in rankings [13].

We address the problem of evaluating ranked lists based on group fairness [23] given a target distribution over attribute values (or groups) in each attribute set. Consider a hypothetical situation in which we want a group-fair ranking of people, say, scholarship applicants. Suppose that the applicants are classified into four classes that represent their income levels (i.e., our attribute values), and that, ideally, we want 75% of the ranking to represent the lowest income group and the other 25% to represent the second-lowest income group. As the user scans the ranked list of applicants, the list yields a series of achieved distributions over the four classes based on the group membership of the item (i.e., income group of the applicant) at each rank. To measure the similarity between each achieved distribution and the gold distribution, we consider utilising ordinal quantification measures, NMD (Normalised Match Distance, a normalised version of Earth Mover’s Distance [68]) and RNOD (Root Normalised Order-aware Divergence [57, 58]), in addition to a nominal quantification measure, JSD (Jensen-Shannon Divergence) [41]. Ordinal quantification measures take the ordinal nature of the attribute values into account and may be appropriate for some applications. For example, consider the situation shown in Figure 1. While nominal quantification measures such as JSD say that Systems A and B are equally effective, NMD and RNOD say that B is better, as B is leaning more towards the lower-income groups (25% is given to Group 3 rather than Group 4).2

Our main contributions are as follows. 3

1) We present a simple and versatile framework for evaluating ranked lists in terms of group fairness and relevance in which the groups (i.e., possible attribute values) can be either nominal or ordinal in nature. We clarify the novel features of this framework compared with prior work.

2) We demonstrate that when our framework is applied to a binary hard group membership setting (e.g., each ranked item is either positive or negative), our Group Fairness and Relevance (GFR) measures can easily quantify the overall polarity of each ranked list.

---

1While item exposure does not necessarily imply user attention, we shall not differentiate between the two hereafter.

2To make this article self-contained, these quantification measures are formally defined in Section 3.2. Applying Equations (25) (JSD), (26) (NMD) and (29) (RNOD) from Section 3.2 to the distributions in Figure 1 with |C| = 3 attribute values is trivial.

3An early version of the present study was made publicly available on arxiv.org on April 1, 2022 [60].
Evaluating Ranked Lists in terms of Group Fairness and Relevance

Fig. 2. The GFR (Group Fairness and Relevance) evaluation framework. In the figure, it is assumed that the Decay and Utility components are based on ERR (Expected Reciprocal Rank) [15, 54]. DistrSim stands for Distribution Similarity. Here, the highest relevance grade is assumed to be 3 (Perfect), and its gain value is set to \((2^3 - 1)/2^3 = 7/8\).

(3) By utilizing an existing diversified search test collection and treating each intent as an attribute value, we demonstrate that our framework can also handle soft group membership (i.e., each ranked item can have multiple attribute values) and that the GFR measures are highly correlated with a diversified IR measure in this context as well.

(4) Using real data from a Japanese local search service, we demonstrate how our framework enables researchers to study intersectional group fairness [25] based on multiple attribute sets.

We also show that the similarity function for comparing the achieved and target distributions over the attribute values should be chosen carefully when the attribute values are ordinal.

Figure 2 illustrates our evaluation framework. Briefly, our GFR framework is composed of a decay function that represents the distribution of a user population over a Search Engine Result Page (SERP), a utility function that reflects the overall relevance of a part of the SERP for each user group, and a distribution similarity (DistrSim) function that, for each attribute set, compares the distribution achieved by the part of the SERP for each user group with the gold distribution. As was discussed earlier, we employ JSD, NMD, or RNOD as an instance of DistrSim. More details will be provided in Section 3.

Our experiments can be reproduced by using this dataset available at https://waseda.box.com/GFR20221202targz.

2 PRIOR WORK

Sections 2.1 to 2.6 discuss prior art that are highly relevant to our own group fairness evaluation framework. Here, we briefly mention prior work that these subsections do not cover. The group-fair ranking measures proposed in Kuhlman et al. [37] assume the existence of a gold fair ranking, and are computed based on concordant and discordant pairs of ranked items by comparing the gold and system rankings. Their work focussed on binary attribute sets (i.e., protected vs. non-protected groups). See also Kuhlman et al. [36] for a comparative study of measures under the binary setting. Raj and Ekstrand [48] compared single-ranking measures of Yang and Stoyanovich [69] and of Sapiezynski et al. [64] (which we discuss in Sections 2.1 and 2.2, respectively) as well as measures designed for distributions and sequences of rankings [7–9, 21, 65]. The latter class of measures is beyond the scope of our work, as we are interested in evaluating a single ranked list.
when the target distribution over attribute values for each attribute set is given, either top-down (e.g., requiring a uniform distribution over all attribute values) or as a result of some bottom-up derivation (e.g., requiring statistical parity [23] based on statistics from the target corpus). Also beyond our scope are the following lines of research. Beutel et al. [6] propose to evaluate group fairness in the context of personalised recommendation by collecting pairwise item preferences from each user (see also Narasimhan et al. [46]); Kirnap et al. [35] estimate fair ranking measure scores from incomplete group membership labels.

It should be noted that the detailed discussion of existing fairness evaluation measures in Sections 2.1 to 2.5 is provided to clarify how our own framework adopted some of their best practices and extended them in the context of our particular group fairness objective. More specifically, unlike previous work, our framework accommodates distributions over ordinal attribute values as well as intersectional group fairness [25] involving ordinal attribute values. The primary purpose of the present study is to demonstrate the intuitiveness and versatility of our evaluation framework rather than to claim that ours is “better” or “more reliable” than the existing ones. We use consistent notations throughout this article to clarify how our framework builds on previous work.

2.1 Normalised Discounted KL Divergence

Geyik et al. [26] considered two approaches to evaluating a ranked list of items (e.g., people), in which the ranked list is expected to reflect as faithfully as possible a given target distribution over possible attribute values (e.g., female, male, other) in an attribute set (e.g., Gender), or over combinations of attribute values from multiple attribute sets (e.g., Gender × Age Groups, where “x” denotes a Cartesian product of two sets). Let \( A = \{ a_i \} \) denote an attribute set. Their first proposal is a set retrieval measure for the top-\( k \) search results, and is computed for a particular attribute value. For a given query, let \( p_r(a_i) \) denote the desired proportion of items having attribute value \( a_i \) in the ranked list, such that \( \sum_i p_r(a_i) = 1 \). That is, \( p_r \) is the gold probability mass function over \( A \). Let \( L \) denote a ranked list to be evaluated. Let \( p_{L,k}(a_i) \) be the actual proportion of items having attribute value \( a_i \) within the top-\( k \) results of \( L \), such that \( \sum_i p_{L,k}(a_i) = 1 \). That is, \( p_{L,k} \) is the achieved probability mass function over \( A \). The Skew at rank \( k \) of SERP \( L \) for \( a_i \) is defined as

\[
Sk(L, k, a_i) = \log_e \frac{p_{L,k}(a_i)}{p_r(a_i)}.
\]

Although Geyik et al. [26] point out that situations in which \( p_r(a_i) = 0 \) should be avoided, note that \( p_r(a_i) = 0 \) may well happen in practice, especially if combinations of multiple attribute sets are considered (e.g., Gender = “male” AND Age = “x > 90” AND ...). They also propose to utilise \( \max_i Sk(L, k, a_i) \) and \( \min_i Sk(L, k, a_i) \) to discuss the quality of the top-\( k \) results with respect to the attribute set \( A \). However, it is clear that the skew-based measures focus only on the worst-case and best-case attribute values. In summary, skew-based measures are not adequate for our purpose because (a) they cannot handle ranked retrieval, (b) they do not consider every attribute value in \( A \) (when \( |A| > 2 \)), and (c) \( p_r(a_i) = 0 \) can cause inconveniences.

The second proposal by Geyik et al. [26] was to slightly modify a ranked retrieval measure of Yang and Stoyanovich [69], called rKL. The modified measure, which Geyik et al. [26] refer to as Normalised Discounted KL divergence (NDKL), utilises the Kullback-Leibler Divergence (KLD) to compare the achieved and gold distributions.

\[
NDKL(L) = \frac{\sum_{k=1}^{\left| L \right|} \left( \sum_i p_{L,k}(a_i) Sk(L, k, a_i) \right) / \log_2(k + 1)}{\sum_{k=1}^{\left| L \right|} 1/\log_2(k + 1)}.
\]

Similarly, Pitoura et al. [47] proposed to utilise \( \max_i |p_{L,k}(a_i) - p_r(a_i)| \).
Evaluating Ranked Lists in terms of Group Fairness and Relevance

Note that NDKL overcomes Limitations (a) and (b) mentioned earlier, but not (c) since it is based on Skew. Yet another inconvenience with KLD is that (d) it is unbounded. Draws et al. [22] and the TREC 2021 Fair Ranking Track\(^5\) have also adopted JSD instead: this overcomes Limitations (c) and (d). However, as we have mentioned in Section 1, JSD can only treat the attribute values as nominal.

Geyik et al. [26] (and Ghosh et al. [28, 29]) argue that intersectional group fairness [25] can be handled by considering combined attribute values from multiple attribute sets, for example, SkinType × Gender [12]. However, we argue that this may not be the best approach to take if both nominal and ordinal attribute values need to be considered and if it seems appropriate to take the ordinal nature of the attribute values into account. For example, consider a Cartesian product of Gender (nominal) and Age Groups (ordinal), and a combined attribute value Gender = "female" AND Age = "x < 20." Which of the following two combined attribute values is closer to the above, Gender = "male" AND Age = "20 ≤ x < 40" (Gender: incorrect; Age Groups: not far off) or Gender = "female" AND Age = "40 ≤ x < 60" (Gender: correct; Age Groups: far off)? Similar problems arise when multiple ordinal groups are combined. Hence, for applications for which ordinal quantification measures seem more appropriate than nominal ones such as JSD, we propose to compare the achieved and gold distributions for each attribute set at a time and finally aggregate the scores across the multiple attribute sets.

NDKL adopted the log-based discounting scheme of nDCG (normalised Discounted Cumulative Gain) [32, 54] because "it is more beneficial for an item to be ranked higher, it is also more important to achieve statistical parity at higher ranks." The discounting scheme can be interpreted as reflecting user attention over a ranked list. However, Ghosh et al. [28] and Sapiezynski et al. [64] argue that the log-based decay is too fat-tailed for modelling user attention. The next section reviews their work.

2.2 Expected Cumulative Exposure

Inspired by the work of Sapiezynski et al. [64], which considered user attention over search results, Ghosh et al. [28] presented a group fairness measure called Attention Bias Ratio (ABR). Let \(F(L, k, a_i) = 1\) if the item at rank \(k\) in ranked list \(L\) has attribute value \(a_i\), and let \(F(L, k, a_i) = 0\) otherwise. The Mean Attention (MA) score of \(a_i\) for \(L\) is defined as

\[
MA(L, a_i) = \frac{\sum_{k=1}^{\|L\|} F(L, k, a_i) \text{Attention}_p(k)}{\sum_{k=1}^{\|L\|} F(L, k, a_i)},
\]

where \(\text{Attention}_p(k) = 100p(1 - p)^{k-1}\), which is essentially the decay function of Rank-Biased Precision (RBP) for a given patience parameter value \(p = 1 - p\) [45]. Note that this decay depends entirely on the document rank; it considers neither relevance nor fairness of the documents seen so far. This limitation applies to NDKL (Equation (2)) as well. Hence, if relevance assessments are available, we use the cascade-based decay of Expected Reciprocal Rank (ERR) [14] as in Biega et al. [7, 8] and Diaz et al. [21].

Ghosh et al. [28] define ABR as

\[
ABR(L) = \frac{\min_{a_i \in A} MA(L, a_i)}{\max_{a_i \in A} MA(L, a_i)}.
\]

Thus, ABR quantifies the disparity between the attribute values with lowest and highest mean attention scores. It is clear that Limitation (b) mentioned in Section 2.1 applies to ABR as well.

\(^5\)https://fair-trec.github.io/docs/Fair_Ranking_2021_Participant_Instructions.pdf.
In Equation (3), note that $F(L, k, a_i)$ is a group membership flag, representing hard group membership. However, in the original work of Sapiezynski et al. [64], group membership is formulated as a probability mass function over the attribute values (i.e., soft group membership). That is, let $G(L, k, a_i)$ be the probability that the item at rank $k$ in $L$ has attribute value $a_i$, such that $\sum G(L, k, a_i) = 1$. If $G(L, k, a_i)$ is available for each $k$, then it can replace $F(L, k, a_i)$ in Equation (3). For ranked list $L$, Sapiezynski et al. [64] compute a probability distribution over $A$ based on Expected Cumulative Exposure (ECE).

$$ECE(L, a_i) = \sum_{k=1}^{|L|} G(L, k, a_i) \, \text{Attention}(k),$$  \hspace{1cm} (5)

$$p_{L}^{ECE}(a_i) = \frac{ECE(L, a_i)}{\sum_{a_i \in A} ECE(L, a_i)}. \hspace{1cm} (6)$$

Note that Equation (5) generalises the numerator of Equation (3), and that we have omitted the suffix $p$ in Equation (5). The latter omission is because, as we shall discuss in Section 2.3, a version of ECE adopted at the TREC 2021 Fair Ranking Track replaces the RBP-based attention with an nDCG-based one.

Sapiezynski et al. [64] propose to compare $p_{L}^{ECE}$ with the gold probability mass function $p_*$ by assuming that both $E_L$ and $p_*$ are binomial distributions, and conduct a form of statistical significance test with a test statistic threshold to discuss whether a ranked list is fair or not. In contrast, we are more interested in quantifying the degree of group fairness of a ranked list rather than binary classification and do not rely on any distributional assumptions. Like the work of Sapiezynski et al. [64], our framework also handles soft group membership, which is important not only for situations in which each ranked item can take multiple attribute values but also for situations in which the group membership of each item needs to be estimated with some degree of uncertainty. We shall report on an experiment involving soft group membership in Section 4.2 by leveraging an existing diversified search test collection.

### 2.3 A Version of Attention-Weighted Rank Fairness Measure

Since TREC 2021, the TREC Fair Ranking Track has been using a measure based on a version of ECE (see Section 2.2); the measure is called Attention-Weighted Rank Fairness (AWRF). In our experiments for demonstrating the versatility of our GRF framework, we include a version of AWRF (and a measure that combines nDCG with it) for reference.

The general form of AWRF can simply be expressed as follows,

$$AWRF(L) = 1 - \text{JSD}(p_{L}^{ECE}, p_*),$$  \hspace{1cm} (7)

where $\text{JSD}$ is the Jensen-Shannon Divergence [41] for comparing the two distributions. Recall that JSD solves Limitations (c) and (d) of KLD (see Section 2.1). A mathematical definition of JSD will be provided in Section 3.2 as our own framework also utilises JSD as an option. Note that JSD is a divergence; therefore, AWRF as defined earlier is a distribution similarity measure. For handling intersectional group fairness, the TREC Fair Ranking Track follows the attribute value combination approach suggested in prior work: that is, the gold distribution $p_*$ is defined over a Cartesian product of multiple attribute sets (e.g., Geographic Location $\times$ Gender). This point will be discussed in more detail in Section 4.3.1 when we discuss an actual implementation of this approach.

---

6Ghosh et al. [29] acknowledge that their method “does not take into account partial group membership.”
Our implementation of AWRF uses a variant of Equation (5), given by

\[ ECE(L, a_i) = \frac{\sum_{k=1}^{L} I(k) G(L, k, a_i)}{\sum_{k=1}^{L} G(L, k, a_i) \cdot Attention(k)}, \]

where \( I(k) = 1 \) if the item at rank \( k \) is known to have a positive group membership for at least one attribute value, and \( I(k) = 0 \) otherwise. The TREC Fair Ranking Track ignores “unknown items” in a similar manner \[24\]. For AWRF, the introduction of this flag is actually very important, which becomes clear if we consider the following example. Suppose that the gold distribution is uniform, and that every unknown item (i.e., one that is not known to match to any of the attribute values) is considered nonrelevant. Furthermore, consider a SERP that contains unknown items only. If we assume that such a SERP is fair (since it does not favour any particular attribute value) and, therefore, that its achieved distribution is also uniform, then an AWRF score of 1 would be given to this completely nonrelevant SERP, which probably is not what we want. Hence, we use \( I(k) \) to filter out unknown items, as shown in Equation (8).

In practice, introducing \( I(k) \) alone is not sufficient because for any completely nonrelevant SERP, the denominator of Equation (6) will be zero and, therefore, AWRF will not be computable. Hence, our precise implementation of AWRF is as follows.

\[ AWRF(L) = \begin{cases} 0 & \text{if } \sum_{k=1}^{L} I(k) = 0 \text{ (i.e., no positive group membership found in } L) \\ 1 - \frac{\text{JS}D(p^ECE_L, p^\ast)}{} & \text{otherwise.} \end{cases} \]

As for relevance-based evaluation, the TREC Fair Ranking Track uses a version of nDCG, which may be expressed as follows:

\[ nDCG(L) = \frac{\sum_{k=1}^{L} g^\ast(L, k) \cdot Attention(k)}{\sum_{k=1}^{L} g(L, k) \cdot Attention(k)}. \]

Here, \( g^\ast(L, k) \) is the gain value of the document at rank \( k \) in SERP \( L \), whereas \( g(L, k) \) is the corresponding gain value in the ideal SERP \[32, 54\]. Finally, the TREC track combines AWRF and nDCG as follows.

\[ AWRF^*nDCG(L) = AWRF(L) \ast nDCG(L). \]

For both Equation (6) (for group fairness evaluation) and Equation (10) (for relevance evaluation), the TREC Fair Ranking Track uses the following version of nDCG-based attention (i.e., decay).

\[ Attention_{TREC}(k) = \frac{1}{\log_2 \max(k, 2)}. \]

While the above is faithful to the original definition of nDCG (with a logarithm base of 2 \[32\]), it is known to have a minor problem \[54\]: it treats documents ranked at 1 and 2 equally by giving both of them “full attention.” That is, no discounting is applied for these two ranks. Hence, in our experiments, we consider AWRF and AWRF^*nDCG by using the following widely used nDCG decay scheme for both Equations (6) and (10).\footnote{Note that the logarithm base becomes irrelevant when the nDCG-based decay is plugged into Equation (6) (for computing the ECE-based probability for AWRF) and Equation (10) (for computing nDCG).}

\[ Attention(k) = \frac{1}{\log_4 (k + 1)}. \]
2.4 Summary of Previous Discussions on Existing Group Fairness IR Measures

Table 1 is a summary of the discussion of existing group fairness IR measures that emphasises how our GFR framework differs from them. It should be noted that the seven viewpoints provided here were chosen to highlight GFR’s novel features and are not meant to provide an exhaustive list of features that distinguish group fairness IR measures. For the latter perspective, the reader is referred to Raj and Ekstrand [48], in which different viewpoints — such as "Does the metric incorporate relevance?" and "How does it respond to edge cases?" — are discussed. Hence, the fact that GFR has a “YES” in every column in Table 1 does not necessarily imply that GFR is superior to the others. Here, we provide a brief remark for each column.

(i) The Skew-based measures cannot handle ranked retrieval.
(ii) Both Skew and ABR focus on the best-case and worst-case attribute values only, even when the attribute set size is larger than two.
(iii) With Skew and NDKL (rKL), zero gold probabilities need to be avoided; for ABR, there is no notion of the gold distribution (hence, “N/A”).
(iv) Even if zero gold probabilities are somehow avoided, Skew and NDKL (rKL) are unbounded.
(v) Geyik et al. [26] (Skew, NDKL) and Ghosh et al. [28] (ABR) argue that attribute values from multiple attribute sets can be combined to handle intersectional group fairness. Ekstrand et al. [23] (AWRF) implemented this Cartesian product approach for Geographic Location and Gender attributes. (Sapiezynski et al. [64] (ECE) do not discuss intersectional group fairness.)
(vi) No prior works distinguish between nominal and ordinal attribute sets.
(vii) When quantifying the group fairness of a SERP, none of the prior works listed here models a population of search engine users who scan the SERP and abandon it at different ranks. Moreover, none considers the possibility that how the users scan the SERP may depend on the relevance of ranked items. The Skew-based measures are for set retrieval; the other measures use either an nDCG-like or RBP-like decay (i.e., attention weight) that depends solely on the document rank.\(^8\)

In Section 3.3, we shall revisit Points (v) to (vii) when we discuss a few fundamental differences between GFR and AWRF in detail.

2.5 Polarity on a Binary Attribute Set

Consider a situation with a single binary attribute set, \(A = \{a_1, a_2\}\) [38, 49]. Suppose that, for each item at rank \(k\) in ranked list \(L\), its bias score \(b(L, k)\) is available [49], whose range is \([-1, 1]\); a negative score means leaning towards \(a_1\), a positive score means leaning towards \(a_2\), and 0 means neutral. Kulshrestha et al. [38] proposed the Output Bias (OB) measure, which is an average-precision-like measure based on bias scores. Whereas OB is applicable to binary attribute sets only, our framework can handle any number of attribute values. If bias scores are available [49] in a binary setting, our framework can leverage them by converting them to group membership probabilities as follows.

\[
G(L, k, a_1) = \frac{1 + b(L, k)}{2}, \quad G(L, k, a_2) = 1 - G(L, k, a_1). \tag{14}
\]

Experimental validation of this method is left for future work.

Gezici et al. [27] also proposed methods to quantify bias in SERPs in a binary setting, for which the objective is to achieve equality of outcome, that is, \(p_*(a_1) = p_*(a_2) = 1/2\). They point out that

\(^{8}\)Diaz et al. [21] discuss an ERR-based model for evaluating stochastic rankings; see also Biega et al. [7, 8].
Table 1. Contrasting GFR with Some Existing Group Fairness IR Measures

| Measure                                | (i) | (ii) | (iii) | (iv) | (v) | (vi) | (vii) |
|----------------------------------------|-----|------|-------|------|-----|------|-------|
| Skew (Section 2.1)                    | NO  | NO   | NO    | NO   | NO  | NO   | NO    |
| rKL/NDKL (Section 2.1)                | YES | YES  | NO    | NO   | NO  | NO   | NO    |
| ABR (Section 2.2)                     | YES | NO   | N/A   | YES  | NO  | NO   | NO    |
| ECE (Section 2.2) / AWRF (Section 2.3) | YES | YES  | YES   | YES  | NO  | NO   | NO    |
| GFR (Section 3.1)                     | YES | YES  | YES   | YES  | YES | YES  | YES   |

(i) handles ranked retrieval; (ii) considers every attribute value; (iii) is free from zero gold probability inconvenience; (iv) does not involve an unbounded criterion; (v) avoids combined attribute values; (vi) distinguishes between nominal and ordinal attribute sets; (vii) a search engine user population is modelled to quantify group fairness.

measures such as NDKL (Equation (2)) cannot tell whether a SERP is biased towards $a_1$ and $a_2$, and propose to compute a (weighted) average of the polarity value $(F(L, k, a_1) - F(L, k, a_2))$ across document ranks, where $F$ is a group membership flag as before (see Equation (3)), but returns 1 only when the document at $k$ belongs to the group in question and is relevant. In Section 4.1, we demonstrate that our GFR framework can easily quantify the overall polarity of each ranked list given a binary attribute set. we also compare it with AWRF in this context.

### 2.6 Diversity Evaluation Measures

Cherumanal et al. [16] applied the measures proposed by Yang and Stoyanovich [69] as well as a diversified search evaluation measure ($\alpha$-nDCG [18]) to evaluate the Touché 2020 argument retrieval runs from CLEF 2020 [10]. They observed that systems are generally ranked differently by fairness, diversity, and relevance measures. Diaz et al. [21] discussed the connection between group fairness measures (for a distribution of ranked lists) and intent-aware diversity measures [1, 14, 54]. On the other hand, for diversity evaluation, Sakai and Song [62, Table 7] demonstrated a few advantages of their $D^\#$-measures over $\alpha$-nDCG and intent-aware measures; Sakai and Zeng [63] reported that an instance of the $D^\#$-measure called $D^\#$-nDCG outperformed intent-aware measures in terms of how the measure agrees with human SERP preferences.

Zehlike et al. [70] remarked that $D^\#$-nDCG can be applied to group-fair ranking evaluation by treating their binary attribute set (protected/non-protected) as two search intents behind the same query. We generalise this idea and use $D^\#$-nDCG as the representative of existing diversity measures in our experiments to see how group-fair and diversified ranking evaluations are related given either hard or soft group membership. $D^\#$-nDCG is the average of intent recall (a.k.a. subtopic recall [71]) and $D$-nDCG [62]. The difference between the standard nDCG (for ad hoc IR) and D-nDCG (for diversified IR) is that the latter is based on the global gain of each document $d$:

$$GG(d) = \sum_{i \in I_q} Pr(i | q)gv_i(d),$$

(15)

where $I_q$ is the set of known intents for topic $q$, $Pr(i | q)$ is the probability that a user who enters $q$ as a query has Search Intent $i$, and $gv_i(d)$ is the gain value of $d$ for Intent $i$. From Equation (15), it can be observed that if the intent probabilities are uniform and the attribute values are mutually exclusive (i.e., hard group membership is defined), the global gain reduces to a single gain value and, therefore, that D-nDCG reduces to the standard nDCG. This is exactly the situation in our first experiment (Section 4.1), as it uses the aforementioned Touché 2020 data: each ranked item (i.e., opinion) is either PRO or CON, but not both.
3 PROPOSED EVALUATION FRAMEWORK

Section 3.1 presents our proposed framework for evaluating ranked lists based on group fairness and relevance. Section 3.2 describes three existing quantification measures that our framework uses as its component for the purpose of comparing the achieved distribution of a SERP to the target distribution, as we have depicted in Figure 1.

3.1 Group Fairness and Relevance

Our premise is that we are given $M$ attribute sets, where the $m$-th attribute set is $A^m (m = 1, \ldots, M)$, with $i$ specifying a particular attribute value $a_i^m (\in A^m)$. For each $A^m$, we are also given a target distribution $p^m$ over its attribute values, such that $\sum_i p^m(a_i^m) = 1$. An example setting with $M = 2$ would be

$$p^\text{Gender}(\text{female}) = p^\text{Gender}(\text{male}) = p^\text{Gender}(\text{other}) = 1/3,$$

$$p^\text{Age}(x < 20) = p^\text{Age}(x \geq 80) = 0.2, \quad p^\text{Age}(20 \leq x < 80) = 0.6.$$  

Given these targets, we are also given a ranked list $L$ to evaluate, where each item at rank $k$ has a group membership probability $G(L, k, a_i^m)$ such that $\sum_i G(L, k, a_i^m) = 1$ for every $k$. (Recall that group membership flag $F$ is a special case of $G$.) If the item at rank $k$ does not correspond to any of the attribute values of $A^m$, we let $G(L, k, a_i^m) = 1/|A^m|$. That is, we assume that the distribution over the attribute values is uniform for that item. Our objective is to quantify how well $L$ aligns with the target distributions and, if relevance assessments are also available, evaluate $L$ in terms of both group fairness and relevance.

For each attribute set, we can evaluate the group fairness of $L$ as

$$GF^m(L) = \sum_{k=1}^{|L|} \text{Decay}^m(L, k) \ DistrSim^m(L, k),$$  

(16)

where $\text{Decay}^m(L, k)$ is a function that represents user attention decay as they go down the ranked list and $\text{DistrSim}^m(L, k)$ compares the achieved distribution $p^m(L, k)$ with the target distribution $p^m$. In the present study, we employ the relevance-based decay of ERR [15, 21] by default, as we view this user model to be more realistic than those of nDCG and RBP, which disregard item relevance. That is,

$$\text{Decay}(L, 1) = P^\text{rel}_{L,1}, \quad \text{Decay}(L, k) = P^\text{rel}_{L,k} \prod_{j=1}^{k-1} (1 - P^\text{rel}_{L,j}) \quad (k > 1).$$  

(17)

where $P^\text{rel}_{L,k} = (2^g - 1)/2^{g_{\text{max}}}$ if the relevance grade of the item ranked at $k$ is $g$ and $g_{\text{max}}$ is the highest relevance grade for the test collection [15]. Note that we have removed the $m$ from Equation (17), as the present study assumes that group membership for a particular attribute set does not affect attention decay. While “group fairness seen so far” may well affect user attention decay just like “relevance seen so far,” this consideration is left for future work. When relevance assessments are unavailable, we employ RBP-based decay instead\footnote{For RBP, we let $\phi = 0.85$ because this setting has been shown to align well with users’ SERP preferences [63] and was the choice in the recent work of Moffat et al. [44].}:

$$\text{Decay}(L, k) = (1 - \phi)\phi^{k-1} \quad \text{with} \quad \phi = 0.85.$$  

(18)

This is equivalent to assuming that $P^\text{rel}_{L,k} = 0.15$ for any $L$ and $k$ when computing Equation (17).
For each rank $k$ in $L$, the achieved distribution $p_{L,k}^m$ for attribute set $A^m$ is defined using the average group membership probability over top $k$ for attribute value $a_i^m \in A^m$.

$$p_{L,k}^m(a_i^m) = \frac{\sum_{j=1}^{k} G(L, j, a_i^m)}{k}.$$  
(19)

As for $DistrSim^m(L, k)$, we consider different functions for comparing $p_{L,k}^m$ with $p_*$ depending on whether the attribute values are nominal or ordinal and, in the latter case, whether considering the ordinal nature of the attribute values makes sense (see Section 1). More specifically, given an achieved and the gold probability mass functions $p$ and $p_*$, we consider the following options$^{10}$:

$$DistrSim(p \parallel p_*) = 1 - \text{Divergence}(p \parallel p_*),$$  
(20)

where $\text{Divergence}(p \parallel p_*)$ is either JSD, NMD, or RNOD, as we shall define in Section 3.2. We use notations such as $\text{DistrSim}^{\text{BSD}}$ where appropriate. Note that JSD should be used for nominal attribute sets unless the attribute set is binary. For binary attribute sets, any of the above divergences can be used as there is no distinction between nominal and ordinal scales. In fact, NMD and RNOD are the same when $A^m = \{a_1, a_2\}$; the proof from Sakai $^{58}$ is duplicated in the Appendix. Hence, we have two options for the binary case: NMD (i.e., RNOD) and JSD.

Given $M$ attribute sets, we predefine a set of weights $w_0, w_1, \ldots, w_M$ such that $\sum_{m=0}^{M} w_m = 1$, and compute the overall score of $L$ as a weighted average. We call it the GFR (Group Fairness and Relevance) score.

$$GFR(L) = w_0 \text{Relevance}(L) + \sum_{m=1}^{M} w_m \text{GF}^m(L).$$  
(21)

Here, $\text{Relevance}(L)$ is a relevance-based score; we let $w_0 = 0$ if relevance assessments are unavailable. In either case, the present study considers only unweighted versions of Equation (21) and leaves the question of how $w_m$’s should be set for future work. Our view is that examining each component of GFR (i.e., the relevance component and the GF component for each attribute set) is more important than relying entirely on tuned GFR scores when evaluating IR systems; we argue that paying attention to detail, rather than an overall summary, is particularly important in fairness evaluation.

As for the choice of $\text{Relevance}(L)$, we consider ERR and $\text{iRBU}$ (intentswise Rank-Biased Utility) $^{59, 63}$ because these measures also rely on the realistic decay function given by Equation (17) and, therefore, enable us to rewrite Equation (21) as follows.

$$GFR(L) = \sum_{k=1}^{L} \text{Decay}(L, k) \left( w_0 \text{Utility}(L, k) + \sum_{m=1}^{M} w_m \text{DistrSim}^m(L, k) \right).$$  
(22)

Here, $\text{Utility}(L, k) = 1/k$ for ERR and $\text{Utility}(L, k) = \phi^k$ for $\text{iRBU}$.\textsuperscript{11} Whereas ERR is suitable for navigational searches, $\text{iRBU}$ is a measure that behaves surprisingly similarly to nDCG $^{59, 63}$ and is more geared towards informational searches. In Section 3.3.1, we shall explain the user model behind Equations (16) and (22) in order to contrast its principle with that of AWRF.

\textsuperscript{10}Equation (20) presupposes that $\text{Divergence}(p \parallel p_*)$ has the $[0,1]$ range; unbounded divergences such as KLD are not applicable here.

\textsuperscript{11}For $\text{iRBU}$, we let $\phi = 0.99$ in the present study as this setting has been shown to align well with users’ SERP preferences $^{63}$. This is a parameter inherited from RBP but is used for computing SERP utility rather than decay $^{3}$.
At the TREC Fair Ranking Track, AWRF and nDCG are multiplied (Equation (11)) rather than linearly combined. Following this approach, we also consider the following alternative.

\[
GFR(L) = \begin{cases} 
\text{Relevance}(L) \prod_{m=1}^{M} GFM^m(L) & \text{if relevance assessments are available} \\
\prod_{m=1}^{M} GFM^m(L) & \text{otherwise.}
\end{cases}
\]

(23)

3.2 Quantification Measures

To make this article self-contained, this section formally defines JSD, NMD, and RNOD, which can be used as Divergence\( (p \parallel p_*) \) in Equation (20).

Let \( C \) denote the set of ordinal classes, represented by consecutive integers for convenience. Let \( p_i \) denote the estimated probability for Class \( i \), so that \( \sum_{i \in C} p_i = 1 \). Similarly, let \( p_i^* \) denote the gold probability. We also denote the entire probability mass functions by \( p \) and \( p^* \), respectively.

As we have discussed in Section 2.1, JSD\([41]\) overcomes Problem (c) (handling situations in which \( p_i^* = 0 \)) and Problem (d) (unboundedness) of KLD.

\[
KLD(p \parallel p^*) = \sum_{i \in C \text{ s.t. } p_i > 0} p_i \log_2 \frac{p_i}{p_i^*}.
\]

(24)

\[
JSD(p, p^*) = \frac{KLD(p \parallel p^M) + KLD(p^* \parallel p^M)}{2},
\]

(25)

where \( p_i^M = (p_i + p_i^*)/2 \). Note that, unlike KLD, JSD is symmetric.

It is easy to observe from the simple summation across classes in Equation (24) that neither KLD nor JSD can consider the ordinal nature of classes. In other words, nominal quantification measures such as KLD and JSD disregard whether Classes \( i \) and \( j \) are close to each other or far apart \([57, 58]\). This is why the present study also considers two ordinal quantification measures, NMD and RNOD, as described later. To the best of our knowledge, we are the first to utilise ordinal quantification measures in the context of fair information access evaluation.

Let \( cp_i = \sum_{k \leq i} p_k \), and \( cp_i^* = \sum_{k \leq i} p_k^* \). Then, NMD is given by \([55]\):

\[
NMD(p, p^*) = \sum_{i \in C} |cp_i - cp_i^*|/|C| - 1.
\]

(26)

This is a normalised form of Earth Mover’s Distance (also known as \textit{Wasserstein} or \textit{Mallows Distance}) \([40, 68]\).

Finally, RNOD is defined as follows. First, let the Distance-Weighted sum of squares for Class \( i \) be:

\[
DW_i = \sum_{j \in C} \delta_{ij}(p_j - p_j^*)^2, \quad \delta_{ij} = |i - j|.
\]

(27)

\( DW_i \) was designed to quantify the overall error from the viewpoint of a particular gold class \( i \): it tries to measure how much of its probability \( p_i^* \) has been misallocated to other classes \( j \in C(j \neq i) \) by assuming that the difference between \( p_j \) and \( p_j^* \) is directly caused by a misallocation of part of \( p_i^* \). The weight \( \delta_{ij} \) is designed to penalise the misallocation based on the distance between the ordinal classes.

Let \( C^* = \{i \in C | p_i^* > 0 \} \). That is, \( C^*(\subseteq C) \) is the set of classes with a non-zero gold probability. Order-aware Divergence is defined as

\[
OD(p \parallel p^*) = \frac{1}{|C^*|} \sum_{i \in C^*} DW_i.
\]

(28)
RNOD is then defined as
\[
RNOD(p \parallel p^*) = \sqrt{\frac{OD(p \parallel p^*)}{|C| - 1}}.
\] (29)

Although a symmetric version of RNOD called RSNOD is available (based on symmetric order-aware divergence given by \(SOD(p, p^*) = (OD(p \parallel p^*) + OD(p^* \parallel p))/2\) [55], we do not consider RSNOD in our experiments because (a) symmetry is not required for quantification evaluation and group-fair ranking evaluation and (b) experiments with several datasets have shown that introducing symmetry is not beneficial in terms of system ranking consistency (i.e., robustness of the system rankings to the choice of test data) [57, 58].

As proven in the Appendix, NMD and RNOD are equivalent when \(|C| = 2\) (i.e., when there are only two classes).

3.3 How Does GFR Differ from AWRF in Principle?

The TREC Fair Ranking Track has been using AWRF (with an nDCG-like decay) for evaluating single-ranked lists since TREC 2021 [24]. The GFR framework has been adopted in the ongoing NTCIR-17 FairWeb-1 task. Before comparing them empirically in the remainder of this article, this section discusses two fundamental differences between the two approaches in detail. The differences are (1) whether a search engine user population model motivates the quantification of group fairness (Section 3.3.1), and (2) how intersectional group fairness is handled, in particular for nominal and ordinal attribute sets (Section 3.3.2).

3.3.1 Search Engine User Population Model. Figure 2 (see Section 1) succinctly presents the search engine user model behind the GFR framework. It shows that GFR is composed of a decay function that models the behaviour of a user population that scans the SERP, a utility function that reflects the SERP’s relevance, and a DistrSim function that reflects the SERP’s Group Fairness. When only the decay and the utility components are combined, this yields a family of measures known as Normalised Cumulative Utility (NCU) [61, 63], which subsumes widely used measures such as Average Precision and ERR. In fact, the left half of Figure 2 shows exactly what ERR does when the highest relevance grade is \(g_{\text{max}} = 3\) and the SERP has 1-relevant and 3-relevant documents at ranks 2 and 4, respectively (see Equation (17)). In this particular example based on ERR, 1/8 of the user population (looking at the same SERP) is assumed to abandon the SERP at rank 2, and the utility of the SERP for this first user group is 1/2. Then, 49/64 of the user population is assumed to abandon the SERP at rank 4, and the utility of the SERP for this second user group is 1/4. Thus, NCU quantifies the expected utility for the user population by viewing the decay function as the abandoning probability distribution over the SERP ranks.

It is easy to see from Figure 2 that the GFR framework generalises the idea of NCU. First, if only the decay component and the DistrSim component are combined, we have the GF measure (Equation (16)). Based on the SERP abandonment model, we quantify the group fairness experienced by the first and second user groups separately, and compute the expected group fairness experienced by the user population. Similarly, when the decay, utility, and DistrSim components are all incorporated, we have the GFR measure (Equation (22)) in which the experience of each user group

---

12https://fair-trec.github.io/docs/Fair_Ranking_2022_Participant_Instructions.pdf.
13http://sakailab.com/fairweb1/.
14In our current instantiation, the nonrelevant document at rank 3 in Figure 2 contributes equally to each attribute value and therefore affects the achieved distribution for the user group that abandon the SERP at rank 2. This is because the top ranked document is considered to be a part of user experience for that user group. In contrast, nonrelevant documents do not affect the achieved distribution for AWRF (See Equation 8).
is assumed to be a linear combination of the relevance-based utility and how a part of the SERP is close to each of the gold group-fair distribution. That is, GFR represents expected user experience, in which the user experience reflects both relevance and group fairness for each user who abandons the SERP at a certain rank.

It is clear that this approach is quite different from those of AWRF and AWRF*nDCG. As Equation (7) shows, AWRF compares a single distribution achieved by the entire SERP with the gold distribution, in which the achieved distribution is based on the ECE (Equation (5)), which is based entirely on the item exposure point of view. That is, the search engine user viewpoint sits entirely in the nDCG component and not in AWRF. The final AWRF*nDCG thus combines a score based on the items’ viewpoint and one based on the user’s viewpoint, implicitly assuming that these two are independent of each other. The product version of GFR (Equation (22)) also multiplies the expected utility (the relevance component) and each expected experience that is purely based on one particular attribute set.

We do not claim that the GFR model is “better” than existing approaches such as AWRF. We merely state the fact that our fundamental model is clearly different from them and that it is a sensible approach not yet explored.

3.3.2 Handling Intersectional Group Fairness, Especially for Nominal and Ordinal Attribute Sets. In Section 2.1, we pointed out that if the ordinal nature of attribute values should be taken into account (see Figure 1), combining multiple attribute sets can be problematic, as it is difficult to define the distance between a combined attribute value (e.g., a particular combination of Age Groups (ordinal) AND Gender (nominal)) and another. Unless this problem is somehow overcome, existing group fair IR evaluation methods such as AWRF can use only nominal quantification measures such as JSD for comparing the similarity between a combined distribution of a SERP with a combined gold distribution. In Section 4.3.1, we shall revisit this issue with an actual example from our experiments.

Our approach of computing a distribution similarity for each attribute set and finally taking a linear combination also facilitates the analysis of what is happening per attribute set. For example, if two attribute sets are considered and a combined GF score is computed for each search engine, this can easily be visualised on a two-dimensional graph in which the axes represent the component GF scores and the combined GF values can be represented as contour lines, as we shall demonstrate in Section 4.3.1. A similar analysis is not as straightforward with the existing approaches that rely on combined attribute values: for example, a certain AWRF score on a Cartesian product “Age Groups × Gender” does not tell us whether the SERP is doing well in terms of Age Groups but suffers in terms of Gender or whether it is the other way around. We argue that GFR is an approach worth exploring from this viewpoint as well.

4 EXPERIMENTS

This section demonstrates the versatility of our group fairness evaluation framework through three case studies with real data. Hereafter, we consider evaluating the top \(|L| = 10\) items of any given ranked list. All relevance-based measures are computed based on an exponential gain value setting. That is, a gain value of \(gv = 2^g - 1\) is given to each \(g\)-relevant document (\(g = 0, 1, 2 \ldots\)). While the purpose of the following experiments is to demonstrate the intuitiveness and versatility of our GFR framework, we also include AWRF and AWRF*nDCG in our experiments for reference, as these are the measures that have been used in the TREC Fair Ranking Track since 2021.

15 If a relevance-based measure is also involved and GFR scores are computed, a three-dimensional graph might be useful.
4.1 Ranking Pros and Cons: Quantifying the Polarity with One Binary Attribute Set

As a case study of a ranking task with a binary attribute set, we follow Cherumanal et al. [16] and utilise theTouché 2020 Data from CLEF 2020 [10]. The participants in this task were asked to retrieve relevant arguments from a focused crawl of arguments originating from debate portals for each given query on some controversial topic, for example, “Is human activity primarily responsible for global climate change?” [10]. The task used the args.me corpus [2], a collection of 387,676 arguments, each tagged with either PRO (\(a_1\)) or CON (\(a_2\)). We use Version 1 of args.me, with the Version 1 qrels file from Touché 2020 (containing 2,964 topic-document pairs, covering 49 topics),\(^17\) and the 21 submitted runs. The qrels file offers graded relevance on a 6-point scale: \(g = 1, \ldots, 5\) along with \(-2\) (non-arguments). We treat the non-argument documents as nonrelevant (\(g = 0\)).

4.1.1 Uniform Gold Distribution over PRO and CON.

First, let us consider a flat setting, where the target distribution is \(p_\ast(\text{PRO}) = p_\ast(\text{CON}) = 0.5\). Although the runs were evaluated only in terms of relevance at Touché 2020,\(^18\) we computed the GFR scores using GF\(^{\text{SD}}\) (group fairness), ERR, and iRBU (relevance). GFR measures based on Equation (21) (linear combination) and Equation (23) (product) are denoted by ERR+GF\(^{\text{SD}}\), ERR*GF\(^{\text{SD}}\), and so on.\(^19\) In addition, we computed AWRF, nDGC, and AWRF*nDGC as defined in Section 2.3. Furthermore, we computed D\(^\ddagger\)-nDGC using the NTCIREVAL toolkit\(^20\) by treating PRO and CON as two search intents behind a query and treating the combination of the relevance level and the PRO/CON label for each document as a per-intent relevance label. As for intent probabilities, we also use a uniform distribution: \(Pr(\text{PRO} \mid q) = Pr(\text{CON} \mid q) = 0.5\) (see Equation (15)). Since we are dealing with hard group membership in this experiment (i.e., each document is either PRO or CON, never both), D\(^\ddagger\)-nDGC reduces to an average of intent recall and the standard nDGC (see Section 2.6).

Table 2 compares the run rankings of group fairness–only and relevance-only measures with D\(^\ddagger\)-nDGC under the flat setting using Kendall’s \(\tau\)\(^34\) with 95% confidence intervals\(^42\). Table 3 shows a similar comparison for measures that consider both relevance and group fairness, again, with D\(^\ddagger\)-nDGC. The following observations can be made from these tables.

- GF\(^{\text{SD}}\) is moderately correlated with ERR, iRBU, nDGC, and D\(^\ddagger\)-nDGC (\(\tau = 0.433–0.662\)). AWRF shows a similar trend, with slightly higher \(\tau\)s (\(\tau = 0.638–0.810\)). These results suggest that group fairness–only evaluation is related to but different from ad hoc IR evaluation and from diversity evaluation, at least in a hard group membership setting with a binary attribute set.\(^21\)
- GF\(^{\text{SD}}\) and AWRF yield moderately correlated but different rankings (\(\tau = 0.567, 95\% \text{ CI} [0.317, 0.743]\)). Thus, not surprisingly, these two pure group fairness measures are not equivalent.
- The four GFR measures (ERR+GF\(^{\text{SD}}\), etc.) are generally highly correlated with D\(^\ddagger\)-nDGC (\(\tau = 0.743–0.867\)). AWRF*nDGC is very highly correlated with D\(^\ddagger\)-nDGC (\(\tau = 0.971 95\% \text{ CI} [0.946, 0.984]\)) because the latter in this particular experiment is the average of intent recall and standard nDGC (see Section 2.6). In summary, the fairness and relevance measures are highly correlated with a diversity measure in this hard group membership experiment.

---

\(^{16}\)See https://webis.de/data/args-me-corpus.html and https://doi.org/10.5281/zenodo.3274636.

\(^{17}\)The qrels file contains 50 topics, but Topic 25 has no relevant documents.

\(^{18}\)The follow-up task, Touché 2021, computed nDGC based on rhetorical quality in addition to that based on relevance\(^11\).

\(^{19}\)We also experimented with GF\(^{\text{SM}}\) (which equals GF\(^{\text{NO}}\) for binary attributes), but we did not find this beneficial over GF\(^{\text{SD}}\).

\(^{20}\)http://research.nii.ac.jp/ntcir/tools/ntcireval-en.html (version 200626).

\(^{21}\)This high-level observation is in line with the results of Cherumanal et al. [16] who also used the Touché 2020 data to compare NKDL (group fairness), nDGC (relevance), and \(\alpha\)-nDGC (diversity).
### Table 2. System Ranking Correlations (Kendall’s τ with 95% CIs) for the 21 Touché 2020 Runs: Relevance-Only and Group Fairness–Only Measures, with D♯-nDCG

|         | AWRF | ERR  | iRBU | nDCG | D♯-nDCG |
|---------|------|------|------|------|---------|
| GF<sub>JSD</sub> | 0.567 | 0.433 | 0.662 | 0.567 | 0.643   |
|         | [0.317, 0.743] | [0.143, 0.651] | [0.448, 0.804] | [0.317, 0.743] | [0.421, 0.792] |
| AWRF   | -    | 0.638 | 0.714 | 0.733 | 0.810   |
|         |      | [0.414, 0.789] | [0.523, 0.837] | [0.552, 0.848] | [0.671, 0.894] |
| ERR    | -    | -    | 0.771 | 0.848 | 0.790   |
|         |      |      | [0.610, 0.871] | [0.733, 0.916] | [0.639, 0.882] |
| iRBU   | -    | -    | -    | 0.886 | 0.867   |
|         |      |      |      | [0.796, 0.938] | [0.764, 0.927] |
| nDCG   | -    | -    | -    | -    | 0.924   |
|         |      |      |      |      | [0.862, 0.959] |

### Table 3. System Ranking Correlations (Kendall’s τ with 95% CIs) for the 21 Touché 2020 Runs: Group Fairness and Relevance Measures, with D♯-nDCG

|         | iRBU<sup>+</sup>GF<sub>JSD</sub> | ERR<sup>+</sup>GF<sub>JSD</sub> | iRBU<sup>+</sup>GF<sub>JSD</sub> | AWRF<sup>+</sup>nDCG | D♯-nDCG |
|---------|---------------------------------|-------------------------------|-------------------------------|-----------------------|---------|
| ERR     | 0.724                           | [0.538, 0.843]               | [0.930, 0.980]               | [0.812, 0.943]       | 0.867   |
|         | [0.962]                         | [0.455, 0.807]               | [0.945, 0.871]               | [0.711, 0.880]       | [0.764, 0.927] |
| iRBU    | 0.686                           | [0.482, 0.819]               | [0.629, 0.878]               | [0.748, 0.921]       | 0.871   |
|         | [0.567]                         | [0.402, 0.783]               | [0.523, 0.837]               | [0.567, 0.854]       | [0.946, 0.984] |
| ERR<sup>*</sup>GF<sub>JSD</sub> | -                              | [0.482, 0.819]               | [0.629, 0.878]               | [0.748, 0.921]       | 0.743   |
| iRBU<sup>*</sup>GF<sub>JSD</sub> | -                              | -                             | [0.523, 0.837]               | [0.567, 0.854]       | 0.971   |
| AWRF<sup>*</sup>nDCG | -                              | -                             | -                             | -                     | [0.946, 0.984] |

- The four GFR measures are highly correlated with AWRF<sup>*</sup>nDCG (τ = 0.714–0.895). They are also generally highly correlated with one another (τ = 0.629–0.962).

Next, to compare the statistical stability of the evaluation measures, we utilise discriminative power [51, 52, 54], as it is one of the most widely used method for comparing IR measures (See, for example, Anelli et al. [4], Ashkan and Metzler [5], Chuklin et al. [17], Clarke et al. [20], Golbus et al. [30], Kanoulas and Aslam [33], Leelanupab et al. [39], Lu et al. [43], Robertson et al. [50], Valcarce et al. [67])). Reasonable statistical stability is a necessary condition for a reliable evaluation measure, although this is by no means a sufficient condition. For a given significance level (e.g., α = 0.05), the discriminative power of an evaluation measure with respect to a test collection with a set of N runs is the proportion of significantly different run pairs among the N(N − 1)/2 pairs according to a statistical significance test, preferably a multiple comparison procedure [56].

However, as relying on a particular significance level promotes dichotomous thinking [31, 56] (i.e., just caring about whether a difference is significant without considering the effect size or practical significance), it is preferable to examine the entire set of p-values obtained. More specifically, sorting the N(N − 1)/2 p-values in decreasing order gives us a discriminative power curve; one that

---

<sup>22</sup>Originally, Sakai [51, 52] used the bootstrap test without correction for multiple comparisons.
Fig. 3. Discriminative power curves based on the randomised Tukey HSD test (Touché 2020 runs).

is relatively close to the origin represents a measure that often gives us conclusive results under a given experimental setting due to its statistical stability.

Figure 3 shows the discriminative power curves of the measures for the 21 Touché runs (210 run pairs) based on a randomised Tukey HSD test \[56\] with 5,000 trials conducted with the \(\text{Discpower}\) tool.\[23\] Table 4 slices the figure in half to show the number of statistically significant differences found at the 5% significance level for each measure. It also shows the minimum difference in means observed among all run pairs that were found statistically significant \[53\]. The following observations can be made.

- The most striking feature of the discriminative power results is that AWRF far underperforms other measures. In particular, it substantially underperforms GF\(^{\text{JSD}}\), even though both are pure group fairness IR measures. Recall the fundamental differences between GF and AWRF discussed in Section 3.3.1; it is clear that GF’s user model, which assumes that different users abandon the ranked list at each rank where a relevant document is found, helps boost its discriminative power. The third column of Table 4 suggests that, for AWRF, a mean difference of about 0.24 is needed to achieve statistical significance at \(\alpha = 0.05\) with this dataset.

- Among the relevance measures, nDCG outperforms ERR and iRBU. This is because the ERR-based decay used in ERR and iRBU is top heavy. For example, when a highly relevant document is found near the top of the ranked list, it is assumed that few users will continue to go further down the list. This is not a new finding \[59\]. Thanks to the nDCG component, it can be observed that AWRF*nDCG enjoys a high discriminative power despite the just described lack of discriminative power of AWRF. Similarly, \(\text{D}^\#\text{nDCG}\), which is an average of nDCG and intent recall in this experiment, shows a relatively high discriminative power.

- ERR, iRBU, and the GFR measures based on these two relevance measures perform more or less similarly to \(\text{D}^\#\text{nDCG}\). The method of combining a relevance measure and a GF measure (i.e., linear combination vs. multiplication) seems to have very little impact on discriminative power in this experiment.

The strikingly low discriminative power of AWRF is rather worrying, as such measures will not be able to provide researchers with statistically conclusive results if used without the nDCG component. Upon inspection, we found that the 20 statistically significant differences detected by

\[23\]http://research.nii.ac.jp/ntcir/tools/discpower-en.html.
Table 4. Discriminative Power at the $\alpha = 0.05$ Significance Level, Obtained from Figure 3

| Measure          | Disc. power | Min. diff. observed |
|------------------|-------------|---------------------|
| AWRF*nDCG        | 99/210 = 47.1% | 0.13                |
| nDCG             | 98/210 = 46.7% | 0.13                |
| D\#-nDCG         | 78/210 = 37.1% | 0.12                |
| ERR*GFJSD        | 77/210 = 36.7% | 0.14                |
| ERR              | 76/210 = 36.2% | 0.18                |
| ERR+GFJSD        | 73/210 = 34.8% | 0.14                |
| iRBU             | 71/210 = 33.8% | 0.19                |
| iRBU*GFJSD       | 70/210 = 33.3% | 0.17                |
| iRBU+GFJSD       | 69/210 = 32.9% | 0.15                |
| GF-JSD           | 63/210 = 30.0% | 0.14                |
| AWRF             | 20/210 = 9.5%  | 0.24                |

"Min. diff. observed" is the minimum difference found to be statistically significant among all run pairs, as an estimate of the true minimum difference required to achieve statistical significance [53].

AWRF shown in Table 4 all tell the same story: among the 21 runs, one particular run statistically significantly underperforms all others. In contrast, the significance test results of GFJSD subsume the 20 run pairs and suggest more: for example, it says that 14 runs are statistically tied in the top performing group, outperforming 4 runs. While these differences do not tell us which measure is "correct," we note that the results of D\#-nDCG are quite similar to the GF-based results. More specifically, the top 15 runs according to D\#-nDCG in terms of statistical significance is a superset of GF’s top 14 runs.

Figure 4 visualises how the iRBU-based and GFJSD-based run rankings are correlated ($\tau = 0.662$, 95% CI [0.448, 0.804] as shown in Table 2). We stress that it is important to visualise the runs in this way to complement a list of runs ranked by GFR scores (Equation (21)), so that we can see how the GF and relevance components are contributing to GFR.\(^{24}\) The dotted lines represent contour lines in terms of iRBU+GFJSD. Figure 5 visualises the per-topic iRBU and GFJSD scores for the lowest performer indicated in Figure 4: as shown with a balloon in Figure 5, we can easily spot SERPs that are relatively poorly balanced between group fairness and relevance.

We have thus demonstrated that our GFR framework is applicable to a hard group membership setting with a binary attribute set for which both group fairness and relevance need to be considered. In this setting, the GFR measures are generally highly correlated with D\#-nDCG (which is the average of intent recall and straight nDCG in this case). In addition, we have observed that AWRF substantially underperforms GFJSD in terms of discriminative power.

4.1.2 100% PRO and 100% CON Gold Distributions. We now demonstrate how our framework can quantify the polarity of runs, that is, whether the runs are biased towards PRO or towards CON, and by how much. Instead of the flat setting that we considered earlier, let us consider a 100% PRO setting ($p_*(\text{PRO}) = 1$, $p_*(\text{CON}) = 0$) and a 100% CON setting ($p_*(\text{PRO}) = 0$, $p_*(\text{CON}) = 1$). Let $GF_{\text{PRO}}(L)$ and $GF_{\text{CON}}(L)$ denote a GF score for ranked list $L$ computed under the two settings, respectively. Then, $\Delta GF(L) = GF_{\text{PRO}}(L) - GF_{\text{CON}}(L)$ is a direct measure of the\(^{24}\) Similar practices have been used in the NTCIR INTENT tasks (plotting relevance against diversity) [66] and more recently in the TREC Fair Ranking Tracks (plotting relevance against (un)fairness) [7, 8, 24].
Fig. 4. Visualising the mean group fairness and relevance scores of the 21 Touché runs (over 49 topics).

Fig. 5. Visualising the per-topic scores of the lowest performer from Figure 4. The Kendall’s $\tau$ between iRBU and $G_F^{JSD}$ for this run is 0.867 (95% CI [0.810, 0.908], $n = 49$).

polarity of $L$: a positive score implies an overall bias towards PRO, and so on. Note that replacing the target distribution affects only the DistrSim part of Equation (16). For reference, we also consider a 100% PRO setting and a 100% CON setting for AWRF (Equation (7)), and compute $\Delta_{AWRF}(L) = AWRF_{PRO}(L) - AWRF_{CON}(L)$.

Figure 6 compares the mean $\Delta G_F^{JSD}$ scores of the 21 Touché 2020 runs; mean $\Delta_{AWRF}$ scores are also shown for comparison. Both measures suggest that the majority of the runs are PRO biased. Both measures point to a run called DreadPirateRoberts-2 as the most PRO biased one. On the other hand, the figure also shows that $\Delta G_F^{JSD}$ and $\Delta_{AWRF}$ disagree on the polarity for 4 runs, as
indicated by the balloons. For example, while Aragorn-3 is the most CON-biased according to $\Delta GF$ (although the run is close to neutral), it is slightly on the PRO-biased side according to $\Delta AWRF$.

To illustrate the mechanism of how $\Delta GF$ and $\Delta AWRF$ can behave differently, we examine DreadPirateRoberts-2 closely. While this run is the most PRO biased according to both measures, it can be observed from Figure 6 that $\Delta GF$ is more decisive than $\Delta AWRF$. Figure 7 visualises the per-topic $\Delta GF$ and $\Delta AWRF$ scores for DreadPirateRoberts-2. It can be observed that many $\Delta GF$ scores are close to 1, while the trend is less clear for $\Delta AWRF$. The balloon in the figure indicates the maximum absolute difference between $\Delta GF$ and $\Delta AWRF$ across all topics: this is Topic = 23 (“Should euthanasia or physician-assisted suicide be legal?”).

Table 5 shows why $\Delta GF = 0.9666$ (i.e., heavily PRO biased) while $\Delta AWRF = -0.0234$ (i.e., slightly CON biased) for the above topic-run combination. As shown in the bottom half of the table, the $\Delta$s are obtained by subtracting the 100% CON-setting score from the 100% PRO-setting score. For this particular ranked list, only the top 5 documents are relevant; therefore, the other 5 documents do not contribute to the score calculation. To compute AWRF, the information shown in the Membership column is plugged into Equations (6) and (8), and a single achieved distribution for the entire SERP is obtained as shown at the bottom of the table: $p^{ECE}$ (PRO) = 0.4852, $p^{ECE}$ (CON) = 0.5148. As the achieved probability is slightly higher for CON, the AWRF (Equation (7)) for the 100% CON setting is also slightly higher than that for the 100% PRO setting, hence, the small negative $\Delta AWRF$ value shown in Figure 7. In contrast, as the top half of Table 5 shows, an achieved distribution (Equation (19)) is compared with the gold distribution five times to compute a GF score (Equation (16)), because GF assumes that there are five different user groups at ranks 1 to 5, each experiencing their own group fairness based on the portion of the SERP that they saw. More importantly, because there is a 5-relevant (i.e., highest relevance level) item at rank 1, the Decay column shows that the ERR-based decay model assumes that about 97% ($= (5^5 - 1)/5^5$) of the user population

Fig. 6. Quantifying the polarity of the 21 Touché 2020 runs (x-axis) using Mean $\Delta GF^{SD}$ and Mean $\Delta AWRF$ (y-axis) over 49 topics.
Fig. 7. Comparison of per-topic Δ scores for DreadPirateRoberts-2 (most PRO-biased run according to ΔGF<sup>SD</sup>).

**Table 5.** How ΔGF<sup>SD</sup> and ΔAWRF are Computed for DreadPirateRoberts-2 with Topic 23 (see Figure 7)

| Rank | Relevance level g | Decay (Equation (17)) | Membership (PRO/CON) | Achieved (Equation (19)) (PRO/CON) | Max diff for Topic 23 | ΔGF = 0.9666 | ΔAWRF = -0.0234 |
|------|-------------------|-----------------------|----------------------|-------------------------------------|-----------------------|----------------|-----------------|
| 1    | 5                 | 0.9688                | 1/0                  | 1/0                                 | GF<sup>SD</sup>       | 0.9890         | 0.0224          |
| 2    | 4                 | 0.0146                | 0/1                  | 0.5/0.5                             |                       |                |                 |
| 3    | 4                 | 0.0078                | 0/1                  | 0.3333/0.6667                       |                       |                |                 |
| 4    | 5                 | 0.0085                | 1/0                  | 0.5/0.5                             |                       |                |                 |
| 5    | 4                 | 0.0001                | 0/1                  | 0.4/0.6                             |                       | 0.6042         | 0.7635          |

We will access just the top-ranked item and then abandon the SERP. Since this top-ranked item belongs to PRO, and GF emphasises this fact, the GF score for the 100% PRO setting is extremely high whereas that for the 100% CON setting is extremely low, hence, the large positive ΔGF score shown in Figure 7. In summary, the discrepancy between ΔGF<sup>SD</sup> and ΔAWRF for this example arises from the fact that GF heavily emphasises the top-ranked item when it is extremely relevant.

We have thus demonstrated that ΔGF scores can quantify the polarity of each run as well as each ranked list in a straightforward manner. In addition, we have demonstrated that mean ΔGF scores may occasionally disagree with mean ΔAWRF scores due to the ERR-based decay function, which assumes that many users abandon the SERP at top ranks with highly relevant items and, therefore, few users continue examining the SERP further.

### 4.2 Ranking Web Pages: Soft Group Membership with One Attribute Set

We now demonstrate that our framework can handle soft group membership, that is, probabilities \( G(L, k, a_i) \) rather than flags \( F(L, k, a_i) \) such as PRO/CON. To this end, we utilise a diversified search...

---

25If a linear gain value setting is used instead of the exponential one, the probability would reduce to \( 5/(5 + 1) = 83\% \).
dataset from the NTCIR-9 INTENT Japanese subtask data [66]. We chose the NTCIR data over the TREC diversity task data [19] because (a) an NTCIR INTENT task dataset (with associated runs) contains 100 topics (with 3–24 intents per topic) whereas a TREC diversity dataset contains only 50; and, more importantly, (b) unlike the TREC data, the NTCIR data contains intent probabilities based on assessors’ majority votes. In our experiments, we directly utilise these intent probabilities to define the target distribution in the group fairness context by treating the search intents for each topic as attribute values. For example, Topic 0127 “Mikuniya” (a Japanese proper name) is an ambiguous topic because it can be a famous Japanese tea shop, a restaurant, a hot springs resort, and so on. These are all different entities; the tea shop intent has a 37% probability according to the INTENT data. We use these intent probabilities directly as the gold probability for group fairness evaluation. Thus, while search result diversification aims to satisfy many users with different intents behind the query Mikuniya, we view the problem in a group fairness context in which we want to make sure that we are giving a fair exposure to each entity named Mikuniya.

In the INTENT data, a single document may be relevant to multiple intents; therefore, soft group membership needs to be handled. For example, for “Mikuniya,” there are five documents that are relevant to as many as six intents. We define the soft group membership for an item based on its per-intent gain values: if there are three intents and the item has 3, 1, 0 as its per-intent gain values, the group membership is distributed across the intents as 3/4, 1/4, 0. To compute the ERR-based decay (Equation (17)) and the ad hoc measures (ERR, iRBU, and nDCG), we utilised the per-topic relevance grades available from Sakai and Zeng [63], which had been derived from the official per-intent relevance assessments. We then computed GF^SD-based GFR measures since the intents should be treated as nominal groups. We also computed D♯-nDCG, the official diversity measure used in the INTENT task, using the intent probabilities and the per-intent relevance assessments as shown in Equation (15). In addition, we computed AWRF and AWRF*nDCG for reference.

Using Kendall’s τ, Tables 6 and 7 compare the rankings of the 21 INTENT runs according to different measures. Recall that, unlike Section 4.1, we are dealing with soft group membership with 3 to 24 intents (i.e., attribute values) per topic. The following observations can be made.

- GF^SD is highly correlated with ERR, iRBU, nDCG, and D♯-nDCG (τ = 0.771–0.889). The τs are higher compared with Table 2 (binary hard group membership). AWRF shows a similar trend (τ = 0.779–0.922).
- The two pure group fairness measures GF^SD and AWRF are very highly correlated with each other (τ = 0.941, 95% CI [0.885, 0.970]).
- The four GFR measures are highly correlated with D♯-nDCG (τ = 0.830–0.922), as is AWRF*nDCG (τ = 0.856, 95%CI [0.731, 0.925]). These results suggest that even in this more general soft group membership setting, fairness-and-relevance ranking (for stakeholders of the ranked items) and search result diversification (for search engine users) may be two sides of the same coin, or at least, of two similar coins.
- The four GFR measures are also highly correlated with AWRF*nDCG (τ = 0.882–0.895); they are also very highly correlated with each other (τ = 0.908–0.980).

While both the Touché experiment and this INTENT experiment suggest that the fairness-and-relevance measures highly resemble D♯-nDCG, it should be noted that we have assumed that the target distribution for group fairness (see Equation (20)) and the probability distribution of intents given a topic (see Equation (15)) are one and the same. In practice, they may well differ, in which case GFR (or AWRF) and D♯-nDCG may possibly give us substantially different results.

Figure 8 shows the discriminative power curves of the measures for the 18 INTENT runs (153 run pairs, Tukey HSD test with 5,000 trials). Table 8 slices the figure in half to show the number
Table 6. System Ranking Correlations (Kendall’s $\tau$ with 95% CIs) for the 18 INTENT Runs: Relevance-Only and Group Fairness-Only Measures, with $D^\#$-nDCG

|        | AWRF     | ERR      | iRBU     | nDCG     | $D^\#$-nDCG |
|--------|----------|----------|----------|----------|-------------|
| GF$_{JSD}$ | 0.941 [0.885, 0.970] | 0.837 [0.699, 0.915] | 0.784 [0.610, 0.886] | 0.771 [0.589, 0.878] | 0.889 [0.790, 0.943] |
| AWRF | - 0.843 [0.709, 0.918] | - 0.791 [0.622, 0.890] | - 0.882 | - 0.777, 0.939) | - 0.850, 0.960] |
| ERR | - - 0.895 [0.801, 0.946] | - - 0.961 | - [0.924, 0.980] | - 0.830 | - 0.687, 0.911] |
| iRBU | - - - | - - | - [0.924, 0.980] | - | - 0.830 | - 0.687, 0.911] |
| nDCG | - - - | - - | - - | - | - | - |

Table 7. System Ranking Correlations (Kendall’s $\tau$ with 95% CIs) for the 18 INTENT Runs: Group Fairness and Relevance Measures, with $D^\#$-nDCG

|        | iRBU +GF$_{JSD}$ | ERR +GF$_{JSD}$ | iRBU *GF$_{JSD}$ | AWRF *nDCG | $D^\#$-nDCG |
|--------|-----------------|-----------------|-----------------|------------|-------------|
| ERR +GF$_{JSD}$ | 0.980 [0.960, 0.990] | 0.922 [0.850, 0.960] | 0.961 [0.924, 0.980] | 0.895 [0.801, 0.946] | 0.882 [0.777, 0.939] |
| iRBU +GF$_{JSD}$ | - 0.915 | - 0.941 | - 0.895 | - 0.790, 0.943 | - 0.830 |
| ERR *GF$_{JSD}$ | - - 0.837 [0.837, 0.957] | - - 0.885 [0.885, 0.970] | - - 0.895 | - - 0.801, 0.946 | - - 0.867, 0.911 |
| iRBU *GF$_{JSD}$ | - - - | - - | - - | - - | - - |
| AWRF *nDCG | - - - | - - | - - | - - | - - |

of statistically significant differences found at the 5% significance level for each measure, with the minimum difference in means observed among all run pairs that were found to be statistically significant [53]. The overall trends are similar to the Touché results (Figure 3 and Table 4). More specifically:

- AWRF underperforms GF$_{JSD}$ again, although it does better compared with the Touché experiment (Table 4). Again, for pure group fairness evaluation, GF’s user model that considers the relevance of each ranked item seems to boost its discriminative power.
- Among the relevance-only measures, nDCG outperforms iRBU and ERR as discussed in Section 4.1.1. Because of this, AWRF*nDCG also enjoys a high discriminative power. $D^\#$-nDCG (a diversity measure) also performs well: recall that it relies on D-nDCG (Equation (15)) rather than straight nDCG in this soft group membership experiment.
- iRBU, ERR, and the four GFR measures based on these two relevance measures perform more or less similarly to one another, with ERR doing less well compared with the Touché experiment. As for the method of combining a relevance measure and a GF measure, multiplication seems more promising at least for iRBU in this experiment: iRBU*GF$_{JSD}$ achieves a 44.4% discriminative power while iRBU+GF$_{JSD}$ achieves only 30.7%.
Fig. 8. Discriminative power curves based on the randomised Tukey HSD test (INTENT runs).

Table 8. Discriminative Power at the $\alpha = 0.05$ Significance Level, Obtained from Figure 8

| Measure            | Disc. power | Min. diff. observed |
|--------------------|-------------|---------------------|
| AWRF*$nDCG$       | 75/153 = 49.0% | 0.08                |
| $nDCG$            | 69/153 = 45.1%  | 0.09                |
| iRBU*GF$^{SD}$   | 68/153 = 44.4%  | 0.09                |
| D$^\#$-nDCG      | 60/153 = 39.2%  | 0.09                |
| $GF^{SD}$         | 53/153 = 34.6%  | 0.09                |
| ERR*GF$^{SD}$    | 53/153 = 34.6%  | 0.06                |
| ERR+GF$^{SD}$    | 53/153 = 34.6%  | 0.08                |
| iRBU              | 52/153 = 34.0%  | 0.12                |
| iRBU+GF$^{SD}$   | 47/153 = 30.7%  | 0.10                |
| AWRF              | 46/153 = 30.1%  | 0.10                |
| ERR               | 44/153 = 28.8%  | 0.08                |

“Min. diff. observed” is the minimum difference found to be statistically significant among all run pairs, as an estimate of the true minimum difference required to achieve statistical significance [53].

By examining the statistical significance test results of the top measures in Table 8 (excluding nDCG) more closely, we noted the following.

- According to AWRF*$nDCG$, the top performers are runs called MSINT-D-J-3 and MSINT-D-J-1: they outperform 8 other runs.
- According to iRBU*GF$^{SD}$, the top performers are MSINT-D-J-3, MSINT-D-J-1, MSINT-D-J-2, uogTr-D-J-1, and uogTr-D-J-2: they outperform 8 other runs.
- According to D$^\#$-nDCG, the top performers are MSINT-D-J-3, MSINT-D-J-1, MSINT-D-J-2, and MSINT-D-J-4: they outperform 7 other runs.

Thus, while the fairness-and-relevance measures behave similarly to a diversity measure, each measure tells a slightly different story.

We have thus demonstrated that our framework can handle soft group membership and that the GFR measures (as well as AWRF) are highly correlated with D$^\#$-nDCG in this experimental setting. In addition, we have shown that AWRF underperforms $GF^{SD}$ in terms of discriminative power, although this can be solved by combining it with nDCG.
4.3 Ranking Local Shops and Restaurants: Intersectional Group Fairness

Our third case study involves two attribute sets, one with nominal attribute values and the other with ordinal attribute values (but without relevance data). In Section 4.3.1, we demonstrate that (i) unlike prior work, our framework can consider the ordinal nature of attribute values if required; and (ii) intersectional group fairness of rankings can be examined without directly combining different attribute sets. For reference, we also discuss how AWRF compares with our GF framework in this setting. Finally, in Section 4.3.2, we closely examine how the choice of the DistrSim function (Equation (20)) for ordinal attribute groups may impact the GF-based evaluation outcome in some cases.

4.3.1 One Nominal and One Ordinal Attribute Sets, and Intersectional Group Fairness. For this experiment, we constructed a dataset based on a query log from a popular Local Shop and Restaurant Search service for smartphone users in Japan. Given a query, this local search service returns a ranked list of items (shops and restaurants) based on various features, including the relevance to the query, the proximity of the item to the user’s location, and user ratings (i.e., review scores). In our query log, each ranked item has a flag indicating whether it is a chain store owned by a company and, if it is, the name of the company that owns it. For some queries, a small number of companies that own many chain stores may dominate the ranking. Hence, as an example of imposing a group fairness requirement based on a nominal attribute set (with hard group membership), we require, for each query, that the ideal ranking should provide the same exposure to all relevant companies. More specifically, the gold distribution is defined as a uniform distribution overall companies that appear in the top 20 ranking (based on the current Local Search results) for that query, where each shop or restaurant that is not a chain store is treated as a distinct company. In order to demonstrate that our framework can evaluate group fairness from this viewpoint, we first obtained a random sample from a one-year local search query log (from September 2020 to September 2021), and then filtered it so that each ranking contains at least one company with multiple chain stores listed in the top 20. This gave us a set of 418 queries (or topics) for our experiment.

Our query log also contains a mean 5-point scale user rating score and a review count for each item. If an item has \(n (> 0)\) reviews, the mean rating is the average over \(n\) user ratings. If there is no review, the mean rating is set to zero. We believe that imposing a group fairness requirement based on review count is of practical significance because this statistic probably reflects the level of exposure of each item in the past, and items with low past exposure may deserve more future exposure. Hence, as an example of handling an ordinal attribute set (with hard group membership), we consider a group fairness constraint based on this view. In the aforementioned query log (before filtering by chain store information), 45% of the ranked items had zero reviews (Group 1), 22% had 1 to 10 reviews (Group 2), 23% had 11 to 100 reviews (Group 3), and the remaining 10% had over 100 reviews (Group 4). We utilise this distribution as the gold distribution over the four review count groups for all topics to demonstrate how this search application can be evaluated in terms of statistical parity [23].

We evaluate the following four runs (i.e., ranking schemes) to demonstrate how our framework can handle a nominal attribute set and an ordinal attribute set at the same time and thereby enable us to quantify intersectional group fairness.

---

26 This service is provided by LINE Corporation, Japan, and the query log was provided to the authors by Naver Corporation, Korea.
27 The exact probabilities we used in our calculations (based on the statistics from our query log) are \(p_1(a_1) = 0.452239\), \(p_1(a_2) = 0.220319\), \(p_1(a_3) = 0.227721\), \(p_1(a_4) = 0.0997214\).
Base. This represents the actual rankings returned by our current local search engine. The search engine leverages various features to produce the rankings as mentioned earlier, but it suffices to treat it as a black box for the purpose of the present experiment.

Rating. This reranks the top 20 search results by mean ratings from user reviews so that items with higher ratings are prioritised. This may possibly hurt the Review Count–based group fairness because the mean ratings are substantially positively correlated with review counts ($\tau = 0.617, 95\% CI [0.609, 0.625], n = 9,329$) according to our data. The correlation means that the items with higher mean ratings are often those that have already enjoyed good exposure to users and therefore received many reviews. Hence, by promoting highly rated items, Rating may also promote items with many reviews, defying the gold distribution which says that 45% of the SERP should belong to Group 1 (items with zero reviews), and so on. As a matter of fact, Rating certainly denotes items with zero reviews at least, since zero-review items are treated as zero-rating items in our experiments. On the other hand, it should be noted that even an item with only one review can have a mean rating of 5.0 (i.e., maximum); that is, ranking by mean rating is not exactly the same as ranking by review count.

Base-UC. Filter the top 20 items of the Base run so that each company (which can own multiple chain stores) appears no more than once. UC stands for “Unique Companies.” This is designed to improve the Company-based group fairness.

Rating-UC. Similar to Base-UC, except that the input to the filtering step is the Rating run. This should also improve the Company-based group fairness of Rating.

Since relevance assessments are missing in this experiment, we use the RBP decay (Equation (18)) to compute GF scores. For discussing Company-based group fairness (in which the number of nominal groups for the uniform gold distribution varies across topics), we compute GF$^\text{JSD}$. For discussing the Review Count–based group fairness (with a statistical parity-based gold distribution over four ordinal groups common to all topics), we compute GF$^\text{NMD}$, and GF$^\text{RNOD}$. We shall denote these measures as $c\text{GF}^\text{JSD}$ (Company-based GF measure using JSD), $r\text{GF}^\text{NMD}$ (Review Count–based GF measure using NMD), and so on. We also average a Company-based score and a Review Count–based score as a special case of Equation (21) with $w_0 = 0$ (i.e., doing without relevance) and $w_1 = w_2 = 0.5$ (where $i = 1, 2$ represent the Company and Review Count attribute sets, respectively); we denote them by $c\text{GF}^\text{JSD} + r\text{GF}^\text{NMD}$, and so on. We also consider multiplying a Company-based score and a Review Count–based score and denote them as $c\text{GF}^\text{JSD} \times r\text{GF}^\text{NMD}$, and so on.

As a reference, we also consider the following AWRF variants: Company-based AWRF (denoted as cAWRF), Review Count–based AWRF (denoted as rAWRF), and AWRF that combines the Company-based attribute values and Review Count–based attribute values (denoted as c×rAWRF).

As a reference, we also consider the following AWRF variants: Company-based AWRF (denoted as cAWRF), Review Count–based AWRF (denoted as rAWRF), and AWRF that combines the Company-based attribute values and Review Count–based attribute values (denoted as c×rAWRF). Figure 9 illustrates how a gold distribution is defined for computing c×rAWRF: Part (a) shows a uniform gold distribution for the Company attribute set (with nominal attribute values) for a topic that happens to have four companies to consider. Part (b) is the gold distribution for the Review Count attribute set (with ordinal attribute values), which is used for all topics. Part (c) is the resultant gold distribution over the Cartesian product of the two attribute sets: note that since the Company attribute values are nominal (i.e., the x-axis does not represent a meaningful order), so are the combined attribute values. For each topic, c×rAWRF compares a SERP’s achieved combined distribution with a gold distribution such as this one, unlike our approach, which compares the SERP with each of the component gold distributions one at a time. Our view is that our approach is more intuitive; regardless of our view, however, the fact remains that only nominal quantification measures such as JSD are applicable to combined distributions such as the one shown in Part (c), since the x-axis does not represent a meaningful order.
Fig. 9. How c×rAWRF defines a gold distribution over the Cartesian product of the Company attribute set (nominal) and the Review Count attribute set (ordinal) for a topic with four distinct companies to consider.

Table 9 shows the mean GF, GFR, and AWRF scores of the 4 runs averaged over the 418 topics. Based on these results, Table 10 summarises the significance test results for all 13 measures: 6 measures computed on one attribute set (either Company based or Review Count based), and 7 measures computed on the two attribute sets. The following observations can be made.

(i) The run rankings by the two Company-based measures (cGF_{JSD} and cAWRF) are the same, with Base-UC and Rating-UC tied at the top, and so are the significance test results. According to these measures, the two UC runs statistically significantly outperform Base and Rating. This means that the Unique-Company filtering step does improve the Company-based group fairness regardless of the baseline run, which is as expected.

(ii) When we compare the four Review Count–based measures, the rankings by rGF_{JSD} and rGF_{RNOD} are the same, with the same conclusions in statistical terms (with five statistically significantly different run pairs). While rAWRF also obtains five statistically significantly different run pairs, the ranking and conclusions reached are slightly different: only rAWRF says that Base statistically significantly outperforms Rating-UC.

(iii) As for rGF_{NMD}, it ranks the 4 runs quite differently compared with the other Review Count–based measures; it says that Rating-UC statistically significantly outperforms Base, which conflicts directly with the aforementioned significance test result with rAWRF. Moreover, note that rGF_{NMD} is the only Review Count–based measure that rates Rating higher than Base in terms of mean scores. This seems rather counterintuitive, as we have observed that the mean ratings (i.e., the sort key used for obtaining Rating) are substantially positively correlated with review counts and, therefore, that Rating probably tends to hurt the Review Count–based group fairness.

(iv) All three rGF measures (but not rAWRF) say that Base-UC statistically significantly outperforms Base. This suggests that the Unique-Company filtering step improves not only the Company-based group fairness but also the Review Count–based group fairness. This seems intuitive since giving a chance to different companies should result in diverse review counts.

(v) Looking at the measures based on the two attribute sets, it can be observed that, with the exception of the NMD-based measures, they all rank the runs identically and reach the same conclusions with six statistically significant differences. That is, every run pair
Table 9. Mean GF, GFR, and AWRF Scores of the 4 Local Search Runs (Over 418 Topics)

| Measure          | Base  | Base-UC | Rating | Rating-UC |
|------------------|-------|---------|--------|-----------|
| (a) One attribute set |       |         |        |           |
| cGF$_{JSD}$      | 0.404 | 0.479   | 0.401  | 0.479     |
| cAWRF            | 0.708 | 0.846   | 0.698  | 0.846     |
| rGF$_{JSD}$      | 0.495 | 0.545   | 0.445  | 0.506     |
| rGF$_{NMD}$      | 0.542 | 0.575   | 0.551  | 0.584     |
| rGF$_{RNOD}$     | 0.460 | 0.500   | 0.422  | 0.470     |
| rAWRF            | 0.728 | 0.745   | 0.663  | 0.704     |
| (b) Two attribute sets |       |         |        |           |
| cGF$_{JSD}$+rGF$_{NMD}$ | 0.449 | 0.521   | 0.423  | 0.493     |
| cGF$_{JSD}$+rGF$_{RNOD}$ | 0.201 | 0.270   | 0.181  | 0.255     |
| cGF$_{JSD}$+rGF$_{NMD}$ | 0.473 | 0.527   | 0.476  | 0.532     |
| cGF$_{JSD}$+rGF$_{RNOD}$ | 0.220 | 0.281   | 0.222  | 0.287     |
| cGF$_{JSD}$+rGF$_{RNOD}$ | 0.432 | 0.490   | 0.411  | 0.475     |
| c$x$\times$r$AWRF | 0.354 | 0.394   | 0.325  | 0.381     |

The highest and lowest mean scores for each measure are **bolded** and *underlined*, respectively.

is statistically significant, with **Base-UC** being the top performer and **Rating** being the worst performer. As for cGF$_{JSD}$+rGF$_{NMD}$ and cGF$_{JSD}$+rGF$_{RNOD}$, their run rankings and the statistical significance test results are the same as those of rGF$_{NMD}$.

(vi) The combination method for GF measures (i.e., linear or product) has no impact on the rankings and the significance test results.

Observations (iii) and (v) imply that the choice of the DistrSim function (Equation (20)) for GF measures matters at least in some situations. In Section 4.3.2, we shall examine the results more closely to discuss the impact of the choice of DistrSim.

Figure 10 provides a visual summary of our local search experiment using GF measures by plotting the rGF$_{RNOD}$ scores (Review Count–based, ordinal attribute values) against the cGF$_{JSD}$ scores (Company-based, nominal attribute values) for the 4 runs that we considered. The green dotted lines represent contours according to cGF$_{JSD}$+rGF$_{RNOD}$. The figure visualises the fact that the Unique-Company filtering step improves the ranking in terms of both Company-based and Review Count–based group fairness (purple arrows) and that reranking by mean rating hurts the Review Count–based group fairness only (red arrow). From the viewpoint of companies that may potentially suffer in terms of exposure, the figure tells us the following:

- Small companies (those owning few chain stores) cannot enjoy much exposure if the SERP is occupied by large companies (comparison along the x-axis).
- Companies with low mean ratings cannot enjoy much exposure if companies with high mean ratings are prioritised in the SERP (comparison along the y-axis).
- Therefore, under a search algorithm that allows high exposure to large companies with high mean ratings (e.g., the **Rating** run), small companies with low mean ratings will suffer most.

We have thus demonstrated that **our framework enables researchers to study intersectional group fairness, even when both nominal and ordinal attribute sets are involved.**

4.3.2 On the choice of DistrSim for GF with Ordinal Attribute Values. Tables 9 and 10 show that rGF$_{NMD}$ behaves differently from the other Review Count–based measures, and that rGF$_{JSD}$ and rGF$_{RNOD}$ agree with each other even in terms of statistical significance. Hence, to closely investigate the impact of the choice of the DistrSim function for a GF measure, we first filtered...
Table 10. Randomised Tukey HSD test results ($\alpha = 0.05$) for Table 9

| Measure | Conclusions |
|---------|-------------|
| cGF$^{\text{JSD}}$, cAWRF, rGF$^{\text{NMD}}$, cGF$^{\text{JSD}}$+rGF$^{\text{NMD}}$, cGF$^{\text{JSD}}$*rGF$^{\text{NMD}}$ | Rating-UC $\gg$ Base, Rating |
| Base-UC $\gg$ Base, Rating |

(a) 4 pairs with a statistically significant difference

| Measure | Conclusions |
|---------|-------------|
| rGF$^{\text{JSD}}$, rGF$^{\text{RNOD}}$ | Base-UC $\gg$ Rating-UC, Base, Rating |
| Rating-UC $\gg$ Rating |
| Base $\gg$ Rating |

(b) 5 pairs with a statistically significant difference

| Measure | Conclusions |
|---------|-------------|
| cGF$^{\text{JSD}}$+rGF$^{\text{JSD}}$ | Base-UC $\gg$ Rating-UC, Rating |
| Rating-UC $\gg$ Base, Rating |
| Base $\gg$ Rating |
| Rating-UC $\gg$ Rating |

(c) 6 pairs with a statistically significant difference

| Measure | Conclusions |
|---------|-------------|
| cGF$^{\text{JSD}}$+rGF$^{\text{JSD}}$, cGF$^{\text{JSD}}$+rGF$^{\text{RNOD}}$, cGF$^{\text{JSD}}$*rGF$^{\text{RNOD}}$, c$\times$rAWRF$^\dagger$ | Base-UC $\gg$ Rating-UC ($p = 0.0006\spadesuit$, 0.0078$\spadesuit$, 0.0024$\spadesuit$, 0.026$\heartsuit$, 0.0426$\dagger$) |
| Base-UC $\gg$ Base, Rating |
| Rating-UC $\gg$ Base, Rating |
| Base $\gg$ Rating ($p = 0.0004\spadesuit$, 0.0014$\heartsuit$) |

"$\gg$" means "statistically significantly better than."

For each comparison, $p$-value $\approx 0$ except where indicated explicitly; for example, in terms of cGF$^{\text{JSD}}$+rGF$^{\text{JSD}}$, Base-UC statistically significantly outperforms Rating-UC with $p = 0.0006$.

the 418 topics and obtained those in which one of the three rGF measures disagreed with the other two; there were 102 such topics. Within this query set, rGF$^{\text{NMD}}$ disagreed with the other two for 83 topics (let us call them Category A), rGF$^{\text{JSD}}$ disagreed with the other two for 16 topics (Category B), and rGF$^{\text{RNOD}}$ disagreed with the other two for the remaining 3 topics (Category C). These results also show that rGF$^{\text{NMD}}$ tends to be the outlier. It is known that RNOD lies between JSD and NMD in terms of how it behaves as a quantification measure [57, 58]. Our results show that the GF measures inherit their properties, as the GF measures are essentially weighted averages of DistrSim scores obtained at each rank (See Equation (16)).

To examine how these discrepancies between the three rGF measures arise, we selected one topic each from Categories A, B, and C with relatively large score discrepancies and examined the results closely. Figure 11 visualises the distributions over the review count groups achieved at rank $k = 10$ by Base and Rating for the three selected topics together with the common gold distribution. Table 11 shows the corresponding DistrSim scores as well as the final rGF scores. Table 11(a) shows that for Topic 416, rGF$^{\text{NMD}}$ disagrees with rGF$^{\text{JSD}}$ and rGF$^{\text{RNOD}}$ precisely because DistrSim$^{\text{NMD}}$ disagrees with DistrSim$^{\text{JSD}}$ and DistrSim$^{\text{RNOD}}$. However, Figure 11(a) shows that this behaviour of DistrSim$^{\text{NMD}}$ is rather counterintuitive: since the gold distribution gives the highest probability to the zero-review group (Group 1), the achieved distribution of Base (red) seems better than that of Rating (blue). On the other hand, Table 11(b) shows that for Topic 1469, while rGF$^{\text{JSD}}$ disagrees with the other two, all three DistrSim functions agree that Rating is better than Base at rank 10. (Figure 11(b) shows the distributions.) That is, this discrepancy at the GF score level is due to the RBP-based weighted averaging step of Equation (16). Finally, Table 11(c) shows that for
Fig. 10. Visualising intersectional group fairness using GF measures for the local search runs.

Table 11. DistrSim and rGF Scores for the Cases Shown in Figure 11

| Topic          | DistrSim<sup>NSD</sup>/rGF<sup>NSD</sup> | DistrSim<sup>NMD</sup>/rGF<sup>NMD</sup> | DistrSim<sup>RNOD</sup>/rGF<sup>RNOD</sup> |
|----------------|----------------------------------------|----------------------------------------|----------------------------------------|
| (a) Topic 416  | 0.801/0.572                            | 0.742/0.566                            | 0.712/0.505                            |
| Base Rating    | 0.730/0.566                            | 0.807/0.576                            | 0.668/0.365                            |
| (b) Topic 1469 | 0.765/0.554                            | 0.708/0.553                            | 0.643/0.480                            |
| Base Rating    | 0.801/0.532                            | 0.742/0.614                            | 0.712/0.512                            |
| (c) Topic 823  | 0.574/0.417                            | 0.458/0.454                            | 0.464/0.380                            |
| Base Rating    | 0.521/0.408                            | 0.492/0.423                            | 0.528/0.414                            |

For each comparison of Base and Rating, the higher (i.e., better) score is indicated in **bold**. For each topic, the outlier DistrSim/rGF score (i.e., one that disagrees with the other two) is **underlined**.

Topic 823, while rGF<sup>RNOD</sup> is the outlier at the GF score level, both DistrSim<sup>NMD</sup> and DistrSim<sup>RNOD</sup> (i.e., those that can handle ordinal groups) prefer Rating over Base. That is, DistrSim<sup>NSD</sup> is the actual outlier at the DistrSim level. From Figure 11(c), it can be observed that, for this topic, the two order-aware measures penalise Base heavily as it emphasises Group 4 (items with over 100 reviews) too much, which is quite intuitive. Based on this close analysis, our recommendation for handling ordinal attribute values is to use multiple similarity functions, pay attention to cases in which they disagree, and, if possible, examine which function seems more appropriate. As we have discussed in Section 1, researchers should also be aware that JSD ignores the ordinal nature of groups and therefore may not be appropriate for some applications.

5 CONCLUSIONS

We presented and validated a simple and versatile framework called GFR (Group Fairness and Relevance), for evaluating ranked lists in terms of group fairness and relevance, in which the groups can be either nominal or ordinal in nature. Our experiments also featured a version of AWRF
Evaluating Ranked Lists in terms of Group Fairness and Relevance

Fig. 11. Review count distributions at rank $k = 10$ achieved by Base and Rating for three example topics, with the common gold distribution.

(Attention-Weighted Rank Fairness), used in the TREC Fair Ranking Track. The fundamental differences between GFR and AWRF (or, more generally, existing approaches to group fair IR evaluation) are (a) GFR represents the expected user experience in which the experience for a user group that abandons the SERP at a particular rank depends on both relevance and the group fairness observed so far; and (b) multiple attribute sets are handled separately, where ordinal quantification measures can be used if an attribute set contains ordinal attribute values.

Our main findings from the three experiments that we detailed are as follows.

- By utilising the Touché 2020 data from CLEF 2020, we demonstrated that our GFR framework is applicable to a hard group membership setting with a binary attribute set in which both group fairness and relevance need to be considered. The GFR measures were generally highly correlated with $D^\#\text{-nDCG}$ (which reduces to the average of intent recall and straight nDCG in hard group membership situations), a diversity measure known to correlate well with human SERP preferences [63]. Furthermore, we showed that $\Delta GF$ scores can quantify the polarity of each run as well as each ranked list in a straightforward manner. In addition, AWRF substantially underperformed $\text{GF}^{\text{JSD}}$ in terms of discriminative power.

- By utilising the diversified web search data from the NTCIR-9 INTENT Japanese subtask, we demonstrated that our GFR framework can handle soft group membership, and that the GFR measures are highly correlated with $D^\#\text{-nDCG}$ (i.e., the average of intent recall and $D\text{-nDCG}$ [62]). Again, AWRF underperformed $\text{GF}^{\text{JSD}}$ in terms of discriminative power, although this problem can be avoided by combining it with nDCG.

- Using real data from a Japanese local search service (without relevance assessments), we showed that our GF measures enable researchers to study intersectional group fairness, even when both nominal and ordinal attribute sets are involved. Our recommendation for handling ordinal attribute values is to use multiple similarity functions (e.g., JSD-based and RNOD-based ones), pay attention to cases in which they disagree, and, if possible, examine which function seems more appropriate.

Our future work includes further investigation of the properties of similarity functions in the context of group fair ranking evaluation and investigating the usefulness and limitations of the GFR framework in the context of the aforementioned NTCIR FairWeb task.

One open question for research in fair IR evaluation is how to demonstrate that an evaluation measure can measure what we really want to measure in practice. In the case of ad hoc IR evaluation measures, this is less difficult (even if it is expensive) because it is possible to investigate how such measures align with user perception, user satisfaction, and so on, by assuming that assessors that we hire can serve as surrogates for real search engine users with an information need (e.g.,

ACM Transactions on Information Systems, Vol. 42, No. 1, Article 11. Publication date: August 2023.
[63]). In contrast, it is unclear how a similar approach could validate a fairness measure, since fairness is for various items and their stakeholders experiencing different levels of exposure and not for a single search engine user. Assessors probably will not be able to speak for the stakeholders. Hence, ultimately, a wide range of stakeholders will have to be consulted in order to validate whether a fairness measure really "makes sense." For example, if a fairness measure says that one SERP is better than another, how many of the stakeholders involved (for a particular topic) will agree?

APPENDIX: PROOF THAT RNOD EQUALS NMD WHEN |C| = 2

The content of this section was first presented in a CIKM 2021 workshop paper [58] and is duplicated with revision here for making this article self-contained. As we have mentioned in Section 3.1, NMD and RNOD are equivalent when |C| = 2 (i.e., there are only two classes). The proof is given below.

Given that |C| = 2, note that \( cp_1 = p_1 \) and \( cp_1^* = p_1^* \), and that \( cp_2 = cp_2^* = 1 \), by definition of cumulative probabilities. Hence, from Equation (26),

\[
NMD(p, p^*) = |cp_1 - cp_1^*| + |cp_2 - cp_2^*| = |p_1 - p_1^*| + 0 = |p_1 - p_1^*|.
\]

(30)

On the other hand, note that when |C| = 2, the following holds:

\[
(p_2 - p_2^*)^2 = (1 - p_1 - 1 + p_2^*)^2 = (p_1 - p_1^*)^2.
\]

(31)

To compute RNOD, the following three cases need to be considered.

Case 1. When \( p_1^* > 0 \) and \( p_2^* > 0 \): from Equations (28) and (31),

\[
OD(p || p^*) = \frac{(DW_1 + DW_2)}{2} = \frac{((p_2 - p_2^*)^2 + (p_1 - p_1^*)^2)}{2} = 2(p_1 - p_1^*)^2/2 = (p_1 - p_1^*)^2.
\]

Hence, from Equation (29),

\[
RNOD(p || p^*) = \sqrt{OD(p || p^*)} = |p_1 - p_1^*|.
\]

(33)

Case 2. When \( p_1^* = 1 \) and \( p_2^* = 0 \): from Equations (28) and (31),

\[
OD(p || p^*) = DW_1 = (p_2 - p_2^*)^2 = (p_1 - p_1^*)^2.
\]

(34)

Therefore, Equation (33) holds for this case as well.

Case 3. When \( p_1^* = 0 \) and \( p_2^* = 1 \): from Equation (28),

\[
OD(p || p^*) = DW_1 = (p_1 - p_1^*)^2
\]

(35)

and Equation (33) holds for this case as well.

Hence, we have proven that \( NMD(p, p^*) = RNOD(p || p^*) \). In addition, by following similar steps as above, it is easy to show that:

\[
RNOD(p^* || p) = \sqrt{OD(p^* || p)} = |p_1 - p_1^*|.
\]

(36)

In summary, \( NMD(p, p^*) = RNOD(p || p^*) = RNOD(p^* || p) \) when |C| = 2. Q.E.D.

REFERENCES

[1] Rakesh Agrawal, Gollapudi Sreenivas, Alan Halverson, and Samuel Leong. 2009. Diversifying search results. In Proceedings of the Second ACM International Conference on Web Search and Data Mining (WSDM’09 Barcelona, Spain). Association for Computing Machinery, 5–14.

[2] Yamen Ajjour, Henning Wachsmuth, Johannes Kiesel, Martin Potthast, Matthias Hagen, and Benno Stein. 2019. Data acquisition for argument search: The args.me corpus. In Advances in Artificial Intelligence (KI’19) (Lecture Notes in Computer Science 11793), Christoph Benzmüller and Heiner Stuckenschmidt (Eds.). Springer, 48–59.
Evaluating Ranked Lists in terms of Group Fairness and Relevance

[3] 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 and Development in Information Retrieval (SIGIR’18 Ann Arbor, MI, USA). Association for Computing Machinery, 625–634.

[4] Vito Walter Anelli, Tommaso Di Noia, Eugenio Di Sciascio, Claudio Fomo, and Azzurra Ragone. 2019. On the discriminative power of hyper-parameters in cross-validation and how to choose them. In Proceedings of the 13th ACM Conference on Recommender Systems (RecSys’19 Copenhagen, Denmark). Association for Computing Machinery, 447–451.

[5] Azin Ashkan and Donald Metzler. 2019. Revisiting online personalized search metrics with the user in mind. In Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR’19 Paris, France). Association for Computing Machinery, 625–634.

[6] Alex Beutel, Jilin Chen, Tulsee Doshi, Hai Qian, Li Wei, Yi Wu, Lukasz Heldt, Zhe Zhao, Lichan Hong, Ed H. Chi, and Cristos Goodrow. 2019. Fairness in recommendation ranking through pairwise comparisons. In Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD’19 Anchorage, AK, USA). Association for Computing Machinery, 2212–2220.

[7] Asia J. Beiga, Fernando Díaz, Michael D. Ekstrand, Sergey Feldman, and Sebastian Kohlmeier. 2021. Overview of the TREC 2020 fair ranking track. In The 29th Text Retrieval Conference (TREC’20) Proceedings. NIST. https://trec.nist.gov/pubs/trec29/papers/OVERVIEW.FR.pdf.

[8] Asia J. Beiga, Fernando Díaz, Michael D. Ekstrand, and Sebastian Kohlmeier. 2020. Overview of the TREC 2019 fair ranking track. In The 28th Text Retrieval Conference (TREC’19) Proceedings. https://trec.nist.gov/pubs/trec28/papers/OVERVIEW.FR.pdf.

[9] Asia J. Beiga, Krishna P. Gummadi, and Gerhard Weikum. 2018. Equity of attention: Amortizing individual fairness in rankings. In The 41st International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR’18 Ann Arbor, MI, USA). Association for Computing Machinery, 405–414.

[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. In Experimental IR Meets Multilinguality, Multimodality, and Interaction (Lecture Notes in Computer Science 12260), Avi Arampatzis, Evangelos Kanoulas, Theodora Tsikriki, Stefanos Vrochidis, Hideo Joho, Christina Lioma, Carsten Eickhoff, Aurélie Névéol, Linda Cappellato, and Nicola Ferro (Eds.). Springer, 384–395.

[11] Alexander Bondarenko, Lukas Gienapp, Maik Fröbe, Meriem Beloucif, Yamen Ajjour, Alexander Panchenko, Chris Biemann, Benno Stein, Henning Wachsmuth, Martin Potthast, and Matthias Hagen. 2021. Overview of touché 2020: Argument retrieval. In Experimental IR Meets Multilinguality, Multimodality, and Interaction (Lecture Notes in Computer Science 12880), K. Selçuk Candan, Bogdan Ionescu, Lorraine Goeuriot, Birger Larsen, Henning Müller, Alexis Joly, Maria Maistro, Florina Piroi, Guglielmo Faggioli, and Nicola Ferro (Eds.). Springer, 450–467.

[12] Joy Buolamwini and Timnit Gebru. 2018. Gender shades: Intersectional accuracy disparities in commercial gender classification. In Proceedings of Machine Learning Research, Sorelle A. Friedler and Christo Wilson (Eds.). Vol. 81. PMLR, 77–91.

[13] L. Elisa Celis, Damian Straszak, and Nisheeth K. Vishnoi. 2017. Ranking with fairness constraints. (2017). http://arxiv.org/abs/1704.06840.

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

[15] Olivier Chapelle, Donald Metzler, Ya Zhang, and Pierre Grinspan. 2009. Expected reciprocal rank for graded relevance. In Proceedings of the 18th ACM Conference on Information and Knowledge Management (CIKM’09 Hong Kong, China). Association for Computing Machinery, 621–630.

[16] 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 and Knowledge Management (CIKM’21 Virtual Event, Queensland, Australia). Association for Computing Machinery, 3363–3367.

[17] Aleksandr Chuklin, Pavel Serdyukov, and Maarten de Rijke. 2013. Click model-based information retrieval metrics. In Proceedings of the 36th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR’13 Dublin, Ireland). Association for Computing Machinery, 493–502.

[18] Charles L. A. Clarke, Nick Craswell, Ian Soboroff, and Azin Ashkan. 2011. A comparative analysis of cascade measures for novelty and diversity. In Proceedings of the 4th ACM International Conference on Web Search and Data Mining (WSDM’11 Hong Kong, China). Association for Computing Machinery, 75–84.

[19] Charles L. A. Clarke, Nick Craswell, and Ellen M. Voorhees. 2013. Overview of the TREC 2012 web track. In The 21st Text Retrieval Conference (TREC 2012) Proceedings. NIST. https://trec.nist.gov/pubs/trec21/papers/WEB12.overview.pdf.

[20] Charles L. A. Clarke, Alexandra Vtyurina, and Mark D. Smucker. 2020. Offline evaluation without gain. In Proceedings of the 2020 ACM SIGIR on International Conference on Theory of Information Retrieval (ICTIR’20 Virtual Event, Norway). Association for Computing Machinery, 185–192.
[21] Fernando Diaz, Bhaskar Mitra, Michael D. Ekstrand, Asia J. Biega, and Ben Carterette. 2020. Evaluating stochastic rankings with expected exposure. In Proceedings of the 29th ACM International Conference on Information and Knowledge Management (CIKM'20 Virtual Event, Ireland). Association for Computing Machinery, 275–284.

[22] Tim Draws, Navin Tintarev, and Ujwal Gadiyaru. 2021. Assessing viewpoint diversity in search results using ranking fairness metrics. ACM SIGKDD Explorations Newsletter 23, 1 (2021), 50–58.

[23] Michael D. Ekstrand, Anubrata Das, Robin Burke, and Fernando Diaz. 2021. Fairness and discrimination in information access systems. (2021). https://arxiv.org/abs/2105.05779.

[24] Michael D. Ekstrand, Graham McDonald, and Amifa Raj. 2022. Overview of the TREC 2021 fair ranking track. In The 30th Text Retrieval Conference (TREC’21) Proceedings. NIST. https://trec.nist.gov/pubs/trec30/papers/Overview-F.pdf.

[25] James R. Foulds, Rashidul Islam, Kamrun Naher Keya, and Shimei Pan. 2019. An intersectional definition of fairness. (2019). http://arxiv.org/abs/1807.08362

[26] 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 and Data Mining (KDD’19 Anchorage, AK, USA). Association for Computing Machinery, 2221–2231.

[27] Gizem Gezici, Aldo Lipani, Yucel Saygin, and Emine Yilmaz. 2021. Evaluation metrics for measuring bias in search engine results. Information Retrieval Journal 24 (2021), 85–113.

[28] Avijit Ghosh, Ritam Dutt, and Christo Wilson. 2021. When fair ranking meets uncertain inference. In Proceedings of the 4th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR’21 Virtual Event, Canada). Association for Computing Machinery, 1033–1043.

[29] Avijit Ghosh, Lea Genuit, and Mary Reagan. 2021. Characterizing intersectional group fairness with worst-case comparisons. CoRR abs/2101.01673 (2021). https://arxiv.org/abs/2101.01673.

[30] Peter B. Golbus, Javed A. Aslam, and Carles L. A. Clarke. 2013. Increasing evaluation sensitivity to diversity. Information Retrieval 16 (2013), 530–555.

[31] Lisa L. Harlow, Stanley A. Mulaik, and James H. Steiger. 2016. What If There Were No Significance Tests? (Classic Edition). Routledge.

[32] Kalervo Järvelin and Jaana Kekäläinen. 2002. Cumulated gain-based evaluation of IR techniques. ACM Transactions on Information Systems 20, 4 (2002), 422–446.

[33] Evangelos Kanoulas and Javed A. Aslam. 2009. Empirical justification of the gain and discount function for nDCG. In Proceedings of the 18th ACM Conference on Information and Knowledge Management (CIKM’09 Hong Kong, China). Association for Computing Machinery, 611–620.

[34] Maurice G. Kendall. 1962. Rank Correlation Methods (3rd Edition). Charles Griffin and Company Limited.

[35] Ömer Kirnap, Fernando Diaz, Asia Biega, Michael Ekstrand, Ben Carterette, and Emine Yilmaz. 2021. Estimation of fair ranking metrics with incomplete judgments. In Proceedings of the Web Conference 2021 (Ljubljana, Slovenia). Association for Computing Machinery, 1065–1075.

[36] Caitlin Kuhlman, Walter Gerych, and Elke Rundensteiner. 2021. Measuring group advantage: A comparative study of fair ranking metrics. In Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society (AIES’21 Virtual Event, USA). Association for Computing Machinery, 674–682.

[37] Caitlin Kuhlman, MaryAnn VanValkenburg, and Elke Rundensteiner. 2019. FARE: Diagnostics for fair ranking using pairwise error metrics. In The World Wide Web Conference (WWW’19 San Francisco, CA, USA). Association for Computing Machinery, 2936–2942.

[38] Juhi Kulshrestha, Motahhare Eslami, Johnnatan Messias, Muhammad Bilal Zafer, Saptarshi Ghosh, Krishna P. Gunmadi, and Karrie Karahalios. 2017. Quantifying search bias: Investigating sources of bias for political searches in social media. In Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing (CSCW’17 Portland, Oregon, USA). Association for Computing Machinery, 417–432.

[39] Teerapong Leelanupab, Guido Zuccon, and Joemon M. Jose. 2012. A comprehensive analysis of parameters settings for novelty-biased cumulative gain. In Proceedings of the 21st ACM International Conference on Information and Knowledge Management (CIKM’12 Maui, Hawaii, USA). Association for Computing Machinery, 1950–1954.

[40] Elizaveta Levina and Peter Bickel. 2001. The earth mover’s distance is the mallows distance: Some insights from statistics. In Proceedings 8th IEEE International Conference on Computer Vision (ICCV’01 Vancouver, BC, Canada). IEEE, 251–256.

[41] Jianhua Lin. 1991. Divergence measures based on the Shannon entropy. IEEE Transactions on Information Theory 37, 1 (1991), 145–151.

[42] Jeffrey D. Long and Norman Cliff. 1997. Confidence intervals for Kendall’s tau. British Journal of Mathematical and Statistical Psychology 50 (1997), 31–41.

[43] Xiaolu Lu, Alistair Moffat, and J. Shane Culpepper. 2016. The effect of pooling and evaluation depth on IR metrics. Information Retrieval Journal 19, 4 (2016), 416–445.
Evaluating Ranked Lists in terms of Group Fairness and Relevance

[44] Alistair Moffat, Peter Bailey, Falk Scholer, and Paul Thomas. 2017. Incorporating user expectations and behavior into the measurement of search effectiveness. *ACM Transactions on Information Systems* 35, 3, Article 24 (2017).

[45] Alistair Moffat and Justin Zobel. 2008. Rank-biased precision for measurement of retrieval effectiveness. *ACM Transactions on Information Systems* 27, 1, Article 2 (2008).

[46] Harikrishna Narasimhan, Andrew Cotter, Maya Gupta, and Serena Wang. 2020. Fairwise fairness for ranking and regression. In *Proceedings of the AAAI Conference on Artificial Intelligence* (New York, USA), Vol. 34, AAAI, 5248–5255.

[47] Evaggelia Pitoura, Panayiotis Tsaparas, Giorgos Flouris, Irini Fundulaki, Fanagiotis Papadakos, Serge Abiteboul, and Gerhard Weikum. 2017. On measuring bias in online information. *ACM SIGMOD Record* 46, 4 (2017), 16–21.

[48] Amifa Raj and Michael D. Ekstrand. 2022. Measuring fairness in ranked results. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'22 Madrid, Spain)*. Association for Computing Machinery, 726–736.

[49] Ronald E. Robertson, Shan Jiang, Kenneth Joseph, Lisa Friedland, David Lazer, and Christo Wilson. 2018. Auditing partisan audience bias within Google search. In *Proceedings of the 33rd International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'10 Geneva, Switzerland)*. Association for Computing Machinery, 603–610.

[50] Tetsuya Sakai. 2006. Evaluating evaluation metrics based on the bootstrap. In *Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR’06 Seattle, Washington, Washington, USA)*. Association for Computing Machinery, 525–532.

[51] Tetsuya Sakai. 2007. Alternatives to bpref. In *Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR’07 Amsterdam, The Netherlands)*. Association for Computing Machinery, 71–78.

[52] Tetsuya Sakai. 2012. Evaluation with informational and navigational intents. In *Proceedings of the 21st International Conference on World Wide Web* (Lyon, France). Association for Computing Machinery, 499–508.

[53] Tetsuya Sakai. 2014. Metrics, statistics, tests. In *Bridging Between Information Retrieval and Databases (Lecture Notes in Computer Science 8173)*, Nicola Ferro (Ed.). Springer, 116–163.

[54] Tetsuya Sakai. 2018. Comparing two binned probability distributions for information access evaluation. In *The 41st International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR’18 Ann Arbor, MI, USA)*. Association for Computing Machinery, 1073–1076.

[55] Tetsuya Sakai. 2018. Laboratory experiments in information retrieval: Sample sizes, effect sizes, and statistical power. Springer.

[56] Tetsuya Sakai. 2020. 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)* (Online). Association for Computational Linguistics, 2759–2769.

[57] Tetsuya Sakai. 2021. A closer look at evaluation measures for ordinal quantification. In *Proceedings of the CIKM 2021 Workshops co-located with 30th ACM International Conference on Information and Knowledge Management (CIKM’21)*. https://ceur-ws.org/Vol-3052/paper21.pdf.

[58] Tetsuya Sakai. 2021. On the instability of diminishing return IR measures. In *Advances in Information Retrieval: 43rd European Conference on IR Research, ECIR 2021 Part I (Lecture Notes in Computer Science 12656)*, Djoerd Hiemstra, Marie-Francine Moens, Josiane Mothe, Raffaele Perego, Martin Potthast, and Fabrizio Sebastiani (Eds.). Springer, 572–586.

[59] Tetsuya Sakai, Jin Young Kim, and Inho Kang. 2022. A versatile framework for evaluating ranked lists in terms of group fairness and relevance. (2022). http://arxiv.org/abs/2204.00280.

[60] Tetsuya Sakai and Stephen E. Robertson. 2008. Modelling a user population for designing information retrieval metrics. In *Proceedings of the 2nd International Workshop on Evaluating Information Access (EVLA’08)*. 30–41. http://research.nii.ac.jp/ncir/workshop/OnlineProceedings07/pdf/EVIA2008/07-EVIA2008-SakaiT.pdf.

[61] Tetsuya Sakai and Ruihua Song. 2011. Evaluating diversified search results using per-intent graded relevance. In *Proceedings of the 34th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR’11 Beijing, China)*. Association for Computing Machinery, 1043–1052.

[62] Tetsuya Sakai and Zhaohao Zeng. 2020. Retrieval evaluation measures that agree with users’ SERP preferences: Traditional, preference-based, and diversity measures. *ACM Transactions on Information Systems* 39, 2, Article 14 (2020).

[63] Piotr Sapiezynski, Wesley Zeng, and Ronald E. Robertson. 2019. Quantifying the impact of user attention on fair group representation in ranked lists. In * Companion Proceedings of The 2019 World WideWeb Conference (San Francisco, USA)*. Association for Computing Machinery, 553–562.

[64] Ashudeep Singh and Thorsten Joachims. 2018. Fairness of exposure in rankings. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD’18 London, United Kingdom)*. Association for Computing Machinery, 2219–2228.
[66] Ruihua Song, Min Zhang, Tetsuya Sakai, Makoto P. Kato, Yiqun Liu, Miho Sugimoto, Qinglei Wang, and Naoki Orii. 2011. Overview of the NTCIR-9 INTENT task. In Proceedings of The 9th NTCIR Workshop Meeting (Tokyo, Japan). National Institute of Informatics, 82–105.

[67] Daniel Valcarce, Alejandro, Bellogín, Javier Parapar, and Pablo Castells. 2020. Assessing ranking metrics in top-N recommendation. Information Retrieval Journal 23 (2020), 411–448.

[68] Michael Werman, Shmuel Peleg, and Azriel Rosenfeld. 1985. A distance metric for multidimensional histograms. Computer Vision, Graphics, and Image Processing 32 (1985), 328–336.

[69] Ke Yang and Julia Stoyanovich. 2017. Measuring fairness in ranked outputs. In Proceedings of the 29th International Conference on Scientific and Statistical Database Management (SSDBM’17: Chicago, IL, USA). Association for Computing Machinery.

[70] Meike Zehlike, Francesco Bonchi, Carlos Castillo, Sara Hajian, Mohamed Megahed, and Ricardo Baeza-Yates. 2017. FA’IR: A fair top-k ranking algorithm. In Proceedings of the 2017 ACM on Conference on Information and Knowledge Management (CIKM’17 Singapore, Singapore). Association for Computing Machinery, 1569–1578.

[71] ChengXiang Zhai, William W. Cohen, and John Lafferty. 2003. Beyond independent relevance: Methods and evaluation metrics for subtopic retrieval. In Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR’03 Toronto, Canada). Association for Computing Machinery, 10–17.

Received 6 April 2022; revised 7 December 2022; accepted 14 March 2023