Discover Computing

Research

A model of the relationship between the variations of effectiveness and fairness in information retrieval

Massimo Melucci[1]

Received: 16 August 2022 / Accepted: 4 April 2024
Published online: 23 April 2024

Abstract
The requirement that, for fair document retrieval, the documents should be ranked in the order to equally expose authors and organizations has been studied for some years. The fair exposure of a ranking, however, undermines the optimality of the Probability Ranking Principle and as a consequence retrieval effectiveness. It is shown how the variations of fairness and effectiveness can be related by a model. To this end, the paper introduces a fairness measure inspired in Gini's index of mutability for non-ordinal variables and relates it to a general enough measure of effectiveness, thus modeling the connection between these two dimensions of Information Retrieval. The paper also introduces the measurement of the statistical significance of the fairness measure. An empirical study completes the paper.

Highlights

- Information Retrieval (IR) systems should provide both effective and fair document rankings.
- Effectiveness and fairness are investigated by co-variations.
- Fairness can be achieved without large loss of effectiveness.

Keywords Probability Ranking Principle · Ranking · Equity · Bias · Gini · Recommendation · Equality · Big data · Classification

1 Introduction

An Information Retrieval (IR) system ranks the retrieved documents by a measure of relevance of information carried by a document to the information need of a user who submitted a query or other expressions to the systems. Usually, relevance is a property of a single document and not of the whole ranking as done in [15, 35], for example.

The documents are products of authors and organizations which are members of groups organized according to different conventions such as economic development and perceived prestige; in contexts other than academic search, some groups are even "protected" from discrimination. In presence of groups, an IR system should not only be effective, it should also be fair. However, there many causes of bias [2], the main one being the belief that bias is only a human fact whereas information systems are immune to it [3]. Similar issues were studied or reported in recommendation [8], social science [1], advertising [38].

✉ Massimo Melucci, massimo.melucci@unipd.it | [1]University of Padua IT, Padua, Italy.

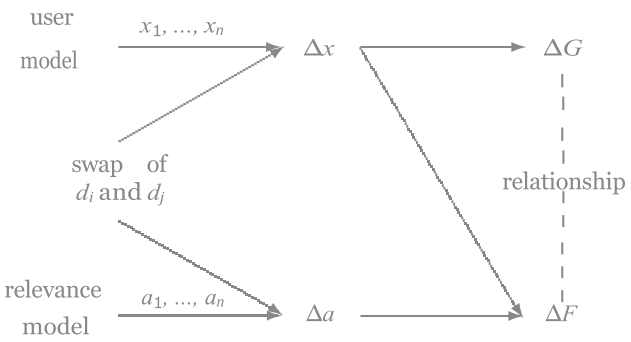


Fig. 1 Suppose a ranking is ranked by decreasing probability of relevance. On the left, a swap between the documents ranked at *i* and *j* is performed where $i < j$. The swap causes the users to access the document which was ranked at *i* before the swap with smaller probability than the probability to access the same document when ranked at *j* after the swap. Similarly, the probability that the information carried by the document ranked at *i* before the swap is relevant is larger than the information carried by the document ranked at *i* after the swap is relevant. The variation of the probability of access is Δx and the variation of the probability of relevance is Δa. These variations determine the variation of the measure of effectiveness, i.e. ΔF and the variation of the measure of fairness, i.e. ΔG, which are then indirectly related through the swap as depicted by the dashed line

The maximization of effectiveness seems to conflict with the maximization of fairness and the amelioration of the trade-off between these two desiderata seems difficult to obtain [36, 41]. The early approaches to reconcile effectiveness and fairness aimed to guarantee a minimal amount of items of protected groups while keeping a minimal level of effectiveness. One approach to check the minimal amount of items of protected groups has been based on statistical hypothesis testing, [42] however, the definition of what fairness is using statistical terms might be difficult; for example, a ranking can be defined as fair if the items of the protected group are represented with a proportion greater than or equal to a certain threshold and it can conversely be defined as biased if the proportion is less than the threshold, however, such a simple approach is affected by group size and the arbitrariness in choosing the threhsold. More complex approaches are necessary for three or more groups and simple counts cannot be related to the common measures of effectiveness [43].

One reason for the difficulty in understanding of how to reconcile fairness and effectiveness in IR might be the lack of a simple, formal description of the relationship between the variations of effectiveness and fairness to predict if and when one desideratum, e.g. maximal effectiveness can at least partially be met without giving up the meeting of the other desideratum, e.g. maximal fairness. It is our opinion that, the understanding of the relationship between the variations of effectiveness and fairness while permuting a ranking requires a model that formally describes the relationship and allows to predict how one quantity varies while changing the other quantity.

The main thrust of this paper is to describe the relationship between effectiveness and fairness. To this aim, the paper provides two functions of ranking: one function measuring effectiveness and one function measuring fairness. There are two models common to the two functions: one model describes how the user accesses a ranking of the documents and the other model describes how the relevance is measured for each document. In technical terms, the user model is a distribution of probability of accessing a document at a certain rank whereas the relevance model can for example be the retrieval function returning the ranking scores. The functions of effectiveness and that of fairness are then related in terms of co-variations. Figure 1 shows an illustration of the model relating the variations of effectiveness and fairness. In summary, the work in this paper provides a treatment of fairness-improving measures as swaps on the original ranked list, a derivation of statistical inference for a fairness measure which is sorely overlooked in the existing literature, and an empirical study of the behavior of the proposed fairness measure, along with its significance test, on the TREC Fair track data.

Some premises are required to find such a relationship and behaviour of effectiveness and fairness, though. One premise is that a notion of fairness should at first be introduced, since the literature on this subject describes at least two notions, i.e. group fairness and individual fairness; see also Section 2.4. In classification, group fairness means that, if the assignment of documents at a certain class were repeated many times, the proportion of documents of each group assigned at the class should approximately be the same. In this paper, group fairness is considered as regards authors and organizations. Individual fairness regards the exposure of single items, i.e. a classifier is fair at the individual level if similar items are equally treated; the problem was addressed in [6] for example. In this paper, individual fairness overlaps if not coincide with optimal ranking by relevance such that a retrieval system places similar documents in nearby ranks.

The documents of the dataset utilized in the empirical study reported in Section 6 are authored by researchers which can be grouped by prestige level indirectly provided by the distribution of the H-indexes, or grouped by economic development level of the country of their organization. A document belongs to more than one group if an author of the document is affiliated with an organization or the organization is located in a country which are respectively different from the organization of another author or from the country of another author's organization. For example, if a document has three authors, each author is affiliated with a distinct organization and an organization is located in a country different from the other organization's countries, then the document belongs to different groups corresponding to the distinct countries.

Furthermore, we in this paper assume that optimal fairness is achieved when each group gets the same total exposure. This is only one fairness target, and it is not obvious that it is the appropriate one, particularly when there are significant differences in group size. There are further notions of fairness such as: equal exposure, equal per-capita exposure, group fairness and individual fairness, exposure proportional to relevance or utility, exposure equivalent to uniform distribution over equivalence class of optimal rankings as reported in some papers such as [6, 7, 15, 18, 19, 34, 39]. The diverse notions of fairness are distinct if not even incompatible in some cases. As a matter of fact, it was shown that, when the notion of fairness comprises different conditions, any risk assessment will at least violate one condition unless the decision problem is trivial [27] or approximation is accepted [21]. However, the incompatibilities between different notions of fairness were found in the context of classification and it is unclear whether and the extent to which they persist in the context of ranking.

Another premise is that, in the context of document ranking, the measure of fairness should depend on the probability of access to a retrieved document, i.e. the user access model and not only on the distribution of the retrieved documents across the groups. The intuition behind this premise is that the groups are exposed to the degree to which the documents can be accessed by the user; the authors and the organizations of a top-ranked document are more exposed than the authors and the organizations of a bottom-ranked document. As a consequence, the measure of fairness reflects a user access model, i.e. a model describing how users access the information provided by the ranked documents. To support the hypothesis that exposure may affect fairness, some analyses suggested that the users tend to adopting the majority viewpoint among the results they examine and as a consequence they tend to changing their attitude [16]. If the information that change the user's attitude comes from the most exposed producers, there might be some bias.

Yet another premise is that a fair ranking can only be obtained by swapping the documents, thus possibly making an effectiveness-optimal ranking sub-optimal. Equivalently, a fair ranking can only be obtained by permutation. The necessity of swapping the documents derives from the assumption that the user scans one retrieved ranking and accesses one document of the list at a time. Although a permuted ranking might be obtained by using other rules such as the one in [11], the permuted ranking is obtained from the optimal ranking by means of a series of swaps. Given that the retrieved documents can be swapped to balance effectiveness and fairness, the measure of fairness is related to that of effectiveness by means of the swaps – when two documents are swapped and therefore the ranks thereof change, the measures of fairness and effectiveness change as well. We explored this relationship both theoretically and empirically and found that fairness and effectiveness can coexist. It is fair to say that, it is possible that the first swap done in the ranking is not the optimal one and it is impossible to know at the moment whether scanning the ranking in a different way would yield different results.

Finally, fairness and its counterpart bias sometimes have been defined with respect to unobservable properties such as recidivism, prestige or wealth which must in turn be inferred from measurements of observable properties [14, 22, 26]. The choice of the observable properties may affect the way fairness is measured and obtained; for example, the inference of prestige depends on indexes of bibliometrics. The arbitrariness of the choice of the observable properties relies on the available experimental data and the investigation of the relationship between effectiveness and fairness should also regard the general quest for a reliable evaluation; some aspects were addressed in machine learning [20] and natural language processing [33].

2 Background

2.1 Ranking by probability of relevance

An IR system ranks retrieved documents by the probability of relevance estimated on the basis of the query content descriptors. If $d_1, ..., d_n$ are the n documents and the $a_i = \Pr(A|\{d_i\})$'s are the relevance model, the documents are ordered by probability of relevance as follows:

$$a_1 \geq a_2 \geq \cdots \geq a_n \,.$$

Not only does the probability of relevance has to be estimated, but the probability of access to a document should also be accurately estimated. To this regard, a *user access model* is defined to describe how the user accesses a ranking. A user access model assumes that the user scans the ranking from the top-ranked document and either moves to the next ranked document or opens the document to assess the relevance to the information need. Once the current document has been opened the scan of the ranking ends. If x_i is the probability that the the user accesses the document at rank i then $0 \leq x_i \leq 1$ and $x_1 + \cdots + x_n = 1$.

Retrieval effectiveness provided by a ranking is measured by the precision expected from accessing the ranking. If the probability of relevance of a document is available, the expected precision of the ranking is defined as

$$F(x) = \sum_{i=1}^{n} a_i x_i \qquad x \in \mathbb{R}^n \,. \tag{1}$$

In particular, when the relevance assessments made for a query are available in terms of indicator function, the measure of effectiveness can exactly be calculated as $\sum_{i=1}^{n} I(d_i \in A)x_i$ of which F is the expected value, i.e.

$$F(x) = E\left(\sum_{i=1}^{n} a_i x_i\right) = \sum_{i=1}^{n} E(I(d_i \in A)x_i = \sum_{i=1}^{n} a_i x_i \,,$$

since $a_i = E(I(d_i \in A))$.

The expected precision is maximized when the documents are ordered by probability of relevance, i.e. when the a's and the x's are perfectly correlated in order; this criterion is also known as the Probability Ranking Principle (PRP). The principle assumes that the optimality of a ranking depends on the quality of estimation of the probability parameters [35].

Expression (1) is a general form of some effectiveness measures described in the pertinent literature. Expected Reciprocal Rank (ERR) introduced in [13] F is an instance of F when $x_i = 1/i$ and the probability of relevance of the i-th document is estimated by the probability that the document is considered as relevant after the previous $i-1$ documents have not. Rank-Biased Precision (RBP) is an instance of F when $x_i = \frac{u^{i-1}(1-u)}{1-u^n}$ and $0 < u < 1$ as described in [30]; if the user skips each document with probability u until the i-th document, which is accessed with probability $1-u$, we have that $x_i \propto u^{i-1}(1-u)$.

2.2 Fairness and the PRP

The groups should be treated equally when the group individuals have to be assigned to a class, rank, or the like, thus leading to equalized odds, statistical parity or group fairness [25]. In classification, a group has equalized odds if the probability that a document of the group can be placed in a class is independent of the membership to the group. In ranking, if the assignment of documents at a certain rank were repeated many times, the probability of access to the documents of each group assigned at the rank should approximately be the same.

The PRP ignores whether the documents belong to the same group. If the principle had to take fairness into account the effectiveness of the ranking would be suboptimal [36]. On the other hand, when the documents are optimally ranked, yet distributed across two or more groups, a group might be unfairly treated because the documents thereof are considered less relevant than the documents of other groups. In general, equalized odds, group fairness and other fairness conditions cannot apply to the ranking functions that meet the PRP.

2.3 Fairness and diversity of rankings

Although some methods to obtain fairness may be used to diversify rankings, fairness differs from diversity [41]. Wang *et al* showed that fairness and diversity are independent goals because the search for fairness may end with little diversified rankings and the search for diversified rankings may end with unfair exposure of groups. The imperfect correspondence between fairness and diversity can be explained by the "diverse nature of diversity" which may refer to the content of a document with respect to its ability to meet the user's information needs, to the content of another document, or to the user's intent underlying a query [43]. Perhaps the most striking difference between

fairness and diversity is the presence, which might be latent, of authors, organizations or in general stakeholders behind a document, all of whom carry some legitimate interest [37], no matter whether all of them match the same aspect of an information need.

2.4 Group vs. individual fairness

According to individual fairness similar individuals of a population should be treated equally. Dwork *et al* addressed [17] the problem of reconciling individual fairness and classification effectiveness using linear programming. The solution consists of a distribution of probability of class membership for each individual, thus allowing to decide whether to assign the individual to a class. The solution maximizes group fairness under the constraint of individual fairness. Dwork *et al* implemented individual fairness as Lipschitz continuity according to which the similarity between the probability distribution of one individual and the probability distribution of another individual should be larger than the similarity between the individuals.

Individual fairness and group fairness might not match because a group may include a subgroup of which the individuals are dissimilar from the individuals of another subgroup. Nevertheless, individual fairness and group fairness are somehow related because the individuals of a group can be viewed as equivalent each other and totally different from the individuals of other groups. Moreover, the degree of similarity can be thresholded to establish which individual belongs to a group. Despite the differences and the relationships, both notions require the knowledge of a similarity function in case of individual fairness and of a group assignment, which is actually a function too, in case of group fairness.

2.5 Variability and mutability

In Statistics, the indexes are classified as either qualitative or quantitative depending on the type of the property observed in individuals such as subjects or documents. If the property observed in individuals is qualitative, it can only be used as group label without the possibility of calculating arithmetic indexes such as means and variances; an example of qualitative property is gender of which no average gender can be calculated. If the property observed in the individuals is instead quantitative, it can be summarized by arithmetic indexes and cannot be treated as group label; an example of quantitative property is the wavelength of electromagnetic radiation. Some properties might be both quantitative and qualitative depending on the purposes of investigation; for example, the year of birth is a number of which minimum and maximum can be calculated yet it can be used as group label to gather all the people born in, say, the 1970s.

When speaking about a qualitative property, the variation of the property is called mutability; when speaking about a quantitative property, the variation is called variability. An index of variability would for example be the standard deviation from the mean of the labels:

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

where $z_k \in \mathbb{R}$ is a numeric label observed in an individual. Clearly, the index of variability depends on the numeric group labels. The dependency of the index of variability on the numeric group labels other than on the frequency distribution follows as a result of the aim of measuring the variation of the labels rather than the variation of the frequency distribution.

The variation of a qualitative property across the individuals must be measured by indexes which take qualitative labels as input and yield an index as number. It follows that, the frequencies of labels and the distribution thereof are the only data that be utilized to compute the indexes of variation.

At first sight, an index of variability is preferable to an index of mutability because the former may also count on numeric labels. However, the dependency on both labels and frequencies makes the index of variability liable to misleading conclusions; for example, if $m = 2$ the index will be close to zero if z_1 is close to z_2 for all f's or if $f_1 = 0$ or $f_1 = n$ for all z's, thus suggesting any variation in any case and making the conclusions regarding fairness dependent on the numeric labels chosen to represent group membership. Instead, the index of mutability only depends on the frequency distribution and is independent of the group labels, thus providing an unequivocal indication of variation.

2.6 Gini's index of mutability

Gini proposed both an index of mutability and an index of variability. The indexes were introduced by Gini in 1912 [23] and their story was told in [12]. Before then, the contribution of the Italian statistician was described in [10]. Gini's index of mutability was mainly used in Economics, Statistics and Social Science. In its original contexts, the index was defined as follows:

$$G = 1 - \sum_{k=1}^{m} y_k^2 \qquad y_k = \frac{f_k}{n} \qquad \sum_{k=1}^{m} f_k = n \tag{2}$$

where m is the number of groups, n is the number of items such as persons or countries, and f_k is the number of items included in group k. The groups might refer to economic development level, e.g. advanced or developing, or to the academic H-index, e.g. high or low. The derivation of G follows.

Suppose 1 is recorded if an individual belongs to group k, 0 otherwise. If an individual is assigned to group k, the squared deviation from the relative frequency of the group k is $\left(1 - y_k\right)^2$. At the same time, the deviation from the relative frequency of a group ℓ other than k is $(0 - y_\ell)^2 = y_\ell^2$. The sum of the deviations yielded by the observation of an individual in group k is then

$$1 - 2y_k + \sum_{\ell=1}^{m} y_\ell^2 .$$

The latter expression is one outcome of a random variable out of the m possible outcomes. An outcome of that random variable has probability y_k which is the probability of an individual in group k. The expected value of that random variable is

$$\sum_{k=1}^{m} \left(1 - 2y_k + \sum_{\ell=1}^{m} y_\ell^2\right) y_k$$

that is

$$G(x) = 1 - \sum_{k=1}^{m} y_k^2 \tag{3}$$

In the event that all the n individuals belong to the group k, i.e. $y_k = 1$and $y_\ell = 0$ for all $k \neq \ell$, one can speak of maximal homogeneity; it can be easily checked that the minimum of G is zero in case of maximal homogeneity. The theoretical maximum of G is $1 - 1/m$ in case of maximal heterogeneity, i.e. when $y_k = 1/m, k = 1, ..., m$. Therefore, G can range between 0 and 1 if divided by $1 - 1/m$.

2.7 Fairness and independence of relevance

A significant assumption that sparked much research over the decades is the *stochastic* independence of the relevance of one document and the relevance of another document. The issues of stochastic independence of relevance are rather complex, since they require the definition of the event space, the random variables and the probability function. These issues were studied since the 1960s, for example in [24], in the 1970s in [35] and further developed [9].

The definition of stochastic independence would impact on the definition of a fairness measure if the latter were dependent on relevance assessments. In the event of the index utilized in this paper, the issues are made easier because the index is independent of the relevance assessments and it refers to the groups rather than to the documents. The probability of relevance is "encapsulated" in the measure of effectiveness and the model presented in this paper regards the co-variations between effectiveness and fairness caused by ranking permutations. The investigation of the relationship between the variations of effectiveness and fairness for *all* the possible measures of effectiveness is out of the scope of this paper; we content ourselves with F, G and the idea of the PRP.

3 The generalization of an index of mutability for measuring fairness

Group fairness requires an index of mutability because fairness is a qualitative phenomenon. As a matter of fact, the group membership observed in the individuals is qualitative, and the group labels are used to tag the individuals. Gini's index of mutability is a function of the distribution of proportions of individuals in each class – it is not a function of the group labels, since these labels are qualitative. Although the group labels are not numeric, the index of mutability can provide a numeric measure of fairness thanks to the availability of the distribution of proportions.

The index of mutability is also interpretable because it can be viewed through the lens of the deviations from the expected proportions caused by the placing of an individual to a group as described in Section 2.6. The interpretability of the index of mutability is not harmed by the absence of numeric group labels, on the contrary, the index of mutability takes advantage of this absence because it is independent of the choice of the group labels. Note that the construct of exposure defined in [15] or attention in [6] is a qualitative index since both exposure and attention measures the proportion of exposure and attention, respectively, of items and subjects.

Suppose a document is assigned to one and only one group; this assumption will be removed later. Whenever the user accesses a document of a ranking, the group of the document is implicitly observed. Let $n > 1$ be the number of observed documents, $m > 1$ be the number of groups, and

$$y_k = \sum_{d_i \in C_k} x_i$$

where C_k be the subset of documents assigned to the group k and $y_k \in \mathbb{R}$ is the probability that the group k is implicitly observed; "implicitly" means that the group is observed when a document authored by authors and organizations of the group is accessed. The succession $y_1, ..., y_m$ is a probability distribution, since $0 \leq y_k \leq 1$ for all $k = 1, ..., m$ and $\sum_{k=1}^{m} y_k = 1$. Note that, the y's are functions of $x = (x_1, ..., x_n)$, therefore

$$y_k \equiv y_k(x) \qquad k = 1, ..., m\,.$$

The original Gini's index is only based on the relative frequencies of observation of group k out of n observations. Relative frequencies are measures of probability, but a user access model x might not necessarily be relative frequencies. Therefore, the use of Gini's index in this paper replaces the relative frequencies with user access models, i.e. probabilities.

Whether G is normalized or not, it can be used as a measure of fairness, since the index estimates the extent to which a user accesses the information provided by different groups. The larger the index, the more varied the range of groups from which information is provided. The case of maximal homogeneity corresponds to minimal fairness and maximal heterogeneity corresponds to maximal fairness.

The assumption that a document can only refer to one group, i.e. all the authors and organization belong to only one group may be unrealistic in some contexts. This assumption is particularly strong in the academic context where the authors of an academic research document may in general be affiliated to different organizations which may be established in different countries. The more the document is authored by authors of a group, the larger the degree to which the document belongs to the group. Consider n distributions $b_1, ..., b_n$ such that $0 \leq b_{i,k} \leq 1$ for all i and k and $\sum_{k=1}^{m} b_{i,k} = 1$ for all i. Specifically,

$$b_{i,k} = \Pr(\text{ an author belongs to group } k \,|\, \text{ a document at rank } i)\,.$$

Consider the probability

$$x_i b_{i,k} \qquad i = 1, ..., n$$

of the event that a user accesses the i-th document with probability x_i and that the document is authored by authors of group k with probability $b_{i,k}$. Therefore, the probability that a user accesses a document authored by authors of group k becomes

$$y_k = \sum_{i=1}^{n} x_i b_{i,k}$$

where

$$\sum_{k=1}^{m} y_k = \sum_{i=1}^{n} x_i \sum_{k=1}^{m} b_{i,k} = 1 .$$

As the y's are functions of $x = (x_1, ..., x_n)$, we have that G is also a function of the x's, i.e.

$$G \equiv G(x) \qquad G(x) = 1 - \sum_{k=1}^{m} \left(\sum_{i=1}^{n} x_i b_{i,k} \right)^2 . \tag{4}$$

Do note that the theoretical maximum of G can hardly be reached because of the user access model which may make the groups of the top-ranked documents much more likely observable than the groups of the other documents; for example, if $x_1 > 1/2$ there will be a group k such that $y_k > 1/2$; as $m > 1$, no group can be observed with probability $1/m$.

The index defined by (4) can also be viewed as a measure of exposure as the index proposed in [15] is. The difference between Diaz et al.'s EE-D and G is conceptual because the former is a sum of squared document exposures and the latter is a sum of squared group exposures; the same difference happens between Biega et al.'s equity of attention [6] and the index of mutability. Moreover, (4) is independent of any relevance assessments and, therefore, only reflects mutability and in this way fairness, thus leaving the measurement of effectiveness to other indexes, such as (1). The independence of (4) on relevance assessments makes the study of the relationship between the variations of effectiveness and fairness possible, since both fairness and effectiveness are viewed as dependent variables as explained in revision 4.

4 The relationship between the variations of effectiveness and fairness

The naïve method to maximize fairness would be the permutation of the ranking to uniformly distribute the probability of access to groups, thus obtaining the maximal heterogeneity yet loosing the optimality of the ranking by probability of relevance. But, the distributions of click-through data are often biased toward the very few top-ranked documents. Consequently, the problem turns the maximization of fairness under the constraint on minimum loss of effectiveness, i.e. how to adjust the optimal ranking to increase fairness enough without significantly decreasing effectiveness.

The basic idea of the method explained in this section consists of starting from an optimal ranking according to the PRP, swapping the items until fairness can improve above a threshold and effectiveness can only decrease below a threshold, thus obtaining a fairer yet less effective ranking than the optimal ranking. As item swapping is repeated until a reasonable balance of fairness and effectiveness is reached, Gini's index of mutability is computed at step t, thus having[1]

$$G^{(t)} = 1 - \sum_{k=1}^{m} {y_k^{(t)}}^2 \tag{5}$$

where

$$y_k^{(t)} = \sum_{d_i^{(t)} \in C_k} x_i b_{i,k}^{(t)}$$

is the probability of accessing a document of group k at step t, $d_i^{(t)}$ is the document ranked at i at step t and $b_{i,k}^{(t)}$ is the probability that $d_i^{(t)}$ belongs to group k. Suppose $d_i^{(t)}$ is replaced by $d_j^{(t)}$. After swapping two documents, the index will change if the swapped documents belong to different groups. The search for an optimal yet fair ranking is thus based on the following variation

$$\Delta G^{(t)} = G^{(t+1)} - G^{(t)} .$$

Consider the swap at step t between $d_i^{(t)}$ and $d_j^{(t)}$. As the swap occurs between ranks i and j, the document ranked at i at step $t+1$was the document ranked at j at step t and the document ranked at j at step $t+1$was the document ranked at i at step t, i.e.

[1] To make notation easy to follow (x) has been removed from $G(x)$ and $y(x)$.

$$d_i^{(t+1)} = d_j^{(t)} \qquad d_j^{(t+1)} = d_i^{(t)} .$$

and

$$b_{i,k}^{(t+1)} = b_{j,k}^{(t)} \qquad b_{j,k}^{(t+1)} = b_{i,k}^{(t)} \qquad b_{h,k}^{(t+1)} = b_{h,k}^{(t)} \qquad h \neq i \wedge h \neq j ,$$

thus obtaining that

$$y_k^{(t+1)} = \sum_{h \neq i \wedge h \neq j} x_h b_{h,k}^{(t)} + x_i b_{j,k}^{(t)} + x_j b_{i,k}^{(t)} .$$

After some passages, the following expression can be obtained:[2]

$$y_k^{(t)^2} - y_k^{(t+1)^2} = \Delta x_{i,j} \Delta b_{i,j,k}^{(t)} \left(2y_k^{(t)} - \Delta x_{i,j} \Delta b_{i,j,k}^{(t)} \right)$$

where

$$\Delta x_{i,j} = x_i - x_j \qquad \Delta b_{i,j,k}^{(t)} = b_{i,k}^{(t)} - b_{j,k}^{(t)} .$$

After summing over k, the variation of the fairness index becomes

$$\Delta G^{(t)} = \Delta x_{i,j} \sum_{k=1}^{m} \Delta b_{i,j,k}^{(t)} \left(2y_k^{(t)} - \Delta x_{i,j} \Delta b_{i,j,k}^{(t)} \right) .$$

To further describe the relationship between fairness and effectiveness, consider again the special case that all the authors of each document belong to one group, i.e., $b_{i,k}^{(t)} = 1$ for a certain group k and $b_{i,\ell}^{(t)} = 0$ for all $\ell \neq k$ and for all t; when all the authors of a document belong to one group the document only belongs to C_k. In particular, $\Delta b_{i,j,k}^{(t)} = 0$ when the swapped documents either belong to the same group k or both documents do not belong to the group k. Otherwise, either $\Delta b_{i,j,k}^{(t)} = 1$ or $\Delta b_{i,j,k}^{(t)} = -1$. In the former case, the group k occurs at rank i and does not at rank j at step t; therefore, the group does not occur at rank i and does at rank j at step $t+1$, i.e. the group has been "downgraded" because the probability of accessing a document at rank j is lower than the probability of accessing a document at rank i ($i < j$). The group would be "upgraded" if $\Delta b_{i,j,k}^{(t)} = -1$. In summary, in the event $i < j$,

$$\Delta b_{i,j,k}^{(t)} = \begin{cases} -1 & C_k \text{ is moved to rank } i \text{ from rank } j \\ 0 & C_k \text{ occurs at both ranks} \\ +1 & C_k \text{ is moved to rank } j \text{ from rank } i \end{cases} .$$

Suppose the documents ranked at i and j belongs to two distinct groups k, ℓ. The variation of G after swapping the documents at ranks i, j becomes:

$$\Delta G^{(t)} = 2\Delta x_{i,j} \left(\Delta y_{i,j}^{(t)} - \Delta x_{i,j} \right) \tag{6}$$

where $\Delta x_{i,j} = x_i - x_j$ and $\Delta y_{i,j}^{(t)} = y_k^{(t)} - y_\ell^{(t)}$. Expression (6) is linked to the variation of expected precision of the ranking before the swapping and the expected precision after the swapping of the documents ranked at i, j, that is

[2] Note that

$$y_k^{(t)^2} - y_k^{(t+1)^2} = \left(y_k^{(t)} - y_k^{(t+1)} \right) \left(y_k^{(t)} + y_k^{(t+1)} \right)$$

and

$$y_k^{(t+1)} = y_k^{(t)} - \Delta x_{i,j} \Delta y_{i,j,k}^{(t)} .$$

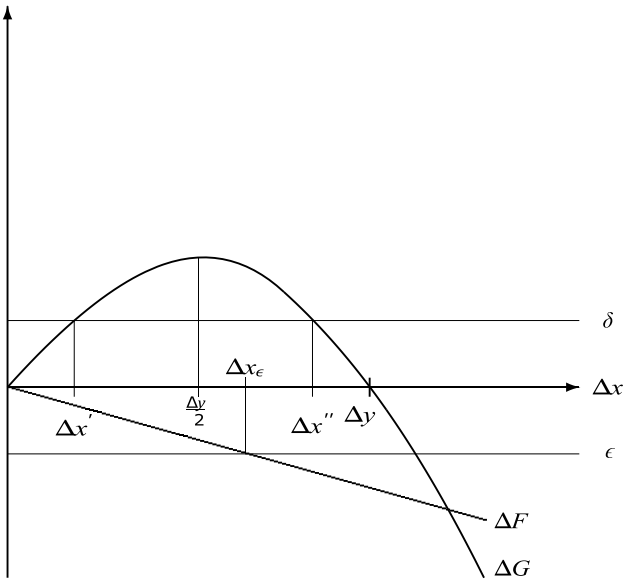


Fig. 2 $\Delta G : \mathbb{R} \to \mathbb{R}$ and $\Delta F : \mathbb{R} \to \mathbb{R}$ are functions of Δx. As ΔG is a second-degree, continuous function on a closed and bounded set of Δx, its extreme value is inside the domain at $\Delta x = \Delta y/2$. ΔF decreases with Δx because the loss of effectiveness becomes large when the difference of probability of relevance between the *i*-th document and the *j*-th document becomes larger

$$\Delta F_{i,j}^{(t)} = -\Delta a_{i,j}^{(t)} \Delta x_{i,j} \qquad \Delta a_{i,j}^{(t)} = \Pr(A|\{d_i^{(t)}\}) - \Pr(A|\{d_j^{(t)}\}) \,. \tag{7}$$

Equation (7) says that effectiveness improves only if the swapped documents are not ordered by probability of relevance. In the rest of this section, the swap between the documents at *i* and *j* at step *t* is assumed and the notation is simplified by removing the subscripts although Δx, Δa and Δy are still functions of *i*, *j*, *t*.

Figure 2 depicts ΔG and ΔF for a certain Δy and as a function of Δx. The figure is built under the assumption that the documents are optimally ordered by probability of relevance, i.e. $\Delta a > 0$. If the ranking is optimal with respect to effectiveness, a negative ΔF is unavoidable and becomes higher as Δx increases until 1. In contrast, ΔG increases until $\Delta x = \Delta y/2$ where it reaches its maximum before decreasing until becoming null at $\Delta x = \Delta y$ and then negative until $\Delta x = 1$.

To obtain a ranking that is fair and effective "enough" at the same time, a threshold of minimum increase of fairness and a threshold of maximum decrease of effectiveness at every swap are required by the method. A swap of d_i and d_j should be performed only if the resulting Δx yields a decrease of F up to a given threshold $\epsilon < 0$ and an increase of G not smaller than a given threshold $\delta \geq 0$.

Figure 2 says that the condition to swap two documents can be written as

$$\Delta x' \leq \Delta x \leq \min\left\{\Delta x_\epsilon, \Delta x''\right\} \qquad \Delta x' \leq \Delta x_\epsilon \qquad \Delta x' \leq \Delta x'' \tag{8}$$

where

$$\begin{aligned} \Delta x' &= \frac{1}{2}\left(\Delta y - \sqrt{\Delta y^2 - 2\delta}\right) \\ \Delta x'' &= \frac{1}{2}\left(\Delta y + \sqrt{\Delta y^2 - 2\delta}\right) \\ \Delta F(\Delta x_\epsilon) &= \epsilon \,. \end{aligned}$$

The search for the "best" Δx is in $O(n^2)$ because every pair of ranks has to be evaluated. Depending on the user access model, the pair of ranks of the "best" Δx might be determined in an algebraic way, thus putting the search in $O(n)$ because one rank can be calculated provided the other rank of the pair.

To analyze the impact of the user access model on the chance of improve fairness, consider the following user access models as an example:

- Exponential user access model where

$$x_i = \frac{1}{i^r H(n,r)} \qquad H(n,r) = \sum_{j=1}^{n} \frac{1}{j^r} \quad n \in \mathbb{N}, n > 1, r \in \mathbb{N}, r > 0\,. \tag{9}$$

The access to the top-ranked documents becomes as more likely as r increases. When r is large this model describes a search tasks like homepage finding, *vice versa*, the search becomes more exhaustive as r decreases. This model was used with $r = 1$in [13].

- Geometric user access model where

$$x_i = \frac{u^{i-1}(1-u)}{1-u^n} \qquad n \in \mathbb{N}, n > 1, u \in \mathbb{R}, 0 < u < 1\,. \tag{10}$$

When u increases, the probability of accessing the top-ranked documents decreases in favor of the bottom-ranked documents. When u is large, this access model describes users willing to explore the whole list and then doing an exhaustive search. This model was used in in [30].

In case of the exponential user access model (9)

$$\Delta x_{i,j} = \frac{i^{-r} - j^{-r}}{H(n,r)}\,.$$

When r increases, Δx approaches 1 if $i = 1$, thus making inequality of (8) more difficult to maintain because the right-hand side is usually less than 1. On the other hand, Δx approaches 0 if $i > 0$, thus similarly making inequality of (8) more difficult to maintain because the left-hand side is usually greater than 0. Therefore, according to the exponential user access model, the swap between two documents occurs with an excessive loss of effectiveness or an insufficient gain in fairness if users tend to access the top-ranked documents and do therefore not browse the ranking.

In case of the geometric user access model (10)

$$\Delta x_{i,j} = \frac{(1-u)\left(u^i - u^j\right)}{u(1-u^n)}\,.$$

When u tends to 0, Δx approaches 0 for all $i < j$, thus making inequality of (8) more difficult to maintain because the left-hand side is usually greater than 0. Therefore, according to the geometric user access model, the swap between two documents occurs with an insufficient gain in fairness if users tend to access the top-ranked documents and do therefore not browse the ranking.

5 Statistical significance of fairness

In the event that G ranges around a given high value, it is difficult to say whether the index value is high and thus state that a ranking is fair. One way to answer the question whether a ranking is fair is to provide a probability that the index value can be obtained from a ranking under the hypothesis that the system has fairly produced the ranking; the ranking can be viewed as fair to a certain degree if the probability of obtaining an index value larger than the observed value is small.

Another case in which one may wonder whether a ranking is fair occurs when one algorithm utilized to retrieve and rank items from one database has to be compared with another algorithm utilized to retrieve and rank items from the same database or with the same algorithm utilized to retrieve and rank items from another database. In such a case, the probability that the two measures of fairness significantly differ can help understand whether the algorithms or the databases show different degrees of fairness. Therefore, one algorithm can be said fairer than another algorithm if the probability of obtaining a difference larger than the difference between the observed values is small under the hypothesis that the algorithms are fair to the same degree.

For the aforementioned reasons, the distribution of probability of G is necessary. The calculation of the probability of observing the values of G requires working out some technical aspects, though. First of all, G is a random variable obtained as transformation of the y's. As G is a transformation of $y_1, ..., y_m$, the calculation of the probability of a certain index value of G requires the definition of a probability distribution for the y's. However, the y's are real numbers and G is a continuous function, although a finite set of y's is observed. Because of the continuity of G, the costly computation of

integration operators is required to calculate the probability distribution; for example, one should resort to complicated mixtures of sums and integrations if the y's would follow a Dirichlet random variable.

Instead of using complicated mixtures of sums and integrations, relative frequencies can be utilized as an approximation of the y's, thus making the calculation of the probability distribution thereof feasible. To this end, the real number ny_k should be expressed as the sum of the integer part n_k and of the decimal part D_k, i.e. $ny_k = n_k + D_k$.[3] After replacing the nq's with the aforementioned sum of the integer part and of the decimal part, the following expression can be obtained:

$$G(y_1, ..., y_m) = G\left(\frac{n_1}{n}, ..., \frac{n_m}{n}\right) - \sum_{k=1}^{m} \frac{D_k}{n}\left(2\frac{n_k}{n} + m\frac{D_k}{n}\right)$$

where the n_k's are obtained by rounding ny_k to the closest integer; for example, $10.499 = 10 + 0.499$ and $1.501 = 2 - 0.499$. If n_k is obtained by rounding ny_k to the closest integer for all $k = 1, ..., m$, we have that each D_k randomly ranges in $[-1/2, 1/2)$. In particular, each D_k can be viewed as a uniform random variable without a prior knowledge about the process generating the y_k's; if some prior knowledge were available, the decimal part of ny_k could follow a Beta random variable of which the uniform variable is a special case. As each D_k is a uniform random variable, one has that $E(D_k) = 0$ and its variance is $1/12$ when the computation of G is repeated many times. It follows that,

$$E\left(\frac{D_k}{n}\left(2\frac{n_k}{n} + m\frac{D_k}{n}\right)\right) = \frac{1}{12n^2}$$

with variance $n^{-2}(1 + m^2/60)/3$. Since the standard deviation is in $O(m/n)$ and $m < n$,

$$G\left(\frac{n_1}{n}, ..., \frac{n_m}{n}\right) + \frac{m}{12n^2}$$

can be a unbiased estimator of $G(y_1, ..., y_m)$ even if n is not large.

The computation of the probability distribution of G also requires the mapping of index values to the events of which the probability can be computed, i.e. the value returned by G has to be mapped to all the frequencies $n_1, ..., n_m$ such that the original index yields z. However, G is non-injective; for example, $G(n_1/n, (n - n_1)/n) = G((n - n_1)/n, n_1/n)$. From the lack of injection it follows that, G^{-1} maps an index value to a subset, i.e.

$$G^{-1}(z) = \left\{ (n_1, ..., n_m) \in \mathbb{N}^m : G\left(\frac{n_1}{n}, ..., \frac{n_m}{n}\right) = z \right\}$$

where $z = G(y_1, ..., y_m)$. Therefore, the probability function of G becomes

$$P\left(G\left(\frac{n_1}{n}, ..., \frac{n_m}{n}\right) = z\right) = P\left(N \in G^{-1}(z)\right) \tag{11}$$

where N is a multinomial random variable with parameters n and $\theta_1, ..., \theta_m$.

Moreover, G is a non-surjective function, since there is an infinity of index values which cannot be mapped from any series $n_1, ..., n_m$. It follows that, the solution set of $G\left(\frac{n_1}{n}, ..., \frac{n_m}{n}\right) = z$ might include non-natural numbers; for example, when $m = 2$ and $z \leq 1/2$, the solutions $u = n(1 - \sqrt{1 - 2z})/2$ and $v = n(1 + \sqrt{1 - 2z})/2$ are not integers when $z = 1/3$. The solutions should therefore be rounded to the closest integer $n_1 = [u]$ and to $n - n_1$, thus obtaining the following approximation of $\Pr(G = z)$:

$$\binom{n}{n_1}\left(\theta^{n_1}(1-\theta)^{n-n_1} + \theta^{n-n_1}(1-\theta)^{n_1}\right).$$

When m grows the complexity of the solutions significantly grows and an approach alternative to the analytical enumeration of the solutions is necessary. A recursive method can be defined to overcome the complexity of finding the series of frequencies in the set $G^{-1}(z)$. To this end, note that the right-hand side of Equation (11) can be rewritten as:

[3] When using the decimal notation, a real number can be written as $n + D = n.\alpha_1\alpha_2...$ where n is the integer part and $D = 0.\alpha_1\alpha_2....$

Fig. 3 The algorithm for computing (11)

gp$(\theta_1, ..., \theta_m, n, m, z)$
$w \leftarrow 0$
if $m > 2$ **then**
  for $h = 0, 1, ..., n$ **do**
    $\eta_k \leftarrow \theta_k/(1-\theta_1)\ \forall k = 2, ..., m$
    $w \leftarrow w + \Pr_{\theta_1}(N_1 = h)$ **gp**$(\eta_2, ..., \eta_m, n-h, m-1, 1-(n/(n-h))^2(1-z))$
  end for
else
  if $z \leq 1/2$ **then**
    $h \leftarrow [n(1-\sqrt{1-2z})/2]$
    $w \leftarrow \Pr_{\theta_1}(N_1 = h) + \Pr_{\theta_1}(N_1 = n-h)$
  end if
end if
return w

$$\sum_{n_1=0}^{n} \Pr(N_1 = n_1|\theta_1)\ \Pr\left(G\left(\frac{n_2}{n}, ..., \frac{n_m}{n}\right) = 1-(1-z)\left(\frac{n}{n-n_1}\right)^2\right)$$

where N_1 is a binomial random variable with parameters n and θ_1. The base step of the recursion is $m = 2$. Figure 3 is the algorithm for computing (11).

The probability that $G = z$ can only provide a measure of the chance of observing z which is in turn a measure of the fairness of the ranking. Nothing can be said as regards the chance of observing larger or small values of z. Information regarding the chance of observing larger or small values of z can be provided by testing the null hypothesis of fairness against the alternative hypothesis of non-fairness. The generalized likelihood-ratio test is the standard means to this end. Using the generalized likelihood-ratio test, the following function is computed:

$$\chi = -2\log\left[n^n \prod_{k=1}^{m}\left(\frac{\theta_k^{(0)}}{n_k}\right)^{n_k}\right] \tag{12}$$

which results from the ratio between $\Pr(G = z)$ under the null hypothesis that, $\forall k = 1, ..., m, \theta_k = \theta_k^{(0)}$ and the alternative hypothesis that $\theta_k = n_k/n$, i.e. the maximum likelihood estimation. When $m > 2$, χ has approximately the chi-square distribution with m degrees of freedom; when n is small, the exact probability can easily be calculated [31]; for example, $\theta_k^{(0)} = 1/m$ if fairness is the null hypothesis which is rejected when $\chi > \chi_{1-\alpha}(m)$ where $\chi_{1-\alpha}(m)$ is the $1-\alpha$-th quantile of the chi-square distribution with m degrees of freedom for a given probability of I-type error α; for example, when $\chi > \chi_{0.95}(3) = 7.81$ the hypothesis that three groups are fairly exposed in a ranking can be rejected with probability of error equal to 5%.

6 Empirical study

The empirical study reported in this section aimed to measure and analyze the reranking method introduced in the previous sections to improve fairness while not decreasing effectiveness much. Specifically, the study was performed to test if a test collection can confirm the hypothesis that the measure of fairness introduced in this paper can significantly increase without the measure of effectiveness significantly decreases. To this end, the impact on effectiveness and fairness was measured and analyzed in order to check the hypothesis that a trade-off between fairness and effectiveness in ranking should be rejected.

The study was performed using the test datasets utilized within two Fair tracks of TREC. The 2019 and 2020 TREC Fair track datasets consist of a document collection, a set of queries, the ranking returned for each query, the document authors, the economic development level and the "prestige" level of the organizations of the document authors, which allow to build the groups. The documents are labelled by binary relevance. Details are available in [4, 5].

In this paper, the relevance judgements used to rank documents and compute effectiveness are assumed to be correct, or at least unbiased. The assumption is important because a fair ranking may be more effective than an unfair

ranking because improving its fairness has corrected for systematic biases in observed or measured effectiveness. Indeed, the Fair TREC data used in this paper likely does have biased effectiveness measures, because relevance is determined by clicks by humans who may be unfair, in response to rankings from a system that is not optimized for fairness.

Note that the datasets used for the experiments make $m = 3$ or $m = 5$ because there are two or four explicit groups plus one implicit group named Others including all the documents for which an explicit group has not been provided. The results for the evaluation queries of the 2019 and 2020 Fair tracks of TREC are summarized in Tables 1 and 2. Note that F depends on the x's, which in turn depends on the parameters; therefore, the values at $t = 1$ are different for different parameter values and access models.

The empirical study considered the retrieved document lists provided by a test collection. Each retrieved document list provided by a test collection was reranked by BM25F with respect to the evaluation queries which originated the list. The document lists are of varying length – from a few documents to some dozens documents depending on the query to which the documents have been retrieved. The retrieved document lists have been embedded in the test collection.

The ranking obtained in this way was then scanned from the top to the bottom ranked document; each document ranking was scanned five times. The documents were scanned more than once because two documents are swapped at each scan, thus giving rise to a new ranking, which has therefore to be scanned again. A maximum of five scans were carried out because it was seen that the ranking stabilizes most of the time.

After each scan of a retrieved document ranking, two documents were swapped, but only if the swap maximized the increase of fairness while this increase was not below a threshold δ and the decrease of effectiveness was not above a threshold ϵ, i.e. the former is the threshold of minimal increase of fairness and the latter is the threshold of maximal loss of effectiveness.

The experiments were repeated for, and the results were averaged over all

$$\delta \in \{0.00, 0.05, 0.10, 0.15, 0.20, 0.25, 0.30\}$$

and

$$\epsilon \in \{0.10, 0.15, 0.20, 0.25, 0.30, 0.35, 0.40, 0.45, 0.50, 0.75\}\ .$$

The documents were not swapped if the increase of fairness was less than δ or the absolute value of the decrease of effectiveness was less than ϵ. The best swap yielded the maximal increase of effectiveness. At each swap, the y's were updated and G and F were computed. In addition, the probability of G and the χ's of Section 5 were computed.

The user access models mentioned in Section 4 were investigated in the empirical study, i.e. the exponential user access model and the geometric user access model. Consider the results obtained from the 2019 TREC Fair track test collection and reported in Table 1; similar results were obtained from the 2020 TREC Fair track test collection and reported in Table 2. In general, the trade-off between fairness and effectiveness was not as strong as expected. The expectation of a strong trade-off between fairness and effectiveness was fueled by the idea that any distancing from the ranking provided by the PRP to increase fairness would have caused a significant loss of effectiveness. In some cases, effectiveness was kept constant if not even increased while fairness increased as well.

In sum, it seems that a few swaps can keep effectiveness constant while increasing fairness if the individual swaps do not hurt effectiveness. In particular, the largest increase of fairness was observed after the first scan whereas the subsequent scans provided smaller increases than the first. Such a small number of swaps is sufficient although $|\epsilon|$ is small or δ is large. This outcome is common to both user access models.

The limited drop in effectiveness may be due to the methodology adopted in building the relevance assessments of the test collections used in this study. This methodology was based on clickthrough data and manual setting of the threshold to transform the proportion of clicks to the binary relevance assessment [4, 5]. The use of clickthrough data and the manual setting of a threshold yielded small retrieved document sets with a relatively high number of relevant documents.

It was found that there is a relationship between the proportion of relevant documents to the number of retrieved documents and the variation of effectiveness due to swapping; no relationship was found between the proportion of relevant documents to the number of retrieved documents and the variation of fairness as measured by G. The following figures summarizes this finding. Figure 4 shows the distribution of the number of retrieved documents whereas Figure 5 shows the distribution of the proportion of relevant documents to the retrieved documents. Figure 6 reports the variation for each grouping and track.

Table 1 The table reports *F* and *G* for the evaluation queries of the 2019 Fair track of TREC

Exponential access model				
r	*t*	*F*	*G*	χ
(a) Economic development level				
1.00	1	0.585	0.563	1.5
	2	0.576	0.659	1.3
	3	0.574	0.662	1.3
	4	0.574	0.662	1.3
	5	0.573	0.662	1.3
2.00	1	0.578	0.384	2.7
	2	0.608	0.708	0.0
	3	0.607	0.712	0.0
	4	0.608	0.712	0.0
	5	0.608	0.712	0.0
5.00	1	0.670	0.062	11.6**
	2	0.702	0.107	10.6*
	3	0.705	0.107	10.6*
	4	0.706	0.107	10.6*
	5	0.706	0.107	10.6*
10.00	1	0.672	0.015	11.8**
	2	0.696	0.051	10.8*
	3	0.697	0.051	10.8*
	4	0.698	0.051	10.8*
	5	0.699	0.051	10.8*
Geometric access model				
u	*t*	*F*	*G*	χ
0.25	1	0.535	0.620	1.3
	2	0.532	0.630	1.4
	3	0.532	0.631	1.4
	4	0.532	0.631	1.4
	5	0.532	0.631	1.4
0.50	1	0.552	0.610	1.2
	2	0.545	0.643	1.3
	3	0.545	0.644	1.3
	4	0.544	0.644	1.3
	5	0.544	0.644	1.3
0.75	1	0.568	0.591	1.2
	2	0.559	0.655	1.2
	3	0.558	0.657	1.2
	4	0.557	0.657	1.2
	5	0.557	0.657	1.2
0.85	1	0.575	0.581	1.3
	2	0.565	0.659	1.2
	3	0.564	0.662	1.2
	4	0.564	0.662	1.2
	5	0.563	0.662	1.2

Table 1 (continued)

Exponential access model				
r	*t*	*F*	*G*	χ
(b) H-index level				
1.00	1	0.585	0.634	5.1
	2	0.577	0.738	3.6
	3	0.574	0.743	3.5
	4	0.573	0.743	3.5
	5	0.573	0.743	3.5
2.00	1	0.578	0.488	6.1
	2	0.624	0.771	1.6
	3	0.623	0.773	1.6
	4	0.623	0.774	1.6
	5	0.623	0.774	1.6
5.00	1	0.670	0.224	12.1*
	2	0.775	0.490	3.3
	3	0.775	0.490	3.2
	4	0.776	0.490	3.3
	5	0.776	0.490	3.3
10.00	1	0.672	0.185	14.3*
	2	0.781	0.461	8.6
	3	0.781	0.461	8.6
	4	0.781	0.461	8.6
	5	0.781	0.461	8.6
Geometric access model				
u	*t*	*F*	*G*	χ
0.25	1	0.535	0.682	4.7
	2	0.533	0.692	4.6
	3	0.532	0.693	4.6
	4	0.532	0.693	4.6
	5	0.532	0.693	4.6
0.50	1	0.552	0.674	4.8
	2	0.545	0.707	4.4
	3	0.544	0.709	4.4
	4	0.544	0.710	4.4
	5	0.544	0.710	4.4
0.75	1	0.568	0.658	4.8
	2	0.559	0.725	3.9
	3	0.557	0.729	3.8
	4	0.556	0.730	3.8
	5	0.556	0.730	3.8
0.85	1	0.575	0.649	5.0
	2	0.566	0.732	3.7
	3	0.563	0.736	3.7
	4	0.562	0.736	3.7
	5	0.562	0.736	3.6

Two documents were swapped at each step. The value of χ was labeled with * (**) if it is significant at $\alpha = 0.05$ ($\alpha = 0.01$). The average was computed over all δ's and ϵ's. The significance of χ means that one can reject the hypothesis that a ranking is fair with probability of wrong rejection equal to α

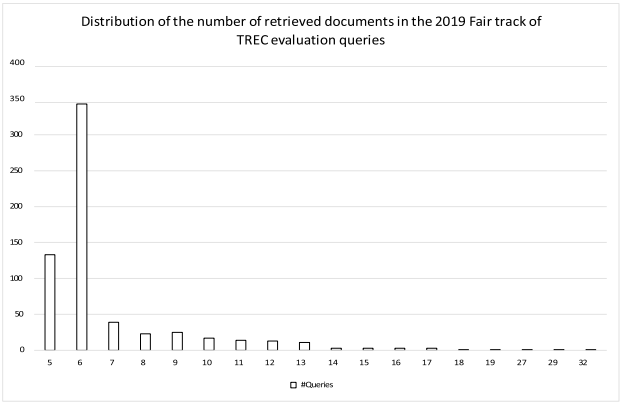
Table 2 The table reports *F* and *G* for the evaluation queries of the 2020 Fair track of TREC

Exponential access model				
r	*t*	*F*	*G*	χ
(a) Economic development level				
1.00	1	0.196	0.671	6.6
	2	0.187	0.879	3.8
	3	0.186	0.896	3.1
	4	0.186	0.896	3.1
	5	0.186	0.896	3.1
2.00	1	0.202	0.471	8.6*
	2	0.259	0.872	1.7
	3	0.259	0.879	1.7
	4	0.259	0.880	1.6
	5	0.260	0.880	1.6
5.00	1	0.210	0.159	14.2**
	2	0.317	0.591	7.5
	3	0.317	0.593	7.4
	4	0.318	0.593	7.4
	5	0.318	0.593	7.4
10.00	1	0.210	0.118	14.4**
	2	0.323	0.556	8.2*
	3	0.323	0.556	8.2*
	4	0.323	0.556	8.2*
	5	0.323	0.556	8.2*
Geometric access model				
u	*t*	*F*	*G*	χ
0.25	1	0.176	0.733	6.0
	2	0.176	0.741	5.9
	3	0.176	0.743	5.9
	4	0.176	0.744	5.9
	5	0.176	0.745	5.9
0.50	1	0.183	0.721	6.1
	2	0.183	0.753	5.7
	3	0.182	0.759	5.7
	4	0.182	0.760	5.6
	5	0.182	0.761	5.6
0.75	1	0.190	0.701	6.3
	2	0.190	0.779	5.5
	3	0.189	0.787	5.4
	4	0.189	0.788	5.3
	5	0.189	0.789	5.2
0.85	1	0.192	0.691	6.4
	2	0.191	0.796	5.4
	3	0.190	0.803	5.1
	4	0.190	0.805	5.1
	5	0.190	0.805	5.0

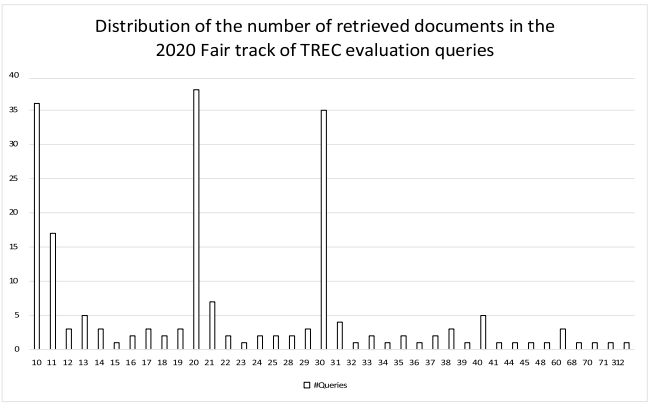
Table 2 (continued)

Exponential access model				
r	*t*	*F*	*G*	χ
(b) H-index level				
1.00	1	0.196	0.601	7.1
	2	0.202	0.681	5.5
	3	0.201	0.687	5.2
	4	0.201	0.688	5.2
	5	0.201	0.688	5.2
2.00	1	0.202	0.395	8.3*
	2	0.254	0.755	2.6
	3	0.254	0.762	2.6
	4	0.256	0.763	2.5
	5	0.255	0.763	2.5
5.00	1	0.210	0.059	20.4**
	2	0.275	0.252	11.9**
	3	0.276	0.253	11.9**
	4	0.279	0.253	11.9**
	5	0.280	0.253	11.9**
10.00	1	0.210	0.014	20.7**
	2	0.271	0.200	12.1**
	3	0.272	0.200	12.1**
	4	0.276	0.200	12.2**
	5	0.277	0.200	12.2**
Geometric access model				
u	*t*	*F*	*G*	χ
0.25	1	0.176	0.662	6.3
	2	0.177	0.671	6.1
	3	0.176	0.673	6.1
	4	0.176	0.674	6.1
	5	0.176	0.675	6.1
0.50	1	0.183	0.651	6.6
	2	0.183	0.683	6.0
	3	0.183	0.689	6.0
	4	0.183	0.690	5.9
	5	0.183	0.691	5.9
0.75	1	0.190	0.631	6.7
	2	0.191	0.711	5.8
	3	0.191	0.718	5.7
	4	0.190	0.720	5.6
	5	0.191	0.721	5.6
0.85	1	0.192	0.621	6.9
	2	0.197	0.727	5.7
	3	0.195	0.735	5.6
	4	0.195	0.737	5.5
	5	0.195	0.738	5.5

Two documents were swapped at each step. The value of χ was labeled with * (**) if it is significant at $\alpha = 0.05$ ($\alpha = 0.01$). The average was computed over all δ's and ϵ's. The significance of χ means that one can reject the hypothesis that a ranking is fair with probability of wrong rejection equal to α

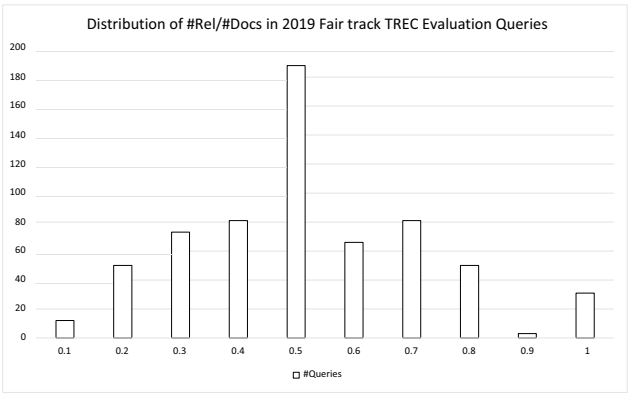


(a) 2019

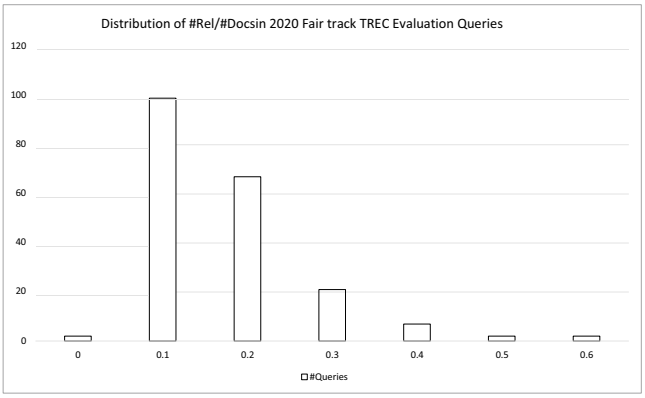


(b) 2020

Fig. 4 The distribution of the number of retrieved documents reports for each ranking length L the number of queries for which the number of retrieved documents is L. **a** refers to the 2019 track whereas **b** refers to the 2020 track



(a) 2019



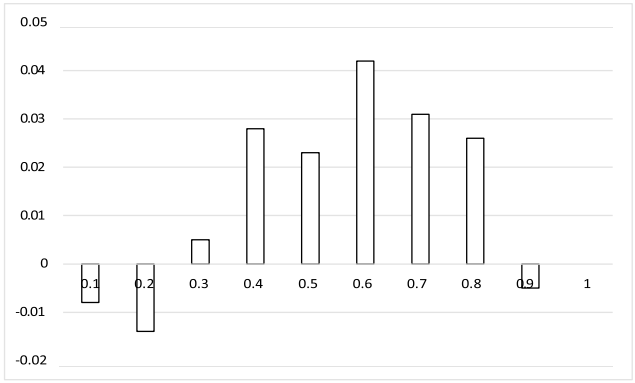
(b) 2020

Fig. 5 The distribution of the proportion of relevant documents to the retrieved documents reports for each proportion P the number of queries for which the number of relevant retrieved documents to the number of retrieved documents is P. **a** refers to the 2019 track whereas **b** refers to the 2020 track

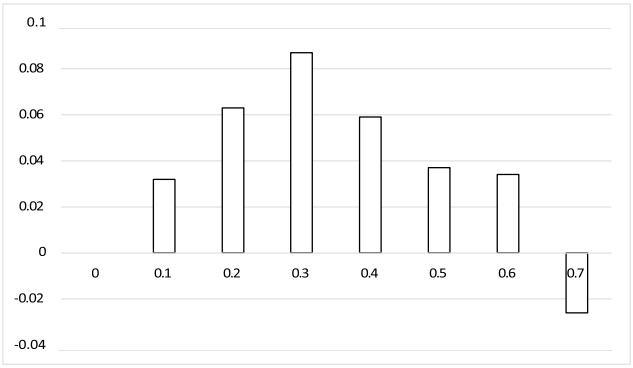
In addition, the average generalized likelihood-ratio test (12) reported in Tables 1 and 2 has been labeled by one or two asterisks if it is above $\chi_{1-\alpha}(m), \alpha \in \{0.01, 0.05\}$ for all t. When the asterisks occur for all t, fairness cannot be obtained in five scans, i.e. after five swaps, thus deciding the rejection of the null hypothesis of fairness with a small probability of wrong decision. If no asterisk occurs, the hypothesis that a retrieved document ranking is fair even before the first swap cannot be rejected. When the asterisk occurs for $t = 1$and disappears for $t > 1$, the first swap was sufficient to transform a biased ranking into a ranking for which fairness cannot be rejected and further swaps do not reverse the ranking to a biased one.

The event of non-rejection of the hypothesis of fairness before any swap was fairly common, since it occurred from a few dozens to a few hundreds of queries depending on the user access model and on the test collection. The non-rejection of fairness before any swap happens for the geometric user access model which describes a search more exhaustive than the search described by the exponential model with large r. An exhaustive search means that user is willing to explore the non-top ranked documents, thus allowing the producers thereof to be more likely exposed than in the event of the exponential user access model. The dependency of the fairness measure and as a consequence of the decision whether to reject on the user access model is due to the definition provided in (4) where the probability that a group is exposed depends on the probability that the documents of the group are accessed and not only on the distribution of document authors across the groups.

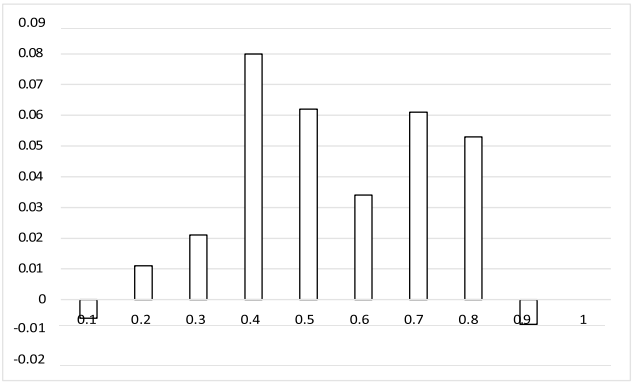
A closer look reveals that the hypothesis of fairness is rejected after one scan for a non-negligible number of queries because all the retrieved documents only belongs to the Other group, thus yielding $G = 0$. Indeed, many documents



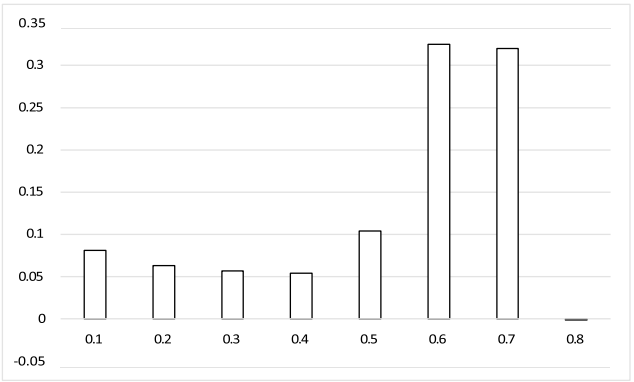
(a) 2019 grouping by economic development level



(b) 2020 grouping by economic development level



(c) 2019 grouping by H-index



(d) 2020 grouping by H-index

Fig. 6 There is a relationship between the proportion of relevant documents to the number of retrieved documents and the variation of effectiveness due to swapping. For each proportion the average variation is reported. The average was computed after joining the data about effectiveness at the first and the last steps for all user access models and for all the values of u, ϵ, δ. The average was computed over all δ's and ϵ's. The variation was then computed between $F^{(1)}$ and $F^{(5)}$

are authored by authors for whom the group membership is unknown. For those cases, the probability that the document belongs to the Other group is one. In such a case, fairness cannot be obtained because the group membership is unknown and G becomes null; for example, when the user access model is exponential with $r = 1$, there are 38 queries out of 635 queries of the 2019 Fair track test collection for which the hypothesis of fairness can be rejected ($\alpha = 0.05$) at the first step, but the rankings of 25 queries out of those 38 have no information about the grouping by economic development level.[4] Table 3 reports the summary for all the groupings and tracks.

As a consequence, the rankings appeared less fair than they might be because G is computed by considering all the groups although some of them do not include any document; in such a case, the index is "artificially" lowered. As a solution, the computation of G should be rescaled after considering the actual number of groups for which the y's are positive, although there might be some issues of comparability.

7 Conclusions

As relevance is the property that makes information useful or necessary to meet the user's information needs, relevance is a well known multidimensional notion which challenged the research in IR for a long time. Historically, the complexity of the user's information needs and problems was the main reason of the multidimensionality of relevance; diversification

[4] Note that no grouping information implies $G = 0$, but the *vice versa* does not hold.

Table 3 For each user access model, a table reports (i) the number of 2020 Fair track queries out of 200 queries for which the hypothesis of rejection was rejected ($\alpha = 0.05$) and (ii) the number of queries out of the aforementioned number for which no information grouping was available, thus leading to $G = 0$

Exponential user		
access model		
r	(i)	(ii)
(a) Economic development level (2019 Fair track)		
1.00	38	25
2.00	76	25
5.00	265	55
10.00	266	129
Geometric user		
access model		
u	(i)	(ii)
0.25	32	25
0.50	34	25
0.75	36	25
0.85	36	25
Exponential user		
access model		
r	(i)	(ii)
(b) H-index level (2019 Fair track)		
1.00	162	38
2.00	0	0
5.00	0	0
10.00	0	0
Geometric user		
access model		
u	(i)	(ii)
0.25	126	38
0.50	139	38
0.75	143	38
0.85	153	38
Exponential user		
access model		
r	(i)	(ii)
(c) Economic development level (2020 Fair track)		
1.00	58	3
2.00	51	3
5.00	33	14
10.00	33	20
Geometric user		
access model		
u	(i)	(ii)
0.25	54	3

Table 3 (continued)

Geometric user		
access model		
u	(i)	(ii)
0.50	61	3
0.75	62	3
0.85	61	3
Exponential user		
access model		
r	(i)	(ii)
(d) H-index level (2020 Fair track)		
1.00	57	3
2.00	59	3
5.00	45	12
10.00	47	20
Geometric user		
access model		
u	(i)	(ii)
0.25	48	3
0.50	52	3
0.75	55	3
0.85	52	3

The difference between the two numbers is as a consequence the number of queries for which the hypothesis of rejection was truly rejected

was indeed studied for some years. As the context where information is produced and consumed has become complicated, the additional complexity of the producers of information requires the addition of other dimensions to the notion of relevance – the fairness of access to information is perhaps among the most crucial dimensions. The approaches to obtaining fairness inherited some techniques from diversification, however, fairness have to consider producers which exhibit different needs.

Similarly to effectiveness, fairness needs a measure, however, the measure of fairness should somehow be related to the measure of effectiveness to study how this dimension relates to relevance. Gini's indexes may provide a measure of fairness, since they may be referred to the way the users access information. In particular, the mutability index is the ground on which a user access model can be implemented to relate the index with the measure of effectiveness. Indeed, the co-variation of the measure of fairness and of the measure of effectiveness can be the basis of a model that relates both measures.

In this paper, it is shown how and the degree to which the variations of fairness relate with the variations of effectiveness. The relationship is a little more complex than an inverse function that when one measure increases the other decreases, and *vice versa*, since there is a point at which both measures decrease. The function between fairness and effectiveness and not only between the variations thereof would also be interesting especially to the aim of finding a ranking principle combining both effectiveness and fairness together along the lines of the use of portfolio theory exposed in [28, 29] applied in IR in [40].

Whether the model that relates the measure of fairness with the measure of effectiveness is an appropriate way to predict the degree to which fairness and effectiveness can coexist depends on the notion of relevance and on that of fairness. Whereas the notion of relevance refers to a property of a single document, the notion of fairness refers to a property of a ranking. As the nature of a single document differs from the nature of a ranking, a model that relates fairness and effectiveness may lack of justification. Although the difference between the nature of a single document and the nature of a ranking might damage a model that relates fairness and effectiveness, the fact that effectiveness and

fairness can coexist seems to suggest further development of a more general theory; some efforts have recently done [32] and further reflections would be useful, though.

As noted in Section 6, the Fair TREC data likely does have biased effectiveness measures, because relevance is determined by eccessively large number of clicks on the top-ranked documents or by humans who may be unfair. If a measure of effectiveness such as Mean Average Precision (AP) is an estimation of the actual effectiveness and therefore affected by bias, and if we decrease the measure of effectiveness to improve fairness it is unclear whether the decreased effectiveness comes from the actual measure or the bias. It may well be that the permuted ranking increases the estimated measure of effective because improving fairness has corrected for systematic biases in the measure of effectiveness.

Acknowledgements I thank the reviewers very much for the insightful and accurate comments and suggestions which helped me improve the final version of the paper.

Author contributions Not applicable.

Funding Open access funding provided by Università degli Studi di Padova within the CRUI-CARE Agreement.

Data availability The manuscript contains empirical evidence obtained from third party material, i.e. the TREC Fair track test collections, and obtained permissions are available on request by the Publisher.

Declarations

Ethical approval Not applicable.

Competing interests The author has no competing interests as defined by Springer, or other interests that might be perceived to influence the results and/or discussion reported in this paper.

References

1. Angwin J, Larson J, Mattu S, Kirchner L. Machine bias. Technical report: ProPublica.org; 2016.
2. Baeza-Yates R. Bias on the Web. Commun ACM. 2018;61(6):54–61.
3. Barocas S, Selbst AD. Big data's disparate impact. 104 California Law Rev 2016;671.
4. Biega AJ, Diaz F, Ekstrand MD, Feldman S, Kohlmeier S. Overview of the TREC 2020 fair ranking track. In Proceedings of TREC, 2020.
5. Biega AJ, Diaz F, Ekstrand MD, Kohlmeier S. Overview of the TREC 2019 fair ranking track. In: Proceedings of TREC, 2019.
6. Biega AJ, Gummadi KP, Weikum G. Equity of attention: amortizing individual fairness in rankings. In: Proceedings of SIGIR, ACM, 2018, pp. 405–414
7. Binns R. On the apparent conflict between individual and group fairness. In: Proceedings of FAccT, 2020.
8. Boratto L, Faralli S, Marras M, Stilo G. Guest editorial of the IPM special issue on algorithmic bias and fairness in search and recommendation. Inf Process Manag. 2022;59(1): 102791.
9. Carbonell J, Goldstein J. The use of MMR, diversity-based reranking for reordering documents and producing summaries. In: Proceedings of SIGIR, 1998;335–336.
10. Castellano V. Corrado Gini - a memoir with the complete bibliography of his works. Metron, XXIV(1-4), 1965.
11. Celis LE, Mehrotra A, Vishnoi NK. Interventions for ranking in the presence of implicit bias. In: Proceedings of FAccT, 2020;369–380.
12. Ceriani L, Verme P. The origins of the Gini index: Extracts from Variabilità e Mutabilità (1912) by Corrado Gini. J Econ Inequal. 2012;10:421–43.
13. Chapelle O, Metlzer D, Zhang Y, Grinspan P. Expected reciprocal rank for graded relevance. In: Proceedings of CIKM, 2009;621–630.
14. Chen J, Kallus N, Mao X, Svacha G, Udell M. Fairness under unawareness: assessing disparity when protected class is unobserved. In: Proceedings of FAccT, 2019;339–348.
15. Diaz F, Mitra B, Ekstrand MD, Biega AJ, Carterette B. Evaluating stochastic rankings with expected exposure. In: Proceedings of CIKM, 2020;275–284.
16. Draws T, Tintarev N, Gadiraju U, Bozzon A, Timmermans B. This is not what we ordered: Exploring why biased search result rankings affect user attitudes on debated topics. In: Proceedings of SIGIR, 2021;295–305.
17. Dwork C, Hardt M, Pitassi T, Reingold O, Zemel R. Fairness through awareness. In: Proceedings of ITCS, . ACM, 2012, 214–226.
18. Ekstrand MD, Das A, Burke R, Diaz F. Fairness in information access systems. Found Trends Inf Retr. 2022;16(1–2):1–177.
19. Friedler SA, Scheidegger C, Venkatasubramanian S. The (im)possibility of fairness: different value systems require different mechanisms for fair decision making. Commun ACM. 2021;64(4):136–43.

20. Friedler SA, Scheidegger C, Venkatasubramanian S, Choudhary S, Hamilton EP, Roth D. A comparative study of fairness-enhancing interventions in machine learning. In: Proceedings of FAccT. ACM, 2019, 329–338.
21. Gao R, Shah C. How fair can we go: Detecting the boundaries of fairness optimization in information retrieval. In: Proceedings of SIGIR, 2019;229–236.
22. Ghosh A, Dutt R, Wilson C. When fair ranking meets uncertain inference. In: Proceedings of SIGIR. ACM, 2021, 1033–1043.
23. Gini C. *Variabilità e mutabilità*. Tipografia di Paolo Cuppin, 1912.
24. Goffman W. A searching procedure for information retrieval. ISAR. 1964;2(2):73–8.
25. Hardt M, Price E, Price E, Srebro N. Equality of opportunity in supervised learning. In: Proceedings of NIPS, 2016.
26. Jacobs AZ, Wallach H. Measurement and fairness. In: Proceedings of FAccT. ACM, 2021, 375–385.
27. Kleinberg J, Mullainathan S, Raghavan M. Inherent trade-offs in the fair determination of risk scores. In: Proceedings of ITCS, 2017;67:43:1–43:23.
28. Markowitz H. Portfolio selection. J Finance. 1952;7(1):77–91.
29. Markowitz HM, Lacey R, Plymen J, Dempster MAH, Tompkins RG. The general mean-variance portfolio selection problem [and discussion]. Philos Trans A Math Phys Eng Sci. 1994;347(1684):543–9.
30. Moffat A, Zobel J. Rank-biased precision for measurement of retrieval effectiveness. ACM Trans Inf Syst. 2008;27:1–27.
31. Mood A, Graybill F, Boes D. Introduction to the theory of statistics. McGraw-Hill, 1974.
32. Oosterhuis H. Computationally efficient optimization of Plackett-Luce ranking models for relevance and fairness. In: Proceedings of SIGIR, 2021;1023–1032.
33. Papakyriakopoulos O, Hegelich S, Serrano JCM, Marco F. Bias in word embeddings. In: Proceedings of FAccT, 2020;446–457.
34. Pessach D, Shmueli E. A review on fairness in machine learning. ACM Comput Surv 2022;55(3).
35. Robertson S. The probability ranking principle in information retrieval. J Doc. 1977;33(4):294–304.
36. Singh A, Joachims T. Fairness of exposure in rankings. In: Proceedings of SIGKDD, 2018;2219–2228.
37. Sonboli N, Smith JJ, Cabral Berenfus F, Burke R, Fiesler C. Fairness and transparency in recommendation: the users' perspective. In: Proceedings of UMAP, 2021;274–279.
38. Sweeney L. Discrimination in online ad delivery: Google ads, black names and white names, racial discrimination, and click advertising. Queue. 2013;11(3):10–29.
39. Verma S, Rubin J. Fairness definitions explained. In: Proceedings of FairWare. ACM, 2018, 1–7
40. Wang J, Zhu J. Portfolio theory of information retrieval. In: Proceedings of SIGIR. ACM, 2009,115–122.
41. Wang L, Joachims T. User fairness, item fairness, and diversity for rankings in two-sided markets. In: Proceedings of ICTIR, ACM, 2021, 23–41.
42. Zehlike M, Bonchi F, Castillo C, Hajian S, Megahed M, Baeza-Yates R. FA*IR: a fair top-k ranking algorithm. In: Proceedings of CIKM, 2017;1569–1578.
43. Zehlike M, Sühr T, Baeza-Yates R, Bonchi F, Castillo C, Hajian S. Fair top-k ranking with multiple protected groups. Inf Process Manag 2021;59(1).