Fairness and Transparency in Ranking*

Carlos Castillo
Universitat Pompeu Fabra
chato@acm.org

Abstract

Ranking in Information Retrieval (IR) has been traditionally evaluated from the perspective of the relevance of search engine results to people searching for information, i.e., the extent to which the system provides "the right information, to the right people, in the right way, at the right time." However, people in current IR systems are not only the ones issuing search queries, but increasingly they are also the ones being searched. This raises several new problems in IR that have been addressed in recent research, particularly with respect to fairness/non-discrimination, accountability, and transparency. This is a summary of some these initial developments.

1 Introduction

Traditionally, the objective of an IR system is to provide maximum utility to searchers. This is achieved by assuming frameworks such as the probability ranking principle [20], which infers the relevance of items and then orders them by decreasing expected relevance. However, when the items to be retrieved are people, and to some extent when they are organizations, communities, and places, the interests of those being searched become a matter of concern. Search engines for finding local business, products, jobs, events, news, and people can have significant effects on the economic, social, career, political, and even affective/reproductive success of those being searched and ranked.

There is abundant literature on fairness and discrimination in Machine Learning (ML) and Data Mining (DM); see, e.g., Hajian et al. [12]. In these fields, fairness is operationalized in various forms that sometimes correspond to different viewpoints in debates around fairness and justice in moral philosophy [4]: ensuring equal outcomes for different groups, or equal accuracy, or equal false positives/negative rates, or equality regarding counterfactual scenarios where items from one group are assumed to belong to the other.

Ensuring fairness, accountability, and transparency in IR has been considered as a key strategic area for the advancement of the field [8]. However, research on these issues in IR is much less developed than in ML/DM. This is a brief summary of some of these initial developments, and it is meant to illustrate general directions in this nascent area.

*From the keynote "Fairness and Transparency in Ranking" at the *Data and Bias (DAB) Workshop* at the ACM International Conference on Information and Knowledge Management (CIKM'18). Turin, Italy, October 2018. Slides at http://chato.cl/slides/DAB_2019_fairness_transparency.pdf

2 What is a Fair Ranking?

Fairness in computing is concerned with matters of equality and justice that have been debated for centuries [4], and there are many competing definitions [16]. In ML/DM, fairness is usually seen as the absence of discrimination [12].

Lippert-Rasmussen [14] defines discrimination by an agent X of subject Y in relation to subject Z as *disadvantageous differential treatment* of Y with respect to Z, based on (X's belief on) subject Y having a certain property P that Z does not have. When P is the property of belonging to a socially salient group, and X has animosity against this group, or the perception that people in this group are inferior or should not intermingle with other groups, this is called *group discrimination*. When group discrimination is based on statistical beliefs held by X based on previous evidence, this is called *statistical group discrimination* or simply *statistical discrimination*.

If X is a computer system based on data, which is the case for most contemporary ML/DM/IR systems, then disregarding animosity on the part of X (or its developers) and considering that statistical beliefs are basically any internal state of the system derived from its input data, we arrive to the conclusion that data-driven systems can engage in statistical discrimination. In IR, this can be due to various issues, including biases in training data, biases in user behavior (e.g., tendency to click on results of a certain kind), and biases in the corpus itself (such as different sections of resumes completed at different rates by men and women [1]).

A fair ranking is a ranking having at least the following characteristics:

1. A **sufficient presence** of items belonging to different groups, particularly groups considered disadvantaged/protected [19], thus avoiding statistical discrimination and hence preventing distributive/allocative harms to members of these groups.
2. A **consistent treatment** of similar items, which ensures individual fairness [30].
3. A **proper representation** of items, particularly disadvantaged/protected groups, that prevents representational harms to members of these groups.

Of these, representational harms are perhaps the least studied category in IR. Representational harms, according to Crawford [7], "occur when systems reinforce the subordination of some groups along the lines of identity." In the context of search this includes, among other issues, sexualized results for queries such as "black women" and stereotyped query completion suggestions for queries such as "(group) are ..." (see Noble [17] for an overview).

Before addressing how to measure fairness with respect to the *sufficient presence* and *consistent treatment* conditions, an observation on the concept of **diversity** is in order. Diversity in search is understood as either seeking that search results do not include results that are too similar to each other (maximizing marginal relevance), or as accounting for uncertainty on the user's intent [10]. Diversity is not the same as fairness, as diversity is concerned with the utility for the searcher and is symmetric, while fairness is concerned with the utility for those searched and is asymmetric, emphasizing not the presence of various groups but ensuring that those in protected groups are effectively included.

3 Measuring Fairness in Rankings

Two main groups of measures for fairness in rankings have been proposed in recent years: attention-based and probability-based.

3.1 Attention-based measures

Attention-based measures seek to quantify either the attention that different items receive from searchers, through proxies such as click-through rates, or the potential attention they might receive, through proxies such as exposure or inferences of the probability that items will be considered relevant (and possibly clicked).

Singh and Joachims [22, 23] define *fairness of exposure* as follows.[1] Let $P_{m\times n}$ represent a probabilistic ranking of m items in n positions, i.e., a doubly stochastic matrix representing the probability of ranking each available item $d_i, i \in [m]$ in position $j \in [n]$. Let v_j for $j \in [n]$ represent the visibility or exposure of position j independently of the specific item being shown in that position. Values for v_j can be obtained empirically through, e.g., eye tracking studies, or assumed to follow a certain form such as logarithmic discount ($v_j \propto 1/\log(j+1)$). Let u_i for $i \in [m]$ represent the relevance or utility of item d_i. Let us further assume items can belong to groups: G_0, representing the majority/advantaged/unprotected group, and G_k, $k \geq 1$ representing the minority/disadvantaged/protected groups.

The exposure of group k given ranking P is defined as the average exposure of its members $\text{Exposure}(G_k|P) = \frac{1}{|G_k|}\sum_{i:d_i\in G_k}\sum_{j=1}^{n} P_{i,j}v_j$. In the same manner, utility is defined as the average utility of its members $\text{U}(G_k) = \frac{1}{|G_k|}\sum_{i:d_i\in G_k} u_i$. Fairness of exposure between two groups, for a given query, is achieved when the ratio between the exposure of items given the ranking results for this query is proportional to the ratio between the utility of items for this query. For instance in the case of groups 0 and 1:

$$\frac{Exposure(G_0|P)}{U(G_0)} = \frac{Exposure(G_1|P)}{U(G_1)} \ .$$

Any deviation from this equality is considered *disparate treatment*. If instead of inferred exposure $P_{i,j}v_j$ the inferred click-through probability $P_{i,j}v_ju_i$ is used, this is considered *disparate impact* [23].

Other ad-hoc measures can be proposed to quantify differences in exposure/attention, including divergence between probability distributions describing the position of items in various groups, or variants of logarithmic discount [26].

3.2 Probability-based measures

Probability-based measures assume a ranking has been generated by a randomized process, such as the one proposed by Yang and Stoyanovich [26], and measure deviations from the expected characteristics of the ranking and those observed. In [26], the process begins by producing separately ordered lists of items, one for every group G_k, sorted by decreasing utility. In the case of a single protected group, the merging of these lists is controlled by a parameter $p \in [0, 1]$ and proceeds as follows. For every position $j \in [n]$, a Bernoulli trial with

[1] For clarity of exposition, we use a unified notation, which is similar but not exactly the same as the one used by the different works being cited.

probability p is performed. If the trial succeeds, the top element from the protected group G_1 is selected; if the trial fails, the top element from the nonprotected group G_0 is selected. This process continues until n elements have been selected. If the list from one of the groups is exhausted before the process finishes, the remaining elements from the other group(s) are selected.

If this process is assumed to generate a fair ranking parameterized by the probability p, which can be seen as representing the *share* of the protected group, then given a ranking a statistical test can be used to determine the probability that the ranking was indeed generated by this procedure. Concretely, if at position j we have seen x elements of the protected group and $j - x$ elements of the nonprotected group, a one-tailed Binomial test can be used to compare the null hypotheses that this list was generated using the procedure above with parameter $p^* = p$, or with $p^* < p$, which would mean the protected group is represented less than what was desired. Zehlike et al. [28] show how this test can be made computationally efficient by avoiding the calculation of the cumulative distribution function of the Binomial distribution and instead using a pre-computed table; and also how to adjust the test sensitivity for multiple hypotheses testing, one at each position $j \in [n]$. Extending this test to the multinomial case, accounting for multiple protected groups, is a work in progress.

4 Creating Fair Rankings

Methods for non-discrimination in DM/ML can be categorized in one of three groups: post-processing, in-processing, and pre-processing [12]. In this section, methods for fairness in IR are presented along the same categories.

4.1 Post-processing methods

Post-processing methods seek to *re-rank* a list of items according to certain constraints. These comprise the majority of the methods for fair ranking proposed so far.

Celis et al.'s [6] approach this as an integer programming optimization problem. The goal is to obtain a permutation matrix $P_{m\times n}$ with $P_{i,j} = 1$ iff item d_i is placed at position j. For every position and every group $k \geq 0$ we have (fairness) bounds $B^{\min}_{k,\ell}$, $B^{\max}_{k,\ell}$ representing the minimum and maximum number of elements of G_k that must be present among the first ℓ positions of the ranking, with $\ell \in [n]$:

$$B^{\min}_{k,\ell} \leq \sum_{1\leq j\leq \ell} \sum_{i\in G_k} P_{i,j} \leq B^{\max}_{k,\ell} \ .$$

Given a utility matrix $U_{i,j}$ indicating the utility of placing item d_i in position j, the goal is to maximize $\sum_{i\in[m],j\in[n]} P_{i,j}U_{i,j}$ subject to the bounds described by $B^{\min}$ and $B^{\max}$. This is an NP-hard problem but an approximate solution can be efficiently generated: if Δ is the maximum number of constrained groups to which an element belongs, an LP relaxation yields a solution in which constraints are violated by at most a factor of $\Delta + 2$ [6].

The framework presented in [6] is quite general. If $\Delta = 1$ the solution is algorithmically simple and consists on scanning the lists top to bottom picking the next element as the highest utility one, among the ones whose addition to the list would not violate the constraints [28].

If no integrality constraints are imposed, the exact solution is a probabilistic ranking and can be obtained by solving the respective LP [23].

The idea of *amortized fairness* introduced by Biega et al. [3] concerns itself with fairness to individuals instead of groups, and considers that fairness to individuals cannot in general be achieved in each and every query, but *across* queries. As in [22], the idea is that attention/exposure should be allocated proportionally to the utility of items. Fairness is achieved by performing an online optimization. The total surplus/deficit of attention each element has received (considering its utility) from the first query processed by a system until the query just before q is maintained, and the online optimization seeks to re-rank the results for query q in such a way to correct the surplus/deficit of attention, subject to the constraint that the utility for the searcher of each query q must be above a certain minimum of quality.

4.2 In-processing methods

In-processing methods consider simultaneously utility and fairness criteria, for instance by performing a joint optimization or by using one criterion as a constraint while optimizing the other.

Zehlike and Castillo [29] describe an extension of ListNet [5], a list-wise learning to rank approach. In ListNet, as in other learning to rank methods, training data takes the form of a series of pairs $(q, y^{(q)})$, where q is a query and $y^{(q)}$ is a vector of size n such that $y_j^{(q)} = i$ means that document d_i should appear in position j for query q. The objective is to find a ranking function f, with $f(q) = \hat{y}^{(q)}$, that minimizes a loss function $L(y^{(q)}, \hat{y}^{(q)})$ for queries q in the training set. This loss function represents the extent to which the ordering of documents induced by f for a query differs from the ordering in which the documents appear in the training set, which is considered the gold standard.

The loss function of a learning to rank method is extended in [29] by considering an extra term accounting for disparate exposure, hence the loss becomes $L(y^{(q)}, \hat{y}^{(q)}) + \gamma U(\hat{y}^{(q)})$, with γ being a parameter that controls the trade-offs between accuracy with respect to the training set and differential exposure. The differential exposure loss $U(\hat{y}^{(q)})$ is defined as the hinge squared loss $\max(0, \text{Exposure}(G_1|\hat{y}^{(q)}) - \text{Exposure}(G_0|\hat{y}^{(q)}))^2$ where G_1 is the protected group, G_0 the nonprotected group, and exposure is defined in the same way as in [23] described previously in the Section 3.1. This leads to a differentiable objective that can be efficiently optimized using gradient descent.

4.3 Pre-processing methods

A pre-processing method for achieving fairness in rankings needs to pre-process training data to reduce the potential impact of bias on it. For instance, if in list $y^{(q)}$ of the training set all elements of a class are ranked below the elements of another class, $y^{(q)}$ could be re-ranked using some of the fairness criteria we have described, before using them to learn a ranking function.

5 Transparency in Ranking

In August 2018, the US President accused search engine Google of being "rigged" against his administration, by displaying only negative coverage in search results for "Trump News" [11].

While frameworks for "black-box" evaluation of political bias in search exist (see, e.g., [13]), this is not the first and probably not the last time search engines are accused of biased search results. Ensuring transparency in the way in which search results are ranked would help build trust from the public in the neutrality (at least in terms of partisan politics) of these platforms.

Transparency in ML/DM/IR systems is important for many reasons: it allows different claims about a system to be tested, it supports ethical compliance, it ensures the objectives of a system are aligned with those of its users, and it makes necessary trade-offs visible [9]. However, transparency is rarely seen in search engines or in general in large commercially-operated platforms on the web. Instead, through legal and technological means, such as business secrets and obfuscation, these platforms are opaque with respect to the way in which they operate [18]. Bias on search and recommendation platforms is also harder to evaluate in comparison to other forms of media: "If the nightly television news does not cover a protest, the lack of coverage is evident ... However, there is no transparency in algorithmic filtering" [25].

The first alarms about transparency in search were raised over 20 years ago with respect to **transparency in advertising**. At that point, search engines mixed paid results with organic search results with limited disclosures about which results were of which kind, and using confusing language such as "premium results" or "featured listings." In the US, this lead to a letter from consumer advocacy groups to the Federal Trade Commission (FTC), which warned search engines about deceptive practices in 2002 [21]. Unfortunately, over the years some observers have noted that ad transparency is becoming more "transparent" and ads blend more into search results than a few years ago [15].

Methods for providing **transparency in search results** are relatively new. One interesting proposal is a series of "Nutritional Labels for Rankings" advanced by Yang et al. [27]. These may include elements such as the different features or "ingredients" used by the ranker, its diversity with respect to various attributes, and whether it passes a series of fair ranking tests. Interestingly, there are many parallels with the "nutrition facts" shown in food products, including the fact that nutritional labels are indeed a solution to the problem of communicating technical, specialized information to a non-expert audience.

Mechanisms for **explainable rankings** are also new. These should produce a justification that allows a user to understand why a list is in a particular ordering and not in another. Ter Hoeve et al. [24] propose a perturbation-based approach, in which what is highlighted to the user is not the feature that plays the most important role in the ranking (e.g., the one with the largest coefficient if the ranking is the result of a linear scoring function), but the feature that if perturbed, would induce the larger changes in the ranking.

6 Conclusions

Just as in DM/ML, we are seeing the emergence of multiple definitions of what is a fair ranking, and different approaches to providing transparency in search. Different definitions and different methods may be used to address different manifestations of bias and discrimination. When evaluating each method, we should first consider what is the problem it is trying to address [2]. Future approaches may involve some of the elements we have seen in the DM/ML literature on fairness and transparency, adapted to ranking problems, and combined with possibly completely new methods that will be IR-specific.

References

[1] K. M. Altenburger, R. De, K. Frazier, N. Avteniev, and J. Hamilton. Are there gender differences in professional self-promotion? an empirical case study of linkedin profiles among recent mba graduates. In *Proc. of ICWSM*, pages 460–463, 2017.

[2] S. Barocas. What is the problem to which fair machine learning is the solution? AI Now Experts Workshop on Bias and Inclusion, 2017.

[3] A. J. Biega, K. P. Gummadi, and G. Weikum. Equity of attention: Amortizing individual fairness in rankings. *arXiv:1805.01788 (pre-print)*, 2018.

[4] R. Binns. Fairness in machine learning: Lessons from political philosophy. In S. A. Friedler and C. Wilson, editors, *Proceedings of the 1st Conference on Fairness, Accountability and Transparency*, volume 81 of *Proceedings of Machine Learning Research*, pages 149–159, New York, NY, USA, 23–24 Feb 2018. PMLR.

[5] Z. Cao, T. Qin, T.-Y. Liu, M.-F. Tsai, and H. Li. Learning to rank: from pairwise approach to listwise approach. In *Proc. of ICML*, pages 129–136. ACM, 2007.

[6] L. E. Celis, D. Straszak, and N. K. Vishnoi. Ranking with fairness constraints. In *Proc. of ICALP*, pages 28:1–28:15, 2018.

[7] K. Crawford. Keynote: The trouble with bias. Keynote at NIPS 2017, https://www.youtube.com/watch?v=fMym_BKWQzk, 2017.

[8] J. S. Culpepper, F. Diaz, and M. D. S. (editors). Report from the Third Strategic Workshop on Information Retrieval (SWIRL). 2018.

[9] F. Doshi-Velez and B. Kim. Towards a rigorous science of interpretable machine learning. *Preprint arXiv:1702.08608*, 2017.

[10] M. Drosou, H. Jagadish, E. Pitoura, and J. Stoyanovich. Diversity in big data: A review. *Big data*, 5(2):73–84, 2017.

[11] H. Gold. Trump slams Google search as 'rigged'. *CNN Money*, August 2017.

[12] S. Hajian, F. Bonchi, and C. Castillo. Algorithmic bias: From discrimination discovery to fairness-aware data mining. In *Proc. of KDD (Tutorials)*, pages 2125–2126. ACM, 2016.

[13] J. Kulshrestha, M. Eslami, J. Messias, M. B. Zafar, S. Ghosh, K. P. Gummadi, and K. Karahalios. Quantifying search bias: Investigating sources of bias for political searches in social media. In *Proc. of CSCW*, pages 417–432. ACM, 2017.

[14] K. Lippert-Rasmussen. *Born free and equal?: a philosophical inquiry into the nature of discrimination*. Oxford University Press, 2014.

[15] G. Marvin. A visual history of google ad labeling in search results. *Search Engine Land*, 2017.

[16] A. Narayanan. Tutorial: 21 definitions of fairness and their politics. Tutorial at FAT* 2018, https://www.youtube.com/watch?v=jIXIuYdnyyk, 2018.

[17] S. U. Noble. *Algorithms of Oppression: How search engines reinforce racism*. NYU Press, 2018.

[18] F. Pasquale. *The black box society: The secret algorithms that control money and information*. Harvard University Press, 2015.

[19] D. Pedreshi, S. Ruggieri, and F. Turini. Discrimination-aware data mining. In *Proc. of KDD*, pages 560–568. ACM, 2008.

[20] S. E. Robertson. The probability ranking principle in IR. *Journal of Documentation*, 33(4):294–304, 1977.

[21] A. Sinclair. Regulation of paid listings in internet search engines: A proposal for ftc action. *Boston University Journal of Science and Technology Law*, 10:353, 2004.

[22] A. Singh and T. Joachims. Equality of opportunity in rankings. In *Workshop on Prioritizing Online Content (WPOC) at NIPS*, 2017.

[23] A. Singh and T. Joachims. Fairness of exposure in rankings. In *Proc. of KDD*, pages 2219–2228, 2018.

[24] M. ter Hoeve, A. Schuth, D. Odijk, and M. de Rijke. Faithfully explaining rankings in a news recommender system. *Preprint arXiv:1805.05447*, 2018.

[25] Z. Tufekci. *Twitter and tear gas: The power and fragility of networked protest*. Yale University Press, 2017.

[26] K. Yang and J. Stoyanovich. Measuring fairness in ranked outputs. In *Proc. of SSDB*, page 22. ACM, 2017.

[27] K. Yang, J. Stoyanovich, A. Asudeh, B. Howe, H. Jagadish, and G. Miklau. A nutritional label for rankings. *arXiv:1804.07890 (pre-print)*, 2018.

[28] M. Zehlike, F. Bonchi, C. Castillo, S. Hajian, M. Megahed, and R. Baeza-Yates. FA*IR: A fair top-k ranking algorithm. pages 1569–1578, 2017.

[29] M. Zehlike and C. Castillo. Reducing disparate exposure in ranking: A learning to rank approach. *Preprint arXiv:1805.08716*, 2018.

[30] R. Zemel, Y. Wu, K. Swersky, T. Pitassi, and C. Dwork. Learning fair representations. In *International Conference on Machine Learning*, pages 325–333, 2013.