Metrics for Explainable Ranking Functions

Abraham Gale and Amélie Marian
Department of Computer Science, Rutgers University
{abraham.gale,amelie.marian}@rutgers.edu

ABSTRACT

Ranking functions are commonly used in a wide variety of applications; these functions vary in complexity from simple point systems, to traditional weighted-sums, to more complex decision trees and results from Learning-to-Rank techniques. As the community realizes the significant societal impacts of the widespread use of algorithms in decisions, there has been a push towards explainability and transparency in decision processes and results. In this paper, we focus on providing tools towards explainability and transparency of ranking functions, with a focus towards making the impacts of the ranking process understandable *a priori*, so that expectations on the outcome are known in advance. To this aim, we first design metrics to assist in making the ranking process transparent to both the decision-makers and the entities being ranked, by assessing the expected importance of each parameter used in the ranking process in the creation of the final ranked outcome, using information about the ranking functions, as well as the underlying distributions of the parameters involved in the ranking. Using our proposed metrics, we investigate algorithms to adjust and translate traditional weighted-sum functions to better reflect the intention of the decision-maker.

1 INTRODUCTION

Rankings are commonly used to make decisions and allocate resources in a wide variety of applications such as school admissions, job applications, public housing allocation, sport competition judging, organ donation lists. Decision-making techniques resulting in rankings of objects using multiples criteria have been studied for centuries [7]. However, these techniques were traditionally developed with the decision-maker's interests and constraints in mind, and did not focus on transparency and explainability of the process for the objects/individuals being affected by the outcome of the rankings.

With today's widespread use of algorithms to make decisions in an information-based society, there has been a realization that the outcomes of these algorithms have significant societal impacts, and that the algorithm designers have a responsibility to address the ethical considerations that arise when applying algorithms to individual, groups, or entities. This has been recognized by several research communities, such as Artificial Intelligence [4, 18], Machine Learning [6], and Data Management [16]. Without transparent and explainable processes, it is not possible to verify whether the outcomes satisfy ethical and fair constraints.

Traditionally, work on ranking algorithms and techniques have either assumed that the ranking function was given and satisfied some required properties, such as monotonicity, and considered the ranking function as an oracle, or have focused on designing complex query functions applicable to specific domains [7, 12]. Little attention has been given to making the ranking function itself transparent. In fact, many techniques preprocess the underlying data being ranked, typically via normalization, so that it has desirable properties for the ranking algorithm. The resulting transformation often murks the data and contributes to making the process opaque.

This paper focuses on the analysis of ranking functions and the relative impact of individual ranking metrics on the overall ranked results in order to understand the impact of the ranking process *a priori*, based on the ranking functions and data distribution. Our goal is to help decision-makers understand the behavior of their ranking functions, and to provide entities being ranked with some transparent and understandable explanation of the ranking process.

The paper makes the following contributions:

- The design of transparent and understandable metrics to clarify the ranking process, by assessing the expected importance of each parameter used in the ranking process in the creation of the final ranked outcome, using information about the ranking functions themselves, as well as observations of the underlying distributions of the parameter values involved in the ranking. (Section 2)
- Using our metrics, we propose heuristics to adjust and translate traditional weighted-sum functions to better take into account the desired importance of each parameter in the final ranking. (Section 3)

2 EXPLAINING RANKING FUNCTIONS

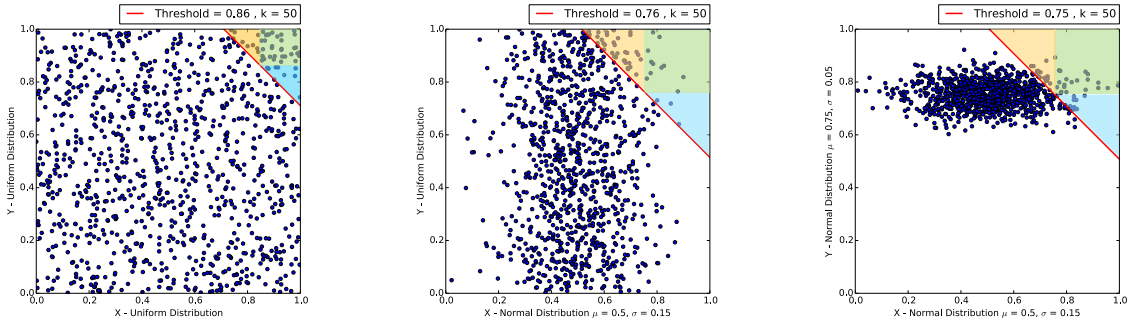
We aim at designing metrics to explain the expected behaviors of ranking functions based on the underlying distributions of the parameters involved in the rankings. To better understand the ranking mechanics, we focus our preliminary analysis on weighted-sum ranking functions, which are widely used in practice, for instance in college rankings, or student admissions.

We define our scoring function f over over a set of P ranking parameters $p_1, ..., p_P$, with weights $W_1, ..., W_P$ such that $\sum_{i=1}^{P} W_i = 1$, over an object o as $f(o) = \sum_{i=1}^{P} W_i * p_i(o)$, where $p_i(o)$ is the value of parameter p_i for object o.

Figure 1 shows the behavior of a simple weighted-sum ranking function over two parameters values $X(o)$ and $Y(o)$ (denoted X and Y for simplicity), $f(o) = 0.5X + 0.5Y$, used to identify the top-50 objects out of 1,000 objects, depending on the underlying distributions of X and Y. We can observe that the score of the top 50^{th} object (defined as the threshold at 50, red line in Figure 1)

EARS'19, July 25, 2019, Paris, France

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM.
https://doi.org/10.1145/nnnnnnn.nnnnnnn



(a) X and Y both follow uniform distributions in $[0, 1]$.

(b) X is normally distributed ($\mu = 0.5$,$\sigma = 0.15$), Y follows a uniform distribution in $[0, 1]$.

(c) X and Y are both normally distributed ($\mu_X = 0.5$, $\sigma_X = 0.15$, $\mu_Y = 0.75$, $\sigma_Y = 0.05$).

Figure 1: *Top-k thresholds (*$score = 0.5X + 0.5Y$*) based on the underlying distribution of values (N=1000, k=50),* X *and* Y *are independent variables.*

varies depending on the underlying distributions of X and Y. This in turn has an impact on the minimum score required in for each dimension (parameter) for an object to qualify as being in the top-50, which we define as the floor value.

Definition 2.1 (Threshold at k). Given a ranking function f, over a set of P ranking parameters $p_1, ..., p_P$, applied to a set of objects O, we compute the threshold value T_k as the k^{th} highest $f(o)$ value for all objects $o \in O$.

In other words, this would be the $f(o)$ value of the worst object that still makes it into the top k. For instance, if 1,000 objects are distributed uniformly in both X and Y as shown in Figure 1(a), the sum $0.5X + 0.5Y$ would follow a triangular distribution:

$$f_{(0.5X+0.5Y)}(x) = \begin{cases} x & 0 \le x \le 0.5 \\ 1 - x & 0.5 \le x \le 1 \end{cases}$$

From which we can trigonometrically estimate the value of the threshold at $k = 50$ ($95^{th} percentile$), T_{50}, of $0.5X + 0.5Y$ as 0.84.

Definition 2.2 (Parameter Floor at k). Given a threshold T_k a parameter p and a ranking function f, the floor at k of p, noted $floor_k(p)$, is the lowest value an object o' can have in p that would still allow for o' to qualify in the top-k assuming all the other values are maximized, that is for $f(o') \ge T_k$.

For instance, the floor at 50 for X if the objects are distributed uniformly in both X and Y as shown in Figure 1(a), would be:

$$floor_{50}(X) = \frac{T_{50} - W_Y}{W_X} = \frac{0.84 - 0.5}{0.5} = 0.68$$

which geometrically corresponds to the intersection between $f(o) = 0.5X + 0.5Y = T_{50} = 0.84$, and $Y = 1$.

Figures 1(a-c) show the threshold values for various underlying distributions of X and Y. Note that the actual T_{50} threshold computed over the data points of Figure 1(a), which is 0.86, slightly differs from the estimated threshold computed above because of variations in the actual data distribution. The computed $floor_{50}(X)$ for the distributions of Figures 1(a-c) are 0.72, 0.52, and 0.5, respectively. For the examples of Figure 1, the values for $floor_{50}(Y)$ are the same as $W_X = W_Y$.

We can use the floor value to define the *disqualifying power* of each parameter of the scoring function.

Definition 2.3 (Disqualifying power of a Parameter at k). Given a parameter floor $floor_k(p)$ for parameter p, the disqualifying power of p at k, $DQ_k(p)$, represents the percentage of objects $o \in O$ for which the value of o for p, $p(o)$ is lower than $floor_k(p)$. Intuitively, $DQ_k(p)$ is the percentile rank of $floor_k(p)$ in p's distribution.

The disqualifying power can be computed from the data, if available *a priori*, or estimated from knowledge of the underlying distribution. For instance, in Figure 1(b), Y is uniformly distributed in $[0, 1]$ and $floor_{50}(Y) = 0.52$, the disqualifying factor of Y at 50, $DQ_{50}(Y)$, is then estimated to be $DQ_{50}(Y) = 0.52$. Similarly, in the same Figure 1(b), X follows a normal distribution ($\mu = 0.5, \sigma = 0.15$), from which we can estimate $DQ_{50}(X) = 0.5517$ ($z\text{-}value = 0.13$). Figure 1(c) exhibits distributions that are not centered on the same values, which result in more variations in disqualifying power between X and Y: $DQ_{50}(X) = 0.5517$ as X follows the same distribution as in Figure 1(b), but $DQ_{50}(Y) \approx 0$. In fact, as we decrease k/N or increase the number of parameters in the ranking function, we are less likely to observe positive disqualifying power values.

We can define the qualifying power similarly as the percentage of objects that are qualified to belong to the top-k, by parameter p alone.

Definition 2.4 (Qualifying power of a Parameter at k). Given a threshold T_k for parameter p, the qualifying power of p at k, $Q_k(p)$, represents the percentage of objects $o \in O$ for which the value of o for p times the weight W_p, $p(o) \times W_p$ is higher than T_k.

In the examples of Figure 1, all qualifying powers are equal to 0, as no object can be part of the top-50 answer if it has a score of 0 in X (resp. Y). As k/N increases, or as the weights W_i are adjusted as we will see in the next section, the qualifying power of individual attributes will become greater than 0.

An interesting observation, from looking at the distributions of Figure 1, is that the two parameters X and Y account differently for tuples that qualify as being part of the top-k, depending on their underlying distribution. For instance in Figure 1(b), only 11 of the top-50 objects have values higher than the threshold $T_{50} = 0.76$ for *both* X and Y (green region). The rest of the 39 objects in the top-50 are qualified because one of their values (X or Y) compensates for a lower value in the other parameter. For these distributions, most of the remaining objects qualify thanks to a high value of Y (35 objects, orange region), whereas only 3 objects qualify thanks to a high value of X (blue region). For these particular distributions of X and Y, we see that for $k = 50$, Y dominates the ranking, despite the underlying scoring function $f = 0.5X + 0.5Y$ giving the same importance to both X and Y. We can compute the relative importance of X and Y in Figure 1(b) for f, as $I_k(X) = (11+3)/50 = 0.28$, and $I_k(Y) = (11 + 35)/50 = 0.92$. We define the importance of a parameter as:

Definition 2.5 (Importance of a Parameter at k). Given a ranking function f, over a set of P ranking parameters $p_1, ..., p_P$, applied to a set of objects $o \in O$, and a threshold value T_k, we compute $I_k(p)$, the importance of a parameter p at k, as the percentage of objects in the top-k answers (i.e, with $f(o) \geq T_k$) such that the value $p(o) \geq T_k$. If we only have distributions and not values this can be expressed by the conditional probability $\Pr(p(o) \geq T_k \mid f(o) \geq T_k)$

In Figure 1(c), we can see that the relative importance of X and Y is more balanced, with 20 objects in the green region, 19 in the orange region, and 11 in the blue region, resulting in importance values: $I_{50}(X) = (20 + 11)/50 = 0.62$, and $I_{50}(Y) = (20 + 19)/50 = 0.78$. The independent uniform distributions of Figure 1(a) result in equal importance for X and Y $I_{50}(X) = I_{50}(Y) = (8 + 34)/50 = 0.84$ with 34 objects in the common green region and 8 objects each in the orange and blue regions.

Importance of a parameter p expresses the percentage of objects that dominate an idealized object o' that would be exactly on the threshold, with all parameter values equal to the threshold, for p. If p's value falls behind this object for many other objects in the top-k answer, it follows that objects are being selected as part of the top-k despite their low values for p. On the other hand, if values of p almost always exceed the value of p for o', we see that p is contributing to these objects' selections, making p an important parameter in the ranking.

3 ADJUSTING RANKING FUNCTIONS

Armed with the knowledge derived from the analysis of ranking functions behavior, we now turn our focus to the design of ranking functions that better fit the needs of the decision-makers while still being understandable *a priori* to the targeted audience (public, entities, or applicants being ranked).

As discussed in the previous Section, the underlying distribution of data has an impact on how much each parameter contributes to the final ranking. Figure 1 illustrated this by showing that, despite sharing the same ranking function, three different distributions of X and Y can lead to significant difference in the importance (Definition 2.5) of each of these parameters in the k top answers. In that example, the ranking processes used an equal weight ranking function, $f(o) = 0.5X + 0.5Y$, however, the final ranking was more influenced by the value of the parameter Y for object o, for the distributions of Figure 1(b) and (c). Intuitively, it seems reasonable to assume that the intention of the decision-maker by using an equal weight ranking function was for both parameters to contribute equally to the top-k outcome. One possibility to achieve that goal is to adjust the ranking function weights W_X and W_Y so that the resulting importance ratio of parameters matches that envisioned by the decision-maker (equal importance in our example).

Figure 2 shows resulting adjustments for the distributions of Figure 1. Visually, we can see that the slope of the ranking function is adjusted so that the number of points in the orange and blue region are equal. In some cases, such as Figure 2(b), this requires adjusting the T_k value. The new threshold line is shown in green and corresponds to $f(o) = 0.78X + 0.22Y$ for Figure 2(b), and $f(o) = 0.56X + 0.44Y$ for Figure 2(c). The corresponding values for $I_{50}(X)$ and $I_{50}(Y)$ become $I_{50}(X) = I_{50}(Y) = 0.72$ for Figure 2(b) and $I_{50}(X) = I_{50}(Y) = 0.7$ for Figure 2(c). The change in weight therefore compensates for the distribution skew and ensures that the intent of the ranking function designer is preserved. The importance of X and Y for the uniform distributions of Figure 1(a) were equal, so the weights are unchanged.

Notice that the adjustment of the ranking function of Figure Figure 2(b) result is some objects being qualified on the X value alone, as $W_X > T_{50} = 0.71$. The qualifying power (Definition 2.4) $Q_{50}(X)$ is then equal to 1 minus the percentile-rank of $(T_{50}/W_X = 0.71/0.78 = 0.91)$ in X distribution: $Q_{50}(X) = 0.0032$.

Figure 3 shows how the importance of X changes as a function of the weight of the parameter X in the ranking function for the data distributions of Figure 2(b).

In the general case, the weights can be adjusted to return desired values of importance by selecting the desired values of importance Ip for each relevant parameter, then creating a function which determines the loss for each set of weights as:

$$(DesiredWeight - \frac{I(X)}{\sum_{p_0}^{p_n} I(p)})^2$$

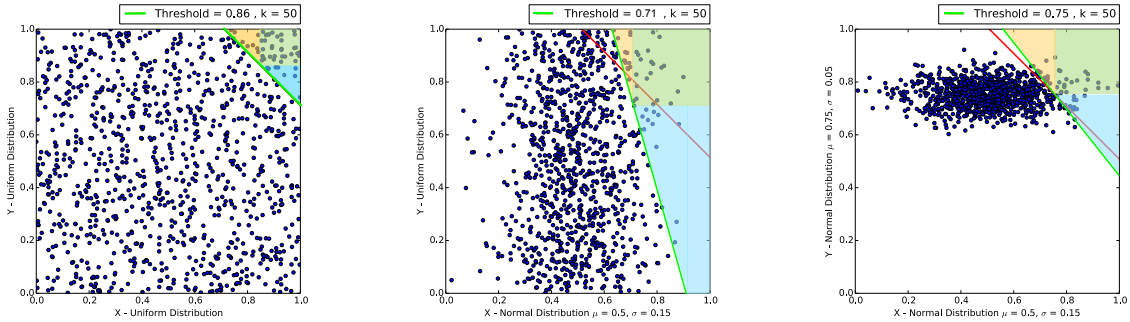
Then a standard optimization algorithm can be run to choose the weights that get closest to the desired importance. In the case where the distributions are given instead of the points, this minimum can often be calculated numerically.

The resulting desired weights would reflect the intention of the decision-maker by taking into account the underlying data distribution and assigning weights accordingly.

Our adjustment function could be combined with additional constraints. Consider a scenario where in addition to X and Y being of comparable desired importance, the decision maker also requires both X and Y to be above a minimum value. For instance, in the distribution of Figure 2(b), if both X and Y are required to be above 0.5 (e.g., a candidate needs a passing grade in both X and Y), then the adjusted ranking function needs to take into account only the points in the upper right quadrant of the figure, and the loss function should be adjusted accordingly.

4 RELATED WORK

Ranking functions have been widely studied in the literature. The Multi Criteria Decision Analysis community focuses on making



(a) X and Y both follow uniform distributions in $[0, 1]$.

(b) X is normally distributed ($\mu = 0.5$,$\sigma = 0.15$), Y follows a uniform distribution in $[0, 1]$.

(c) X and Y are both normally distributed ($\mu_X = 0.5$, $\sigma_X = 0.15$, $\mu_Y = 0.75$, $\sigma_Y = 0.05$).

Figure 2: *Function adjustments of the ranking functions of Figure 1*

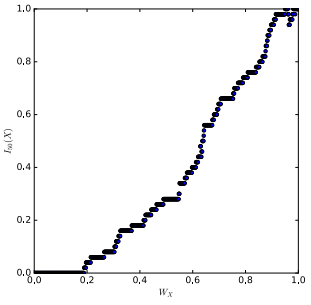


Figure 3: *Importance of X at $k = 50$ as a function of W_X for the data distributions of Figure 2(b)*

decisions from a set of possibly conflicting criteria [7, 19]. The techniques are typically aimed at experts, and provide complex computation, often hidden in black-box algorithms, with little possibility of explanation. Ranking functions are also widely used in Information Retrieval [12, 14]. More recently, the Information Retrieval community has focused on learning-to-rank approaches [10, 11]. However such techniques produce complex ranking functions, that are impossible to explain to a non-expert.

Several measures have been proposed to compare the outcomes of ranking processes. The Spearman ρ [15], and Kendall τ [9] are the most commonly used metrics to compute rank correlation. More recently, distance measures have been proposed by the Database community [5]. These focus on comparing the outputs of ranking processes. In contrast, we focus on the behavior of the ranking functions before the ranking takes place, by analysis the impact of different data distributions on the ranking functions.

Recently, there has been a lot of discussion in the research community and in the media on the impact of algorithms in societal issues and on the inherent bias in many algorithms, including ranking algorithms. Recent work have looked at how to include fairness and diversity into ranking techniques [1, 13, 21]. Our work is complementary to these approaches: by providing tools to explain and understand ranking processes, we can design more ethical ranking functions.

Explainability and transparency have been at the forefront of many works in Artificial Intelligence (e.g., [3]) and Machine Learning (e.g., [20]). This has been driven in part by political regulations that call for "right to explanation" [8]. Work that aim to explain rankings have mostly focused on *a posteriori* explanations of the results. Most of these work focus on feature selection to explain the contribution of each individual features to the final ranking outcome, in a process similar to sensitivity analysis [2, 17]. In contrast, we focus on making the process and parameter importance transparent so that the information is shared *a priori*.

5 CONCLUSIONS AND FUTURE WORK

We proposed a set of metrics to explain the expected behaviors of ranking processes. Using these metrics, we proposed techniques to adjust the weights of ranking functions to better match the intention of the decision-maker. Our work has focused on weighted-sum ranking functions; we plan to investigate a wider range of functions, including step functions, non-linear functions, decision trees and the output of learning-to-rank algorithms. In addition, we plan to extend and deepen our analysis to investigate how these metrics behave in a wider range of situations: varying k, increasing the number of ranking parameters P, and varying the data distributions (atypical distributions, correlation, or partial distribution information). Of course, we may not have access, or knowledge, of the full real-world data distribution. We will need to study the behavior of ranking functions in the presence of atypical distributions, or partial information on the underlying distributions (e.g., coarse histograms).

This work aims at providing an understanding on the impact of individual parameters in the ranking process in order to assist decision-makers in designing their functions to reflect their goals, and to provide explanations to the entities being ranked. In particular, our proposed importance metric, can be seen as an explanation of the relative impact of the parameters on the final ranked outcome, which may be different from the impact on the score itself.

REFERENCES

[1] L Elisa Celis, Damian Straszak, and Nisheeth K Vishnoi. 2017. Ranking with fairness constraints. *arXiv preprint arXiv:1704.06840* (2017).

[2] Shay B Cohen, Eytan Ruppin, and Gideon Dror. 2005. Feature Selection Based on the Shapley Value.. In *IJCAI*, Vol. 5. 665–670.

[3] Mark G Core, H Chad Lane, Michael Van Lent, Dave Gomboc, Steve Solomon, and Milton Rosenberg. 2006. Building explainable artificial intelligence systems. In *AAAI*. 1766–1773.

[4] Virginia Dignum. 2018. Ethics in artificial intelligence: introduction to the special issue. *Ethics and Information Technology* 20, 1 (01 Mar 2018), 1–3. https://doi.org/10.1007/s10676-018-9450-z

[5] Ronald Fagin, Ravi Kumar, and Dakshinamurthi Sivakumar. 2003. Comparing top k lists. *SIAM Journal on discrete mathematics* 17, 1 (2003), 134–160.

[6] FATML [n. d.]. Fairness, Accountability, and Transparency in Machine Learning. ([n. d.]). http://www.fatml.org/.

[7] J. Figueira, S. Greco, and M. Ehrgott. 2005. *Multiple Criteria Decision Analysis: State of the Art Surveys*. Springer. https://books.google.com/books?id=YqmvlTiMNqYC

[8] Bryce Goodman and Seth Flaxman. 2016. European Union regulations on algorithmic decision-making and a" right to explanation". *arXiv preprint arXiv:1606.08813* (2016).

[9] Maurice George Kendall. 1948. Rank correlation methods. (1948).

[10] Hang Li. 2011. A short introduction to learning to rank. *IEICE TRANSACTIONS on Information and Systems* 94, 10 (2011), 1854–1862.

[11] Tie-Yan Liu et al. 2009. Learning to rank for information retrieval. *Foundations and Trends® in Information Retrieval* 3, 3 (2009), 225–331.

[12] Christopher D. Manning, Prabhakar Raghavan, and Hinrich SchÃijtze. 2008. *Introduction to Information Retrieval*. Cambridge University Press.

[13] Filip Radlinski, Robert Kleinberg, and Thorsten Joachims. 2008. Learning diverse rankings with multi-armed bandits. In *Proceedings of the 25th international conference on Machine learning*. ACM, 784–791.

[14] Gerard Salton and Christopher Buckley. 1988. Term-weighting approaches in automatic text retrieval. *Information processing & management* 24, 5 (1988), 513–523.

[15] Charles Spearman. 1904. The proof and measurement of association between two things. *The American journal of psychology* 15, 1 (1904), 72–101.

[16] Julia Stoyanovich, Serge Abiteboul, and Gerome Miklau. 2016. Data, responsibly: Fairness, neutrality and transparency in data analysis. In *International Conference on Extending Database Technology*.

[17] Maartje ter Hoeve, Anne Schuth, Daan Odijk, and Maarten de Rijke. 2018. Faithfully Explaining Rankings in a News Recommender System. *arXiv preprint arXiv:1805.05447* (2018).

[18] Jim Torresen. 2018. A Review of Future and Ethical Perspectives of Robotics and AI. *Frontiers in Robotics and AI* 4 (2018), 75.

[19] Mark Velasquez and Patrick T Hester. 2013. An analysis of multi-criteria decision making methods. *International Journal of Operations Research* 10, 2 (2013), 56–66.

[20] Leanne S Whitmore, Anthe George, and Corey M Hudson. 2018. Explicating feature contribution using Random Forest proximity distances. *arXiv preprint arXiv:1807.06572* (2018).

[21] Ke Yang and Julia Stoyanovich. 2017. Measuring fairness in ranked outputs. In *Proceedings of the 29th International Conference on Scientific and Statistical Database Management*. ACM, 22.