Pearson Rank: A Head-Weighted Gap-Sensitive Score-Based Correlation Coefficient

Ning Gao, Mossaab Bagdouri, and Douglas W. Oard
University of Maryland, College Park
{ninggao, mossaab, oard}@umd.edu

ABSTRACT

One way of evaluating the reusability of a test collection is to determine whether removing the unique contributions of some system would alter the preference order between that system and others. Rank correlation measures such as Kendall's τ are often used for this purpose. Rank correlation measures are appropriate for *ordinal* measures in which only preference order is important, but many evaluation measures produce system scores in which both the preference order and the magnitude of the score difference are important. Such measures are referred to as *interval*. Pearson's ρ offers one way in which correlation can be computed over results from an interval measure such that smaller errors in the gap size are preferred. When seeking to improve over existing systems, we care the most about comparisons among the best systems. For that purpose we prefer head-weighed measures such as τ_{AP}, which is designed for ordinal data. No present head weighted measure fully leverages the information present in interval effectiveness measures. This paper introduces such a measure, referred to as Pearson Rank.

Keywords

Evaluation Metric; Correlation Coefficient

1. INTRODUCTION

In information retrieval evaluation, we often wish to compare the effectiveness of alternative systems by using some single-valued evaluation metric (e.g., F_1 or Mean Average Precision) on some publicly available collections. Judgments of all the items in the collections are required to get the ground-truth evaluation of the systems, which is often infeasible. Hence, sampling or pooling techniques are used in shared tasks (e.g., TREC, CLEF) to choose the documents that will be assessed. Naturally, we want to quantify the adequacy of test collections created this way for assessing the effectiveness of different systems, especially systems that did not participate to the creation of the pool [5]. We might also want to know whether we can approximate the effectiveness of these systems by reducing the number of relevance judgments [1]. We can attempt to answer these questions by measuring the correlation between two ranked lists of system scores (e.g., the reference and the approximated).

When making comparisons in the aforementioned cases, we focus on two considerations. First, we prefer the comparison to be on an interval rather than ordinal scale. In shared tasks, the systems are ranked according to their measured effectiveness on a test collection. In addition to system ranks, it is important to know whether some systems are substantially better than the others. Moreover, for formative evaluation it can be useful to characterize small incremental improvements, even when the new evaluation scores do not alter a system's rank with regard to other systems. For this reason, we prefer a correlation coefficient that is sensitive to the size of the gaps between system scores. Second, we care more about comparisons among the best systems, so we prefer a correlation coefficient that is influenced more by differences near the top of a ranked list of systems. We call such measures head-weighted.

Several statistics have been proposed to quantify the correlation between two ranked lists of scores. Pearson's ρ [4] assumes that the scores are on an *interval* scale (i.e., one in which all score differences of the same size have the same meaning). Mean Average Precision can properly be treated as being on an interval scale because precision is *interval* and expected values computed on *interval* scales are *interval*. Kendall's τ [3] and Yilmaz et al.'s τ_{AP} [6], by contrast, make the weaker assumption of an *ordinal* scale in which score differences are not necessarily informative, but the relative ordering of systems by those scores is informative.

Gao and Oard suggest a head-weighted gap-sensitive correlation coefficient called τ_{GAP} [2], giving greater penalty to larger gaps when a swap occurs, but assigning no penalty for misestimating the gap size if no swap occurs. Consider, for example, the situation depicted in Figure 1, which occurred in an unpublished system-ablation study in which we were studying the reuse of a test collection by systems that did not contribute to a judgment pool. The x-axis is the reference system score (measured in this case as mean precision at rank one) obtained using relevance judgments for the unique contributions of every system, while the y-axis is the score that we estimated for that same system without using any relevance judgments for documents uniquely contributed by that system. As Pearson's $\rho = 0.89$ shows, the estimates exhibit a strong linear correlation with the reference scores. Note, however, the substantial gap in the reference scores between the best and second-best systems; the estimate does not preserve that gap. Pearson's correlation coefficient is dominated by the many other well approximated gaps, however, and does not emphasize the problem. Kendall's $\tau = 0.74$ is affected by several swaps in the preference order, but as the high value of ρ indicates, most of these swaps are among systems with small gaps that should not trouble us much. There are several reversals near the top of the ranked list that affect $\tau_{AP} = 0.64$, although

SIGIR '16, July 17 - 21, 2016, Pisa, Italy

ISBN 978-1-4503-4069-4/16/07...$15.00

DOI: http://dx.doi.org/10.1145/2911451.2914728

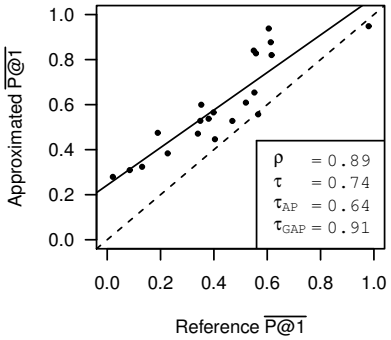


Figure 1: Different correlation coefficient values when the gap between the best system and a lower one is mischaracterized.

as $\tau_{GAP} = 0.91$ shows, these reversals are among systems that had very small differences in the reference condition.

Intuitively, this seems like the wrong answer, since if we were to make a system that was considerably better than the second-best one, we have no reason to believe that the effectiveness score estimates on the y-axis would reflect that. None of the standard measures capture the fact that the difference we should care the most about—the large gap in the reference system scores between the best two systems, is completely mischaracterized by the approximated scores as being of only negligible gap size. Pearson's ρ misses this because it is not head-weighted, τ_{AP} misses this because it is not gap-sensitive, and τ_{GAP} misses it because it is gap-sensitive only when a swap occurs. None of these measures focus on what we care about the most when we have an evaluation measure that is *interval*, which is that the size of the gaps among the best systems be correctly estimated.

This example illustrates the need for a new correlation coefficient that is at the same time head weighted and sensitive to both swapped and unswapped gaps. Section 2 introduces Pearson Rank (ρ_r), our novel correlation coefficient, and shows that it has several desirable properties. Through extensive simulation, Section 3 contrasts some behaviors of ρ_r with those of rank-based correlation coefficients. We conclude in Section 4.

2. PEARSON RANK

This section defines Pearson Rank, demonstrates its satisfaction of desired properties, and contrasts it with other correlations.

2.1 Definition

Let $S = \{s_1, \cdots, s_m\}$ be a list of m items ranked in descending order by their reference scores $X = \{x_1, \cdots, x_m\}$; and $Y = \{y_1, \cdots, y_m\}$ be their approximated scores, after scaling all of the scores so that $x_i, y_i \in [0,1]$. We define the new (asymmetric) Pearson Rank correlation coefficient (ρ_r) X and Y as:

$$\rho_r(Y|X) = \frac{1}{\sum_{i=2}^{m} x_i} \cdot \sum_{i=2}^{m} x_i \cdot \frac{\sum_{j=1}^{i-1}(x_j - x_i)\cdot(y_j - y_i)}{\sqrt{\sum_{j=1}^{i-1}(x_j - x_i)^2 \sum_{j=1}^{i-1}(y_j - y_i)^2}}. \quad (1)$$

Figure 2 is a toy example for calculating ρ_r of two lists, where $X = \{x_1, x_2, x_3\}$ and $Y = \{y_1, y_2, y_3\}$. The solid arrows show the score differences between item pairs considered when calculating ρ_r at x_2. The value of ρ_r at x_2 is:

$$\rho_{r,2} = \frac{x_2}{x_2 + x_3} \cdot \frac{(x_1 - x_2)(y_1 - y_2)}{\sqrt{(x_1 - x_2)^2 (y_1 - y_2)^2}}.$$

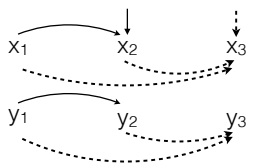


Figure 2: Example for calculating ρ_r correlation coefficient.

The dotted arrows show the score differences between item pairs considered for ρ_r at x_3. The value of ρ_r at x_3 is:

$$\rho_{r,3} = \frac{x_3}{x_2 + x_3} \cdot \frac{(x_1 - x_3)(y_1 - y_3) + (x_2 - x_3)(y_2 - y_3)}{\sqrt{(x_1 - x_3)^2 + (x_2 - x_3)^2}\sqrt{(y_1 - y_3)^2 + (y_2 - y_3)^2}}.$$

The value of ρ_r is the sum of $\rho_{r,i}$ at items from 2 to m = 3.

2.2 Properties

THEOREM 1. *The value of ρ_r is always between -1 and 1. $\rho_r = 1$ if and only if $Y = X$; and $\rho_r = -1$ if and only if $Y = 1 - X$.*

PROOF. Due to the Cauchy-Schwarz inequality, we have:

$$\left(\sum_{j=1}^{i-1}(x_j - x_i)\cdot(y_j - y_i)\right)^2 \le \sum_{j=1}^{i-1}(x_j - x_i)^2 \sum_{j=1}^{i-1}(y_j - y_i)^2, \quad (2)$$

where the two sides are equal if and only if X and Y are linearly dependent (or, in a geometrical sense, they are parallel). Since the scores are scaled $x_i, y_i \in [0,1]$, we have

$$-1 \le \frac{\sum_{j=1}^{i-1}(x_j - x_i)\cdot(y_j - y_i)}{\sqrt{\sum_{j=1}^{i-1}(x_j - x_i)^2 \sum_{j=1}^{i-1}(y_j - y_i)^2}} \le 1, \quad (3)$$

where the upper bound 1 is reached when $Y = X$, and lower bound -1 is reached when $Y = 1 - X$. Therefore,

$$-1 = \sum_{i=2}^{m} \frac{-x_i}{\sum_{i=2}^{m} x_i} \le \rho_r \le \sum_{i=2}^{m} \frac{x_i}{\sum_{i=2}^{m} x_i} = 1. \quad (4)$$

□

THEOREM 2. *Let $X = \{x_1, \cdots, x_{i-1}, x_i, \cdots, x_{p-1}, x_p, \cdots, x_m\}$ be a reference score list; $Y^1 = \{x_1, \cdots, x_i, x_{i-1}, \cdots, x_{p-1}, x_p, \cdots, x_m\}$ and $Y^2 = \{x_1, \cdots, x_{i-1}, x_i, \cdots, x_p, x_{p-1}, \cdots, x_m\}$ be two lists of approximated scores, where Y^1 has only one swapped adjacent item pair s_{i-1} and s_i near the head of the lists; Y^2 has only one swapped adjacent item pair s_{p-1} and s_p near the end of the list. If the score difference between the swapped pairs are identical $(x_{i-1} - x_i) = (x_{p-1} - x_p)$, the ρ_r score for Y^1 with error near the head will be lower than Y^2 with error near the end.*

PROOF. Let $\varepsilon = 1 - \rho_r$ be the loss of a list of approximated scores, therefore, $\varepsilon = \sum_{i=2}^{m} \varepsilon_i$ where ε_i is the loss for each position from 2 to m. For Y^1, we have $\varepsilon_2 = \cdots = \varepsilon_{i-2} = 0$;

$$\begin{aligned}\varepsilon_{i-1} &= \frac{x_{i-1}}{\sum_{i=2}^{m} x_i} \cdot \left(1 - \frac{\sum_{j=1}^{i-2}(x_j - x_{i-1})(y_j - y_{i-1})}{\sqrt{\sum_{j=1}^{i-2}(x_j - x_{i-1})^2 \sum_{j=1}^{i-2}(y_j - y_{i-1})^2}}\right) \\ &= \frac{x_{i-1}}{\sum_{i=2}^{m} x_i} \cdot \left(1 - \frac{\sum_{j=1}^{i-2}(x_j - x_{i-1})(x_j - x_i)}{\sqrt{\sum_{j=1}^{i-2}(x_j - x_{i-1})^2 \sum_{j=1}^{i-2}(x_j - x_i)^2}}\right).\end{aligned} \quad (5)$$

Let $n = x_j - x_{i-1}$ and $c = x_{i-1} - x_i$, the derivative of

$$\frac{\sum_{j=1}^{i-2}(x_j - x_{i-1})(x_j - x_i)}{\sqrt{\sum_{j=1}^{i-2}(x_j - x_{i-1})^2 \sum_{j=1}^{i-2}(x_j - x_i)^2}} \tag{6}$$

could be represented in an alternative form of

$$\frac{\partial\left(\frac{\sum n(n+c)}{\sqrt{\sum n^2 \sum (n+c)^2}}\right)}{\partial n} = \frac{3((1+2n)(1+2n(2+n))+c(3-4n^2(2+n)))}{(\sqrt{n}((2+n)(1+2n)(3+2n))^{3/2})}, \tag{7}$$

which is positive when both n and c are positive. Therefore, the value of equation (6) increases with increasing i; and thus the value of ε_{i-1} decreases with increasing i. Similarly,

$$\begin{aligned}\varepsilon_i &= \frac{x_i}{\sum_{i=2}^{m} x_i} \cdot \left(1 - \frac{\sum_{j=1}^{i-1}(x_j - x_i)(y_j - y_i)}{\sqrt{\sum_{j=1}^{i-1}(x_j - x_i)^2 \sum_{j=1}^{i-1}(y_j - y_i)^2}}\right) \\ &= \frac{x_i}{\sum_{i=2}^{m} x_i} \cdot \frac{2(x_{i-1} - x_i)^2}{\sqrt{\sum_{j=1}^{i-1}(x_j - x_i)^2 \sum_{j=1}^{i-1}(x_j - x_{i-1})^2}}.\end{aligned} \tag{8}$$

As can be seen, the value of ε_i decreases with increasing i. The penalty for $\varepsilon_{t\in[i+1,m]}$

$$\begin{aligned}\varepsilon_{t\in[i+1,m]} &= \frac{x_t}{\sum_{i=2}^{m} x_i} \cdot \left(1 - \frac{\sum_{j=1}^{t-1}(x_j - x_t)(y_j - y_t)}{\sqrt{\sum_{j=1}^{t-1}(x_j - x_t)^2 \sum_{j=1}^{t-1}(y_j - y_t)^2}}\right) \\ &= \frac{x_t}{\sum_{i=2}^{m} x_i} \cdot \left(1 - \frac{\sum_{j\in[1,n],j\neq(i-1,i)}(x_j - x_t)^2 + 2(x_{i-1} - x_t)(x_i - x_t)}{\sqrt{\sum_{j=1}^{t-1}(x_j - x_t)^2 \sum_{j=1}^{t-1}(x_j - x_t)^2}}\right) \\ &= \frac{x_t}{\sum_{i=2}^{m} x_i} \cdot \frac{(x_{i-1} - x_i)^2}{\sum_{j=1}^{t-1}(x_j - x_t)^2}\end{aligned} \tag{9}$$

depends only t. Therefore, the value of $\varepsilon = \varepsilon_{i-1} + \varepsilon_i + \sum_{i+1}^{m} \varepsilon_t$ is larger when the swapped error appears in the head of the list. □

THEOREM 3. *Let $X = \{x_1, \cdots, x_{i-1}, x_i, \cdots, x_m\}$ be a reference score vector, and $Y = \{x_1, \cdots, x_i, x_{i-1}, \cdots, x_m\}$ be an approximation list with one adjacent swapped pair x_{i-1} and x_i, then the value of ρ_r is inversely related to the score difference between the swapped pair $(x_{i-1} - x_i)$.*

PROOF. The derivative of $\frac{\sum n(n+c)}{\sqrt{\sum n^2 \sum (n+c)^2}}$ with respect to c

$$\frac{\partial\left(\frac{\sum n(n+c)}{\sqrt{\sum n^2 \sum (n+c)^2}}\right)}{\partial c} = \frac{3c(1+n)(1+2n)(1-n^2)}{((1+n)(1+2n)(1+6c+6c^2+3n+6cn+2n^2))^{1.5}} \tag{10}$$

is negative. Therefore, from equations (5) and (7), we can conclude that the value of ε_{i-1} is positively related to the value of $c = x_{i-1} - x_i$. The derivative of $\frac{2c^2}{\sqrt{\sum n^2 \sum (n+c)^2}}$ with respect to c

$$\frac{\partial\left(\frac{2c^2}{\sqrt{\sum n^2 \sum (n+c)^2}}\right)}{\partial c} = \frac{12cn^2(n+1)(2n+1)(6c^2+9cn+9c+4n^2+6n+2)}{[n^2(2n^2+3n+1)(6c^2+6cn+6c+2n^2+3n+1)]^{1.5}} \tag{11}$$

is positive. Therefore, from equation (8) and (11) we can conclude that ε_i is positively related to $c = x_{i-1} - x_i$. Similarly, equation (9) also shows that the value of ε_t is positively related to $c = x_{i-1} - x_i$. As a conclusion, when excluding the effect of item weights $(x_i / \sum_{i=2}^{m} x_i)$, the approximation list is penalized more when swapping item pairs with larger score differences. □

Table 1: Property comparisons.

	Ordinal	Interval	Head Weighted	Symmetric
ρ	yes	yes	no	yes
τ	yes	no	no	yes
τ_{AP}	yes	no	yes	no
τ_{GAP}	yes	yes	yes	no
ρ_r	yes	yes	yes	no

2.3 Analysis

Table 1 shows the properties of our proposed ρ_r and the other correlation measures. As can be seen, all the measures can be applied with *ordinal* effectiveness measures, but only τ_{GAP}, ρ and ρ_r leverage the fact that typical information retrieval effectiveness measures are *interval*. τ_{GAP}, however, is a hybrid that treats the reference scores as *interval* but the approximated scores as *ordinal*. τ_{AP}, τ_{GAP} and ρ_r are head-weighted measures giving more weights to the items at the head of the list. As a result, all the head-weighted measures are asymmetric, yielding different results if the roles of the reference and the approximation are swapped. If a symmetric measure is desired, the results in both directions could be averaged. For example, we could define $\rho_r(X,Y) = (\rho_r(X|Y) + \rho_r(Y|X))/2$.

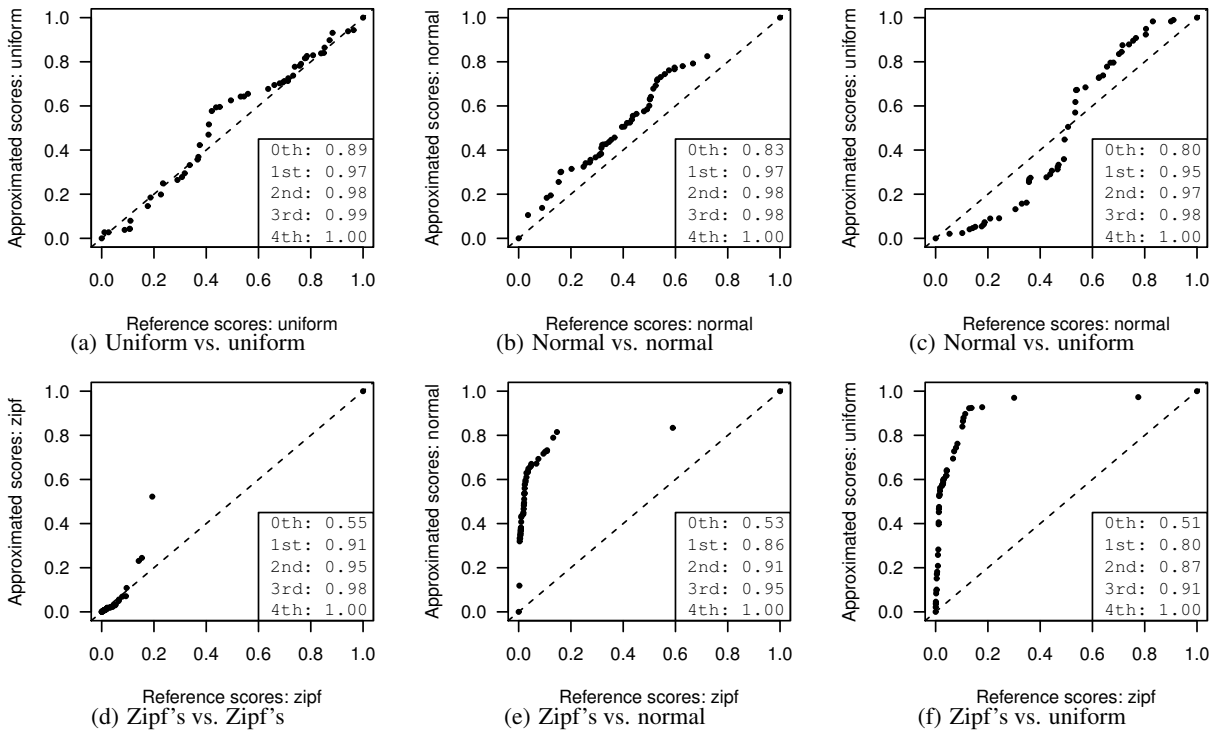
3. SIMULATION EXPERIMENTS

In Section 2.2 we proved some properties of ρ_r when a swap occurs between two items. Another aspect we explore next is its reaction to score changes that maintain the ranking of the systems.

One way of characterizing the behavior of ρ_r is to simulate cases in which the rank of each system is held invariant but the gap sizes are allowed to differ in some systematic way. When we do this, the rank correlation coefficients we have considered (τ, τ_{AP}, τ_{GAP}) show a perfect correlation of 1. Some approximations are better than others, however, and we can use simulation to characterize the behavior of ρ_r in such cases. We assume that the reference and approximated scores follow some distributions. To implement this condition, we sample N systems from the distribution of the reference scores $\mathscr{X}$, and sort them in a descending order of scores. We scale these scores to the interval $[0,1]$ by subtracting the minimum score before dividing by the difference between the maximum and minimum scores, yielding a vector X. In a similar manner, we generate a vector of the ranked scaled approximated scores Y, from a distribution $\mathscr{Y}$. We compute the ρ_r for this pair of vectors, and repeat this process 100,000 times.

We study three different distributions, but others could have been considered as well. In the uniform distribution ($\mathscr{U}_{[0,1]}$), the gaps between different systems are equal (in expectation). In the normal distribution ($\mathscr{N}_{(0.5,1)}$), a few systems would have a score close to each end of the ranked list, while the majority of the systems would be located around the middle of the score distribution. With Zipf's distribution ($\mathscr{Z}_{(2^{31}-1,2)}$), a few systems would have a score close to the head of the ranked list, while most of the systems would be clustered towards its tail. Each scatter plot in Figure 3 corresponds to the median ρ_r, for a pair of distributions, out of the 100,000 pairs of score vectors corresponding to 50 simulated systems. We also indicate in that figure the different quartiles of the values of ρ_r.

We first observe that ρ_r is more likely to have a high value when both of the reference and approximated scores follow the uniform distribution (Figure 3(a)). In fact, the differences between the scores are equal in expectation. The next highest ρ_r values appear to be those of the normal distribution (Figure 3(b)). The gaps there are bigger (in expectation) closer to either end. Thus, the penalty (which is weighted towards the head of the list) is expected to be



(a) Uniform vs. uniform (b) Normal vs. normal (c) Normal vs. uniform

(d) Zipf's vs. Zipf's (e) Zipf's vs. normal (f) Zipf's vs. uniform

Figure 3: Pearson Rank quartile values for simulated scores of 50 systems, and actual simulated scores for the median case.

larger than that of the uniform distribution. Figure 3(c) shows lower values of Pearson Rank. In this case, the reference scores are drawn from the normal distribution, which means there is (in expectation) a high gap towards either end of the reference scores. However, these gaps are lost in the approximated scoring space, as successive scores are expected to be equally distant.

We now turn to the bottom row of Figure 3, where at least one of the axes follows Zipf's distribution. In this distribution, we expect that large gaps would appear only close to the head of the ranked list. Figure 3(d) shows that ρ_r can have a low value of 0.55 and a median of 0.95, even when both the reference and approximated scores are drawn from an identical distribution. This can be explained by the high variance of the gaps near the head of the scores. In Figure 3(e) we observe even lower ρ_r scores (e.g., the median value is 0.91). Although, both of the normal and Zipf's distributions have a tendency for high gaps close to the head of the score list, those of the Zipf's are (in expectation) much larger. Thus, the gaps appear to get lost when the Zipf's scores are "converted" into normal ones. Finally, the worst ρ_r values are shown in Figure 3(f), when very large gaps are expected to appear at the head of the reference scores, while those of the approximated scores are expected to be much smaller. In this case, we observe the lowest ρ_r value of 0.51, and a median of 0.87.

4. CONCLUSION

We have proposed a novel head-weighted gap-sensitive score-based correlation coefficient ρ_r. By construction, ρ_r gives more weight to the items with higher reference scores. We have also shown that ρ_r more severely penalizes the swaps occurring near the head of the list, and those with larger reference score gaps. Simulation experiments illustrate the sensitivity of ρ_r to score values, even when (as happened by construction in our simulation) rank-based metrics fail to detect any difference between reference and approximated scores due to the consistent ranking of all systems.

ACKNOWLEDGMENT

This work was made possible by NSF award 1065250, and NPRP grant # NPRP 6-1377-1-257 from the Qatar National Research Fund (a member of Qatar Foundation). The statements made herein are solely the responsibility of the authors.

References

[1] B. Carterette. Robust test collections for retrieval evaluation. In *SIGIR*, pages 55–62, 2007.

[2] N. Gao and D. Oard. A head-weighted gap-sensitive correlation coefficient. In *SIGIR*, pages 799–802, 2015.

[3] M. G. Kendall. A new measure of rank correlation. *Biometrika*, pages 81–93, 1938.

[4] K. Pearson. Note on regression and inheritance in the case of two parents. *Proceedings of the Royal Society of London*, 58(347-352):240–242, 1895.

[5] E. M. Voorhees. Variations in relevance judgments and the measurement of retrieval effectiveness. *Information processing & management*, 36(5):697–716, 2000.

[6] E. Yilmaz et al. A new rank correlation coefficient for information retrieval. In *SIGIR*, pages 587–594, 2008.