When Rank Order Isn't Enough: New Statistical-Significance-Aware Correlation Measures

Mucahid Kutlu[⋆], Tamer Elsayed[⋆], Maram Hasanain[⋆], Matthew Lease[♦]
[⋆]Qatar University [♦]University of Texas at Austin
{mucahidkutlu,telsayed,maram.hasanain}@qu.edu.qa
ml@utexas.edu

ABSTRACT

Because it is expensive to construct test collections for Cranfield-based evaluation of information retrieval systems, a variety of lower-cost methods have been proposed. The reliability of these methods is often validated by measuring rank correlation (e.g., Kendall's τ) between known system rankings on the full test collection vs. observed system rankings on the lower-cost one. However, existing rank correlation measures do not consider the statistical significance of score differences between systems in the observed rankings. To address this, we propose two statistical-significance-aware rank correlation measures, one of which is a head-weighted version of the other. We first show empirical differences between our proposed measures and existing ones. We then compare the measures while benchmarking four system evaluation methods: pooling, crowdsourcing, evaluation with incomplete judgments, and automatic system ranking. We show that use of our measures can lead to different experimental conclusions regarding reliability of alternative low-cost evaluation methods.

KEYWORDS

Rank Correlation; Evaluation; IR System Ranking.

ACM Reference Format:
Mucahid Kutlu, Tamer Elsayed, Maram Hasanain, Matthew Lease. 2018. When Rank Order Isn't Enough: New Statistical-Significance-Aware Correlation Measures. In *The 27th ACM International Conference on Information and Knowledge Management (CIKM '18), October 22–26, 2018, Torino, Italy.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3269206.3271751

1 INTRODUCTION

Test collections provide the cornerstone for Cranfield-based batch evaluation of information retrieval (IR) algorithms [11], allowing empirical A/B testing of new IR systems and thus playing an important role in the development of more effective systems. However, improvements that are not statistically significant may result in misleading (or at least inaccurate) conclusions. Statistical significance testing in system evaluation is therefore deemed crucial for achieving meaningful advancements [6, 39].

CIKM '18, October 22–26, 2018, Torino, Italy

ACM ISBN 978-1-4503-6014-2/18/10…$15.00
https://doi.org/10.1145/3269206.3271751

Building a high-quality test collection is a costly process which encouraged research on developing low-cost evaluation methods such as selecting documents to be judged [31], crowdsourcing [30, 33], predicting system performances with incomplete judgments [49] and judgment-free system evaluation [44]. One of the main ways to evaluate these low-cost methods is to compare systems ranking on the full test collection (usually TREC test collections) with that produced by the low-cost evaluation method. Kendall's τ [28] is the most popular ranking correlation measure used to perform the comparison (e.g., [2, 7, 30, 31, 49]). While Kendall's τ provides a very intuitive score (i.e., $\frac{1-\tau}{2}$% of system pairs are relatively-ranked differently in the two rankings), it has drawbacks for IR system evaluation [5, 40]. One major drawback is that all system pairs are treated equally, ignoring their ranks and performance differences.

Due to shortcomings of τ, other rank correlation measures were proposed. Yilmaz et al. [50] proposed τ_{AP} which gives more weight to swaps at higher ranks. Gao and Oard [19] proposed τ_{GAP} that extends τ_{AP} such that performance difference between pairs is also considered. Gao et al. [18] proposed ρ_r, a head-weighted version of Pearson correlation [36]. However, none of these consider statistical significance in performance difference among systems. Cormack and Lynam [14] proposed applying τ only on pairs that exhibit statistical significance of difference in the true ranking. Carterette [5] proposed a rank distance measure d_{rank} that imposes a higher penalty when ranking of significantly different pairs in the true ranking is swapped in the predicted ranking, but ignores statistical significance of difference in scores in the predicted ranking.

To our knowledge, none of the proposed measures in the literature considered whether the statistical significance of difference in performance scores between pairs in the predicted ranking is concordant with that in the true ranking or not. However, in order to achieve a reliable evaluation in cases where rank correlation is computed between two lists of items for which we can compute statistical significance of difference, item pairs should be concordant in terms of *both* ranking and statistical significance of difference. To understand what problems can be introduced when ignoring statistical significance, consider a simple example of ranking 3 IR systems (a, b, c). Let their true ranking using evaluation method **E** be $<a, b, c>$ such that performance differences among all pairs are statistically significant. Assume that using another evaluation method **E'** (different metric or different judgments, etc.), they are ranked as $<a, b, c>$, but this time none of the differences among pairs is statistically significant. That is, **E'** actually could not "distinguish" between the 3 systems. However, τ, τ_{AP}, τ_{GAP}, d_{rank} would all return perfect correlation scores or almost-perfect (e.g.,

ρ_r). Trusting high rank correlation scores, an IR researcher evaluating IR systems with **E'** is likely to have inaccurate conclusions about statistical significance of the experimental results.

As a further motivating example, Section 3 shows an analysis of 11 TREC test collections in terms of statistical significance of difference among system pairs. We found that a noticeable percentage of system pairs (22% to 37%) are actually not exhibiting statistically-significant differences, suggesting that rank correlation measures ignoring statistical significance of differences could be misleading in estimating the quality of low-cost test collections.

In this work, we propose two statistical-significance-aware rank correlation measures tested in the context of evaluating low-cost evaluation methods. The first measure, named τ_{Sig}, considers *both* system rankings and statistical significance of differences among systems. The other measure, named τ_{SigH}, is a head-weighted version of τ_{Sig}, in which mistakes at higher ranks are penalized more than those in lower ranks, while still considering both rankings and statistical significance of differences. The two measures give the user the liberty to control the importance of statistical significance in the correlation score. Towards that end, the measures include two weighting parameters: α for the penalty on rank-concordant pairs whose score difference is statistically significant in one ranking but not the other, and β for the penalty on rank-discordant pairs whose score difference is not statistically significant in either ranking.

In our experiments, we evaluate the impact of the two parameters on the correlation score and compare the proposed measures against existing ones. The results suggest that τ_{Sig} and τ_{SigH} capture discordant pairs in terms of statistical significance of difference that existing measures are not able to. We also re-evaluate low-cost evaluation methods in four different areas. We find that the conclusions of the experiments can change based on the correlation measure used in evaluation, suggesting that measures covering multiple aspects of rank correlation can produce more reliable results.

Contributions. The contributions of this work are three-fold:

- We analyze statistical significance of differences among runs that participated in 11 TREC test collections and show that differences among many run pairs are indeed not statistically-significant, which can be problematic when a rank correlation analysis is performed with measures that are not aware of statistical significance, e.g., Kendall's τ.
- We propose two new rank correlation measures, namely τ_{Sig} and τ_{SigH}, taking both ranking and statistical significance of differences among pairs into consideration.[1]
- We re-evaluate low-cost IR evaluation methods proposed by prior work in 4 different areas including pooling, crowdsourcing, prediction of system performance with incomplete judgments, and automatic system ranking. We find that the conclusions of experiments could change depending on the correlation measure used in evaluation.

The remainder of the paper is organized as follows. We first summarize related work on main correlation measures used for IR evaluation in Section 2. Section 3 presents an analysis of statistical significance of differences among system pairs in several TREC collections. Proposed correlation measures are introduced in Section 4. We show empirical differences between proposed and existing rank correlation measures in Section 5. Section 6 covers re-evaluation of low-cost IR evaluation methods. Finally, final remarks and directions for future work are presented in Section 7.

[1]Implementation is available at http://qufaculty.qu.edu.qa/telsayed/code/correlation/.

2 RELATED WORK

Rank correlation measures are often used to validate reliability of proposed methods for low-cost evaluation of IR systems [20, 30, 31, 49]. They are also used for other comparisons as well, such as centrality scores on graphs [46] and search results [16, 17, 29, 45]. When comparing two ranked result lists, a document may exist in one of the rankings but not in the other. However, we focus on correlation between rankings in which both rankings have the same set of items (e.g., systems).

While a re-usable test collection should enable precise measurement of IR evaluation metrics (e.g., MAP) for A/B testing of baseline vs. newly proposed search algorithms, a lower bar for A/B testing is merely to verify the better algorithm scores higher, regardless of the magnitude of the difference. Generalizing to multiple systems, we might measure rank correlation between the "true" ranking of participant systems in a shared task evaluation (e.g., TREC), according to an evaluation metric and standard process (e.g., pooling using NIST judgments), vs. the possibly-incorrect "predicted" ranking of systems induced by an alternative (lower-cost) process.

Given two low-cost evaluation methods A and B vs. a third, baseline method C, it is typical to compare the rank correlation of A vs. C to that of B vs. C to assess whether A or B yields higher correlation. However, most prior work has not measured whether the difference in rank correlation between low-cost evaluation methods A and B is actually statistically significant. For an exception, see [27], which uses the t statistic for triangle significance testing following Hotelling [25].

In the remainder of this section, we summarize main correlation measures used in IR evaluation, focusing on rank correlation ones. A comparison of these measures is given in **Table 1**.

2.1 Correlation *without* Statistical Significance

Kendall's τ [28] is an easily interpretable measure and is the most popular method used in IR evaluation. Simply, $\frac{1-\tau}{2}$% of the pairs are ranked in reverse order in the rankings of interest. While being easily interpretable, it ignores variance in performance differences among pairs and treats all swaps equally regardless of their ranks.

There are a number of studies proposing variants of Kendall's τ with different weighting schemes. The idea of weighted τ is first proposed by Shieh [41]. Melucci [34] proposed a more flexible weighted scheme such that users can define any weight for each rank. Yilmaz et al. [50] proposed a *head-weighted* version of τ, meaning that swaps in higher ranks are penalized more than swaps in the lower ranks in the true ranking. Gao and Oard [19] further improved τ_{AP} such that swaps between pairs with large performance differences are penalized more than swaps between pairs with low performance difference. Related to Kendall's τ, Voorhees [47] proposed a rank correlation measure that estimates the probability of a discordant pair, ignoring ties and concordant pairs.

Assigning more weights to agreements in top ranks (i.e., head-weighted correlation) has also been studied in other rank correlation

Table 1: Rank correlation measures. SSA stands for *Statistical-Significance-Aware*.

Coefficient	Ordinal	Interval	Head-weighted	Symmetric	True Ranking SSA	Predicted Ranking SSA	Range
ρ [36]	-	✓	-	✓	-	-	[-1,1]
τ [28]	✓	-	-	✓	-	-	[-1,1]
τ_{AP} [50]	✓	-	✓	-	-	-	[-1,1]
τ_{GAP}[19]	✓	✓	✓	-	-	-	[-1,1]
ρ_r [18]	✓	✓	✓	-	-	-	[-1,1]
τ_{DP} [14]	✓	-	-	✓	✓	-	[-1,1]
d_{rank}[5]	✓	-	-	-	✓	-	[0,∞]
τ_{Sig}	✓	-	-	✓	✓	✓	[-1,1]
τ_{SigH}	✓	-	✓	-	✓	✓	[-1,1]

measures [26, 32]. Gao et al. [18] proposed ρ_r, a head-weighted version of Pearson [36] considering ranks and performance scores of systems together. Henzgen and Hüllermeier [24] proposed a rank correlation measure for fuzzy orderings in which the positions of items with small score differences are considered equal.

The main difference of our proposed coefficients with the aforementioned coefficients is that none of them considers statistical-significance of differences among system pairs.

2.2 Correlation *with* Statistical Significance

To the best of our knowledge, Cormack and Lynam [14] were the first to propose a rank correlation measure that considers statistical significance. They adapt τ such that only statistically different systems are considered in the calculation. Sakai [38] utilized *discriminative power* (DP) to compute the percentage of the runs that have been discriminated statistically using a particular evaluation method. Even though his method is not a rank correlation measure, it has the similar intuition with Cormack and Lynam [14], as mentioned by Carterette [5]. Therefore, we refer to Cormack and Lynam's coefficient as τ_{DP}.

Carterette [5] introduced d_{rank} which has the following features: 1) penalizes swaps based on the differences between pairs, 2) gives penalty if similar items in true ranking are separated in the predicted ranking, and 3) eliminates variance due to systems and assumes a fixed population of systems. d_{rank} is quite different than others due to being a distance measure, not a correlation measure. Being a distance measure, d_{rank} is ≥ 0 and lower scores mean better correlations. One disadvantage of d_{rank} is that it is not easily interpretable due to having no theoretical upper bound. To overcome this issue, Carterette also provided a statistical hypothesis test with a p-value.

τ_{DP} and d_{rank} do not consider statistical significance in the predicted ranking. If a system pair is significantly different in the true ranking but not in the predicted ranking, it is not penalized as long as the rankings are concordant. These coefficients take only statistical significance in the true ranking into consideration. However, in our proposed methods, both ranking and statistical significance are considered together and the correlation score can get penalized if the pairs are discordant in terms of statistical significance but concordant in terms of ranking. To the best of our knowledge, no other rank correlation measure considers this issue.

3 STATISTICAL SIGNIFICANCE AT TREC

TREC test collections provide a valuable resource for developing and evaluating IR systems, and also a test environment to evaluate the effectiveness of IR evaluation methods such as crowdsourcing [30, 33], pooling documents to be judged [31], predicting performance of systems using incomplete judgments [2, 49] among others.

A popular way to show the effectiveness of a proposed method is to compare the ranking of systems using the proposed evaluation method with the ground-truth ranking. In this correlation analysis, one of the most popular rank correlation measures is Kendall's τ. However, as mentioned before, τ does not take statistical significance into account and treats all system pairs equally. This can potentially lead to drawing inaccurate conclusions on the quality of proposed methods. For instance, assume we have two evaluation methods such that one of them causes swaps between significantly different pairs (i.e., ranking them in the reverse order wrt. ground-truth ranking) while the other causes same amount of swaps but only between not-significantly different pairs. We would like to use the latter because the other causes more severe evaluation problems. However, we cannot distinguish these two methods using τ or any other coefficient that ignores performance differences between system pairs. If there are many not-significantly different pairs in the ground-truth ranking, the performance of evaluation methods can get penalized heavily due to swapping similar systems even though we might not care about swaps between those pairs.

We investigate the statistical significance of difference among system pairs in 11 TREC test collections including ad-hoc search task of TREC5-10 and Web Track 2010-2014. For each test collection, we first rank the runs based on *mean average precision* (MAP). Subsequently, given the list of per topic average precision for each system, we apply paired t-test for each system pair [42]. We report the percentage of system pairs that are not significantly different (i.e., p-value ≥ 0.05) in **Table 2**, denoted as NS pairs. We observe that many of the system pairs do not show statistically-significant differences in AP (e.g., 37.3% of system pairs in TREC-10).

To further analyze the observed differences among system pairs, we cluster the pairs such that all systems in a cluster are not-significantly different. Because statistical significance between pairs is not a transitive relation, we run a greedy approach that clusters systems starting from the first ranked system to the last ranked system. A brief summary on the clusters is shown in 4^{th} and 5^{th} columns of Table 2. The number of clusters acquired is generally

Table 2: TREC test collections used in experiments.

Collection	Runs	NS Pairs	Clusters	Max Cluster Size
TREC-5 [21]	61	36.4%	9	16
TREC-6 [21]	74	35.9%	12	19
TREC-7 [21]	103	26.8%	14	16
TREC-8 [21]	129	27.6%	15	20
TREC-9 [23]	104	31.1%	13	18
TREC-10 [48]	97	37.3%	13	21
WT2010 [8]	56	34.2%	10	12
WT2011 [9]	62	34.6%	8	14
WT2012 [10]	48	37.2%	7	13
WT2013 [12]	61	36.7%	10	17
WT2014 [13]	30	22.3%	11	7

low for a test collection despite having high number of runs. To visualize the clustering results, we show the clusters of TREC-8 as an example in **Figure 1**. As seen from the figure, there are only 5 clusters covering systems ranked between 8 and 94.

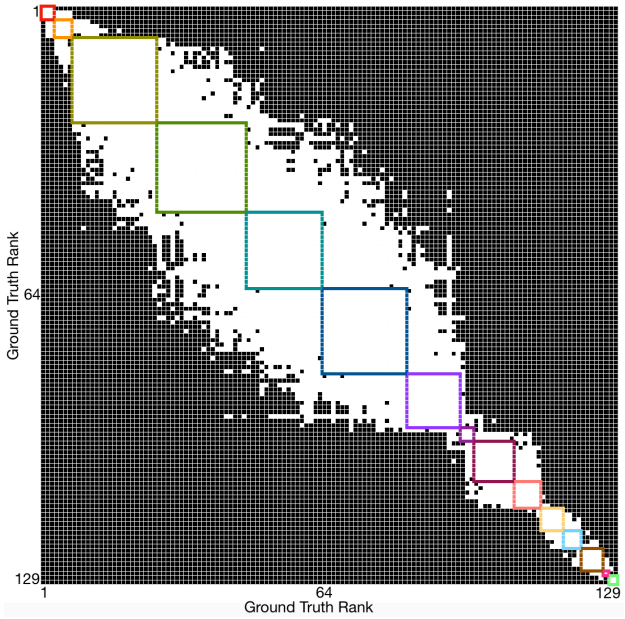


Figure 1: TREC-8 runs grouped based on statistical significance of AP scores. Black dots represent the significantly-different pairs and white dots represent the not-significantly-different pairs. Colored squares represent the clusters in which all pairs are not-significantly different to each other. X and Y axes show the ranks of systems based on MAP scores.

Overall, despite the high number of runs in TREC collections, many of them are actually not-significantly different. As mentioned earlier, low-cost evaluation methods can easily rank these similar systems in reverse order wrt. true ranking and get penalized when τ or τ_{AP} is used for rank correlation. This can potentially cause inaccurate conclusions drawn from the experiments. For example, evaluation methods causing swaps between significantly different pairs can be evaluated as outperforming methods that cause swaps only in not-significantly different pairs.

4 NEW RANK CORRELATION MEASURES

In this section, we propose two new rank correlation measures which incorporate statistical significance of differences between system scores. In Section 4.1, we introduce τ_{Sig}, which extends Kendall's τ. In Section 4.2, we introduce a head-weighted variant τ_{SigH}, which extends τ_{AP} [50].

4.1 Significance-aware Measure: τ_{Sig}

Given a set of n items, there are $\binom{n}{2}$ unique combinations (i.e., pairs) of items. Kendall's τ [28] compares two rankings over n items and computes the number of *concordant* pairs (C) (i.e., ranked in the same order in both rankings) vs. the number of *discordant* pairs (D) (i.e., ranked in reverse order). Positive $C - D$ indicates correlation, while negative $C - D$ indicates inverse correlation. This sum is normalized by the number of unique pairs, thus $\tau \in [-1, 1]$. The formula of Kendall's τ is given in **Equation 1**:

$$\tau = \binom{n}{2}^{-1} (C - D) \tag{1}$$

As mentioned earlier , τ does not consider the magnitude of the item values, only their relative ordering. Assuming these values are evaluation metric scores (e.g., MAP scores of alternative IR systems), this further means that τ does not consider the statistical significance of differences between system scores, as has been discussed in prior work [5, 40].

We now introduce τ_{Sig}, which extends τ to incorporate statistical significance of differences between system scores. Assume R_1 and R_2 are two rankings (of real-valued scores) to be compared. Let A and B be two systems included in each ranking, and let $P(R1_{A,B}, R2_{A,B})$ denotes a penalty according to how R_1 and R_2 score A and B. We define τ_{Sig} per **Equation 2**:

$$\tau_{Sig}(R_1, R_2) = \binom{n}{2}^{-1} \sum_{i=1}^{n} \sum_{j=i+1}^{n} 1 - P(R1_{i,j}, R2_{i,j}) \tag{2}$$

Next, we proceed to define the penalty function $P(R1_{A,B}, R2_{A,B})$ according to five possible cases, where α and β are parameters.

Case 1. (A, B) is concordant and the difference in scores is statistically significant in either both or neither ranking. $P(\cdot) = 0$ (no penalty).

Case 2. (A, B) is concordant but their score difference is statistically significant in one ranking and not the other. $P(\cdot) = \alpha$.

Case 3. (A, B) is discordant but the score difference is not statistically significant in either ranking. $P(\cdot) = \beta$.

Case 4. (A, B) is discordant and the score difference is statistically significant in one ranking but not both. $P(\cdot) = \alpha + \beta$.

Case 5. (A, B) is discordant and the score difference is statistically significant in both rankings. $P(\cdot) = 2$ (maximum penalty).

More precisely, assume A and B are scored in R_1 as A_1 and B_1, and in R_2 as A_2 and B_2. Assume that R_1 ranks A higher than B. If the difference in their scores is statistically significant, we denote this as $A_1 > B_1^*$, otherwise we denote it as $A_1 > B_1$. R_2 may similarly

score A and B as: (1) $A_2 > B_2$, (2) $A_2 > B_2^*$, (3) $B_2 > A_2$, or (4) $B_2 > A_2^*$. Thus we have $2 \times 4 = 8$ possible scenarios. If R_1 ranks B higher than A, we have another 8 possible scenarios. **Table 3** enumerates these 16 cases. Note its symmetry.

Table 3: Penalty function $P(\cdot)$ used in τ_{Sig} and τ_{SigH}.

Ranking R_1	**Ranking** R_2			
	$A_2 > B_2^*$	$A_2 > B_2$	$B_2 > A_2$	$B_2 > A_2^*$
$A_1 > B_1^*$	0	α	$\alpha + \beta$	2
$A_1 > B_1$	α	0	β	$\alpha + \beta$
$B_1 > A_1$	$\alpha + \beta$	β	0	α
$B_1 > A_1^*$	2	$\alpha + \beta$	α	0

Note that Kendall's τ can be viewed as a special case of τ_{Sig} as shown below.

Theorem 1. *τ_{Sig} is reduced to Kendall's τ when $\alpha = 0$ and $\beta = 2$.*

Proof. When $\alpha = 0$ (i.e., we ignore statistical significance) and $\beta = 2$ (i.e., we adopt same penalty as τ for discordant pairs), $1 - P(\cdot)$ will always be 1 for concordant pairs and -1 for discordant pairs regardless of statistical significance. According to Equations 1 and 2, this makes $\tau_{Sig} = \tau$. □

We further impose constraints that $\alpha \geq 0$, $\beta \geq 0$, and $\alpha + \beta \leq 2$, motivated as described below.

Theorem 2. $\tau_{Sig} \in [-1, 1]$.

Proof. With $\alpha, \beta \geq 0$ and $\alpha + \beta \leq 2$, it is trivial that $P(\cdot) \in [0, 2]$. Because τ_{Sig} simply sums $1 - P(\cdot)$ over all $\binom{n}{2}$ unique pairs, normalized by their count, then $P(\cdot) \in [0, 2] \rightarrow \tau_{Sig} \in [-1, 1]$. □

Theorem 3. *τ_{Sig} is symmetric, i.e., $\tau_{Sig}(X, Y) = \tau_{Sig}(Y, X)$, where X and Y are two rankings to be compared.*

Proof. $P(\cdot)$ is defined symmetrically, shown precisely in Table 3, and informally in the five enumerated cases preceding it. □

4.2 Significance-aware Head-weighted Measure: τ_{SigH}

Assume R_1 is a ground-truth ranking and R_2 is a (possibly noisy) predicted ranking which we wish to assess via its rank correlation to R_1. Yilmaz et al. [50] proposed τ_{AP}, a *head-weighted* version of τ in which swaps at higher ranks in R_1 are penalized more than swaps at its lower ranks. Let I denotes the item at rank $i \in R_1$, and let C_i denotes the number of items correctly ranked higher than I in R_2. **Equation 3** defines τ_{AP} as follows.

$$\tau_{AP} = \frac{2}{n-1} \sum_{i=2}^{n} \frac{C_i}{i-1} - 1 \tag{3}$$

Similar to τ, τ_{AP} does not consider the magnitude of the item values, only their relative ordering. Assuming these values are evaluation metric scores (e.g., MAP scores of alternative IR systems), τ_{AP} also does not consider the statistical significance of differences between system scores.

To extend τ_{AP} to incorporate statistical significance of differences in system scores, we replace its C_i term with a new M_i term, defined as the total weight above rank $i \in R_2$ wrt. I. In other words, instead of giving weight 1 to each concordant pair and 0 to each discordant pair above the rank cutoff threshold, we instead use the weight $1 - P(\cdot)$ (as with τ_{Sig}), with penalty function $P(\cdot)$ as defined in Table 3. The formula of τ_{SigH} is given in **Equation 4**.

$$\tau_{SigH} = \frac{1}{n-1} \sum_{i=2}^{n} \frac{M_i}{i-1} \tag{4}$$

Note that τ_{AP} is not symmetric (i.e., $\tau_{AP}(X, Y) \neq \tau_{AP}(Y, X)$) because it distinguishes R_1 vs. R_2. This is also true for τ_{SigH}.

Theorem 4. *τ_{SigH} is reduced to τ_{AP} when $\alpha = 0$ and $\beta = 2$.*

Proof. Equation 3 can be written as Equation 5 by simple mathematical manipulation.

$$\tau_{AP} = \frac{1}{n-1} \sum_{i=2}^{n} \frac{2C_i - (i-1)}{i-1} \tag{5}$$

As defined above, C_i is the number of correctly ranked items above i in the predicted ranking (with respect to the item i in the true ranking). Therefore, there are $i - 1 - C_i$ items that are not correctly ranked above i in the predicted ranking. When $\alpha = 0$ and $\beta = 2$, the weight of each correctly and incorrectly ranked items in M_i calculation is 1 and -1 respectively. Therefore, $M_i = C_i - (i - 1 - C_i) = 2C_i - (i - 1)$. Replacing $2C_i - (i - 1)$ with M_i in Equation 5 indicates that $\tau_{AP} = \tau_{SigH}$. □

Theorem 5. $\tau_{SigH} \in [-1, 1]$.

Proof. $P(\cdot) \in [0, 2]$ as shown before. Because M_i simply sums $1 - P(\cdot)$ over $(i - 1)$ pairs, divided by $(i - 1)$, then $\frac{M_i}{(i-1)} \in [-1, 1]$. Summing those terms normalized by their count yields a τ_{SigH} value $\in [-1, 1]$. □

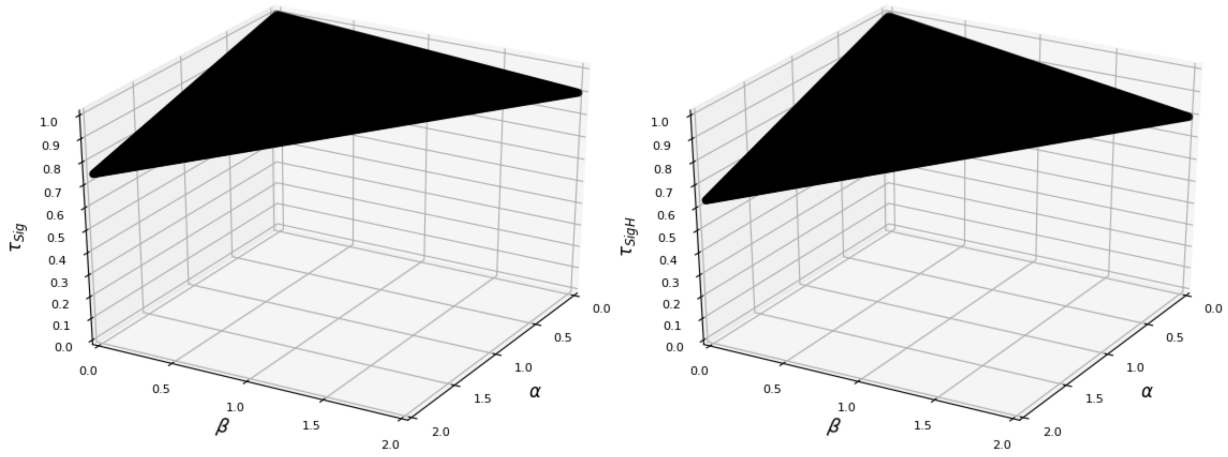
5 EXPERIMENTAL EVALUATION

In this section, we show empirical differences between proposed rank correlation measures and existing ones. First, we show the impact of changing α and β parameters on the value of the measures (Section 5.1). Next, we compare proposed measures with existing ones in literature by conducting experiments on simulated data. The aim is to show how different correlation measures behave when the rankings of systems are exactly the same but the statistical significance of difference in scores between some system pairs change (Section 5.2). In all experiments, we used all system runs submitted to TREC-8 except two runs that have the same results with other two runs (i.e., 127 runs in total). For all experiments including those in Section 6, statistical significance testing is conducted using paired t-test with a p-value threshold of 0.05.

5.1 Impact of Changing Measure Parameters

In this section, we address the following research question:
RQ1: *How do changes in α and β affect the values of the proposed correlation measures?*

To answer this question, we conducted an experiment where we varied the values of α and β parameters and studied the impact on both τ_{Sig} and τ_{SigH}. We first ranked the runs based on *MAP*. Then, we randomly selected a pair of runs that are consecutively ranked



(a) τ_{Sig} **(b)** τ_{SigH}

Figure 2: Impact of α and β parameters on τ_{Sig} and τ_{SigH}

and swapped their AP scores for each topic. We repeated this process on the modified ranking without repetition until we reached a ranking that yields 0.9 τ correlation score (a traditionally-accepted threshold for acceptable correlation between two IR system rankings [47]) with the true (original) ranking. In the resultant ranking, there are 1008 discordant pairs (out of 8001) in terms of statistical significance and τ_{AP} is 0.794. Subsequently, we changed α and β from 0 to 2 with 0.01 increments such their sum does not exceed 2, and calculated τ_{Sig} and τ_{SigH}. The results are shown in **Figure 2**.

When α and β are 0 (i.e., only discordant pairs in terms of both ranking order and statistical significance are penalized), τ_{Sig} and τ_{SigH} reach 0.996 and 0.989 respectively, yielding the maximum observed values. The minimum values of τ_{Sig} and τ_{SigH} were 0.744 and 0.629 respectively, achieved when β is 0 and α is 2 (i.e., swaps between not-significantly different pairs are ignored but any change in statistical difference of pairs is penalized with the maximum penalty). These results show that in two rankings with 0.9 τ correlation score, there can be many discordant pairs in terms of statistical difference (in our case, 1008/8001=12.6% of the pairs) which can deeply impact the correlation score between two rankings depending on how much we care about statistical significance of differences in underlying scores.

5.2 Impact of Changing Statistical Significance

In this section, we address the following research question:

RQ2: *How do different rank correlation measures (including proposed ones) behave if we fix the rankings but change the underlying system scores (and thus potentially the statistical significance of differences in those scores)?*

To answer this question, we conducted another experiment on simulated data. We first ranked the 127 runs based on their *MAP* scores. Then, we randomly selected N consecutively-ranked pairs. For each selected system pair (i, j), we changed the AP scores for all topics for the system with the lower score using a Gaussian distribution with mean set to $(MAP_i + MAP_j)/2$ and standard deviation set to half of the mean. This ensures that the order of the systems is preserved, but the statistical significance of differences in their scores might change. Next, we counted the number of pairs where the statistical significance has changed over the entire ranking and also computed correlation scores with respect to the true (original) ranking using various measures. We repeated this process 1,000 times and also experimented with N ranging from 1 to 63 (spanning the entire range from a single pair to the maximum number of disjoint consecutive pairs), yielding $63 \times 1,000 = 63,000$ cases. We grouped the cases resulting in the same number of changes in statistical significance and computed the average correlation score for each group. Groups with less than 5 cases were discarded. For our proposed measures, we set α and β to 1 and 0 respectively for both τ_{Sig} and τ_{SigH}. The results are shown in **Figure 3**.

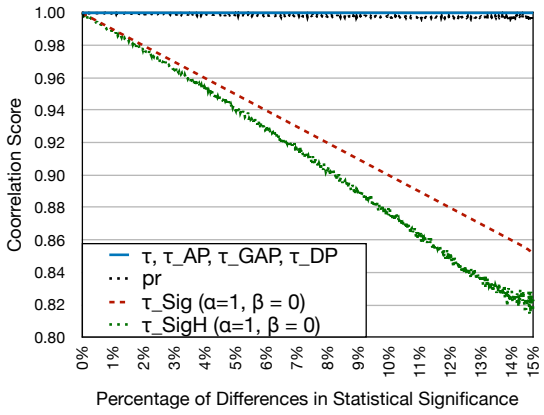
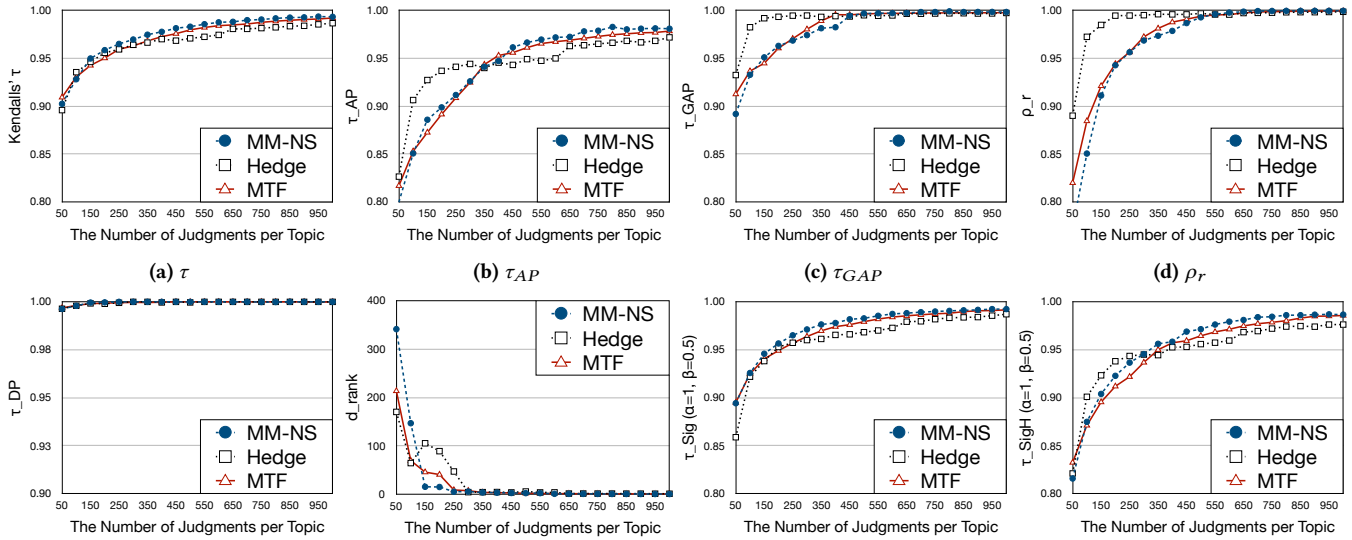


Figure 3: Correlation when the rankings are same but statistical significance among pairs are changed

Even though the relative ranking of the systems is preserved, the score changes we imposed on the selected pairs resulted in changes in statistical significance among 0%-15% of the total number of pairs. However, none of Kendall's τ, τ_{AP}, τ_{GAP}, τ_{DP}, or d-rank



(a) τ **(b)** τ_{AP} **(c)** τ_{GAP} **(d)** ρ_r

(e) τ_{DP} **(f)** d_{rank} **(g)** τ_{Sig} ($\alpha = 1, \beta = 0.5$) **(h)** τ_{SigH} ($\alpha = 1, \beta = 0.5$)

Figure 4: Evaluation of Hedge, MM-NS and MTF pooling methods on TREC-5 based on different rank correlation measures.

were affected; their values were all fixed at 1 (as shown in the figure) except d-rank that was fixed at 0 (not shown in the figure), as they only consider relative ranking with no consideration to statistical significance.

ρ_r score changes very slightly due to differences in the mean scores, ranging from 0.9972 to 1.0. Different from all other measures, τ_{Sig} and τ_{SigH} are able to catch the changes in statistical significance of differences among systems and their values decrease as the differences among the two rankings in terms of statistical significance increase; τ_{Sig} value decreases linearly while τ_{SigH} oscillates due to taking the rank of the pairs into consideration.

6 RE-ASSESSING LOW-COST EVALUATION RELIABILITY VIA RANK CORRELATION

As discussed in Section 2, rank correlation is often used to validate the reliability of proposed low-cost IR system evaluation strategies [20, 30, 31, 49]. In this section, we consider four low-cost evaluation strategies: (1) pooling methods (Section 6.1); (2) crowdsourcing for collecting relevance judgments (Section 6.2); (3) system evaluation with incomplete set of judgments (Section 6.3); and (4) automatic evaluation without relevance judgments (Section 6.4). In each case, we compare our proposed τ_{Sig} and τ_{SigH} measures vs. existing ones, as well as the conclusions one would draw from each measure.

In all the experiments reported in this section, we set $\alpha = 1$ and $\beta = 0.5$ for our proposed measures. This means that swaps between pairs that do *not* exhibit statistical significance in difference are penalized by 0.5. If the statistical significance of *concordant* pairs do *not* match in both rankings, we penalize by 1, i.e., half of the maximum penalty. As for the discordant pairs when the score difference is statistically significant in one of the rankings but not both, we penalize each by 1.5. This penalty scheme allows us to have a wide range of penalty weights including 0, 0.5, 1, 1.5 and 2.

6.1 Evaluating Pooling Methods

Losada et al. [31] recently proposed seven pooling methods adapting algorithms for multi-armed bandit problem [37]. They compare their proposed methods against existing pooling methods including *Move-To-Front* (MTF) [15] and *Hedge* [3]. The authors share all their code and other implementation details[2], allowing us to reproduce their results. In one of their experiments, they compare MTF, Hegde, and their non-stationary version of MaxMean (MM-NS) method. They vary the number of documents to be judged and rank the systems using the resultant judgment set of each pooling method. Subsequently, the rankings are compared against the ground-truth ranking in which all judgments are used, using Kendall's τ (see Figure 4 of [31]).

We compare these three pooling methods over TREC-5 using the authors' implementation and compute 8 different rank correlation measures: Kendall's τ, τ_{AP}, τ_{GAP}, ρ_r, d-rank, τ_{DP}, τ_{Sig} and τ_{SigH}. Because MTF and MM-NS methods are stochastic, we generate 3 different pools for these methods and report average results.[3] We change the number of judged documents per topic from 50 to 1,000.

Figure 4 shows results. The conclusions of the experiment vary across correlation measures. Based on Kendall's τ, MTF is seen to yield slightly better correlation than others using 50 judgments per topic (MTF: 0.9092, Hedge: 0.8958, MM-NS:0.9023), Hedge is slightly better than others with 100 judgments per topic, and MM-NS outperforms others in rest of the cases. However, based on τ_{Sig}, MM-NS is consistently better than MTF and Hedge. For instance, with 100 judgments per topic, Hedge achieves 0.9354 τ and 0.8587 τ_{Sig} while MM-NS achieves 0.928 τ and 0.8942 τ_{Sig}. With 100 judgments per topic, MM-NS cannot capture statistical significance of difference in concordant pairs correctly (i.e., Case 2 in Section 4.1) in 277 pairs while Hedge causes the same error in 306 pairs. While

[2] https://tec.citius.usc.es/ir/code/pooling_bandits_ms.html

[3] Losada et al. [31] do not report if they run multiple trials for these methods.

Table 4: Correlation between system rankings with NIST and crowd judgments over WT2014

	τ	τ_{AP}	τ_{GAP}	ρ_r	τ_{DP}	d_{rank}	τ_{Sig} (α=1,β=0.5)	τ_{SigH}(α=1,β=0.5)
MAP_{NIST} vs. $statAP_{NIST}$	0.905	0.876	0.98	0.966	1.0	2.443	0.873	0.810
MAP_{NIST} vs. $statAP_{CROWD}$	0.937	0.921	0.99	0.973	0.997	2.529	0.854	0.790

τ does not give any penalty for this case and treats them as fully concordant pairs, τ_{Sig} gives α penalty for each. In addition, the frequency of Case 3 error (i.e., discordant pairs but difference between pairs is not statistically significant) in MM-NS and Hedge are very similar (170 vs. 154). However, τ penalizes each by 1, while τ_{Sig} is more tolerant and penalizes by 0.5 (since $\beta = 0.5$). That justifies why MM-NS outperforms Hedge based on τ_{Sig} but not on τ.

Based on head-weighted measures that are not aware of statistical significance of differences (i.e., τ_{AP}, τ_{GAP} and ρ_r), Hedge outperforms others with few number of judgments per topic and MM-NS is the best performing method with 400 or more judgments per topic. However, based on τ_{SigH}, we observe a slightly different pattern: MTF yields slightly better correlation than others with 50 judgments per topic and the difference in correlation between Hedge and others is smaller wrt. τ_{AP}, τ_{GAP} and ρ_r when the number of judgments per topic is 100-300.

Based on τ_{DP}, all methods have very high correlation scores (0.995+) even with 50 judgments per topic, indicating that all of them are able to rank significantly-different pairs in true ranking accurately. However, it makes it harder to distinguish the effectiveness of different pooling methods.

Overall, when we consider both ranking and statistical significance of difference, MM-NS seems the best method for pooling. While Hedge appears to detect better systems with fewer judgments than MTF and MM-NS, we notice that it causes more mistakes in terms of statistical significance. Our proposed measures are able to capture different types of information (i.e., ranking, statistical significance of differences, and position of concordant/discordant pairs), yielding different conclusions than other measures in many cases of the experiments.

6.2 Evaluating Systems via Crowdsourcing

A variety of studies have investigated crowdsourcing relevance judgments as a method of more efficient, albeit potentially noisy, data labeling [1]. Recently, McDonnell et al. [33] described a *rationale*-based crowdsourcing method reported to achieve high accuracy in binary relevance judgments with respect to NIST judgments for around 700 documents. In a follow-up study [30], the authors release WebCrowd25K[4] in which they collect crowd judgments using the same method for WT2014, with 100 documents per topic selected via statAP stratified sampling [2]. The authors rank the systems in three ways: 1) MAP_{NIST}, when all NIST judgments are used to calculate AP scores; 2) $statAP_{NIST}$, when AP scores are estimated using statAP method with NIST judgments of the sampled documents (100x50=5000 judgments in total); and 3) $statAP_{CROWD}$, when AP scores are estimated using statAP method with aggregated crowd judgments of the sampled documents. The authors report τ and τ_{AP} scores in comparison of each ranking pair.

[4]http://qufaculty.qu.edu.qa/telsayed/datasets/webcrowd25k/

Using their data, we also rank the WT2014 participating systems in same three ways as above and calculate rank correlation using a wider range of rank correlation measures, including those we propose in this work. Results are shown in **Table 4**.

Taking MAP_{NIST} rankings as ground truth, we expect $statAP_{NIST}$ should yield better rankings than $statAP_{CROWD}$ because crowd judgments sometimes disagree with NIST. Ranking coefficients that consider statistical significance (i.e., τ_{DP}, d_{rank}, τ_{Sig} and τ_{SigH}) affirm this expected finding. However, all rank correlation measures ignoring statistical significance of underlying scores (i.e., τ, τ_{AP}, τ_{GAP} and ρ_r) would lead to the opposite conclusion.

As for the reliability of evaluating IR systems via crowdsourcing, while we contravene the prior study [30] wrt. finding $statAP_{NIST}$ does yield better rankings than $statAP_{CROWD}$, results in Table 4 shows the crowd results are still quite reliable.

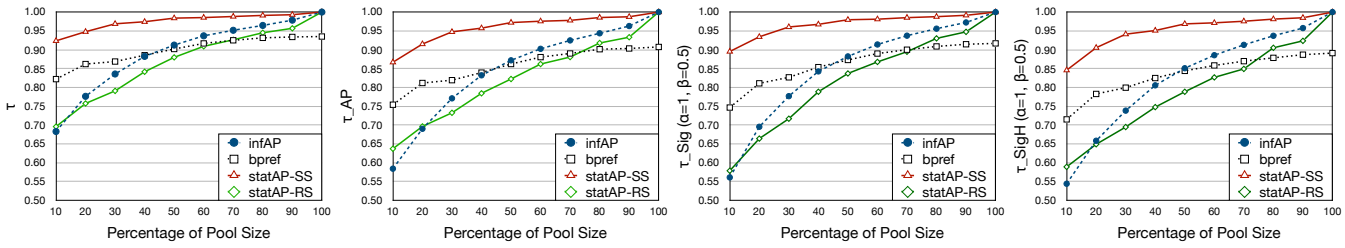
6.3 Evaluation Systems with Incomplete Judgments

A variety of measures have been proposed for evaluating IR systems using incomplete judgments, such as *bpref* [4], inferred AP (*infAP*) [49], and *statAP* [2]. Kendall's τ was used in the above studies as the rank correlation measure to evaluate the proposed methods. In this section, we re-evaluate these methods using a wider range of rank correlation measures.

We use TREC-8 in our experiments here since it was commonly used in the original authors' experiments. We de-duplicate the runs of TREC-8 because d_{rank} adds jitter to its values when there is a duplicate run. We use `trec_eval` software to compute *infAP* and *bpref*, which adopts Soboroff's revised *bpref* [43].

Yilmaz and Aslam [49] compared *infAP* and *bpref* via Kendall's τ on TREC-8 runs, varying the number of judgments from 50% to 100%. We follow a similar setup, randomly sampling N% of the judgments and ranking runs based on *infAP* and *bpref* using sampled judgments. *statAP* consists of two components: stratified sampling and estimation of the AP metric based on the weights assigned to each document. For fair comparison, we computed the results in two different ways: (1) *statAP w. stratified sample* (*statAP-SS*), where we sample N% of the judgments using *statAP*'s stratified sampling method and rank runs accordingly, and (2) *statAP w. random sample* (*statAP-RS*), where we compute *statAP* score using the same randomly-sampled set used with *bpref* and *infAP*. We repeat this process 10 times and also vary N from 10 to 100, computing average rank correlation scores for each measure. **Figure 5** presents results of Kendall's τ, τ_{AP}, τ_{Sig} and τ_{SigH} correlation measures.

Based on all correlation measures (including our proposed measures and also unreported ones that we omit due to space limitation), *statAP* with its stratified sampling consistently yields a more reliable system evaluation than others. This is a case when the correlation measures clearly agree, yielding a unified conclusion.



(a) τ **(b)** τ_{AP} **(c)** τ_{Sig} **(d)** τ_{SigH}

Figure 5: Performance of low-cost evaluation metrics with varying pool sizes.

Table 5: Comparison of automatic ranking methods on TREC-5 and TREC-10. Best performing method per correlation measure is in **bold**.

	TREC-5				TREC-10			
	RS	**Score**	**Score.EM**	**Vote.EM**	**RS**	**Score**	**Score.EM**	**Vote.EM**
τ	0.419	0.474	0.475	**0.482**	**0.609**	0.571	0.571	0.582
τ_{AP}	0.281	0.343	**0.345**	0.334	**0.357**	0.341	0.340	0.351
τ_{GAP}	0.295	0.348	0.348	**0.350**	0.219	0.233	0.233	**0.244**
ρ_r	0.072	0.244	0.246	**0.248**	-0.114	**-0.051**	**-0.051**	-0.053
τ_{DP}	0.768	**0.814**	**0.814**	**0.814**	**0.888**	0.879	0.880	0.884
d_{rank}	27.74	**24.05**	**24.05**	24.582	820.6	621.3	**609.3**	668.7
τ_{Sig}	0.352	0.429	0.431	**0.434**	**0.528**	0.512	0.513	0.518
τ_{SigH}	0.203	0.283	**0.284**	0.282	0.300	0.318	0.319	**0.329**

We also observe some differences in comparing *statAP-RS* vs. others across measures. For instance, when pool size is 70%, *bpref* and *statAP-RS* have similar performance based on τ (0.925 vs. 0.926) but *bpref* outperforms *statAP-RS* based on τ_{SigH} (0.87 vs. 0.848).

Furthermore, we note that τ_{SigH} manages to better distinguish between the four evaluation measures (i.e, with *larger* gaps in correlation scores) when the pool size is 70% and above, compared to the other measures. As it considers all of rank-order, statistical significance of differences, and also head-weighted swaps, it exploits more features of the tested rankings.

6.4 Automatic System Ranking Methods

The task of automatic ranking (AR) [44] is to rank IR systems without relevance judgments. Rank correlation (e.g., τ) is computed between the *predicted* ranking of systems vs. the *actual* ranking given human judgments [20, 35, 44]. We re-implemented four AR methods: Random Sampling (RS) [44], Score, Score.EM, and Vote.EM [20]; all are among the best performing methods as reported in [20]. We evaluated the methods over two TREC collections (TREC-5 and TREC-10) for the ad-hoc search task using 8 rank correlation measures computed between the estimated and true ranking of systems based on *MAP*. We repeated RS method 50 times per collection (as it uses random sampling) and report the average correlation. We show results over both collections in **Table 5**.

Over TREC-5, Vote.EM outperforms others according to τ, τ_{GAP}, ρ_r and τ_{Sig}, while Score.EM is the best performing method according to τ_{AP} and τ_{SigH}. Over TREC-10, according to τ, τ_{AP}, τ_{DP} and τ_{Sig}, the best performing method is RS, but Vote.EM is the second best. However, according to τ_{GAP}, ρ_r, d_{rank} and τ_{SigH}, RS is the worst, showing that the conclusion of the results can be completely different based on different correlation measures.

Analyzing how the AR methods ranked the top 10 systems according to the ground truth over TREC-10, we observe that the top systems are ranked very low in the predicted rankings, which is a typically-observed issue in many AR methods [22]. For the RS method in particular, the ranking of top systems is much lower in the predicted ranking on average compared to Score.EM, Score, and Vote.EM. Such analysis explains why correlation scores for head-weighted measures are lower, in general, compared to non-head-weighted ones (**Table 5**).

7 CONCLUSION AND FUTURE WORK

Statistical significance testing is crucial to acquire reliable conclusions from evaluation of IR systems. Therefore, we need IR evaluation methods that do not make mistakes in ranking system pairs with performance differences that are statistically significant. Thus, the way we evaluate IR evaluation methods should also consider statistical significance. However, existing rank correlation measures either do not consider this important aspect at all or considers it only in the true ranking but not in the predicted ranking.

In this work, we first analyzed runs from 11 TREC test collections and showed that many TREC runs were actually not statistically-different, suggesting that experiments on TREC collections with ranking measures that are not aware of statistical significance can cause inaccurate evaluation of methods. We then introduced two new rank correlation measures, τ_{Sig} and τ_{SigH}, that consider both ranking and statistical significance. The proposed measures provide two parameters to control the weight of statistical significance and

rank swaps. We also showed that τ_{Sig} and τ_{SigH} can be reduced to τ and τ_{AP}, respectively, with certain parameter setting. Finally, we re-benchmarked IR evaluation methods in 4 different areas including pooling, crowdsourcing, evaluation with incomplete judgments, and automatic system ranking, using various rank correlation measures including τ_{Sig} and τ_{SigH}. Results showed that the choice of correlation measure affects conclusions drawn from the evaluations, suggesting that the use of measures incorporating multiple aspects of correlation leads to more reliable conclusions.

In the future, we plan to investigate the impact of the parameters of our proposed measures to provide a suggested parameter setting for specific evaluation scenarios. We would also like to conduct our re-benchmarking experiments on a larger set of collections and on a more diverse set of evaluation methods.

ACKNOWLEDGMENTS

This work was made possible by NPRP grant# NPRP 7-1313-1-245 from the Qatar National Research Fund (a member of Qatar Foundation). The statements made herein are solely the responsibility of the authors.

REFERENCES

[1] Omar Alonso, Daniel E Rose, and Benjamin Stewart. 2008. Crowdsourcing for relevance evaluation. In *ACM SIGIR Forum*, Vol. 42. 9–15.
[2] Javed A. Aslam and Virgil Pavlu. 2007. A practical sampling strategy for efficient retrieval evaluation. *Technical Report* (2007).
[3] Javed A. Aslam, Virgiliu Pavlu, and Robert Savell. 2003. A Unified Model for Metasearch, Pooling, and System Evaluation. In *Proceedings of the Twelfth International Conference on Information and Knowledge Management (CIKM '03)*. 484–491.
[4] Chris Buckley and Ellen M Voorhees. 2004. Retrieval evaluation with incomplete information. In *SIGIR '04*. 25–32.
[5] Ben Carterette. 2009. On rank correlation and the distance between rankings. In *Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval (SIGIR '09)*. 436–443.
[6] Ben Carterette. 2017. But Is It Statistically Significant?: Statistical Significance in IR Research, 1995-2014. In *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '17)*. 1125–1128.
[7] Ben Carterette, James Allan, and Ramesh Sitaraman. 2006. Minimal test collections for retrieval evaluation. In *SIGIR '06*. 268–275.
[8] Charles L Clarke, Nick Craswell, Ian Soboroff, and Gordon Cormack. 2010. Overview of the TREC 2010 Web track. In *TREC 2010*.
[9] Charles L Clarke, Nick Craswell, Ian Soboroff, and Ellen M. Voorhees. 2010. Overview of the TREC 2011 Web track. In *TREC 2011*.
[10] Charles L. Clarke, Nick Craswell, and Ellen M. Voorhees. 2012. Overview of the TREC 2012 Web Track. In *TREC 2012*.
[11] Cyril W Cleverdon. 1959. The evaluation of systems used in information retrieval. In *Proceedings of the international conference on scientific information*, Vol. 1. National Academy of Sciences Washington, DC„ 687–698.
[12] Kevyn Collins-Thompson, Paul Bennett, Fernando Diaz, Charles L A Clarke, and Ellen M Voorhees. 2013. TREC 2013 Web Track Overview. In *TREC 2013*.
[13] Kevyn Collins-Thompson, Craig Macdonald, Paul Bennett, Fernando Diaz, and Ellen M Voorhees. 2015. TREC 2014 web track overview.
[14] Gordon V Cormack and Thomas R Lynam. 2007. Power and bias of subset pooling strategies. In *SIGIR '07*. 837–838.
[15] Gordon V Cormack, Christopher R Palmer, and Charles LA Clarke. 1998. Efficient construction of large test collections. In *SIGIR '98*. 282–289.
[16] Giorgio Maria Di Nunzio and Gianmaria Silvello. 2015. A Graphical View of Distance Between Rankings: The Point and Area Measures.. In *IIR*.
[17] Ronald Fagin, Ravi Kumar, and Dakshinamurthi Sivakumar. 2003. Comparing top k lists. *SIAM Journal on discrete mathematics* 17, 1 (2003), 134–160.
[18] Ning Gao, Mossaab Bagdouri, and Douglas W Oard. 2016. Pearson rank: a head-weighted gap-sensitive score-based correlation coefficient. In *SIGIR '16*.
[19] Ning Gao and Douglas Oard. 2015. A head-weighted gap-sensitive correlation coefficient. In *SIGIR '15*. 799–802.
[20] Ning Gao, William Webber, and Douglas W. Oard. 2014. Reducing reliance on relevance judgments for system comparison by using expectation-maximization. In *ECIR '14*.
[21] Donna Harman. 2011. *Information retrieval evaluation*. Number 19 in Synthesis lectures on information concepts, retrieval, and services. Morgan & Claypool, San Rafael, Calif.
[22] Claudia Hauff, Djoerd Hiemstra, Leif Azzopardi, and Franciska De Jong. 2010. A case for automatic system evaluation. In *ECIR '10*. 153–165.
[23] David Hawking. 2000. Overview of the TREC-9 Web Track. In *TREC 9*.
[24] Sascha Henzgen and Eyke Hüllermeier. 2015. Weighted rank correlation: a flexible approach based on fuzzy order relations. In *ECML PKDD '15*. 422–437.
[25] Harold Hotelling. 1940. The selection of variates for use in prediction with some comments on the general problem of nuisance parameters. *The Annals of Mathematical Statistics* 11, 3 (1940), 271–283.
[26] Ronald L Iman and WJ Conover. 1987. A measure of top–down correlation. *Technometrics* 29, 3 (1987), 351–357.
[27] Hyun Joon Jung and Matthew Lease. 2012. *Evaluating Classifiers Without Expert Labels*. Technical Report. University of Texas at Austin. arXiv:1212.0960.
[28] Maurice G Kendall. 1938. A new measure of rank correlation. *Biometrika* 30, 1/2 (1938), 81–93.
[29] Ravi Kumar and Sergei Vassilvitskii. 2010. Generalized distances between rankings. In *Proceedings of the 19th international conference on World wide web (WWW '10)*. 571–580.
[30] Mucahid Kutlu, Tyler McDonnell, Yassmine Barkallah, Tamer Elsayed, and Matthew Lease. 2018. Crowd vs. Expert: What Can Relevance Judgment Rationales Teach Us About Assessor Disagreement?. In *SIGIR 2018*. 805–814.
[31] David E Losada, Javier Parapar, and Alvaro Barreiro. 2017. Multi-armed bandits for adjudicating documents in pooling-based evaluation of information retrieval systems. *Information Processing & Management* 53, 5 (2017), 1005–1025.
[32] Tahani A Maturi and Ezz H Abdelfattah. 2008. A new weighted rank correlation. *Journal of mathematics and statistics*. 4, 4 (2008), 226–230.
[33] Tyler McDonnell, Matthew Lease, Mucahid Kutlu, and Tamer Elsayed. 2016. Why is that relevant? Collecting annotator rationales for relevance judgments. In *HCOMP '16*. 139–148.
[34] Massimo Melucci. 2009. Weighted rank correlation in information retrieval evaluation. In *AIRs '09*. 75–86.
[35] Rabia Nuray and Fazli Can. 2006. Automatic ranking of information retrieval systems using data fusion. *Information Processing & Management* 42, 3 (2006).
[36] Karl Pearson. 1895. Note on regression and inheritance in the case of two parents. *Proceedings of the Royal Society of London* 58 (1895), 240–242.
[37] Herbert Robbins. 1985. Some aspects of the sequential design of experiments. In *Herbert Robbins Selected Papers*. Springer, 169–177.
[38] Tetsuya Sakai. 2007. Alternatives to bpref. In *SIGIR '07*. 71–78.
[39] Tetsuya Sakai. 2014. Statistical reform in information retrieval?. In *ACM SIGIR Forum*, Vol. 48. ACM, 3–12.
[40] Mark Sanderson and Ian Soboroff. 2007. Problems with Kendall's tau. In *SIGIR '07*. 839–840.
[41] Grace S Shieh. 1998. A weighted Kendall's tau statistic. *Statistics & probability letters* 39, 1 (1998), 17–24.
[42] Mark D Smucker, James Allan, and Ben Carterette. 2007. A comparison of statistical significance tests for information retrieval evaluation. In *Proceedings of the sixteenth ACM conference on Conference on information and knowledge management (CIKM '07)*. 623–632.
[43] Ian Soboroff. 2006. Dynamic test collections: measuring search effectiveness on the live web. In *Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval (SIGIR '06)*. ACM, 276–283.
[44] Ian Soboroff, Charles Nicholas, and Patrick Cahan. 2001. Ranking Retrieval Systems Without Relevance Judgments. In *SIGIR '01*.
[45] Luchen Tan and Charles LA Clarke. 2015. A family of rank similarity measures based on maximized effectiveness difference. *IEEE Transactions on Knowledge and Data Engineering* 27, 11 (2015), 2865–2877.
[46] Sebastiano Vigna. 2015. A weighted correlation index for rankings with ties. In *Proceedings of the 24th international conference on World Wide Web (WWW '15)*. 1166–1176.
[47] Ellen M. Voorhees. 2000. Variations in relevance judgments and the measurement of retrieval effectiveness. *Information Processing & Management* 36, 5 (2000).
[48] Ellen M Voorhees and Donna Harman. [n. d.]. Overview of TREC 2001. In *TREC 2001*.
[49] Emine Yilmaz and Javed A Aslam. 2006. Estimating average precision with incomplete and imperfect judgments. In *Proceedings of the 15th ACM international conference on Information and knowledge management (CIKM '06)*. 102–111.
[50] Emine Yilmaz, Javed A. Aslam, and Stephen Robertson. 2008. A new rank correlation coefficient for information retrieval. In *Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval (SIGIR '08)*. 587–594.