Rank Correlation and Distance Between Rankings

Ben Carterette
Center for Intelligent Information Retrieval
Computer Science Department
University of Massachusetts Amherst
Amherst, MA 01003
carteret@cs.umass.edu

ABSTRACT

Rank correlation statistics are useful for determining whether a there is a correspondence between two measurements, particularly when the measures are of less interest than their ranks. Kendall's τ in particular has found use in Information Retrieval as a "meta-evaluation" measure: it has been used to compare evaluation measures, evaluate system rankings, and evaluate predicted performance. However, correlations are not intended to be used as evaluation measures: in the meta-evaluation domain, correlations between systems confound relationships between measurements, practically guaranteeing that τ will be highly positive regardless of any correlation between measures. We introduce an alternative measure of distance between rankings that has the additional benefit of a natural significance test for the hypothesis that two rankings are the same. Duplicating some classic evaluation experiments produces some surprising results. Like τ, our rank distance measure has application to IR outside of meta-evaluation; we also show that it is competitive as an objective function for learning a ranking in some "toy" problems.

1. INTRODUCTION

Ranking is a ubiquitous problem in Information Retrieval. Retrieval systems rank documents by estimated relevance; evaluations like the Text Retrieval Conference (TREC) rank systems by evaluation measures; systems rank queries by predicted difficulty; users rank systems by preference. Evaluating a ranking requires some measure of comparison between two rankings. In the case of ranking documents, there are a wide variety of evaluation measures that implicitly compare a ranking of documents to a perfect ranking; for the other tasks listed above, rank correlation measures, in particular Kendall's τ, have become *de facto* standards.

Kendall's τ is appealingly intuitive: given two different rankings of the same m items, count the number of pairs that are concordant—two items in the same order in both rankings—and discordant—two items in swapped order. If P is the concordance count and Q the discordance count,

$$\tau = \frac{P-Q}{P+Q}.$$

τ ranges from -1 to 1, with 1 meaning the two rankings are identical and -1 meaning one is in reverse of the other. A τ of 0 means that 50% of the pairs are concordant and 50% discordant. Every value of τ maps directly to a percentage of concordant pairs (assuming no ties).

But correlation measures are not evaluation measures, and care must be exercised when they are used that way. Correlation between variables can be confounded by correlation in the sample over which the variables are measured. Suppose we want to measure the rank correlation between two evaluation measures. If we calculate τ between rankings of a system that is nearly perfect, a middle-of-the-road system, and a system with a fatal bug, it does not matter how similar the measures are. If they capture anything at all about performance, they are guaranteed to be highly correlated. Correlation measures are meaningful when samples are drawn independently and with identical sampling distributions (i.i.d.), but when they are not—as retrieval systems are not—the interpretation becomes unclear.

Some recent work has revealed strange behavior by Kendall's τ when comparing rankings of systems [1, 11, 14, 15]. At a high level, much of this can be explained by correlations between systems: when they are ranking the same documents for the same topics, they will be so highly correlated that it is unclear whether τ has any meaning whatsoever. In Section 2 we explain the high-level problem in more detail, along with other hurdles to interpreting a τ correlation.

If our samples are not independent and identically distributed, a measure of distance between rankings should take into account the likelihood of the particular rankings. In Section 3 we develop such a measure, along with a significance test for the hypothesis that two rankings are the same. In Section 4 we compare our rank distance measure to Kendall's τ over simulation experiments as well as some classic experiments from IR evaluation studies. The work in these three sections is in the domain of meta-evaluation, or evaluating rankings of systems. In Section 5, we consider how the ideas in this paper might relate to rankings of documents.

2. THE TROUBLE WITH KENDALL'S TAU

Let us start with an example. Suppose we want to know how well a ranking of systems by mean average precision (MAP) correlates to a ranking of systems by precision@10.

We have three retrieval systems (call them A, B, and C) and four topics. We calculate average precision (AP) and precision@10 for each system and each topic, resulting in two 4×3 matrices:

$$AP = \begin{bmatrix} 0.283 & 0.481 & 0.516 \\ 0.017 & 0.399 & 0.544 \\ 0.075 & 0.300 & 0.277 \\ 0.183 & 0.616 & 0.662 \end{bmatrix} \quad p@10 = \begin{bmatrix} 0.8 & 0.8 & 0.8 \\ 0.2 & 0.7 & 0.5 \\ 0.3 & 0.5 & 0.5 \\ 0.7 & 1.0 & 1.0 \end{bmatrix}$$

where the matrix columns are the values for systems A, B, C respectively. Averaging the columns gives MAP and mean precision@10:

$$MAP = \begin{bmatrix} 0.139 & 0.449 & 0.500 \end{bmatrix}'$$
$$p@10 = \begin{bmatrix} 0.50 & 0.75 & 0.70 \end{bmatrix}'$$

The ranking of systems by MAP is $C > B > A$ while the ranking of systems by precision@10 is $B > C > A$. Since one pair out of three is swapped, the τ correlation is $\frac{2-1}{3} = \frac{1}{3}$.

First, this points out the coarseness of τ [11]: with only a few systems, a few swaps can have a big effect, even if they are immaterial. On the other hand, with many systems, swaps that are important to note can have little effect.

What makes a swap important? In this example, B and C are quite close in both MAP and precision@10 while A and B are distant in both, yet swapping either pair yields the same τ of $\frac{1}{3}$. If we look closer at B and C we see that not only are their MAPs and precisions close to each other, they are close for a reason: the two systems ranked almost exactly the same documents in almost exactly the same order. A also ranked some of the same documents, but not nearly as many as B and C. It is unlikely that any ranking would ever place A between B and C; in fact, it is much more likely that A would be ranked above both B and C! Kendall's τ does not reflect this situation, as the latter ranking would have a τ of either $-\frac{1}{3}$ or -1, depending on how B and C are ordered, but the former would have a τ of $\frac{1}{3}$ or $-\frac{1}{3}$.

If $C > A > B$ and $B > A > C$ are effectively impossible, then the distribution of τ values over pairs of rankings does not match the distribution it is intended to have. Any conclusions we draw are flawed—and we have not yet considered the relevance judgments. If A is simply a bad system by any reasonable measure, then it is a virtual certainty that $C > B > A$ or $B > C > A$. The only effectively possible values of τ are 1 and $\frac{1}{3}$, but if we are not aware of that, then our interpretation of a τ of $\frac{1}{3}$ will be flawed.

If a statistic is to have any meaning, it must take on a range of values with some distribution and offer an interpretation of what the estimated statistic means in relation to that distribution. The correlations between these systems obliterates the natural distribution of τ and hopelessly confounds any correlation between measurements.

Correlation between systems is endemic. There were 103 systems submitted to the TREC-7 ad hoc track. Calculating correlation of AP between all 5,253 possible pairs, only 4.5% have less than 30% correlation. 6.4% have less than 30% correlation in precision@10. And there is a good reason for it: the systems are retrieving the same documents that have the same relevance judgments to the same queries. Many systems are going to rank the same documents; even when they don't, their decisions are nevertheless based on similar information about the topics and corpus.

The high positive correlation between systems practically guarantees that τ between any two measures will be positive and high, regardless of the actual correlation between the measures. Without an understanding of how much of the correlation is due to systems and how much to the measures, a single measurement of τ is virtually meaningless.

This problem is not unknown, and solutions have been proposed. Cormack and Lynam proposed estimating power and bias of pairwise system comparisons by effectively calculating τ over only pairs with significant differences [6] (Sakai [13] independently did something similar). This would seem to be susceptible to the multiple comparisons problem, that in performing k independent hypothesis tests at level α, the probability that at least one produces a false positive is $1-(1-\alpha)^k$, which for large enough k means that some discordant swaps will be wrong. Furthermore, it ignores pairs that are so similar to each other that they should not be separated, like B and C in our example. Another solution, adapted by Aslam et al. in various works (e.g. [2, 3]), is to use root mean square error rather than τ. RMSE is sufficient but not necessary: while low RMSE usually implies high τ, the reverse is not generally true[1]. Melucci proposed that τ be replaced by Kolmogorov-Smirnov's D [11]; we discuss this below.

2.1 Sampling Error in Tau

Dependence between systems is not the only problem with using τ this way. Since it is calculated over a sample, it has sampling error. Kendall [9] showed that without an understanding of the distribution of measurements, the tightest 95% confidence interval that can be achieved is

$$c.i. = \frac{\tau \pm 1.96\sqrt{\frac{2}{m}}\sqrt{1 + \frac{2 \cdot 1.96^2}{m} - \tau^2}}{1 + \frac{2 \cdot 1.96^2}{m}}$$

where τ is calculated over our sample of systems (ignoring, for the moment, that topics also constitute a sample). For our example above, if $m = 3$ and $\tau = \frac{1}{3}$, the confidence interval is $(-0.741, 0.928)$! Even for large m, the confidence interval for τ is wide: for 100 systems with a τ of 0.85, the 95% confidence interval on τ is $(0.636, 0.943)$. This is far too wide to be able to say that a τ of 0.85 is any greater than a τ of, say, 0.75 ($c.i. = (0.512, 0.881)$). The intervals can be tightened if information about the distribution of measurements is used; Kendall [9] has details. The coarseness of τ along with wide confidence intervals may explain the results of Sanderson & Soboroff [14].

But this discussion is somewhat academic, as it assumes that the system sample is drawn i.i.d. from some population. Retrieval systems are not i.i.d. samples, of course: any set of systems is subject to technological limits, fashion, experimenter bias, training data available, and so on. Furthermore, it ignores a second source of sampling error: that from averaging measurements over multiple topics.

Given a sample of topics and a set of non-i.i.d. systems, we have two options: we can treat our systems as a fixed population, or we can treat them as an i.i.d. sample from some unknown population. Work using τ has almost always done the former. In one counterexample, Sakai [12] explains a significance test for τ that is only meaningful if systems are assumed to be a sample; he then produces a bootstrap estimate for the standard error of τ that treats the systems

[1]We note that these works have the specific goal of estimating evaluation measures, for which RMSE is well-suited.

as a fixed population and only considers error due to the sample of topics. The bootstrap estimate is not related to the significance test since they are based on completely different assumptions about the systems and topics.

If we do wish to treat systems as a sample, we can develop a bootstrap estimate of the confidence interval of τ resampling both systems and topics. A bootstrap procedure for estimating a confidence interval works as follows [18]: first, sample m systems and n topics from the set with replacement. Take the two measurements for each of the sampled systems over the sampled topics. Calculate the correlation between the rankings by the two measurements. After B different samples, find the 2.5% and 97.5% quantiles of the resulting set of τ correlations. This represents the 95% confidence interval of τ.

2.2 Significance of Tau

Kendall's τ can also be seen as the test statistic for a hypothesis test [9, 12, 11]. The default, or null, hypothesis is that there is no correlation between the two lists, i.e. $\tau = 0$. It will be rejected if the proportion of swapped pairs is significantly more or significantly less than half. It is important to understand that significance is evaluated over the population of *systems* (or in general whatever is being ranked), not topics, so again the fact that our sample of systems is not i.i.d. means this is of questionable utility.

Since correlation between systems guarantees a high τ, it is also almost a guarantee that the null hypothesis that there is no correlation will be rejected [11]. Looking through IR evaluation literature, the only experiments we have found that would *not* reject the null are comparisons between rankings over full qrels to rankings over a handful of qrels. The significance of τ is almost completely noninformative.

3. RANK DISTANCE

Kendall's τ and Spearman's ρ can be seen as proportional to a distance metric between ranked lists. In the case of Spearman's ρ, the distance is the sum of the squares of the difference in ranks at which items appear. Kendall's τ is the total number of pairwise swaps it would take to convert one list into the other. We would like a distance measure that is smaller for rankings that are more likely and larger for rankings that are less likely, where "more/less likely" takes into account the dependence between systems.

Consider the Mahalanobis distance [18]: if x and y are vectors drawn from a m-variate normal distribution with mean μ and covariance matrix Σ, then the Mahalanobis distance between x and y is defined to be:

$$d(x, y|\Sigma) = \sqrt{(x-y)'\Sigma^{-1}(x-y)}.$$

This is not a distance between rankings, as x and y are not lists of ranks. However, if we map x and y to their ranks and set Σ to the $m \times m$ identity matrix, then the Mahalanobis distance is proportional to Spearman's ρ. In other words, ignoring covariance between items produces Spearman's ρ. This shows explicitly how a traditional rank correlation measure ignores dependence. We cannot get Kendall's τ from the Mahalanobis distance, but τ and ρ are closely related.

Can the Mahalnobis distance be adapted for rankings? It is not as simple as mapping values to ranks, as that does not preserve covariance. Returning to our example:

$$MAP = \begin{bmatrix}0.139 & 0.449 & 0.500\end{bmatrix}' \; p@10 = \begin{bmatrix}0.5 & 0.75 & 0.7\end{bmatrix}'$$

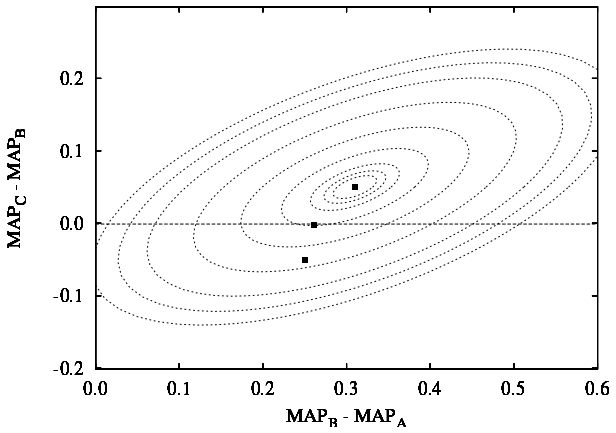


Figure 1: Contours of the bivariate normal distribution $\mu = [0.310\ 0.051]'$, $\Sigma = \left[\begin{smallmatrix}0.013 & 0.005\\ 0.005 & 0.005\end{smallmatrix}\right]$. Points are the mean, the precision values $(0.25, -0.05)$, and the minimum-distance rank-preserving point $(0.261, -0.001)$.

$$AP = \begin{bmatrix}0.283 & 0.481 & 0.516\\ 0.017 & 0.399 & 0.544\\ 0.075 & 0.300 & 0.277\\ 0.183 & 0.616 & 0.662\end{bmatrix} \; p@10 = \begin{bmatrix}0.8 & 0.8 & 0.8\\ 0.2 & 0.7 & 0.5\\ 0.3 & 0.5 & 0.5\\ 0.7 & 1.0 & 1.0\end{bmatrix}$$

$$\Sigma_{AP} = \begin{bmatrix}0.014 & 0.009 & 0.006\\ 0.009 & 0.018 & 0.020\\ 0.006 & 0.020 & 0.026\end{bmatrix} \; \Sigma_{p@10} = \begin{bmatrix}0.09 & 0.04 & 0.06\\ 0.04 & 0.04 & 0.05\\ 0.06 & 0.05 & 0.06\end{bmatrix}$$

Σ_{AP} and $\Sigma_{p@10}$ are the covariance matrices of AP and $p@10$; the diagonals are the variances for systems A, B, C and the off-diagonals covariances.

Clearly precision@10 comes from a different distribution than MAP. If we calculate the Mahalanobis distance from MAP to p@10, it will be huge despite the rankings not being far apart, simply because those values are very unlikely to occur under the distribution with mean MAP and variance Σ_{AP}. We cannot use the Mahalanobis distance directly.

Looking at the differences between measures preserves ordering and puts them closer to the same distribution:

$$AP_\Delta = \begin{bmatrix}0.198 & 0.035\\ 0.382 & 0.146\\ 0.226 & -0.023\\ 0.434 & 0.046\end{bmatrix} \; p@10_\Delta = \begin{bmatrix}0.0 & 0.0\\ 0.5 & -0.2\\ 0.2 & 0.0\\ 0.3 & 0.0\end{bmatrix}$$

$$MAP_\Delta = \begin{bmatrix}0.310 & 0.051\end{bmatrix}' \; p@10_\Delta = \begin{bmatrix}0.25 & -0.05\end{bmatrix}'$$

$$\Sigma_{AP_\Delta} = \begin{bmatrix}0.013 & 0.005\\ 0.005 & 0.005\end{bmatrix} \; \Sigma_{p@10_\Delta} = \begin{bmatrix}0.04 & -0.02\\ -0.02 & 0.01\end{bmatrix}$$

but still not close enough. We would like to find some vector of values that comes from the distribution of AP_Δ but preserves the ordering of precision@10, and then calculate the distance between MAP_Δ and that vector.

Unfortunately, there are infinitely many such values. Figure 1 shows a contour plot of the joint distribution of $MAP_B - MAP_A$ and $MAP_C - MAP_B$. The point in the center is the mean, $(.310, .051)$. Precision@10 is at $(.25, -0.05)$. Every point within the box $MAP_C - MAP_B < 0, MAP_B - MAP_A > 0$ preserves the ordering of precision values. Which one do we choose?

The contour lines represent multivariate normal confidence intervals: the innermost contour is the boundary of the 5% c.i.; the outermost 95%. Points occurring within more distant contours have greater Mahalanobis distance from the mean. Distance in principle goes to ∞ in the limit (though of course in reality MAP is bounded at 1, so distance is somehow bounded as well). One obvious choice is the point that preserves the relative ordering of precision@10 and has minimum distance from the mean. This is the point at $(.261, -0.001)$ in Figure 1.

The minimum-distance rank-preserving point is attractive because the distance of this point is directly affected by how likely two systems are to swap: the closer together two systems are, or the less covariance they have, the closer this point will be to the mean and therefore have less effect on the distance. In other words, as systems become more likely to swap, the minimum distance rank-preserving point gets closer to the mean, and therefore rank distance will get closer to zero.

Formally, given $n \times m$ matrix X with column means x, we define the rank distance from x to vector y to be the smallest Mahalanobis distance between x and a third vector θ that preserves the rank ordering of y, i.e.

$$d_{rank}(y; x|\Sigma_X) = \min_{\theta} \sqrt{(\Delta x - \Delta\theta)' \Sigma_X^{-1} (\Delta x - \Delta\theta)} \quad (1)$$
$$\text{s.t. } \tau(\theta, y) = 1$$

where x, y, θ, and the columns of X are sorted by decreasing x, Σ_X is the covariance matrix of $X\Delta'$, and Δ is an $n-1 \times n$ matrix that transforms a vector into a vector of adjacent differences: row i contains a 1 in position i and a -1 in position $i+1$. Note that we have found a use for Kendall's τ as a constraint on θ! Ties (which we have ignored to this point) can be handled by this constraint, using one of the modifications for ties proposed by Kendall [9]. If we ignore the square root, the objective function is convex [4] (there is another good reason for ignoring the square root that we will see in the next section). For relatively small n and m, solving this is fast.

Rank distance is not symmetric: $d_{rank}(p@10; MAP|\Sigma_{AP}) \neq d_{rank}(MAP; p@10|\Sigma_{p@10})$. Though it may seem counterintuitive, it is correct: a measure with high overall variance has more rankings that are "close" in some sense, while a measure with lower overall variance is more likely to be unique and therefore far from other rankings. This means that when calculating d_{rank}, we need to keep in mind which vector is the baseline and which is the test.

The caveat to this is that IR evaluation measures are typically not normally distributed over queries. Mahalanobis distance, and therefore rank distance, assume that the vectors come from a normal distribution with the specified covariance matrix. Overall, we suspect this is not a big problem. In Section 4.2 we investigate whether the values of θ produced by the minimization routine are reasonable.

3.1 A Rank Distance Hypothesis Test

Unlike Kendall's τ, which is always between -1 and 1, our rank distance measure produces a real number in the range $[0, \infty)$ whose meaning is not always clear. Knowing whether the difference is significant is important.

Contrary to the τ test's hypothesis that two rankings have no correlation, our null hypothesis will be that the two rankings are the same, i.e. that $d_{rank} = 0$. If the hypothesis is rejected, we can conclude that the rankings are different. If our rank distance measure is any good, this test should be far more informative than the τ test.

Mahalnobis distance can be seen as a multivariate generalization of the z-score $\frac{x-y}{\sigma}$, where σ is the standard deviation of a normal distribution from which x and y are drawn. Over a sample of n (x, y) pairs, the z-statistic is $\frac{x-y}{\sigma/\sqrt{n}}$. This is the statistic used to determine significance in a paired t-test. Applying the same idea to the Mahalanobis distance produces the T^2 statistic:

$$T^2 = \left(\sqrt{(x-y)'\Sigma^{-1}(x-y)}/\sqrt{n}\right)^2$$
$$= n(x-y)'\Sigma^{-1}(x-y)$$

T^2 is the test statistic in a multivariate generalization of the t-test known as *Hotelling's T^2 test* after Harold Hotelling, who discovered it [8]. Hotelling showed that T^2 has a well-understood distribution:

$$T^2 \sim \frac{(n-1)m}{n-m} F_{m,n-m}$$

where $F_{m,n-m}$ is the F distribution with degrees of freedom m (in our case, the number of systems) and $n-m$ (the number of topics minus the number of systems). Thus determining whether a Mahalanobis distance is significant is as easy as looking up a value in an F distribution table.

We can easily apply this to d_{rank}: using $m-1$ because taking adjacent differences reduces the degrees of freedom by 1, the test statistic is $\frac{n-m+1}{(n-1)(m-1)} n d_{rank}^2$. If this statistic is greater than the value of $F_{m-1,n-m+1}(\alpha)$, where α is the level of significance, then we can reject the null hypothesis: the rankings are not the same.

Again, the fact that IR evaluation measures are not normally distributed is a factor. We believe the test will tend to reject at a higher rate than it should. Unlike the standard use of hypothesis tests in IR, though, rejecting is the more conservative action in this case: believing that two rankings are different when they are actually the same is less harmful than believing that two rankings are the same when they are actually different. Melucci makes this argument as well, suggesting the use of Kolmogorov-Smirnov's D and its associated test that two rankings are the same [11]. Like τ, however, D fails to model dependence between items being ranked.

4. EXPERIMENTS

We first examine the d_{rank} measure in detail, comparing it to Kendall's τ. We then use it to re-examine some well-known results in IR evaluation.

4.1 Experimental Data

The standard datasets used in evaluation studies are the complete retrieval results of systems submitted to TREC tracks over the years, particularly the *ad hoc* track systems. For the ad hoc track, submitted systems retrieve up to 1,000 documents for each of 50 topics; the data consists of the full ranked list for every topic along with the relevance judgments (*qrels*) for those topics.

A constraint on the data is that d_{rank} and the associated hypothesis test require $n > m$, i.e. the number of topics be greater than the number of systems. (This can be relaxed for d_{rank}, but not for the hypothesis test.) Some of the standard collections used in these types of studies, e.g. TREC-5

through -8 ad hoc systems. cannot be used since they have more than 50 systems over 50 topics. Therefore we have restricted our data to the TREC-3 ad hoc collection (40 systems over 50 topics) and the TREC-13 Robust collection (110 systems over 249 topics).

4.2 Properties of Rank Distance

First, we wanted to see how rank distance changed as adjacent systems swapped. Starting with the 40 TREC-3 systems ordered by MAP, we randomly picked one system to exchange with a neighbor with greater MAP (if it had no such neighbors, no change was made). This would ensure that the ranking of systems would go from perfect to perfectly random to perfectly inverted after enough trials, and the τ correlation would decrease monotonically and linearly.

Figure 2(a) shows τ decreasing and d_{rank} changing as the number of swapped pairs goes to 1000. Overall, the correlation between the two is high (-0.718 Pearson correlation), but drops as the number of swaps increases: the correlation over the first 100 swaps is -0.85, but -0.46 over the last 100 swaps. This is because there are few ways that τ can be, say 0.99, many ways it can be 0.75, and many more ways it can be 0. Depending specifically on which pairs are swapped, these can have very different rank distances. Significance is achieved very quickly: after only a dozen swaps, the ranking is significantly different from the true ranking by our hypothesis test (for $m = 40, n = 50$, the critical value is 2.967). This is because all pairs are equally likely to swap, including those that are significant. τ, by contrast, does not lose significance until it is slightly under 0.2.

For the second experiment we attempted to model swaps more realistically. A pair would swap with its neighbor with high probability (0.95) if they were not significantly different by a paired, two-sided t-test at $\alpha = 0.95$, but with low probability (0.05) if they were. Figure 2(b) shows τ and d_{rank} changing in this experiment. Note that both have long sequences of not changing as significant pairs are selected and not swapped. Significance is achieved much more slowly and the distances are overall an order of magnitude lower, suggesting that swapping non-significant pairs has less of an affect on rank distance. The two measures are less correlated over the same range of τ: for $\tau \in [1, 0.75]$, the correlation for the first experiment is -0.88, but -0.76 for the second.

We looked at some of the biggest increases and decreases in rank distance in both experiments. In the first experiment, the single biggest increase (from 25.6 to 83.5) occurred at trial 436, when a swap caused a cluster of good systems to be separated by a bad system. The biggest drop (from 64.2 to 27.0) occurred when a single swap brought two pairs of similar systems that had become separated back together. There is also a big drop from trials 260–263; this happened because consecutive swaps brought pairs of very similar separated systems back together.

In the second experiment, the single biggest increase (from 8.9 to 20.8) occurred with the swap of a significant pair that also caused a greater disparity between another significant pair. The biggest drop (from 20.2 to 12.6) undid one of those swaps, resulting in a situation similar to our example from Section 2 in which $C > A > B$ is actually less likely that $C > B > A$.

Though they are irrelevant to our interpretation of d_{rank}, we also looked at the values of θ that the minimization procedure found. We would like them to "look like" MAP values; if they do not, then the normality assumption is a serious problem. Figure 2(c) compares the distribution of θs to the distribution of MAPs; they appear to be quite similar, and a goodness-of-fit test cannot reject the hypothesis that they are the same.

depth	judgments	τ (95% c.i.)	d_{rank}, p-value
1	4,789	0.67 (0.56, 0.76)	3.39, p=0.00
5	16,478	0.83 (0.75, 0.88)	1.97, p=0.00
10	28,569	0.89 (0.83, 0.93)	1.48, p=0.00
25	60,099	0.96 (0.92, 0.97)	0.84, p=0.71
50	101,626	0.98 (0.96, 0.99)	0.46, p=1.00

Table 2: Kendall's τ and rank distance to official ranking when evaluating over shallow pools.

4.3 Comparing Evaluation Measures

It is well-known that different evaluation measures correlate highly. Voorhees and Harman give τ correlations between rankings by different evaluation measures for the TREC-7 data [17]. By our argument in Section 2, this should not be surprising.

Table 1(a) shows τ correlations between rankings of 110 Robust systems over 249 topics by traditional TREC evaluation measures. It also shows the standard errors determined by the bootstrap procedure described in Section 2.1. There is a fair amount of overlap in the c.i.s, for example τ(P10,P30) and τ(P10,R-prec) overlap a great deal.

Table 1(b) shows the rank distances between the same rankings. As explained above, rank distance is not symmetric: it depends on which measure is chosen as the "baseline". This table presents distance from the measure on the row (the baseline) to the measure on the column and the p-value of the hypothesis test[2]. The table shows that rankings by bpref and MAP are indistinguishable from each other; that while a ranking by MAP is not significantly different from a ranking by R-prec, the converse is not true; that P30 and R-prec are not significantly different from each other (likely because these topics were chosen for having few relevant documents); and that a ranking by MRR is very different from anything else. The rank distances are correlated to the τs, but not perfectly, indicating that they are capturing different qualities of the rankings. However, all of the τs are significant, while the nonsignificant rank distances are enlightening.

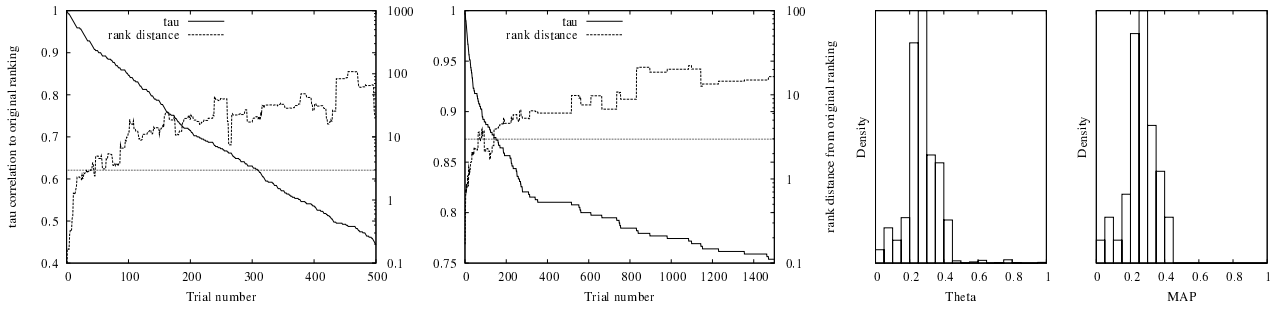
4.4 Incomplete Relevance Judgments

4.4.1 Shallow Pools

It is commonly understood that evaluation over shallow judgment pools correlates highly to evaluation over a deep pool. Again, this should not be surprising given what we now know about rank correlation methods.

Table 2 shows τ correlation and rank distance between a baseline ranking of Robust systems by official MAP and the ranking by MAP calculated over a shallow pool. While τ reaches nearly 0.9 with a depth 10 pool, the rank distance test rejects the hypothesis that the rankings are the same at that point. It takes a deeper pool before the hypothesis test fails to reject.

[2]For $n = 249, m = 110$, the critical value for rejection at $\alpha = 0.05$ is 1.02.



(a) Random adjacent swaps. (b) Swaps determined by significance. (c) Distributions of θ and MAP.

Figure 2: Comparison between τ and rank distance as adjacent pairs in a ranking are randomly swapped.

	P10	P30	R-prec	MAP	MRR	bpref
P10	1	0.84 (0.76, 0.89)	0.81 (0.71, 0.86)	0.80 (0.70, 0.85)	0.73 (0.60, 0.79)	0.80 (0.71, 0.86)
P30		1	0.90 (0.82, 0.92)	0.89 (0.81, 0.91)	0.63 (0.48, 0.71)	0.89 (0.81, 0.91)
R-prec			1	0.92 (0.86, 0.94)	0.61 (0.57, 0.70)	0.93 (0.89, 0.95)
MAP				1	0.58 (0.44, 0.68)	0.96 (0.91, 0.97)
MRR					1	0.60 (0.46, 0.70)
bpref						1

(a) Kendall's τ correlation between rankings by different evaluation measures averaged over 249 topics.

	P10	P30	R-prec	MAP	MRR	bpref
P10	0	1.67, p=0.00	2.01, p=0.00	1.89, p=0.00	2.73, p=0.00	1.71, p=0.00
P30	1.65, p=0.00	0	0.97, p=0.14	1.06, p=0.02	4.73, p=0.00	0.86, p=0.59
R-prec	2.16, p=0.00	0.97, p=0.14	0	0.78, p=0.92	4.30, p=0.00	0.67, p=1.00
MAP	3.23, p=0.00	1.51, p=0.00	1.56, p=0.00	0	6.08, p=0.00	0.75, p=0.96
MRR	6.24, p=0.00	12.69, p=0.00	7.97, p=0.00	8.22, p=0.00	0	9.36, p=0.00
bpref	3.27, p=0.00	1.29, p=0.00	1.16, p=0.00	0.69, p=1.00	5.73, p=0.00	0

(b) Rank distance from the ranking by the measure on the row to the ranking by the measure on the column, with hypothesis test p-values.

Table 1: Comparisons between ranking by common evaluation measures over the 110 Robust systems.

4.4.2 *Incomplete Judgments*

Buckley and Voorhees introduced the *bpref* measure for evaluation with incomplete test collections [5]. To test it, they compared bpref, MAP, R-prec, and P10 over increasingly incomplete sets of relevance judgments. We duplicated their experiment on the Robust collection, calculating d_{rank} and the rank test p-value in addition to τ between the official ranking of systems by each measure and the ranking of systems by the same measure calculated over the reduced set. (For full details of the methodology we refer the reader to the original paper.)

Figure 3(a) shows τ correlation between a measure calculated with a reduced qrels set and the same measure over all the qrels. bpref seems to be more robust to missing judgments, having a higher τ with very incomplete sets; R-precision seems to be least robust to missing judgments. This agrees with the results of Buckley and Voorhees.

Figure 3(b) tells a slightly different story. This shows rank distance decreasing as the qrels becomes more complete. In this case, bpref is *least* robust when judgments are most incomplete. However, it very quickly becomes the most robust. The other three are almost equally susceptible to missing judgments, with MAP being slightly more robust overall and R-prec being slightly less robust with the most incomplete set.

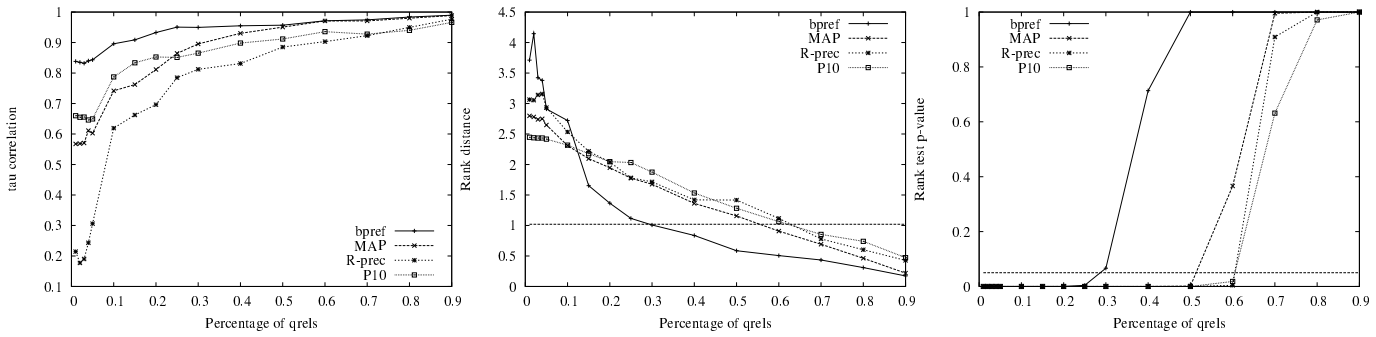
Finally, Figure 3(c) shows the p-value of the rank hypothesis test increasing as qrels completeness increases. Recall that a low p-value means that the ranking is significantly different; higher p-value means more similar to the "true" ranking. Here we see that bpref very quickly "loses" significance between 25% and 30% qrels, confirming the quick decline in the rank distance plot. The other three measures lose significance in the interval $0.5 - 0.7$, meaning that more than half the full collection was needed before the rankings were indistinguishable. All four measures very quickly transition from "significant" to "not significant"; this suggests there is a sort of "phase transition" due to the number of judgments, at which point the variance constrains the ranking enough that it is unlikely to change drastically.

This suggests that, in contrast to what τ shows, bpref is less good for very small collections, but very good for mid-sized collections. In Section 3 we discuss the possible reasons, which tie into the theme of this paper.

4.4.3 *No Relevance Judgments*

Soboroff et al. examined ranking systems by, in effect, random judgments [15]. They randomly sampled subsets of documents to be labeled relevant. They then used those pseudo-relevance judgments ("pseudo-rels") to evaluate and rank systems. They found a positive and significant τ correlation to the true ranking.

We duplicated the experiment on the Robust systems.



(a) Kendall's τ. (b) Rank distance. (c) Rank test p-value.

Figure 3: Reduced qrels. Horizontal lines indicate significance; values above the line in (b) and below the line in (c) are significant at $\alpha = 0.05$. Every value in (a) is significant.

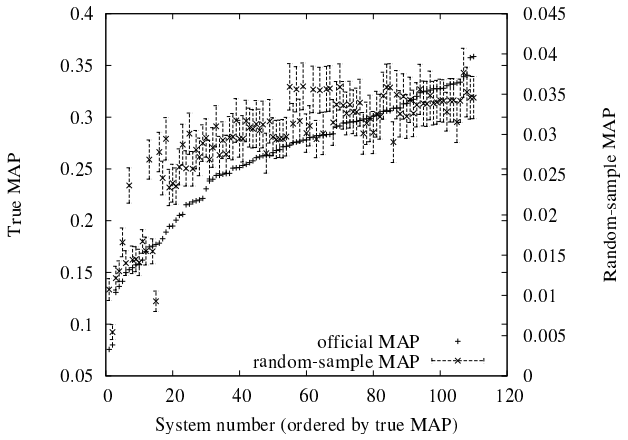


Figure 4: Rankings of Robust systems by official MAP and random sample MAP. The τ correlation is 0.656; the rank distance is 2.40.

Over 50 sets of pseudo-rels that were assembled (sampling from the pool without duplicates), the average rank distance was 2.40 with a standard deviation of 0.013, while the average τ correlation was 0.656 with a standard deviation of 0.021. Although the rank distance was relatively low, it was significant; the hypothesis that the ranking by pseudo-rels was equivalent to the ranking by the qrels was rejected ($p \approx 0$). By contrast, the τ correlation was also significant, indicating that the ranking was highly correlated—in this case a noninformative result.

The low rank distance here compared to some of the above experiments can be explained by the underlying reason for correlation between systems: they are retrieving the same documents for the same topics (Aslam et al. make this argument as well [1]; this is also well-known in metasearch [10]). Pseudo-rels tend to keep systems that are similar close together and systems that are very different far apart, and that is the kind of performance that rank distance rewards. It does not reward it enough to lead us to believe the ranking is correct, though. Figure 4 shows the 110 Robust systems' MAPs and random-sample MAPs sorted in increasing order of true MAP. The figure gives some idea of what a rank distance of 2.4 might look like.

5. DISCUSSION AND IMPLICATIONS

Kendall's τ correlation or variants appear in other contexts than meta-evaluation. In the TREC Robust track, τ is used to compare predicted performance to actual performance [16]. This seems to be a fairly appropriate use of τ—the sample space is topics instead of retrieval systems, and while topics are not i.i.d. samples, we have seen circumstantial evidence that they can be treated as if they are.

5.1 bpref

The bpref measure can be seen as a variant of tau, comparing a ranking of relevant and nonrelevant documents to a perfect ranking in which all relevant documents are ranked above all nonrelevant documents. A concordant pair is a relevant document ranked above a nonrelevant document; a discordant pair is the reverse.

bpref makes three concessions to the inefficiency of τ:

1. it abstracts away details of items to just their relevance labels (of course, all IR evaluation measures do this to some extent);
2. it ignores ties: pairs of relevant documents and pairs of nonrelevant documents are not counted as either concordant or discordant;
3. as implemented in `trec_eval` v8.1, when $R < N$, it counts only as many nonrelevant documents as relevant documents (note that this is different from bpref as described in the original paper).

By abstracting away details and ignoring ties, correlation between pairs of top-ranked relevant documents and pairs of top-ranked nonrelevant documents can only have a limited effect: while there will certainly be sets of documents for which a ranking is in a sense pre-ordained by certain retrieval models (similar to our example in Section 2), if they all have the same label they will not count. If they have different labels, their effect on bpref is reduced by other sets with the same labels but less correlation.

If the abstraction factors out correlations completely, then only counting an equal number of relevant and nonrelevant documents ensures that a "random" ranking of $\min\{2|R|, 2|N|\}$ judged documents will have an expected bpref of 0.5. Higher bpref therefore definitely means that the system did a better job of ranking relevant documents above nonrelevant documents than one with a lower bpref.

Using a re-centered cosine similarity that is equivalent to Pearson's correlation, we found that the expected correlation between a randomly-chosen relevant and nonrelevant document is indeed close to 0 in the TREC-3 corpus. If this continues to hold beyond simple linear correlation of bags-of-words, this explains why bpref does so well when qrels are degraded randomly.

5.2 Learning to Rank

Our rank distance measure can also be used as an objective function for learning to rank. If it is competitive with other learning to rank algorithms, that lends weight to the idea that it is correct.

Say we are given matrix X with column means x and feature matrices $\Phi_1, \Phi_2, ...$ We want to learn weights β_i on Φ_i such that the ranking by the column means of $\beta_0 + \beta_1\Phi_1 + \beta_2\Phi_2 + ...$ are as close as possible to the ranking by x. In our examples above, X might be AP values, x MAP values, and Φ_i features such as the number of relevant documents retrieved by rank 10 by each system for each topic, a prediction of the performance of each system on each topic, or anything else.

More formally, let Y be the matrix $\beta_0 + \beta_1\Phi_1 + ...$, and let y be the column means of Y. Note that $y = \beta_0 + \beta_1\phi_1 + ...$, where ϕ_i is the column means of Φ_i. We want to find

$$\arg\min_{\beta} d_{rank}(\beta_0 + \sum \beta_i\phi_i; x|\Sigma) + C\beta'\beta$$

where Σ is, as in Eq. 1, the covariance matrix of $X\Delta'$ and $C\beta'\beta$ is a regularization term that acts as a smoother and also allows for higher-dimensional sets of features.

We tested this with a "toy" learning to rank problems with our TREC data: learning a ranking by MAP given recall at rank 1000 and MRR, i.e. the percent of relevant documents retrieved and the (reciprocal) rank of the first relevant document. To find β, we used numerical optimization methods built into `R`. We used the TREC-3 set for training and tested on the Robust set. We compared our results to the ranking SVM of Joachims [7] (with linear kernel).

The result: our objective function has higher τ correlation (0.819 for the SVM to 0.859 for d_{rank}) and lower rank distance (5.08 for the SVM to 2.29 for d_{rank}) on the Robust systems. It is doubtful that the differences are significant. But the fact that it performs at least as well as the ranking SVM suggests that it is a correct measure of distance.

It is not immediately clear how to apply this to document retrieval. There are two challenges: (1) there is generally not a single correct total ordering of documents to train against; (2) it requires some estimate of correlation between documents. Nonetheless, this seems a promising direction for future work.

6. CONCLUSION

We have presented a new measure of rank distance that, unlike Kendall's τ, captures something about the likelihood of a particular ranking occurring and has the additional advantage of a natural hypothesis test. While it tracks τ enough to be believable, it also clearly disagrees with τ about some well-known results. Though it requires a distribution assumption, it appears to be relatively robust to its violation (on inspection).

There are some clear directions for future work. Our argument against τ suggests an argument *for* bpref; it would be interesting to re-examine some recent work that was critical of bpref in that light. The fact that our rank distance measure is also competitive as an objective function for learning to rank lends weight to its correctness and also points to additional work in applying it to learning rankings of documents.

Acknowledgments

This work was supported in part by the Center for Intelligent Information Retrieval and in part by Microsoft Live Labs. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author and do not necessarily reflect those of the sponsor.

7. REFERENCES

[1] J. Aslam and R. Savell. On the effectiveness of evaluating retrieval systems in the absence of relevance judgments. In *Proceedings of SIGIR*, pages 361–362, 2003.

[2] J. A. Aslam, V. Pavlu, and E. Yilmaz. A statistical method for system evaluation using incomplete judgments. In *Proceedings of SIGIR*, pages 541–548, 2006.

[3] J. A. Aslam and E. Yilmaz. Inferring document relevance via average precision. In *Proceedings of SIGIR*, pages 601–602, 2006.

[4] S. Boyd and L. Vandenberghe. *Convex Optimization*. Cambridge University Press, 2004.

[5] C. Buckley and E. M. Voorhees. Retrieval evaluation with incomplete information. In *Proceedings of SIGIR*, pages 25–32, 2004.

[6] G. V. Cormack and T. R. Lynam. Power and bias of subset pooling strategies. In *Proceedings of SIGIR*, pages 837–838, 2007.

[7] T. Joachims. Optimizing search engines using clickthrough data. In *Proceedings of KDD*, pages 133–142, 2002.

[8] R. A. Johnson and D. W. Wichern. *Applied Multivariate Statistical Analysis*. Prentice-Hall, 1982.

[9] M. Kendall. *Rank Correlation Methods*. Griffin, London, UK, fourth edition, 1970.

[10] J. H. Lee. Analyses of multiple evidence combination. In *Proceedings of SIGIR*, pages 267–276, 1997.

[11] M. Melucci. On rank correlation in information retrieval evaluation. *SIGIR Forum*, 41(1):18–33, 2007.

[12] T. Sakai. Evaluating evaluation metrics based on the bootstrap. In *Proceedings of SIGIR*, pages 525–532, 2006.

[13] T. Sakai. Alternatives to bpref. In *Proceedings of SIGIR*, pages 71–78, 2007.

[14] M. Sanderson and I. Soboroff. Problems with kendall's tau. In *Proceedings of SIGIR*, pages 839–841, 2007.

[15] I. Soboroff, C. Nicholas, and P. Cahan. Ranking Retrieval Systems without Relevance Judgments. In *Proceedings of SIGIR*, pages 66–73, 2001.

[16] E. Voorhees. Overview of the TREC 2005 Robust Retrieval Track. In *TREC 2005 Notebook*, 2005.

[17] E. M. Voorhees and D. K. Harman, editors. *TREC: Experiment and Evaluation in Information Retrieval*. MIT Press, 2005.

[18] L. Wasserman. *All of Statistics*. Springer, 2006.