Toward Estimating the Rank Correlation between the Test Collection Results and the True System Performance

Julián Urbano
Universitat Pompeu Fabra
Barcelona, Spain
urbano.julian@gmail.com

Mónica Marrero
National Supercomputing Center
Barcelona, Spain
monica.marrero@bsc.es

ABSTRACT

The Kendall τ and AP rank correlation coefficients have become mainstream in Information Retrieval research for comparing the rankings of systems produced by two different evaluation conditions, such as different effectiveness measures or pool depths. However, in this paper we focus on the *expected* rank correlation between the mean scores observed with a test collection and the *true*, unobservable means under the same conditions. In particular, we propose statistical estimators of τ and AP correlations following both parametric and non-parametric approaches, and with special emphasis on small topic sets. Through large scale simulation with TREC data, we study the error and bias of the estimators. In general, such estimates of expected correlation with the true ranking may accompany the results reported from an evaluation experiment, as an easy to understand figure of reliability. All the results in this paper are fully reproducible with data and code available online.

Keywords

Evaluation; Test Collection; Correlation; Kendall; Average Precision; Estimation

1. INTRODUCTION

The Kendall τ [3] and AP [8] rank correlation coefficients are widely used in Information Retrieval to compare rankings of systems produced by different evaluation conditions, such as different assessors [6], effectiveness measures [4] or topic sets [1]. One reason for this success is their simplicity: they provide a single score that is easy to understand.

In this paper we tackle the problem of estimating the correlation between the ranking of systems obtained with a test collection and the *true* ranking under the same conditions. Such estimates can make a nice companion to a set of evaluation results, as a single figure of the reliability of the experiment. Voorhees and Buckley [7] proposed to report a similar figure in terms of sensitivity, that is, the minimum difference required between two systems to ensure a maximum error rate in relative comparisons. Common practice nowadays is to report the p-value of a statistical significance test run either for each pair of systems (e.g. t-test) or for the whole set (e.g. ANOVA and F-test). They provide a sense of confidence about individual pairs of systems or about a swap somewhere in the ranking, but they do not give a general idea of how similar the observed ranking is to the truth.

We propose parametric and non-parametric approaches to estimate the τ and AP correlations. Through large scale simulation with TREC data, we show that they have very low bias and small error even for mid-sized collections.

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

ISBN 978-1-4503-4069-4/16/07…$15.00

DOI: http://dx.doi.org/10.1145/2911451.2914752

2. CORRELATION BETWEEN TWO RANKINGS

Let $A = \langle a_1, \ldots, a_m \rangle$ and $B = \langle b_1, \ldots, b_m \rangle$ be the mean scores of the same set of m systems as observed under two different evaluation conditions, such that a_i and b_i refer to the i-th system. In many situations we are interested in the distance between the two rankings. Considering systems in pairs, a distance can be computed by counting how many pairs are concordant or discordant between the two rankings: a pair is concordant if their relative order is the same in both rankings, and discordant if it is the opposite. Kendall [3] followed this idea to define his τ correlation coefficient

$$\tau = \frac{\#concordants - \#discordants}{total} = 1 - 2\frac{\#discordants}{total}, \quad (1)$$

which evaluates to -1 when the rankings are reversed, $+1$ when they are the same, and 0 when there are as many concordant pairs as there are discordant. Note that the term $\#discordants/total$ can be interpreted as the expected value of a random experiment: pick two arbitrary systems and return 1 if they are discordant, or 0 if they are concordant. The Kendall τ coefficient can thus be interpreted in terms of the probability of discordance.

Yilmaz et al. [8] followed this idea to define a correlation coefficient with the same rationale as Average Precision. It is similar to Kendall τ, but it penalizes more if swaps occur between systems at the top of the ranking, much like AP penalizes more if the non-relevant documents appear at the top of the results. In particular, they considered that one of the rankings, say B, is the true ranking and the other one is an estimate of it. The random experiment is now as follows: pick one system at random from A and another one ranked above it, and return 1 if they are discordant, or 0 if they are concordant. Their AP correlation coefficient can then be defined just as in (1) as follows:

$$\tau_{AP} = 1 - \frac{2}{m-1}\sum_{i=2}^{m}\left(\frac{\#discordants\ above\ i}{i-1}\right). \quad (2)$$

Note that τ_{AP} also ranges between -1 and $+1$.

3. EXPECTED CORRELATION WITH THE TRUE RANKING

The previous section contemplated the case where we compute the correlation between two given rankings A and B. In this section we study the case where we are given a ranking A obtained with the sample of topics in the test collection, and want to estimate its correlation with the true ranking B over the population of topics, which is of course unknown. For simplicity, let us first assume that the systems are already sorted in descending order by their mean score. Let us further define D_{ij} as the random variable that equals 1 if systems i and j are discordant and 0 otherwise, that is, whether they are swapped in the true ranking. Both τ and τ_{AP} can be re-defined from (1) and (2) in terms of D_{ij} alone:

$$\tau = 1 - \frac{4}{m(m-1)}\sum_{i=1}^{m-1}\sum_{j=i+1}^{m} D_{ij}, \quad (3)$$

$$\tau_{AP} = 1 - \frac{2}{m-1}\sum_{i=1}^{m-1}\sum_{j=1}^{i-1}\frac{D_{ij}}{i-1}. \quad (4)$$

Since they are just a linear combination of random variables, their expectations are as in (3) and (4) but replacing D_{ij} with $\mathrm{E}[D_{ij}]$. Note that each D_{ij} is a Bernoulli random variable, so its expectation is just the probability of discordance $\mathrm{E}[D_{ij}] = P(\mu_i - \mu_j < 0) = p_{ij}$. The problem of estimating the correlation with the true ranking thus boils down to estimating the probability that any two systems are swapped. The next subsection presents four ways of achieving this.

3.1 Estimating the Probability of Discordance

Since each p_{ij} is estimated independently from the other systems, let us simplify notation here to just p. In addition, let $X_1, \ldots, X_n$ be the differences in effectiveness between the two systems and for each of the n topics in the collection. The problem is therefore to estimate $p = P(\mu < 0)$ from these n observations. Recall that systems are assumed to be ranked by mean observed scores, so $\overline{X} > 0$.

In the following we present two parametric estimators based on the Central Limit Theorem (CLT) and then two non-parametric estimators based on resampling.

3.1.1 Maximum Likelihood (ML)

The CLT tells us that $\overline{X}$ is approximately normally distributed with mean μ and variance σ^2/n as $n \to \infty$. Using the *cdf* of the normal distribution we can therefore estimate the probability of discordance. However, our estimates are likely off with small samples (see Section 3.1.2), so we assume $X_i \sim N(\mu, \sigma^2)$ and employ the t distribution to account for the uncertainty in estimating σ^2. Standardizing, we have that $\sqrt{n}(\overline{X} - \mu)/\sigma \sim t(n-1)$, so

$$p = P(\mu < 0) \approx T_{n-1}\left(-\sqrt{n}\frac{\hat{\mu}}{\hat{\sigma}}\right), \quad (5)$$

where T_{n-1} is the *cdf* of the t distribution with $n-1$ degrees of freedom. The estimates $\hat{\mu}$ and $\hat{\sigma}$ are computed via Maximum Likelihood as

$$\hat{\mu} = \overline{X} = \frac{1}{n}\sum X_i, \quad (6)$$

$$\hat{\sigma} = s \cdot C_n, \quad (7)$$

$$s = \sqrt{\frac{1}{n-1}\sum (X_i - \overline{X})^2},$$

$$C_n = \sqrt{\frac{n-1}{2}}\frac{\Gamma((n-1)/2)}{\Gamma(n/2)},$$

where s is the sample standard deviation. The C_n factor [2] ensures that $\mathrm{E}[\hat{\sigma}] = \sigma$. This bias correction is applied because, even though s^2 is an unbiased estimator of σ^2, by Jensen's inequality s is *not* an unbiased estimator of σ.

3.1.2 Minimum Squared Quantile Deviation (MSQD)

The problem when estimating σ from a small sample is that the observations are likely to be concentrated around the mean and seldom occur near the tails. As a consequence, (7) is likely to underestimate the true dispersion in the population. If the sample contains a few dozen observations this is not expected to be a problem, but with very small samples of, say, just 10 topics, it might be.

We propose a new and generic estimator to avoid this problem. Let us consider a distribution function F with parameter θ. A random sample from this distribution is expected to uniformly cover the quantile space, that is, all quantiles are equally likely to appear in the sample. Thus, when we are given a sample we may force them to uniformly cover the quantile space and then select the θ that minimizes the observed deviations. For instance, if our sample contains only one observation, we force it to correspond to the quantile 1/2; if we have two observations then we force them to be the quantiles 1/3 and 2/3. In general, if R_i is the rank of X_i within the sample, it will correspond to the $R_i/(n+1)$ quantile, which is $F^{-1}\left(\frac{R_i}{n+1}; \theta\right)$. The squared quantile deviation of an observation X_i is therefore

$$SQD(X_i; \theta) = \left(F^{-1}\left(\frac{R_i}{n+1}; \theta\right) - X_i\right)^2.$$

The Minimum Squared Quantile Deviation estimator is then the one that minimizes the sum of squared deviations:

$$\hat{\theta}_{MSQD} = \arg\min_{\theta \in \Theta} \sum SQD(X_i; \theta).$$

Let us assume again a normal distribution, so that

$$F^{-1}\left(\frac{R_i}{n+1}; \mu, \sigma\right) = \mu + \sigma\sqrt{2}e_i,$$

$$e_i = \mathrm{erf}^{-1}\left(2\frac{R_i}{n+1} - 1\right).$$

The sum of squared deviations is thus

$$\sum \mu^2 - \cancelto{0}{2\mu\sigma\sqrt{2}e_i} + 2\sigma^2 e_i^2 - 2X_i\mu - 2X_i\sigma\sqrt{2}e_i + X_i^2,$$

and the second term cancels out because $\sum e_i = 0$. To find the μ and σ that minimize this expression, we simply differentiate, equal to 0, and solve. The partial derivatives are

$$\frac{\mathrm{d}SQD}{\mathrm{d}\mu} = \sum 2\mu - 2X_i = 2n\mu - 2\sum X_i \text{ and}$$

$$\frac{\mathrm{d}SQD}{\mathrm{d}\sigma} = \sum 4\sigma e_i^2 - 2\sqrt{2}X_i e_i,$$

and therefore, the estimators are

$$\hat{\mu} = \frac{1}{n}\sum X_i = \overline{X}, \tag{8}$$

$$\hat{\sigma} = \frac{\sqrt{2}\sum X_i \cdot \mathrm{erf}^{-1}\left(2\frac{R_i}{n+1} - 1\right)}{2\sum \mathrm{erf}^{-1}\left(2\frac{R_i}{n+1} - 1\right)^2}. \tag{9}$$

As above, the probability of discordance is estimated with the *cdf* of the t distribution as in (5), but using estimators (8) and (9) instead of (6) and (7).

3.1.3 Resampling (RES)

In both the ML and MSQD estimators above we assumed that scores are normally distributed, but this is clearly not strictly true. A non-parametric alternative is the use of resampling to estimate the sampling distribution of the mean and from there the probability of discordance.

Suppose we draw a random sample $X_1^*, \ldots, X_n^*$ with replacement from our original observations, and compute their sample mean $\overline{X}^*$. This experiment is replicated $T = 1,000$ times, yielding sample means $\overline{X}_1^*, \ldots, \overline{X}_T^*$. By the law of large numbers, the distribution of these sample means converges to the sampling distribution of $\overline{X}$ as $T \to \infty$. The probability of discordance can thus be estimated as the fraction of times that $\overline{X}_i^*$ is negative:

$$p = P(\mu < 0) \approx \frac{1}{T}\sum \mathrm{I}\big[\overline{X}_i^* < 0\big]. \tag{10}$$

3.1.4 Kernel Density (KD)

A potential problem with resampling from the original observations is again that estimates from very small samples are likely off. An alternative is to approximate the true *pdf* via Kernel Density Estimation, and use it to estimate the probability of discordance. The estimated *pdf* has the form

$$\hat{f}(x) = \frac{1}{nh}\sum k\left(\frac{x - X_i}{h}\right),$$

where k is the *pdf* of the kernel and h is the bandwidth. Next, we need to estimate the sampling distribution of the mean, which is basically the distribution of the sum of n variables drawn from $\hat{f}$. For $n = 2$ this requires the evaluation of the self-convolution of $\hat{f}$ as follows:

$$\hat{f}_{X+X}(x) = \frac{1}{n^2h^2}\sum_i\sum_j \int k\left(\frac{x - z - X_i}{h}\right) k\left(\frac{z - X_j}{h}\right)\,\mathrm{d}z,$$

which involves the sum of n^2 terms. In general, for n variables this requires the evaluation of n^n terms, which is clearly unfeasible even for small samples, so instead we resort to Monte Carlo methods. As with the RES estimator, we generate a random sample $X_1^*, \ldots, X_n^*$ from $\hat{f}$ and compute the mean $\overline{X}^*$. After T replications, the probability of discordance is estimated as the fraction of times that $\overline{X}_i^*$ is negative. We set $T = 1,000$ replications and use gaussian kernels.

4. EVALUATION

4.1 Criteria

There are two properties of the correlation estimators that we are interested in, namely error and bias. Error refers to the expected difference between the estimate and the truth. Here we measure absolute error, thus quantifying the expected magnitude of the error when estimating the correlation of a given collection:

$$error = \mathrm{E}\big[|\hat{\tau}(\mathbf{A}) - \tau(\mathbf{A}, \boldsymbol{\mu})|\big].$$

Even if the error is small, it could tend to be in the same direction, that is, over- or underestimating the correlation. Bias refers to this tendency, measured as the expected difference between the estimated and the true correlation:

$$bias = \mathrm{E}\big[\hat{\tau}(\mathbf{A}) - \tau(\mathbf{A}, \boldsymbol{\mu})\big].$$

If the bias is positive it means that the estimator tends to overestimate the correlations. In general, we seek estimators with small error and zero bias.

4.2 Methods, Data and Baselines

From the above definitions it is evident that we need to know the true ranking of systems $\boldsymbol{\mu}$, but this is of course unknown. To solve this problem we resort to the simulation method proposed by Urbano [5]. Given the topic-by-system matrix of scores $\mathbf{B}$ from an existing collection, it generates a new matrix $\mathbf{A}$ with the scores by the same set of systems over a new and random set of topics. There are two important characteristics of this method that are appealing for us. First, the simulated scores are realistic, as they maintain the same distributions and correlations among systems as in the original collection. Second, it is designed to ensure that the expected mean score of a system is equal to the mean score in the original collection, that is, $\mathrm{E}\big[\overline{A}_s\big] = \overline{B}_s$. For us, this means that the true mean scores are fixed to be the mean scores in the original collection, that is, $\mu_s := \overline{B}_s$. This allows us to analyze the error and bias of the estimators with a large number of simulated, yet realistic test collections.

We use the TREC 6, 7 and 8 ad hoc collections as evaluated with Average Precision. As is common practice, we first drop the bottom 25% of results to avoid effects of possibly buggy systems. From each original collection, we simulate $1,000$ new collections of sizes $n = 10, 20, \ldots, 100$ topics, leading to a total of $30,000$ simulated collections. For each of them, we estimate τ and τ_{AP} using each of the estimators defined above, and also compute the true correlations (recall that this is possible because the true system scores are fixed upfront when simulating new collections). Finally, for each correlation coefficient, original collection, topic set size and estimator, we compute expected error and bias.

Two baselines are used to compare our estimators to. They are based on a split-half method that randomly splits the available topic set in two subsets, and then computes the correlations as if one was the truth and the other one the estimate. This is replicated a number of times for different subset sizes, up to a maximum of $n/2$ topics. The observations are then used to fit a model and extrapolate the expected correlation with n topics. This simple estimator is found for instance in [7, 4]. Here we run $2,000$ replicates to fit the model $y = a \cdot e^{b \cdot x}$, and sample topics with and without replacement, leading to baselines SH(w) and SH(w/o).

4.3 Results

Figure 1 shows that the error of the estimators is larger with small collections. This is somewhat expected, because collections with too few topics are unstable and the rankings of systems vary too much to begin with. The error seems to plateau at about 0.025 in all our estimators, though with small collections of just 10 topics they are expected to be off by about 0.065. With the usual 50 topics, the expected error is 0.035. We can finally observe that the typical SH

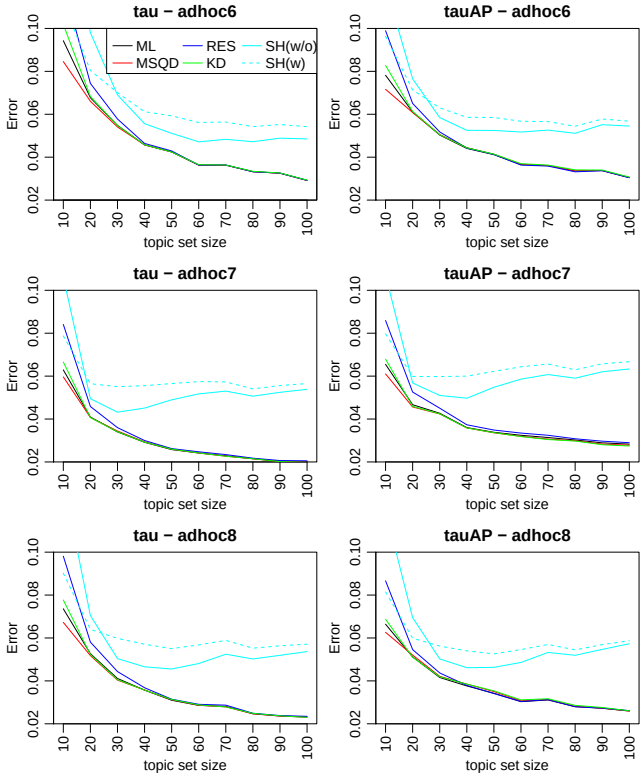

Figure 1: Error of the estimators of τ (left) and τ_{AP} (right) for each of the three original collections.

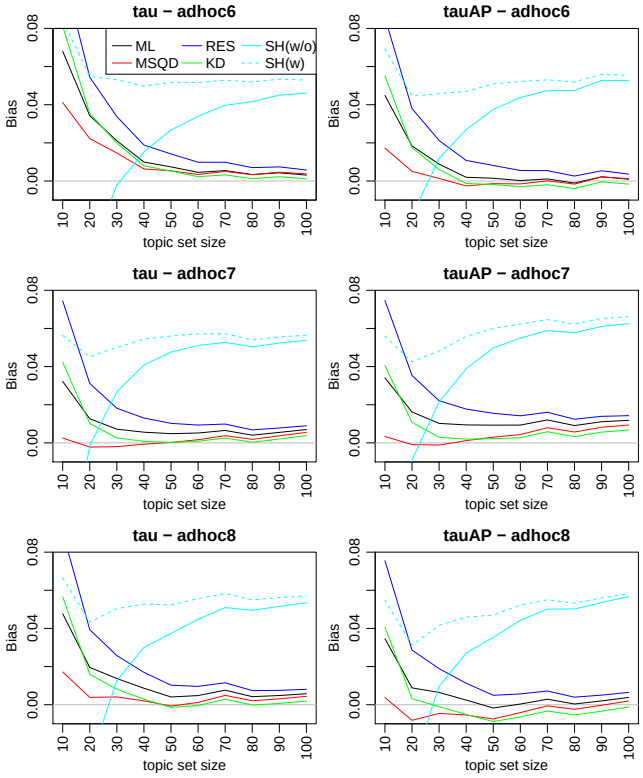

Figure 2: Bias of the estimators of τ (left) and τ_{AP} (right) for each of the three original collections.

estimators are clearly outperformed by all our proposed estimators. In general, with 30–40 topics they behave almost the same, but with small samples MSQD is slightly better.

Figure 2 shows that the correlations tend to be overestimated, especially with small collections, but this time we see clear differences among estimators. MSQD behaves much better than the others, especially with very small collections. With only 10 topics ML outperforms KD because there is just too little data to properly approximate the *pdf*, but with 20 or more topics it does a very good job at approximating the true distribution. ML, on the other hand, assumes a normal distribution and can therefore be less faithful to the data. Even at around 40–50 topics KD gets to slightly outperform MSQD for the same reason. Overall, they seem to plateau at about 0.004, and RES always performs worse than the others. Finally, the SH estimator with replacement has a roughly constant bias of about 0.055. The SH estimator without replacement shows a clearly biased behavior probably due to the choice of model.

5. CONCLUSION

In this paper we present two estimators of the Kendall τ and AP rank correlation coefficients between the mean system scores produced by a test collection and the true, unobservable means. We proposed parametric and non-parametric alternatives, and through large scale simulation with realistic collections we showed that even with small topic sets the estimators have little bias and the errors are generally small with collections of medium size. These estimators may prove useful as an easy to understand indicator of reliability in the results of an evaluation experiment.

In light of the expected error with individual collections, our future work will mainly focus on the development of interval estimates. We also plan to study other estimators of discordance as well as the application of a fully bayesian approach to estimate correlations. All the results in this paper are fully reproducible with data and code available online at `http://github.com/julian-urbano/sigir2016-correlation`.

Acknowledgments. Work supported by the Spanish Government: JdC postdoctoral fellowship, and projects TIN2015-70816-R and MDM-2015-0502. Florentino dimisión.

References

[1] B. Carterette, V. Pavlu, E. Kanoulas, J. A. Aslam, and J. Allan. If I Had a Million Queries. In *ECIR*, 2009.

[2] W. H. Holtzman. The Unbiased Estimate of the Population Variance and Standard Deviation. *Am. J. Psychology*, 1950.

[3] M. G. Kendall. A New Measure of Rank Correlation. *Biometrika*, 1938.

[4] T. Sakai. On the Reliability of Information Retrieval Metrics Based on Graded Relevance. *Inf. Proc. & Mngmnt*, 2007.

[5] J. Urbano. Test Collection Reliability: A Study of Bias and Robustness to Statistical Assumptions via Stochastic Simulation. *Information Retrieval*, 2016.

[6] E. M. Voorhees. Variations in Relevance Judgments and the Measurement of Retrieval Effectiveness. In *SIGIR*, 1998.

[7] E. M. Voorhees and C. Buckley. The Effect of Topic Set Size on Retrieval Experiment Error. In *SIGIR*, 2002.

[8] E. Yilmaz, J. Aslam, and S. Robertson. A New Rank Correlation Coefficient for Information Retrieval. In *SIGIR*, 2008.