A new rank correlation coefficient for information retrieval

Conference Paper · January 2008

DOI: 10.1145/1390334.1390435 · Source: DBLP

3 authors:

Emine Yilmaz
University College London
57 PUBLICATIONS **1,525** CITATIONS

SEE PROFILE

Javed A. Aslam
Northeastern University
138 PUBLICATIONS **4,498** CITATIONS

SEE PROFILE

Stephen E. Robertson
University College London
246 PUBLICATIONS **13,258** CITATIONS

SEE PROFILE

A New Rank Correlation Coefficient for Information Retrieval

Emine Yilmaz* eminey@microsoft.com

Javed A. Aslam†* jaa@ccs.neu.edu

Stephen Robertson* ser@microsoft.com

*Microsoft Research
7 JJ Thomson Avenue
Cambridge CB3 0FB, UK

†College of Computer and Information Science
Northeastern University
360 Huntington Ave, #202 WVH
Boston, MA 02115

ABSTRACT

In the field of information retrieval, one is often faced with the problem of computing the correlation between two ranked lists. The most commonly used statistic that quantifies this correlation is *Kendall's* τ. Often times, in the information retrieval community, discrepancies among those items having high rankings are more important than those among items having low rankings. The Kendall's τ statistic, however, does not make such distinctions and equally penalizes errors both at high and low rankings.

In this paper, we propose a new rank correlation coefficient, *AP correlation* (τ_{ap}), that is based on average precision and has a probabilistic interpretation. We show that the proposed statistic gives more weight to the errors at high rankings and has nice mathematical properties which make it easy to interpret. We further validate the applicability of the statistic using experimental data.

Categories and Subject Descriptors: H.3 Information Storage and Retrieval; H.3.4 Systems and Software: Performance Evaluation

General Terms: Experimentation, Measurement, Algorithms

Keywords: Evaluation, Kendall's tau, Average Precision, Rank Correlation

1. INTRODUCTION

Most of the research in the field of information retrieval depends on ranked lists of items. The output of search engines are ranked list of documents, the search engines themselves are also ranked based on their performance according to different evaluation criteria, the queries submitted to search engines are again ranked based on their difficulty, and so on.

*We gratefully acknowledge the support provided by NSF grant IIS-0534482.

SIGIR'08, July 20–24, 2008, Singapore.

Since most of the research in IR is based on ranked lists of items, it is often the case that we need to compare two ranked lists and report the correlation between them. Two of the most commonly used rank correlation statistics are Kendall's τ [7] and Spearman rank correlation coefficient [15].

The Spearman correlation coefficient is equivalent to the traditional linear correlation coefficient computed on ranks of items [15]. The Kendall's τ distance between two ranked lists is proportional to the number of pairwise adjacent swaps needed to convert one ranking into the other.

Kendall's τ has become a standard statistic to compare the correlation between two ranked lists. When various methods are proposed to rank items, Kendall's τ is often used to compare which method is better relative to a "gold standard". The higher the correlation between the output ranking of a method and the "gold standard", the better the method is concluded to be. Pairs of rankings whose Kendall's τ values are at or above 0.9 are often considered "effectively equivalent" [13], at least empirically.

For example, Soboroff et al. [12] propose a new method for system evaluation in the absence of relevance judgments and use Kendall's τ to measure the quality of their method. Buckley and Voorhees [2] propose a new evaluation measure, bpref, to evaluate retrieval systems and use Kendall's τ to show that this measure ranks systems similar to average precision. Similarly, Aslam et al. [1] propose a new method for evaluating retrieval systems with fewer relevance judgments. They compare their method with the depth pooling method by comparing the Kendall's τ correlation of the rankings of systems obtained using both methods with the actual rankings systems to show that their method is better than the depth pooling method. The Kendall's τ statistic is also used to compare the rankings of queries based on their estimated difficulty with the actual ranking of queries [14]. Melucci et al. [8] provides an analysis of places where Kendall's τ is used in information retrieval.

In most of the places where Kendall's τ is used, authors aim for a Kendall's τ value of 0.9 and conclude that their method produces "good" rankings if they obtain a τ value greater than this threshold [17, 3, 9].

Although Kendall's τ seems to be a reasonable choice for comparing two rankings, there is an important problem with this statistic, at least in the context of IR. Kendall's τ equally penalizes errors that occur at any part of the list. In other words, it does not distinguish between the errors that occur towards the top of the list from the errors towards

the bottom of the list [8]. However, in almost all cases in information retrieval we care more about the items that are ranked towards one end of the list (either top or bottom). For example, in TREC, the goal is often to correctly identify the best systems, i.e., those at the top of a list ranked by performance. When the goal is to predict query difficulty, it is typically more important to identify the most difficult queries. Similarly, when comparing the outputs of two search engines, the differences towards the top of the two rankings typically matter more than the differences towards the bottom of the rankings.

As an example of the aforementioned problem, consider 8 different retrieval systems. Let's assume that their actual ranking is <1 2 3 4 5 6 7 8>. Suppose that there are two alternate methods to evaluate these systems, and one would like to compare the relative quality of these two methods. Let's assume that when the first method is used, the systems are ranked <4 3 1 2 5 6 7 8> and when the second method is used, they are ranked <1 2 3 4 8 7 5 6>. The former ranking has the first four systems in inverse order compared with the actual ranking, while the latter the last four system in inverse order. The Kendall's τ correlation of each rankings with the actual ranking of the systems is the same (in both cases equal to 0.6429). Hence, based on the Kendall's τ values, the two methods are equivalent in terms of how they rank the systems. Note, however, that in many IR contexts it is much more important to get the "top half" of the list "right" than the "bottom half". Thus, we might well much prefer the latter ranking as compared to the former.

In the real world, we are often times faced with ranked lists in which there are many mistakes in terms of the rankings of the top (best) items. Figure 1 and Figure 2 show two such cases for TREC 8. In the figures, the leftmost plots show the mean average precision (MAP) values of systems using depth-1 and depth-6 pooling versus the actual MAP values. It can be seen in Figure 1 that the Kendall's τ correlation between the rankings of systems induced by depth-1 MAP values and the actual rankings of systems is 0.733. Hence, based on the Kendall's τ value, there is a positive correlation between the two rankings. However, if we only consider the top (best) 10 systems in TREC 8 (middle plot), it can be seen that these systems are ranked almost in reverse order among themselves. On the other hand, if we only focus on the worst 10 systems (right plot), it can be seen that these systems are ranked almost perfectly with respect to each other. Hence, even though the depth-1 pooling method is quite poor at identifying the best systems, since Kendall's τ does not make any distinction between errors towards the top versus the errors towards the bottom, the overall Kendall's τ value is still 0.733. The same behavior can be seen in Figure 2.

Many researchers are aware of this flaw in using Kendall's τ [8] and have tried to use some alternatives to Kendall's τ. Voorhees [14] makes use of a new measure for estimating query difficulty by gradually removing the items of interest and comparing how the average values of items change with respect to average values of actual items. However, this method cannot be used to measure correlation between ranked lists. To compare rankings of retrieval systems, Wu et al. [16] use an accuracy measure which is the total number of items that are common in the top-n of both lists, divided by n, where n is the number of items one is interested in. This approach has the problem that the best possible accuracy measure is obtained as long as the top-n of both lists contain the same items, even if the top items are ranked in reverse order in two lists.

Shieh [11] recently devised an extension of Kendall's τ where the errors in different ranks are penalized by different weights. However, this requires assigning arbitrary weights to these errors beforehand and defining such weights is not easy. Similarly, Fagin et al. [5] proposed an extension to Kendall's τ for comparing top-k lists. Their extension is also based on defining arbitrary penalties when there are errors in rankings. Furthermore, their approach still gives equal weights to errors within the top-k lists and is not very applicable for comparing the entire ranked lists while giving more weight to the errors at the top. Haveliwala et al. [6] used the Kruskal-Goodman τ statistic (a statistic very similar to Kendall's τ) to compute correlations in the regions they are interested in (e.g. the top). However, this approach also suffers from the same problems as the former method in that it is not very applicable for comparing all the items in ranked lists at once while giving more weight to the errors at the top.

In this paper, we first show that the problem of evaluating the correlation between two ranked lists is analogous to the problem of evaluating the quality of a search engine, concluding that similar ideas can be used in both cases. We propose a new rank correlation coefficient, *AP correlation* (τ_{ap}), that is based on average precision and has a probabilistic interpretation similar to Kendall's τ, while giving more weight to the errors nearer the top of the list, as in AP. The proposed statistic has two nice properties: (1) When ranking errors are randomly distributed across the list, the AP correlation value is equal to Kendall's τ, in expectation. (2) If there are less (more) errors towards the top of the list, the AP correlation value is higher (lower) than the Kendall's τ value, as desired. These two properties make the AP correlation coefficient easy to interpret (by comparing with Kendall's τ) and use. We further demonstrate the applicability of the statistic through experimental data, and we demonstrate with a real example how AP correlation can point out errors that might otherwise be incorrectly ignored when using Kendall's τ.

2. KENDALL'S TAU

The Kendall's τ measure is one of the most commonly used measures employed to compute the amount of correlation between two rankings. Given two lists of length N, let C be the total number of concordant pairs (pairs that are ranked in the same order in both rankings) and D the total number of discordant pairs (pairs that are ranked in opposite order in the two rankings). Then, the Kendall's τ value between the two lists is defined as

$$\tau = \frac{C - D}{N(N-1)/2}$$

Note that given a ranked list with N objects, there are $\binom{N}{2} = N(N-1)/2$ pairs of items among them (the denominator in the formula). Hence, the Kendall's τ value has the following nice probabilistic interpretation: Consider the following experiment,

1. Pick a pair of items at random.
2. Return 1 if the pair is ranked in the same order in both lists; otherwise, return 0.

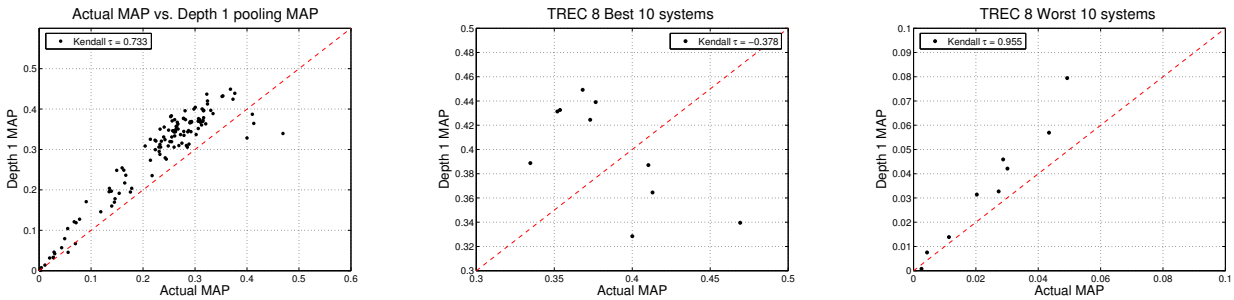

Figure 1: (Left) TREC 8 depth-1 pooling MAP vs. actual MAP. (Middle) TREC 8 depth-1 pooling MAP vs. actual MAP for best 10 systems. (Right) TREC 8 depth-1 pooling MAP vs. actual MAP for worst 10 systems.

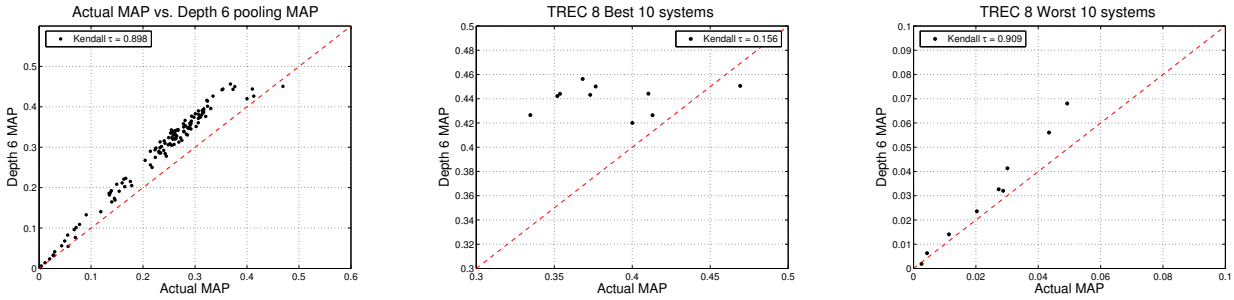

Figure 2: (Left) TREC 8 depth-6 pooling MAP vs. actual MAP. (Middle) TREC 8 depth-6 pooling MAP vs. actual MAP for best 10 systems. (Right) TREC 8 depth-6 pooling MAP vs. actual MAP for worst 10 systems

Let p be the expected outcome of this random experiment, i.e., the probability that a randomly chosen pair of items is ranked in the same order in both lists. Note that $p = C/[N(N-1)/2]$, and thus Kendall's τ is

$$\tau = p - (1-p) = 2p - 1.$$

As such, Kendall's τ is a monotonic function of the probability that a randomly chosen pair is ordered identically in both rankings.

Note that if the two rankings are identical ($p = 1$), then the Kendall's τ value is 1, while if the two rankings perfectly *disagree* ($p = 0$), then the Kendall's τ value is -1, and if the two rankings are statistically independent ($p = 1/2$), then the Kendall's τ value is 0.

In information retrieval, many different evaluation measures have been developed to assess the quality of the output of a search engine, typically a ranked list of documents. Since the documents retrieved towards the top of the list are more important than the others, evaluation measures that give more importance to the documents retrieved towards the top of the ranking have been proposed. Note that the problem of evaluating the correlation between two ranked lists (when the top items are more important than the others) is a similar problem to the problem of evaluating the quality of a search engine. As such, one can use similar ideas to the ones used in evaluating the quality of a search engine in order to compare rankings.

Average precision is perhaps the most commonly used evaluation measure. It has a nice probabilistic interpretation that assigns a meaning to the measure. In the following section, we show that average precision can be defined based on preferences. Based on this preference based version of average precision and the probabilistic interpretation of the measure, we then propose a new rank correlation coefficient (*AP correlation*) that also has a nice probabilistic interpretation and that distinguishes between the errors made towards the top of a ranked list.

3. AVERAGE PRECISION

Average precision can be defined as the average of the precisions at relevant documents, where the precisions at unretrieved documents are assumed to be zero. It is also known to be an approximation of the area under the precision-recall curve.

Since documents that are retrieved at the top of a list are more important than the documents towards the bottom, average precision assigns more weight to the errors made towards the top of a ranking than the errors towards the bottom.

An intriguing property of average precision is that these weights are assigned such that the value of the measure has a nice probabilistic interpretation.

Recently, Yilmaz and Aslam [17] proposed a new probabilistic interpretation of average precision and used this interpretation to estimate the value of average precision when the relevance judgments are incomplete. The average precision according to this interpretation can be computed as the expected outcome of a random experiment,

1. Pick a *relevant* document at random. Let r be the rank of this document.

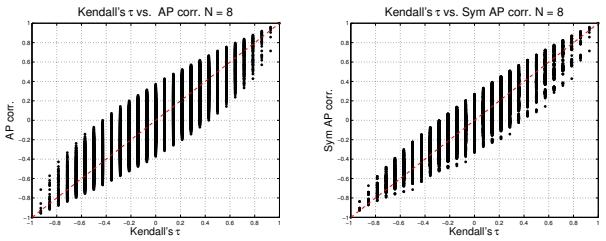

Figure 3: (left) AP correlation coefficient vs. Kendall's τ for all possible permutations of a list of length 8 (right) Symmetrized AP correlation coefficient vs. Kendall's τ for all possible permutations of a list of length 8.

2. Pick a second document at or above rank r, at random.

3. Return the binary relevance of this second document.

Average precision is the expected outcome of this random experiment: the expected outcome of Steps 2 and 3 is the precision at rank r (containing a relevant document), and in expectation, Step 1 corresponds to the average of these precisions.

Note that the aforementioned probabilistic interpretation of average precision holds in the case of binary relevance judgments (i.e., a document can be either relevant or non-relevant). It can be altered to hold in the case of preferences as well (i.e., one document is more relevant than the other and so on) as follows:

1. Pick a relevant document at random. Let r be the rank of this document.

2. Pick a second document at or above rank r, at random.

3. Return 1 if the second document is *at least* as relevant as the first; return 0 otherwise. In other words, return 1 if the documents are in the correct *relative order*; otherwise, return 0.

Average precision is the expected outcome of this random experiment. Note that the above definition of average precision gives us a way to compute average precision of a ranked list of documents given the actual ranked list of documents.

As mentioned before, in information retrieval, we often face the problem of comparing two ranked lists of items where we care more about the top items than the items towards the bottom and Kendall's τ does not distinguish between such cases. Average precision itself is such a top-heavy measure. Hence, one could use a rank correlation statistic based on average precision. In the following section, we propose a new rank correlation statistic based on the probabilistic interpretation of average precision, and we show how this new statistic compares to the Kendall's τ statistic.

4. AP CORRELATION COEFFICIENT

One can alter the probabilistic interpretation of average precision [17] slightly to obtain a new rank correlation statistic that gives more importance to the items towards the top of the list. The main idea behind the AP correlation coefficient (τ_{ap}) is that since the rankings at the top part of the list are more important, given an item, one mainly cares if this item is ranked correctly with respect to the items above this current item. Hence, the correlation is based on computing the probability that each item is ranked correctly with respect to the items above this current item and averaging over all items.

Let *list1* and *list2* be two lists of items of length N and suppose *list2* is the actual ranking of items and *list1* is a ranking of items whose correlation with the actual ranking (*list2*) we would like to compute. Consider the following random experiment:

1. Pick any item from *list1*, other than the top ordered item, at random.

2. Pick another item from this list that is ranked above the current item, at random.

3. Return 1 if this pair of documents are in the same relative order as in *list2*; otherwise, return 0.

In mathematical terms, the expected outcome of this random experiment can be written as follows:

$$p' = \frac{1}{N-1} \cdot \sum_{i=2}^{N} \frac{C(i)}{(i-1)}$$

where $C(i)$ is the number of items above rank i and correctly ranked with respect to the item at rank i in *list1*.

Note that this random experiment is very similar to the random experiment upon which Kendall's τ is based; the only difference is that instead of comparing an item with any other randomly chosen item, it is only compared with items above.

It is easy to see that the value of p' falls between 0 and 1, where 1 means that all items ranked by *list1* are ranked in the same order as the items ranked by *list2* and 0 means that all items ranked above another item are ranked incorrectly according to *list2* (complete negative correlation). Note that if there are items that are not shared by both lists then p' is computed based only on the common items.

We define the AP correlation coefficient as a function of the expected outcome of the above random experiment, in much same way that Kendall's τ is defined as a function of the outcome of an analogous random experiment, so that its value will fall between -1 and $+1$, the range of values commonly used by correlation coefficients. The AP correlation τ_{ap} is defined as

$$\tau_{ap} = p' - (1-p') = 2p' - 1 = \frac{2}{N-1} \cdot \sum_{i=2}^{N} \left(\frac{C(i)}{i-1} \right) - 1.$$

4.1 AP Correlation Coefficient vs. Kendall's tau

Given the two rank correlation statistics, Kendall's τ and AP correlation coefficient, one might wonder how the value of the two statistics compare with each other for different types of lists. In this section, we show how these two statistics compare with each other when the distribution of error changes in lists.

THEOREM 1. *When the errors are uniformly distributed over the list, Kendall's τ and AP rank correlation coefficient are equivalent.*

PROOF. Suppose you are given an ordered list of items, *list1* and a second list with the actual ranking of items.

Suppose that the probability of obtaining a correct pair in *list1* is p and all errors are uniformly distributed over the list.

First consider the traditional Kendall's τ value. Based on the given setup, the probability that any randomly picked pair is concordant or discordant is p or $1-p$ respectively. Hence, when the errors are uniformly distributed over the list with probability $1-p$, the value of Kendall's τ is $2p-1$.

Now consider the AP rank correlation coefficient value for this setup. Since the errors are uniformly distributed over the entire list with probability $1-p$, at rank i, the expected number of concordant items above rank i with respect to the item at rank i is $p(i-1)$. Hence, the value of AP rank correlation coefficient can be computed as:

$$\begin{aligned}\tau_{ap} &= \frac{2}{N-1}\cdot\sum_{i=2}^{N}\frac{C(i)}{i-1}-1\\ &= \frac{2}{N-1}\sum_{i=2}^{N}\frac{p(i-1)}{i-1}-1\\ &= 2p-1\\ &= \tau\end{aligned}$$

□

Therefore, when the error is uniformly distributed with probability $1-p$ over the entire list, the values of Kendall's τ and AP rank correlation coefficient are the same and are both equal to $2p-1$. As mentioned before, when the two rankings of items are completely independent of each other, the Kendall's τ value is 0. This corresponds to the case where $p=0.5$ and empirically verifies the theorem above, since in this case both the Kendall's τ and the AP rank correlation coefficient are 0.

THEOREM 2. *If the probability of error is increasing with rank (more errors towards the bottom of the list when compared to the top of the list), then the Kendall's τ is always less than the AP rank correlation coefficient. Similarly, if the probability of error is decreasing with the rank, then the Kendall's τ is always less than the AP rank correlation coefficient.*

PROOF. Suppose you are given an ordered list of items, *list1* and a second list with the actual ranking of items. Suppose that the probability of obtaining a correct pair in *list1* is varying with the rank and for an item at rank i, the probability that the items ranked above item i are ranked correctly with respect to item i is p_i.

First, consider the value of the traditional Kendall's τ in this setup. The expected number of items that are ranked above item i and are ranked correctly with respect to item i is $p_i * (i-1)$. For each rank i, we can compute the total number of concordant items above this rank with respect to the item at rank i and sum these values to obtain the total number of concordant items in the list. Therefore the expected Kendall's τ value is:

$$E[\tau] = \frac{2\cdot\sum_{i=2}^{N}p_i(i-1)}{N(N-1)/2}-1$$

Now consider the AP rank correlation coefficient value in the same setup. Using the same idea for Kendall's τ, the AP rank correlation coefficient value can be written as:

$$\begin{aligned}E[\tau_{ap}] &= \frac{2}{N-1}\cdot\sum_{i=2}^{N}\frac{p_i(i-1)}{i-1}-1\\ &= \frac{2}{N-1}\cdot\sum_{i=2}^{N}p_i-1\end{aligned}$$

To demonstrate how the Kendall's τ and AP rank correlation coefficient change with respect to each other with the probabilities p_i, the difference between the AP rank correlation coefficient and the Kendall's τ is computed:

$$\begin{aligned}E[\tau_{ap}-\tau] &= \frac{2}{N-1}\cdot\sum_{i=2}^{N}p_i-\frac{2\cdot\sum_{i=2}^{N}p_i(i-1)}{N(N-1)/2}\\ &= \frac{2}{N-1}\left(\sum_{i=2}^{N}p_i-\frac{2\sum_{i=2}^{N}p_i(i-1)}{N}\right)\\ &= \frac{2}{N(N-1)}\left(\sum_{i=2}^{N}(N-2i+2)\,p_i\right)\end{aligned}$$

The constant factor $2/(N(N-1))$ can be ignored. We can expand the remaining formula and rewrite it in a simpler form as, $((N-2)p_2+(N-4)p_3+(N-6)p_4+\cdots+2p_{N/2}+0p_{N/2+1}-2p_{N/2+2}\cdots-(N-4)p_{N-1}-(N-2)p_N$.

This summation can be written as

$$E[\tau_{ap}-\tau]\propto\sum_{i=2}^{N/2}(N-2i+2)(p_i-p_{N-i+2})$$

Note that this summation is always greater than zero when the probabilities are decreasing and is always smaller than zero when the probabilities are increasing. Hence, when the probability of error in the list increases (or decreases) as we go lower in the ranking, the AP rank correlation coefficient is always larger (or smaller) than Kendall's τ. □

Note that the above proof shows that if the probability of error is decreasing by rank (more errors at the top), then $E[\tau-\tau_{ap}]<0$ (or vice versa). However, this does not imply that if $E[\tau-\tau_{ap}]<0$ then it is always the case that the probability of error is decreasing by rank. In fact, one can identify some particular distributions of error where this statement does not hold. However, in the next sections, we show using real data that it is often the case that the lists for which $\tau_{ap}\approx\tau$, the errors are quite random over the entire list and for the ones $\tau_{ap}>\tau$, there are fewer mistakes towards the top of the list.

The left plot in Figure 3 shows the value of AP correlation coefficient vs. Kendall's τ for all possible permutations of a list of length 8. It can be seen that for a given value of Kendall's τ, there are many lists that have a different τ_{ap} and vice versa. In what follows, we show that for lists for which $\tau_{ap}\approx\tau$, the errors are quite random within the list and for those for which $\tau_{ap}>\tau$, there are fewer mistakes towards the top of the list as proved. Furthermore, we show that the proposed statistic can produce more desirable assessments of correlations among rankings, in that it produces higher values when the top ranked items are correctly identified and lower values when they are not.

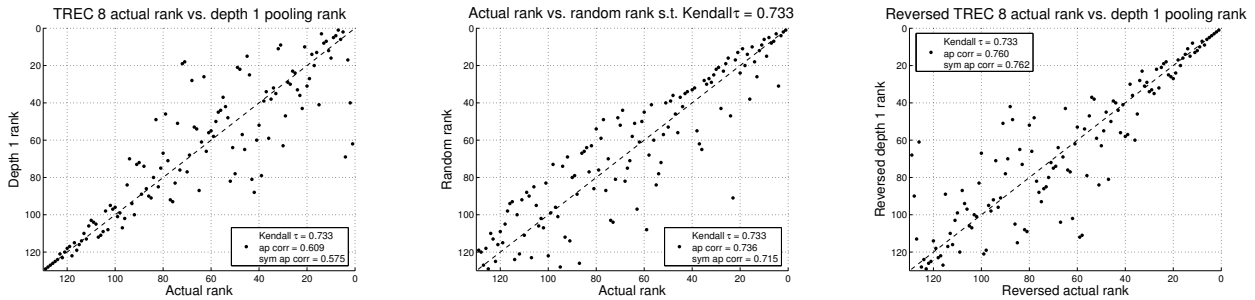

Figure 4: (Left) TREC8 depth1 pooling ranking of systems vs. actual rankings of systems. (Middle) A list of size equal to number of systems submitted to TREC8 with the same Kendall tau value as depth-1 pooling. (Right) TREC8 ranking of systems vs. actual rankings of systems when the rankings are reversed (best systems become worst and worst systems become best)

4.2 Symmetric AP Correlation Coefficient

Note that the AP correlation coefficient is not a symmetric statistic. It assumes that there is an actual ranked list (*list2*) of items and an estimated ranked list (*list1*) of items and one would like to compute the correlation among these two lists ($\tau_{ap}(list1|list2)$: AP correlation of list1 given list2). However, if we were to compute the AP correlation of list2 given list1, we obtain a different value for the statistic. In information retrieval, it is often the case that we would like to compare the correlation between an estimated ranked list given an actual ranked list (e.g. actual rankings of systems). Hence, in such situations, AP correlation can be used.

However, in some cases one would simply like to compute the correlation among two ranked lists where we do not have the notion of the "actual" rankings. In such cases, a symmetrized version of the statistic (*symmetrized AP correlation coefficient*) could be used. The symmetrized AP correlation coefficient can be computed based on the same idea used in defining the symmetrized version of the KL distance measure [4]:

$$symm\tau_{ap}(list1, list2) = (\tau_{ap}(list1|list2)+\tau_{ap}(list2|list1))/2$$

Hence, the symmetrized AP correlation coefficient is the average of the AP correlation coefficients when each list is treated as the actual list.

Right plot in Figure 3 shows the value of AP correlation coefficient vs. Kendall's τ for all possible permutations of a list of length 8. It can be seen that the shape of the curve is similar to the shape for AP correlation coefficient (Figure 3).

5. EXPERIMENTAL RESULTS

We have shown theoretically that when the errors are randomly distributed the entire list, $\tau_{ap} = \tau$, when there are more errors towards the top of the list, $\tau_{ap} < \tau$, and when there are fewer errors towards the top of the list, $\tau_{ap} > \tau$. What this means is that this new measure can correctly order correlations among items (lists with fewer errors towards the top have higher AP correlations and so on.)

Figure 4 demonstrates this behavior in practice. The figure shows three different rankings with identical Kendall's τ values. The leftmost plot illustrates the ranking of systems according to depth1 pooling versus the actual ranks of systems in TREC 8 (the top right corner of the plots refer to the top ranked systems). The rightmost plot shows the rank of the systems in TREC 8 based on depth-1 pooling when the rankings are reversed (worst systems become the best and best systems become the worst). Note that the rightmost plot corresponds to rotating the leftmost plot by 180°. The middle plot corresponds to a random ranked list that has random errors along the entire list such that the Kendall's τ of the list is equal to the Kendall's τ of depth-1 pooling on that TREC. It can be seen that when the errors are randomly distributed over the list (middle plot), the AP correlation coefficient is approximately equal to Kendall's τ as expected in expectation. In the leftmost plot, where there are more errors towards the top of the list than the bottom, it can be seen that the AP correlation is less than Kendall's τ. Similarly, when the top systems are ranked mostly correctly (rightmost plot), the value of the AP correlation coefficient increases and is higher then the Kendall's τ value. Hence, even though all three plots are equivalent in terms of their Kendall's τ value, the proposed correlation coefficient correctly identifies the distinction among all these three cases, as desired.

The fact that Kendall's τ gives equal weight to all errors in the rankings also has practical implications for IR research. For example, as mentioned before, in information retrieval, pairs of rankings whose Kendall's τ values are at or above 0.9 are often considered effectively equivalent [13]. As also mentioned by Sanderson et al. [10], this threshold has implicitly or explicitly been used in many places to reach conclusions regarding the quality of rankings. Work by Carterette et al. [3], Sanderson et al. [9], and Yilmaz et al. [17] are just a few examples of the many such research papers.

In Sanderson et al. [9], for example, the authors claim that since manual runs identify most of the relevant documents, instead of judging the entire depth pool, one could only judge the outputs of manual runs and use these judgments to evaluate the quality of retrieval systems. In order to test their claim, they take a manual run out, form a qrel by using only judgments for documents that are retrieved by this run and are also contained in the actual TREC qrels. Then, using these qrels, they evaluate the quality of retrieval systems by mean average precision and compare the induced rankings of systems with the actual rankings using Kendall's τ. They report that for most of the manual runs, the τ correlations are at or above 0.9 and therefore conclude that qrels formed from the output of a single manual run can be effectively used to evaluate the quality of retrieval systems.

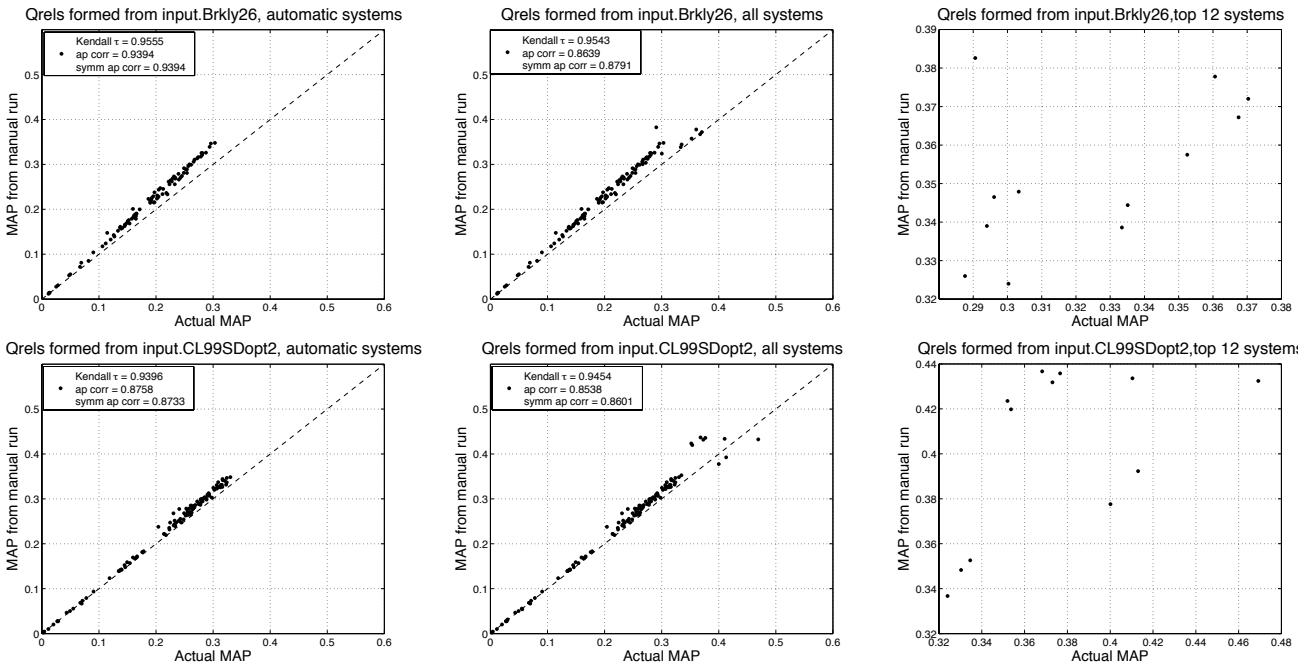

Figure 5: Rankings of systems induced by qrels formed from manual runs (top) Brkly26 in TREC 7 and (bottom) CL99SDopt2 in TREC 8, compared to actual rankings.

Figure 5 shows an example of these experiments. The upper and lower plots show how qrels formed from manual runs Brkly26 from TREC 7 and CL99SDopt2 from TREC 8 evaluate retrieval systems in these TRECs, respectively. The x axis in these plots is the actual MAP value for the systems and the y axis shows the MAP value of the systems as using these qrels.

The leftmost plots show the rankings of systems for only the automatic runs submitted to these TRECs whereas the middle plot contains rankings for all systems (manual and automatic). The rightmost plots focus only on the top 12 systems (sorted by the actual mean average precision value) in these TRECs and show how the manual qrels rank these top systems as compared to their actual rankings.

First consider the plots for TREC 7. By looking just at the Kendall's τ values, one would conclude that qrels formed from this manual run are very effective at evaluating systems (both when only automatic runs are included — τ value of 0.9555 — and when all runs are included — τ value of 0.9543). Although this may be a reasonable conclusion when only automatic runs are considered, the conclusion is much less valid when all runs are considered, since the topmost systems are not ranked correctly (middle plot). As shown by the rightmost plot, when all runs are included the best system according to these qrels is actually the 11th best system, the 2nd best system is actually the 3rd best system, and so on. Furthermore, although the correlations in the middle plot looks worse than the correlations in the left plot, the τ values are very similar. However, the AP correlation (or symmetrized AP correlation) correctly takes these changes into account and correctly distinguishes these two cases, correctly pointing out that the correlation in the middle plot is worse than in the left plot.

Similar behavior can be seen for qrels formed from CL99SDOpt2 in TREC 8. In this case, even when only automatic runs are considered (left plot), the rankings of the top-most systems are not perfect. Therefore, the AP correlation is not so high even for this case where Kendall's τ is 0.9396. When all runs are included (middle plot), the Kendall's τ value is 0.9454 although the correlations among the top systems become even worse (top system according to the actual qrels are identified as the 4th best system using these qrels, and the best system according to these qrels is actually the 7th best system if actual qrels were used as seen by the the rightmost plot). Surprisingly, the Kendall's τ value in this case is higher than the Kendall's τ value when only automatic runs are included. Once again, the AP correlation correctly distinguishes these cases.

Both of these examples show that even though the Kendall's τ statistic shows very good correlations, the qrels formed from a single manual system cannot be reliably used to evaluate the quality of all other systems (including other manual systems); the AP correlation statistic points out this issue more reliably than Kendall's τ. This problem with the Kendall's τ statistic was also discussed by Sanderson and Soboroff in a similar context [10].

To further demonstrate the applicability of the AP correlation statistic against Kendall's τ, we generated 100 random lists of length 50 that have a Kendall's τ value of 0.9 (using dynamic programming), and we scatter plot all such lists versus the actual ranking (1 2 ... 50), in a manner similar to the middle plot of Figure 4 superimposing all 100 scatter plots on top of each other. In this way, we can see the trend of error, i.e., which part of the ranking is the error likely to occur at. The left plot in Figure 6 shows the result of this experiment. It can be seen that for a list with Kendall's τ value of 0.9, the errors are likely to occur at any part of the list, hence the top items may not be identified correctly. Similarly, to see the trend of error in the lists that have an AP correlation value of 0.9, we generated 100 random lists

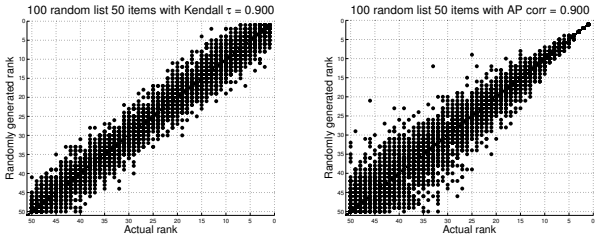

Figure 6: (Left) 100 randomly generated lists of length 50 that have a Kendall's τ value of 0.9. (Right) 100 randomly generated lists of length 50 that have a AP correlation value of 0.9.

of length 50 that have an AP correlation value of 0.9. The right plot in Figure 6 shows the result of this experiment. It can be seen that the lists that have an AP correlation of 0.9 are very unlikely to have errors towards the top of the list, and most of the errors tend to be towards the bottom, reinforcing our conclusion that AP correlation coefficient can be effectively used in the cases where we are mostly interested in identifying these top ranked items.

6. CONCLUSIONS

We propose a new rank correlation coefficient that has a nice probabilistic interpretation and more heavily weights the errors towards the top of the ranking. The statistic has the nice property that (1) if errors are uniformly distributed over the entire ranking, the statistic is equal to Kendall's τ in expectation and (2) if there is more error towards the top/bottom of the list, the value of the statistic is lower/higher than Kendall's τ. We theoretically and experimentally show through TREC data the applicability of this measure and conclude that in the context of information retrieval, where the goal is often to identify the best systems while still having a good overall ranking, this statistic can be better used to compute the correlation between two rankings.

The statistic as defined in this paper is not symmetric, hence it is not a metric. Although we show how this statistic can be extended to be symmetric, throughout the paper we mainly focus on the behavior of the nonsymmetric version as it is simpler to analyze. The overall behavior of the symmetric version of the statistic is the subject of future work. Furthermore, the Kendall's τ statistic has an associated probability distribution, enabling one to do hypothesis testing. In the future, we are planning to investigate the distribution of the proposed statistics to enable making better inferences about the values of these statistics.

7. REFERENCES

[1] J. A. Aslam, V. Pavlu, and R. Savell. A unified model for metasearch, pooling, and system evaluation. In O. Frieder, J. Hammer, S. Quershi, and L. Seligman, editors, *Proceedings of the Twelfth International Conference on Information and Knowledge Management*, pages 484–491. ACM Press, November 2003.

[2] C. Buckley and E. M. Voorhees. Retrieval evaluation with incomplete information. In *Proceedings of the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 25–32, New York, NY, USA, 2004. ACM Press.

[3] B. Carterette and J. Allan. Incremental test collections. In *CIKM '05: Proceedings of the 14th ACM international conference on Information and knowledge management*, pages 680–687, New York, NY, USA, 2005. ACM Press.

[4] T. Cover and J. Thomas. *Elements of Information Theory*. Wiley, 1991.

[5] R. Fagin, R. Kumar, and D. Sivakumar. Comparing top k lists. In *SODA '03: Proceedings of the fourteenth annual ACM-SIAM symposium on Discrete algorithms*, pages 28–36, Philadelphia, PA, USA, 2003. Society for Industrial and Applied Mathematics.

[6] T. H. Haveliwala, A. Gionis, D. Klein, and P. Indyk. Evaluating strategies for similarity search on the web. In *WWW '02: Proceedings of the 11th international conference on World Wide Web*, pages 432–442, New York, NY, USA, 2002. ACM.

[7] M. Kendall. A new measure of rank correlation. *Biometrica*, 30(1–2):81–89, 1938.

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

[9] M. Sanderson and H. Joho. Forming test collections with no system pooling. In *SIGIR '04: Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 33–40, New York, NY, USA, 2004. ACM.

[10] M. Sanderson and I. Soboroff. Problems with kendall's tau. In *SIGIR '07: Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 839–840, New York, NY, USA, 2007. ACM.

[11] G. S. Shieh. A weighted kendall's tau statistic. *Statistics & Probability Letters*, 39:17–24, 1998.

[12] I. Soboroff, C. Nicholas, and P. Cahan. Ranking retrieval systems without relevance judgments. In *Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 66–73, New Orleans, Louisiana, USA, Sept. 2001. ACM Press, New York.

[13] E. M. Voorhees. Evaluation by highly relevant documents. In *Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 74–82. ACM Press, 2001.

[14] E. M. Voorhees. Overview of the TREC 2004 robust retrieval track. In *Proceedings of the Thirteenth Text REtrieval Conference (TREC 2004)*, 2004.

[15] D. D. Wackerly, W. Mendenhall, and R. L. Scheaffer. *Mathematical Statistics with Applications*. Duxbury Advanced Series, 2002.

[16] S. Wu and F. Crestani. Methods for ranking information retrieval systems without relevance judgments. In *SAC '03: Proceedings of the 2003 ACM symposium on Applied computing*, pages 811–816, 2003.

[17] E. Yilmaz and J. A. Aslam. Estimating average precision with incomplete and imperfect judgments. In *Proceedings of the Fifteenth ACM International Conference on Information and Knowledge Management*. ACM Press, November 2006.