Problems with Kendall's Tau

Mark Sanderson
Department of Information Studies
University of Sheffield
Western Bank, Sheffield, S10 2TN, UK
+44 114 22 22 648
m.sanderson@shef.ac.uk

Ian Soboroff
National Institute of Standards and Technology
Gaithersburg, MD 20899, USA
ian.soboroff@nist.gov

ABSTRACT
This poster describes a potential problem with a relatively well used measure in Information Retrieval research: Kendall's Tau rank correlation coefficient. The coefficient is best known for its use in determining the similarity of test collections when ranking sets of retrieval runs. Threshold values for the coefficient have been defined and used in a number of published studies in information retrieval. However, this poster presents results showing that basing decisions on such thresholds is not as reliable as has been assumed.

Categories and Subject Descriptors
H.3.3 [**Information Search and Retrieval**]: Systems and Software --- performance evaluation.

General Terms
Measurement, Experimentation

Keywords
Kendall's Tau.

1. INTRODUCTION
A great deal of information retrieval research is concerned with ranking of objects, either documents ranked by a retrieval system, sets of system *runs* ranked using a test collection (Voorhees, 1998), even the difficultly of topics ranked by predictive measures (Yom-Tov, et. al. 2005). A wide range of measures have been created to assess the quality of document ranking (e.g. P@10, MAP, etc), for topics ranked for difficulty or runs ranked against each other, the de-facto standard is *Kendall's Tau rank correlation coefficient* (τ).

$$\tau = \frac{P - Q}{\sqrt{(P + Q + T) * (P + Q + U)}} \quad (1)$$

P, No. correctly-ordered pairs T, No. ties in 1st ranking
Q, No. incorrectly ordered pairs U, No. ties in 2nd ranking

For any two rankings, Kendall's τ is a linear function of the number of pairs of items which are in different orders in the two rankings (Kendall, 1938). This function is constrained such that τ=1 if the two rankings are in the same order, and τ=-1 if they are inverted. There are a number of τ variations, we use the one defined in (1), which handles ties in a ranking.

The role for τ in test collections is to compare how similarly two test collections rank a set of runs. Voorhees was the first to use the coefficient in this way comparing two versions of TREC ad hoc collections each using relevance judgments from different sets of assessors (1998). Voorhees considered $\tau \geq 0.9$ to indicate that two test collections were equivalent. In a later paper (2001) she stated.

...evaluation schemes that produce correlations of at least .9 should be considered equivalent since it is not possible to be more precise than this. Correlations less than .8 generally reflect noticeable changes in the rankings, not simply inversions among neighbors, and suggest that the evaluation schemes have different emphases.

These thresholds have been explicitly re-used in a number of works including Sanderson & Joho (2004), Carterette & Allan (2005), Yilmaz & Aslam (2006) and subsequent papers from Voorhees (e.g., Buckley & Voorhees, 2004). Other works (e.g. Lee et al, 2002) also treat coefficients over 0.9 as important though without explicit reference to Voorhees's work. However, as pointed out by Bland and Altman (1986, p.308)

...correlation depends on the range of the true quantity in the sample. If this is wide, the correlation will be greater than if it is narrow.

In other words, when comparing the way that test collections rank runs, if the range of scores assigned to each of the runs (being ranked) is wide, τ will tend to have a higher coefficient than if the range of scores is narrow. Such qualities of correlation coefficients have long been known about, what is less well known is how much variation will occur when using τ to compare test collections. If only small variations in τ are found, there may be no problem worth considering. This poster presents the design and results of an experiment that tests how much correlation varies given sets of runs with different score ranges. The implications of the results are discussed along with avenues for future work.

2. EXPERIMENTAL DESIGN
In order to test the variation of τ on run sets of different ranges, the run data for the adhoc track of TRECs 6-8 and the web track of TREC 9 were downloaded from the TREC web site. For each of the years of TREC, automatic runs[1] were ranked against each other using full TREC relevance judgments and using judgments formed from the top 100 relevant documents retrieved for each topic in the 25% best performing manual runs (similar to the experiment described in Sanderson & Joho, 2004). Runs were ranked using Mean Average Precision (MAP). The τ between the

SIGIR'07, July 23–27, 2007, Amsterdam, The Netherlands.
ACM 978-1-59593-597-7/07/0007.

[1] Automatic runs were those runs where a topic was processed fully automatically by a retrieval system.

two ranks of runs in each year of TREC was measured. The following table shows this value in each of the four TRECs as well as the number of automatic and manual runs used.

TREC	Manual runs	Automatic runs	τ on full run set
6	4	57	**0.96**
7	4	86	**0.97**
8	3	116	**0.96**
9	3	92	**0.95**

The table shows that a test collection using qrels created just from the output of a few manual runs ranks automatic runs similarly to the full TREC relevance judgments. To measure the τ of runs over a smaller score range, the automatic runs in each year of TREC were sorted by their MAP and split in half: top 50% runs and bottom 50% runs. The table below shows τ for each of these reduced score range sets. As can be seen, τ is either the same or less than τ measured on the full set and for the top 50% of runs on TREC-9, τ is below the 0.9 threshold.

TREC	full run set	Top 50%	Bot. 50%	Av.
6	**0.96**	0.92	0.96	**0.94**
7	**0.97**	0.95	0.93	**0.94**
8	**0.96**	0.91	0.95	**0.93**
9	**0.95**	0.89	0.94	**0.92**

Further sub-dividing the runs into quarters further restricts the score range: results are shown in the table below. With the exception of two values in the sixteen shown, τ is lower again. In addition, the average τ of the 25% sized runs is always lower than the τ for the full runs and all four averages are under the threshold of 0.9. If one were to judge qrels based on runs with such a reduced score range instead of one conveying a fuller range, one might conclude that using a few manual runs to form relevance judgments is potentially problematic; the opposite conclusion drawn from the results shown in the first Table.

TREC	100%	Top 25%	Top-mid 25%	Bot.-mid 25%	Bot. 25%	Av.
6	**0.96**	0.80	0.89	0.89	0.98	**0.89**
7	**0.97**	0.93	0.87	0.83	0.93	**0.89**
8	**0.96**	0.82	0.84	0.83	0.97	**0.87**
9	**0.95**	0.82	0.71	0.86	0.89	**0.82**

It is notable that the τ for top performing automatic runs (with the exception of data from TREC-7) is lower than the τ from the bottom 25% of automatic runs. It is tempting to think that this result is showing that top performing automatic runs are harder to rank than bottom performing runs. However, one cannot say this with certainty as the range of score values in the top 25% run set is different from the bottom 25%: at this early stage of our research we have not been able to estimate the degree of influence of different score ranges on τ, only that its presence can substantially affect τ.

Note, on their own, these experiments fail to show definitively that score ranges are influencing τ, as when the run sets are halved or quartered, two variables in the experiment are changed: the score range *and* cardinality of the sets. We eliminated the possibility that set size is the cause of the change in τ with a second experiment (not shown due to lack of space). In it, reduced run sets were formed by random selection of runs from a full set. Repeating such experiments multiple times, it was found that on average τ measured on these smaller run sets (which on average had the same score range of the full sets) was the same as the τ measured on the full set. From this we concluded that τ was not affected by set size, but by the range of scores across the runs composing a set.

3. DISCUSSION

From the results, we conclude that using thresholds for correlation coefficients when comparing test collections is potentially problematic. The variation of τ due to changes in the range of scores in run sets can be so large that coefficients measured on sets with narrow score ranges can be substantially different from coefficients measured on sets with wider scores. It would appear that absolute thresholds used with τ should be applied with great care, or preferably avoided.

Determining what would be a better approach for comparing test collections with each other is work left for future consideration. In addition we believe that it will be worthwhile exploring in more detail the use of τ in past information retrieval research and examining if the observed variations in τ (shown here) have unknowingly influenced published results.

4. ACKNOWLEDGMENTS

Thanks to Andrew Holmes who pointed us to an important reference that led us to conduct this investigation. This work was supported in part by the Tripod project: IST-FP6-045335.

5. REFERENCES

Bland, J.M., Altman, D.G. (1986) Statistical methods for assessing agreement between two methods of clinical measurement. *Lancet*, i, 307-310

Buckley, C. & Voorhees, E.M. (2005) Retrieval evaluation with incomplete information, *Proc. of ACM SIGIR*, 25-32

Carterette, B. & Allan, J. (2005) Incremental test collections, *Proc. of ACM SIGIR*, 680-687

Kendall, M. (1938) A New Measure of Rank Correlation, *Biometrika*, 30, 81-89.

Lee, S., Myaeng, S.H., Kim, H., Seo, J.H., Lee, B., Cho, S. (2002) Characteristics of the Korean Test Collection for CLIR in NTCIR-3, *Working Notes of NTCIR*

Sanderson, M. & Joho, H. (2004) Forming test collections with no system pooling, *Proc. of ACM SIGIR*, 33-40

Voorhees, E.M. (1998) Variations in relevance judgments and the measurement of retrieval effectiveness, *Proc. of ACM SIGIR*, 315-323

Voorhees, E.M. (2001) Evaluation by highly relevant documents, *Proc. of ACM SIGIR*, 74-82

Yilmaz, E. & Aslam, J.A. (2006) Estimating Average Precision with Incomplete and Imperfect Judgments *Proc. of ACM CIKM*, 102-111

Yom-Tov, E, Fine, S., Carmel, D., Darlow, A. (2005) Learning to estimate query difficulty: including applications to missing content detection and distributed information retrieval, *Proc. of ACM SIGIR*, 512-519