How Precise Does Document Scoring Need To Be?

submitted for anonymous review

Abstract. We explore the implications of tied scores arising in the document similarity scoring regimes that are used when queries are processed in a retrieval engine. Our investigation has two parts: first, we evaluate past TREC runs to determine the prevalence and impact of tied scores, to understand the alternative treatments that might be used to handle them; and second, we explore the implications of what might be thought of as "deliberate" tied scores, in order to allow for faster search. We show that **Alistair says: we end up knowing something, we hope**.

Keywords: Information retrieval, effectiveness, document scoring, ties

1 Introduction

Batch evaluation techniques are widely used in information retrieval system measurement. Each system that is to be compared generates a ranking, or *run*, for each of a set of topics, with documents included in the run and also ordered within the run on the basis of some computed textual *similarity score* relative to the given query. Example similarity computations include the Okapi BM25 mechanism of Robertson et al. [6] and the language modeling techniques of Ponte and Croft [5]. Those runs are then mapped to numeric *effectiveness values* using a set of relevance judgments and an *effectiveness metric*, which generates a single number as an assessment of the quality, or utility, of that run in the eyes of the user that is presumed to have inspected it. Those effectiveness values are then aggregated in some way across topics to get an overall performance measure; or are used as the basis of a statistical test to (hopefully) determine a question of the form "is System A demonstrably better than System B".

In this paper we consider the consequences of allowing *tied similarity scores* (or just *ties*) in the ranking. One obvious issue is that ties admit a level of ambiguity in the effectiveness values, and hence (potentially) in the outcome of a system versus system comparison, since a group of documents that all share the same computed similarity score could be presented to the user in any permutation that is consistent with the scores being non-increasing. Our first goal is thus to quantify the extent to which past Text Retrieval Conference (TREC) evaluation exercises have been affected by ties, and determine whether the presence of ties may have caused ambiguity to flow through to system scores. In this part of the project we make use of a range of tie-breaking regimes, including the rules embedded in the well-known `trec_eval` program, and conclude that while ties had the potential to be quite significantly disruptive, in practice they did not influence the outcomes of the measurements that were undertaken.

A second and related goal is to then ask whether the deliberate introduction of ties might be useful in some way. For example, a range of ways in which similarity scoring

might be approximated or otherwise quantized have been suggested over the years, including the impact-ordered indexes of Anh and Moffat [1]. If we allow that the retrieval system might gain tangible efficiency benefits from being permitted to assign low precision scores to documents, or even to deliberately choose the precision to which particular documents are scored, then we may end up with large numbers of ties in the runs that the system generates. In this scenario, being able to estimate the extent to which ties can be permitted before there is risk in incorrect effectiveness scores being generated is a necessary precondition. In experiments using submitted TREC runs, we show that **Alistair says: something**.

[[*Macros to be used for leaving messages through the paper:*]] Alicia says: done with `alicia`. **Andrew says: done with** `andrew`**. Alistair says: done with** `alistair`**.**

2 Ties, and Methods for Dealing With Them

Run Order [[*could use the ordering generated by the system, whatever it is, and just say, we presume thatthey knew what they were doing*]]

External Tie-Break Rule [[`trec_eval`*, and similar*]]

Averaging Across Permutations McSherry and Najork [4], and describe basis for their formulations

Limits [[*optimistic and pessimistic, to get bounds*]]

3 Ties in past TREC Runs

TREC Resources Harman [3]

Ties in TREC7

Ties in Other Years

4 Deliberate Score Grouping

We now consider whether the deliberate use of tied scores – which might allow efficiency improvements in the underlying search system – has a discernible effect on retrieval effectiveness.

Score Approximation Scoring documents using modern similarity computations involves non-trivial amounts of arithmetic, especally if phrase components or term proximity components are being used. Regimes such as WAND [2] seek to minimize the number of documents scored, while still giving rise to exactly the same ranking for the top-k documents. That is, every document in the first k positions of the ranking must be in exactly the "right" position relative to the documents ahead of it and behind it in the ranking that is generated. This is a relatively stringent requirement, and

other computation-pruning techniques can also be considered that provide alternative trade-offs.

Now consider the following weaker requirement: that each document must be scored in a manner that guarantees that it is in the correct *band* of the ranking, where the bands are defined geometrically based on a parameter $\rho > 1$. The bands are defined by the sequence $b_i = 1$, and thereafter by $b_{i+1} = \lceil \rho \cdot b_i \rceil$. The i th band spans the ranks from b_i to $b_{i+1} - 1$ inclusive. For example, if $\rho = 2$, then the bands are $[1 \ldots 1]$, $[2 \ldots 3]$, $4 \ldots 7]$, and so on; and if (say) $\rho = 1.62$ (the golden ratio) the bands are $[1 \ldots 1]$, $[2 \ldots 3]$, $[4 \ldots 6]$, $[7 \ldots 11]$, and so on, with widths given by the Fibonacci sequence. The smaller the value of ρ, the smaller each of the groups is, and the closer the approximate ranking is to the "true" and exact ranking. As ρ approaches 1, the retrieval system is obliged to place each of the documents closer and closer to its final "correct" position. That is, for a given value of ρ, we allow the retrieval system to return bands of documents $[b_i \ldots b_{i+1} - 1]$, with equal scores assumed within each band.

Effectiveness Score Changes Given this framework, it is appropriate to ask: to what extent does an allowance for rank-based score imprecision affect retrieval effectiveness? To respond to this question, we make use of TREC resources, taking the same system runs as were already examined in Section 3, and in effect mapping all of the documents ranked in band i to a synthetic score of $1/i$, as if the retrieval system had deliberately not carried out the extra computation that would be needed to fully differentiate between the approximately ρ^i documents in that band.

Our first experiment explores the score variance introduced as a result of score "banding", calculating for a set of runs from the [[*YYYY*]] TREC round the variation in run score that results when a given degree of score imprecision is introduced. In all cases the score difference calculated is the across-permutations computation that was illustrated in Section 2 when applied to the deliberately-tied rankings, compared to the score the same metric achieves on the original ranking. We followed standard protocols and assumed that unjudged documents were not relevant for the purposes of scoring the runs.

Figure 1 shows the results of this **Alistair says: Build the graph, and then discuss what it shows. Hopefully it shows something interesting or at least a little bit unexpected.**

System Comparison Sensitivity Effectiveness measurements are also used to compare systems in a pairwise manner. In the second experiment, we explore the implications that score rounding has on the ability of metrics to differentiate between systems. The normal approach to comparing systems is to take their computed scores across a set of topics, and perform a paired t-test to explore the null hypothesis that the two systems are in fact the same. The process of carrying out the t-test generates a p value; the smaller the p value, the smaller the chance that the two systems being comapred are giving the same performance. To establish significance, a threshold value α is employed, often $\alpha = 0.05$, with $p \leq \alpha$ being regarded as a significant outcome.

To measure the effect that score rounding has on system comparsions, we took the 50 topics of the TREC7 [[*citation*]] collection and the 103 runs associated with it, and computed: (a) a set of p values, generated by comparing each pair of systems using the

metric scores associated with the original set of runs; and then (b) the corresponding set of p values, generated when the same runs are first mapped into group scores, and the all-permutations averaging technique applied.

Results

5 Conclusion

References

[1] V. N. Anh and A. Moffat. Pruned query evaluation using pre-computed impacts. In *Proc. SIGIR*, pages 372–379, 2006.

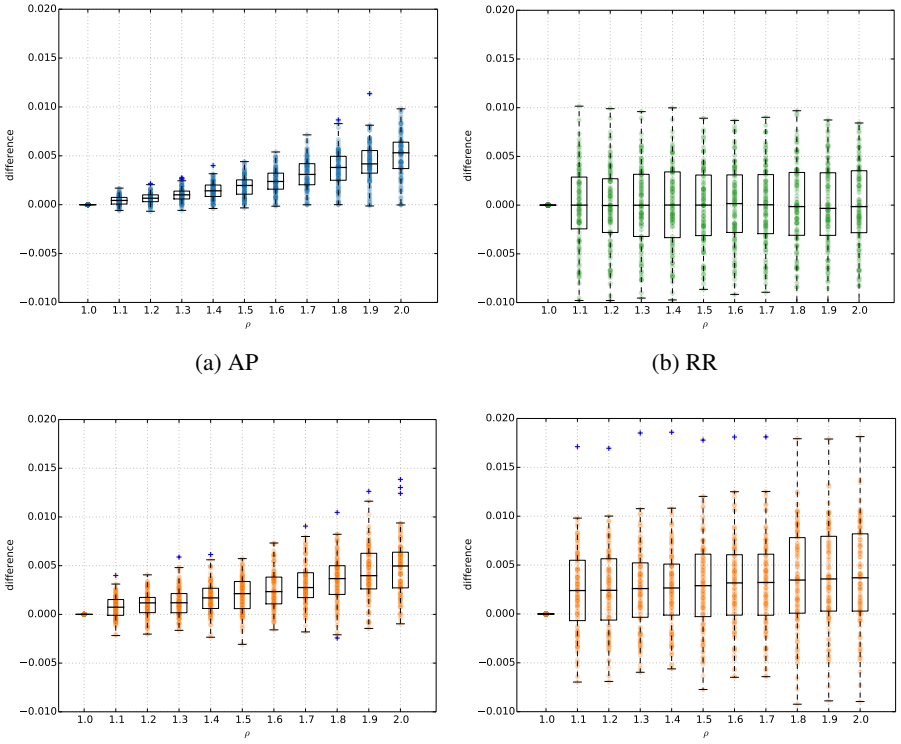
[2] A. Z. Broder, D. Carmel, M. Herscovici, A. Soffer, and J. Zien. Efficient query evaluation using a two-level retrieval process. In *Proc. CIKM*, pages 426–434, 2003.

[3] D. K. Harman. The TREC test collections. In E. M. Voorhees and D. K. Harman, editors, *TREC: Experiment and Evaluation in Information Retrieval*, chapter 2, pages 21–52. MIT Press, 2005.

[4] F. McSherry and M. Najork. Computing information retrieval performance measures efficiently in the presence of tied scores. In *Proc. ECIR*, pages 414–421, 2008.

[5] J. M. Ponte and W. B. Croft. A language modeling approach to information retrieval. In *Proc. SIGIR*, pages 275–281, 1998.

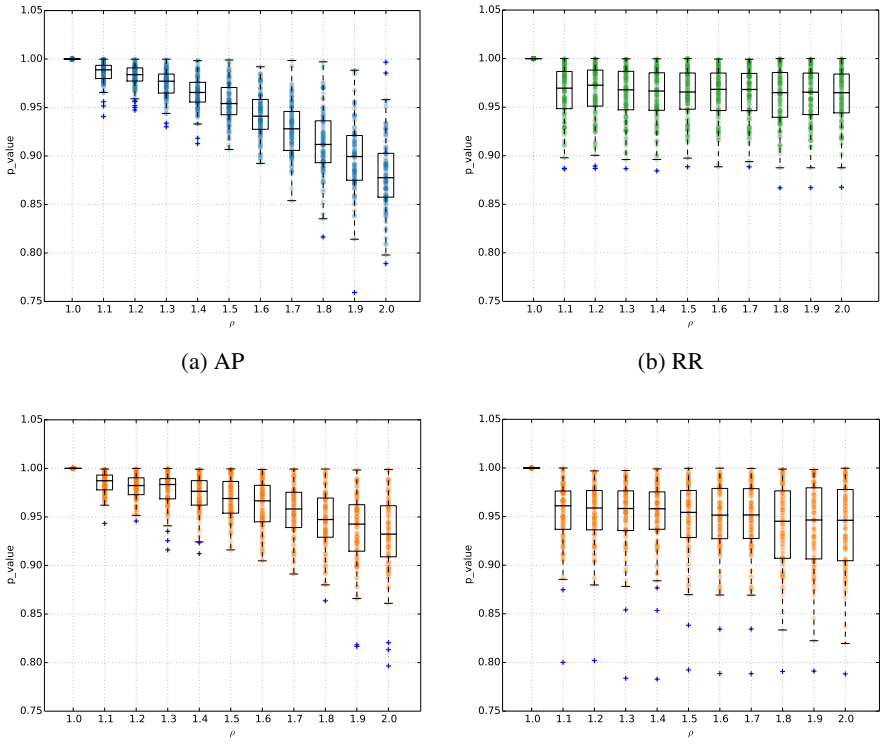
[6] S. E. Robertson, S. Walker, S. Jones, M. Hancock-Beaulieu, and M. Gatford. Okapi at TREC-3. In *Proc. TREC*, pages 109–126, 1994.



(a) AP (b) RR

(c) RBP0.85 (d) RBP0.6

Fig. 1: Variation in metric effectiveness score across a set of 103 runs and 50 topics, 103 points plotted in total in each column), as a function of ρ, for three different retrieval effectiveness metrics. **Alistair says: Only one so far, what is ? In end want: RR, RBP0.85, AP? RR should be relatively unaffected, the others will have broader variance.** Alicia says: Have updated the graphs, not sure if we need RBP(p=0.6). Caption need to be updated...



(a) AP (b) RR

(c) RBP0.85 (d) RBP0.6

Fig. 2: Student t-test values when comparing system scores for original-order runs, and grouped-score runs. A p value of 1.0 indicates that the two runs were identical.

Fig. 3: Variation in p values across system pairs in a set of runs, plotted as a function of ρ, for three different retrieval effectiveness metrics. **Alistair says: RR, RBP0.85, AP? Alistair says:** $\rho = 1.0, 1.1, 1.2, 1.3, \ldots, 2.0$**, or something like that. Alistair says: Box and whisker plot, as** ρ **is smaller, the mean score difference should get closer to zero, and the variance should also be getting smaller. RR should be relatively unaffected, the others will have broader variance. Would be cool in the mean stayed near zero even when** ρ **is relatively large.** Alicia says: I am sure what should be placed here... have not they all been placed in Figure 2?