Score Aggregation Techniques in Retrieval Experimentation

Sri Devi Ravana **Alistair Moffat**

Department of Computer Science and Software Engineering
The University of Melbourne
Victoria 3010, Australia
{sravana, alistair}@csse.unimelb.edu.au

Abstract

Comparative evaluations of information retrieval systems are based on a number of key premises, including that representative topic sets can be created, that suitable relevance judgements can be generated, and that systems can be sensibly compared based on their aggregate performance over the selected topic set. This paper considers the role of the third of these assumptions – that the performance of a system on a set of topics can be represented by a single overall performance score such as the average, or some other central statistic. In particular, we experiment with score aggregation techniques including the arithmetic mean, the geometric mean, the harmonic mean, and the median. Using past TREC runs we show that an adjusted geometric mean provides more consistent system rankings than the arithmetic mean when a significant fraction of the individual topic scores are close to zero, and that score standardization (Webber et al., SIGIR 2008) achieves the same outcome in a more consistent manner.

Keywords: Retrieval system evaluation, average precision, geometric mean, MAP, GMAP.

1 Introduction

Measurement is an essential precursor to all attempts to improve information retrieval (IR) system effectiveness. But to experimentally measure the effectiveness of an IR system is a non-trivial exercise, and requires that a complex sequence of tasks and computations be carried out. These tasks typically involve:

1. Selecting a representative corpus and a set of topics;
2. Creating appropriate relevance judgements that describe which documents are relevant to which topics;
3. For each of the systems being compared, building a set of runs, one for each of the topics;
4. Selecting one or more effectiveness metrics, and applying them to create a set of per-metric per-topic per-system scores;
5. For each effectiveness metric, comparing the scores obtained by one system against the scores obtained by the other systems, to determine if the difference in behavior between the systems can be assessed as being statistically significant; and then, finally,
6. Writing a paper that describes what the new idea was, and summarizing the measured improvement (or not) that was obtained relative to other previous techniques.

A variety of mechanisms have evolved over the years to perform these tasks. For example, because of the high cost of undertaking relevance judgements, steps 1 and 2 have tended to be carried out via large whole-of-community exercises such as TREC and CLEF. To perform step 3, we might build our own software system, or, to again make use of shared resources, we might choose to modify a public system such as Lemur, Terrier, or Zettair[1]. In the area of effectiveness metrics (step 4), the IR community has converged on a few that are routinely reported and can be evaluated via public software such as `trec_eval`. These include precision@10 (denoted here as P@10), R-precision, average precision (denoted as AP), normalized discounted cumulative gain (nDCG), rank-biased precision (RBP), and so on. Common tools and agreed techniques for step 5 are also emerging – there is now a clear community expectation that researchers must indicate whether any claimed improvements are statistically significant using an appropriate test (Cormack & Lynam 2007).

Now consider the last stage, denoted step 6 above. We wish to describe the new system in a context that allows the reader to appreciate the aspects of it that will lead to superior performance; and then need to provide empirical evidence that the hypothesized level of performance is attained. And, inevitably, we seek to do all of that within an eight or ten page limit. In support of our new system we describe the (established) corpus and topics that we have used in any training that we did to set parameters for our system and a baseline or reference system; and we describe the (different) corpus and/or topics that we then used to test the two systems with those parameters embedded. We can also succinctly summarize any statistically significant relationships between the two systems (the step 5 output): the sentence "System *New* was significantly better than System *Old* at the 0.05 level for all of X, Y, and Z" (where X, Y, and Z are effectiveness metrics) takes up hardly any space at all in our paper. Yet such a claim is the holy grail of IR research – provided, of course that System *Old* is a state-of-the-art reference point; that we had implemented it correctly; and that the experiments do indeed lead to the desired level of significance.

We would then like to provide evidence of the magnitude of the improvement we have obtained. To do that, we add a table of numbers to our paper. And here is the question that is at the heart of this work: what numbers? Two systems (at least) have been compared, over (probably) 50 or more topics, using (say) three effectiveness metrics. The experiments thus give rise to at least 300 per-system, per-topic, per-metric scores, and while we have all seen IR papers with that many numbers in them, including that

[1] See `http://www.lemurproject.org/`, `http://ir.dcs.gla.ac.uk/terrier/`, and `http://www.seg.rmit.edu.au/zettair/` respectively.

volume of data is rather less than ideal. Even worse, the effectiveness scores are all derived from the fifty underlying runs, each containing (typically) $1{,}000$ document identifiers. So if we wish to provide sufficient data that follow-up researchers can apply new effectiveness measures to the data, we need to publish $2 \times 50 \times 1{,}000 = 10^5$ "things" as the output of even the simplest two-system evaluation. (For example, TREC has accumulated the actual system runs lodged by the participating groups in over a decade of experimentation, and these now form an invaluable resource in their own right, and have been used in the experiments leading to this paper.)

Given that it is impossible for the data that was used as the input to our statistical testing to be included in our paper, but desirous of including at least some element of numeric output, we inevitably *average* the per-topic per-system per-metric effectiveness scores, to obtain a greatly reduced volume of per-system per-metric scores. We then populate a table in our paper with these numbers (perhaps as few as six of them in a one-collection, two-system, three-metric evaluation); use superscript daggers or bold font to indicate the relationships that are statistically significant, and submit our work for refereeing. This approach is now so prevalent that the phrase "mean average precision" has taken on a life of its own, and a table of "MAP" scores is an unavoidable part of every experimental IR paper, as if MAP was an axiomatic measure in its own right.

In this paper we take a pace back from this process, and ask two very simple questions that arise from the discussion above: what does it mean to "average" effectiveness scores? And, if it is indeed a plausible operation, is the arithmetic mean the most sensible way to do it, or are there other methods that should be considered?

2 Numeric aggregation

If a set of observations describes some phenomenon, it is natural to seek some kind of gross, or *aggregate* statistic that summarizes those observations. The simplest of these central tendencies is the *arithmetic mean*, which, for a set of t observations $\{x_i \mid i \in 1 \ldots t\}$ is computed as:

$$\mathsf{AM} = \left(\sum_{i=1}^{t} x_i \right) / t \,,$$

As examples, consider the four possible sets of $t = 5$ observations that are shown in Table 1. The arithmetic mean of example set S_1 is 0.28.

One point worth noting in connection with the arithmetic mean is that all of the values should be on the same scale – it is not possible to compute the average of five inches, ten centimeters, and 0.001 miles without first converting them to a common framework. Similarly, it is impossible to average three liters, five centimeters, and four kilograms, because they cannot be converted to common units.

Another key aggregation mechanism is the *geometric mean*, defined as the t th root of the product of the t numbers,

$$\mathsf{GM} = \left(\prod_{i=1}^{t} x_i \right)^{1/t} .$$

The geometric mean is more stable than the arithmetic mean, in the sense of being less affected by outlying values. However, when any of the values in a set is zero, the geometric mean over that set is also zero. For IR score aggregation purposes, this introduces the problem that a single "no answers" topic in the topic set might force all of the system scores to be zero. To sidestep this difficulty, Robertson (2006) thus defined an *ϵ-adjusted geometric mean*,

$$\epsilon\mathsf{GM} = \exp\left(\frac{\sum_{i=1}^{t} \log(x_i + \epsilon)}{t} \right) - \epsilon \,,$$

where ϵ is a small positive constant, and the summation, exp, and log functions calculate the product and t th root of the set of t ϵ-adjusted values x_i in a non-underflowing manner. Because it is based on multiplication, in which there is no requirement that the quantities have the same units, it is permissible (although somewhat confusing) to take the geometric mean of, for example, three liters, five centimeters, and four kilograms. In this example, the GM is 3.91 (liters $\cdot$ centimeters $\cdot$ kilograms)$^{1/3}$. Further examples of GM and ϵGM are shown in Table 1.

A variant of ϵGM-AP in which ϵ is applied to AP scores in a thresholding sense rather than in an additive/subtractive sense is now one of the aggregate scores routinely reported by the `trec_eval` program (see `http://trec.nist.gov/trec_eval`) using $\epsilon = 10^{-5}$ (Voorhees 2005):

$$\epsilon\mathsf{GM}_{\texttt{trec\_eval}} = \exp\left(\frac{\sum_{i=1}^{t} \log \max\{x_i, \epsilon\}}{t} \right) .$$

For example, the $\epsilon\mathsf{GM}_{\texttt{trec\_eval}}$ score for sequence S_2 in Table 1 is 0.039, and would be 0.157 with $\epsilon = 0.01$, the value used in the table. Note that when ϵ approaches ∞ the additive/subtractive ϵGM score for a set of numbers (but not the thresholded $\epsilon\mathsf{GM}_{\texttt{trec\_eval}}$ variant) approaches the AM score for the same set. For this continuity reason, we prefer the ϵGM-AP additive/subtractive version to the `trec_eval` version, and have primarily used the former in this paper.

The *harmonic mean* is another central tendency that is typically used to combine rates, and can also be used as a method for score aggregation. It is defined as the reciprocal of the average of the reciprocals,

$$\mathsf{HM} = \frac{t}{\sum_{i=1}^{t} (1/x_i)} .$$

The harmonic mean is undefined if any of the set values are zero, meaning that it is again convenient to make use of an ϵ-adjusted version:

$$\epsilon\mathsf{HM} = \frac{t}{\sum_{i=1}^{t} 1/(x_i + \epsilon)} - \epsilon \,.$$

When ϵ is large, the ϵHM score again converges to the AM score.

The fourth and final central tendency explored in this paper is the *median*, denoted here as MD. The median of a set is the middle value of the set when they are sorted into numeric order: $x_{(t+1)/2}$ when t is odd, and $(x_{t/2} + x_{t/2+1})/2$ when t is even. The median has the benefit of being relatively unaffected by outliers, but the flip side of this is that it is completely insensitive to changes of value that do not affect the set ordering except when it is the middle values that change.

Table 1 shows the application of these four aggregation methods, plus two ϵ-adjusted variants, to four example data sets, with the largest value in each column picked out in bold. It is apparent that the aggregation techniques have different properties, since the four sequences are placed into different "overall" orderings by the various aggregation techniques. There is, of course, no sense in which any of systems S_1 to S_4 in Table 1 is superior to the others (presuming that the five elements in each sequence can

System	Scores					AM	GM	ϵGM	HM	ϵHM	MD
S_1	0.1	0.1	0.3	0.8	0.1	**0.280**	0.189	0.192	0.145	0.148	0.100
S_2	0.0	0.4	0.2	0.4	0.3	0.260	0.000	0.151	–	0.034	**0.300**
S_3	0.1	0.5	0.3	0.2	0.2	0.260	**0.227**	**0.228**	0.197	0.200	0.200
S_4	0.2	0.2	0.3	0.2	0.2	0.220	0.217	0.217	**0.214**	**0.214**	0.200

Table 1: Example sets of values corresponding to different systems applied to $t = 5$ topics, and their calculated central tendencies. In the ϵGM and ϵHM methods, $\epsilon = 0.01$. The values in bold are the largest in each column.

be regarded as paired observations and then a statistical significance test applied), so no answer is possible to the question "which system is better in the sense of having the highest score?" Nevertheless, and despite that patent lack of demonstrable differentiation, as soon as an aggregate score has been computed for the observations generated by some system, it is immediately tempting to then "order" the systems by their aggregate scores – exactly as we have in Table 1 by presenting the sequences in decreasing AM order. In Table 1, system S_1 at face value "outperforms" the other three systems by a quite wide margin.

Robertson (2006) provides an insightful discussion of measures and how they apply to AP and other effectiveness scores, including the relationship between AM-AP and GM-AP. Our discussion here can be seen as extending Robertson's evaluation, through the use of experiments in which aggregation methods are used to represent the overall performance of retrieval systems. In order to understand the effects of using these aggregation methods and their ability to produce consistent system rankings, evaluations were conducted using various TREC collections, and different types of effectiveness measure. Our experiments indicate that ϵGM handles variability in topic difficulty more consistently than does the usual AM aggregation method, and also better than the median MD and harmonic mean HM methods, when a significant fraction of the individual topic scores are close to zero. Also of considerable interest is that the standardized average precision scores of Webber et al. (2008a) achieve the same outcomes, even when coupled with the standard AM aggregation.

3 Topic hardness

The effectiveness of a retrieval system is gauged as a function of its ability to find relevant documents (Sanderson & Zobel 2005). One of the aims of the recent TREC Robust Track is to improve the consistency of retrieval technology by focusing on poorly performing topics – ones for which most of the participating systems score poorly (Voorhees 2003). The GM-AP aggregation method was introduced as part of this effort, in order to de-emphasize the role of high-scoring topics in system comparisons, and to enhance the relative differences amongst low-scoring topics (Voorhees 2005, Robertson 2006). Note, however, that changing to GM-AP has no effect on the significance or otherwise of any pairwise system comparison, since significance is a function of the elemental effectiveness scores, prior to any summary value being computed. The aggregation mechanism relates purely to the gross statistic that is presented as being the overall score for the system.

Mizzaro (2008) makes further observations in this regard. The importance of good relative performance over all topics, and not just excellent performance on one (which is how system S_1 obtains its high AM score in Table 1), and the fact that users remember any delivery of poor results by a system for a topic was also discussed by Mandl et al. (2008). Similarly, Buckley (2004*a*) points out that the topic variability is the main problem when designing an IR system for all user needs, and that a universal IR system should perform well across a range of topics with varying levels of difficulty.

It has been noted that a key expectation (or rather, hope) arising from any form of IR experimentation is that the system performance results based on one topic or one collection should be able to predict system performance on other topics and other collections (Buckley 2004*a*, Webber et al. 2008b). It has also been noted (see Buckley (2004*b*) and Webber et al. (2008a)) that topic variability is at least as great as system variability, and that in the nominal matrix of per-system per-topic scores, there is more commonality of scores across any particular topic than there is across any individual system. Restating this observation another way, the score achieved by a particular system on a given topic tends to be more a function of the topic than of the system.

In a similar vein, Mizzaro & Robertson (2007) argue that GM-AP is a more balanced measure than AM-AP for TREC effectiveness evaluations. This is due to the way in which the arithmetic mean can be influenced by easy topics, for which the system-topic scores are generally high, and bad systems might still get scores that are numerically large. Mizzaro & Robertson also asserted that GM-AP is not overly biased towards the low end of the scale, where the system-topic scores are low, and, equivalently, the topics are hard. Indeed, as was noted by O'Brien & Keane (2007), users prefer strategies and technologies that maximize the amount of information they gain as a function of the interaction cost that they invest.

Observations such as these then raise the question as to how best to measure topic "hardness". In the TREC 2003 Robust Retrieval Track, difficult topics were defined as being those with a low median AP score and at least one high outlier score (Voorhees 2003). Other definitions include computing (Mizzaro 2008)

$$D_t = 1 - \mathit{mean}_t \,, \tag{1}$$

where mean_t is the average of the system-topic scores for topic t; or computing

$$D_t = 1 - \mathit{max}_t \,,$$

where max_t is the maximum score for topic t. For the purposes of the experiment, we took another approach, and defined the difficulty D_t of a topic t to be

$$D_t = \frac{\mathit{max}_t - \mathit{mean}_t}{\mathit{sd}_t} \,, \tag{2}$$

in which standardized z-scores are calculated in the system-topic matrix (Webber et al. 2008a), and the most difficult topic is deemed to be the one for which the most "surprisingly good" score is obtained by one of the systems, with surprise defined in terms of standard deviations above the mean.

4 Methodology

Our purpose in this investigation was to examine the effect that the choice of score aggregation technique had on the outcomes of experiments, and whether the proposed use of

the ϵGM adjusted geometric mean (Robertson 2006) could be experimentally supported in any way. To carry out this study, we devised the following experimental methodology. While we have no basis for proving that what we have computed has "real" meaning, we trust that the reader will find the experiment plausible (and interesting), and will agree that the results we have computed are grounded in practice.

We made use of standard TREC resources – collections, matching topics, and corresponding relevance judgements (see `http://trec.nist.gov`). We also used the official TREC runs, as lodged each year by the participating research groups, and were able to compute retrieval effectiveness scores for each of the submitted systems based on any subset of the topics that we wished to use. Al-Maskari et al. (2008) consider these test collections and support their use in IR experimentation, arguing that they can be used to predict users' effectiveness successfully.

Each of our main experiments proceeded by:

1. Choosing a random subset containing half of the set of t topics.
2. Extracting the rankings for those $t/2$ topics from the s available runs.
3. Using the chosen effectiveness metric and the relevance judgements to calculate a set of $st/2$ per-system per-topic scores.
4. Using the chosen score aggregation technique to compute a set of s per-system scores.
5. Sorting the s systems into decreasing score order, based on the per-system scores.
6. Repeating this process, using the other $t/2$ topics.
7. Taking the two s-item system orderings, and calculating the similarity between them using a mechanism such as Kendall's τ (Kendall & Gibbons 1990).
8. Then repeating this entire sequence $10{,}000$ times, so that $10{,}000$ Kendall's τ scores could be used to represent the self-consistency of the score aggregation technique.

We note that researchers have also applied this methodology in investigations examining other aspects of retrieval performance (Zobel 1998, Sanderson & Zobel 2005).

Part of the purpose of Table 1 was to illustrate the inconsistencies that can arise out of system "orderings" based on aggregate scores, and our experiments in this paper are intended to uncover the extent to which such inconsistencies are an issue in real IR experimentation. If an aggregation computation was "perfect", and if subsets of topics could be equally balanced (whatever that means), the two system orderings would be the same, and the Kendall's τ would be 1.0. Variation in aggregation technique, and variations in subset balance, mean that it is unlikely that Kendall's τ scores of 1.0 can be achieved. But, if the same (large number of) topic subsets are used for all aggregation methods, any consistently-observed difference in Kendall's τ can be attributed to the aggregation method.

5 Testing aggregation

For the initial experiments, the AP metric was coupled with a range of aggregation techniques. Use of AP in IR experimentation is widespread, and while it is normally regarded as being a "system" metric rather than a "user" one, it does still correspond to a (somewhat contrived) user model (Robertson 2008).

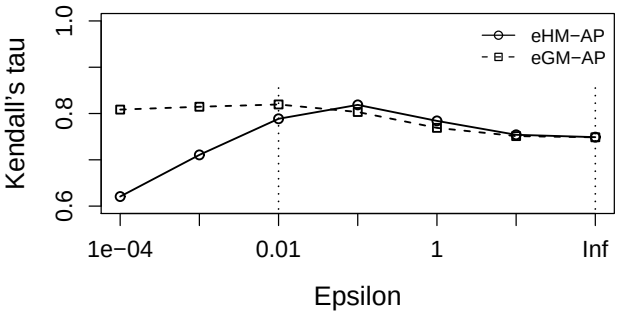

Figure 1: Average system ordering correlations when GM-AP and HM-AP are used as the score aggregation method across topics. In this experiment, the 50 TREC9 Web Track topics were randomly divided into equal-sized subset pairs, and the system rankings generated on those two subsets were compared using Kendall's τ. When the GM-AP and HM-AP parameter ϵ approaches infinity, the resultant system ordering approaches the ordering generated by AM-AP.

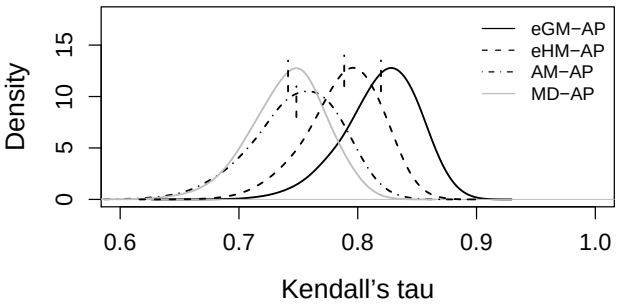

Figure 2: Density distribution of $10{,}000$ Kendall's τ values for ϵGM-AP, ϵHM-AP (both with $\epsilon = 0.01$), AM-AP, and MD-AP. In each experiment the TREC9 Web Track topics are randomly split into two halves, each system is scored using the topic subsets, and then the two resultant system orderings are compared. Three of the four curves represent density cross-sections corresponding to points in Figure 1. The MD-AP cross-section is additional. The vertical dotted lines on the curves indicate the means of the four density distributions.

5.1 Results using average precision

Figure 1 provides a first illustration of the data collected using the experimental methodology described in the previous section. In this graph ϵGM-AP and ϵHM-AP induced system orderings are compared for different values of ϵ, with the average value of Kendall's τ for random pairs of query subset-induced system orderings plotted as a function of ϵ. The line shows the average τ value over $10{,}000$ random splittings of the $t = 50$ TREC9 topics and $s = 105$ TREC9 systems, in an experiment designed to answer the very simple question as to whether either ϵGM-AP or ϵHM-AP should be preferred to AM-AP, and if so, what range of values of ϵ is appropriate.

The shape of the curve in Figure 1 makes it clear that when ϵ is small the average system ordering correlations are higher – that is, that the ϵGM-AP aggregate score (towards the left end of the graph) places the systems into rankings that are more self-consistent than does the conventional AM-AP summarization at the righthand end of the graph (when $\epsilon \rightarrow \infty$). The ϵHM-AP method is also better at ordering the systems than is AM-AP provided that mid-range values are chosen for ϵ. For small ϵ, the quality of the induced system rankings for ϵHM-AP drops markedly.

A major theme of this paper is that plain averages should be treated with caution, and we should heed our

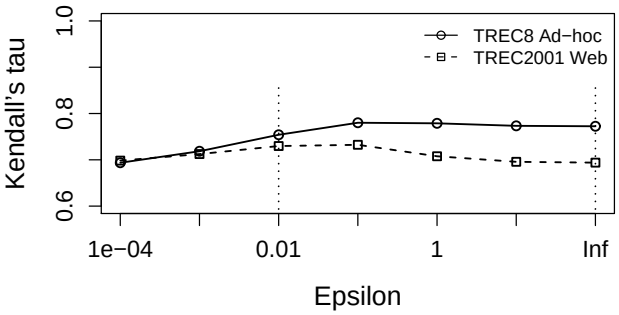

Figure 3: Average system ordering correlations when GM-AP and AM-AP (as $\epsilon \to \infty$) are used as the score aggregation method across topics. The two curves for the TREC8 Ad-Hoc Track ($t = 50$ topics, $s = 129$ systems) and the TREC2001 Web Track ($t = 50$ topics, $s = 97$ systems) can be directly compared with the ϵGM-AP curve in Figure 1.

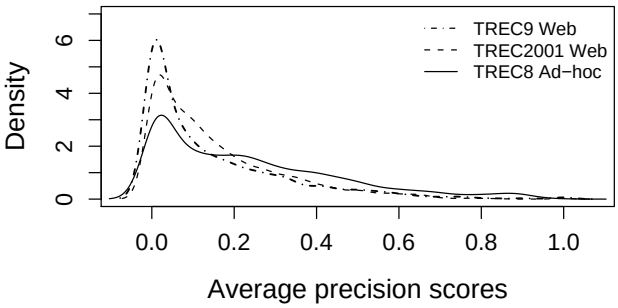

Figure 4: Density distribution of AP scores for TREC9 Web Track data ($t = 50$ topics and $s = 105$ systems); TREC8 Ad-Hoc Track data ($t = 50$ topics and $s = 129$ systems); and TREC2001 Web Track data ($t = 50$ topics and $s = 97$ systems).

own advice in this regard. Figure 2 shows the density distribution of the $10{,}000$ τ values at three of the points plotted in Figure 1, showing the variability of the system orderings when ϵGM-AP and ϵHM-AP (both with $\epsilon = 0.01$) and AM-AP are used to aggregate the per-system per-topic and generate the system orderings. Also shown as a fourth line is the τ density curve for the system similarities generated using the median AP score as the gross system statistic. The density curves are derived from $10{,}000$ random splittings of the 50 TREC9 topics into two 25-topic subsets. The system orderings produced by GM-AP aggregation are consistently more similar than the system orderings induced by AM-AP for the same topic splits, and suggest that GM-AP is the more stable aggregation technique for this data set. Analysis of the paired differences shows that all of the relationships shown are significant at the $p = 0.01$ level, that is, that ϵGM-AP $>$ ϵHM-AP $>$ AM-AP $>$ MD-AP with high confidence.

5.2 Other collections

Having used the TREC9 Web Track data to confirm that ϵGM-AP with $\epsilon = 0.01$ yields more consistent system rankings than does AM-AP, we turned to other TREC data sets in order to determine the extent to which that relationship is a general one. Figure 3 was created via the same experimental methodology as was used for Figure 1, but using the TREC8 Ad-Hoc Track queries and systems ($t = 50$, $s = 129$); and the TREC2001 Web Track queries and systems ($t = 50$, $s = 97$). Neither of these experiments favor ϵGM-AP compared to AM-AP, and for the TREC8 data set, ϵGM-AP with $\epsilon = 0.01$ is significantly less consistent that AM-AP ($p = 0.05$).

Figure 4 helps understand why this difference in behavior arises. It shows the distribution of the per-topic per-system AP scores for the three TREC data sets used in our experiments. The TREC9 collection, topics, and judgements combination generates a high number of low scores compared to the other two data sets, and it is these low scores that the ϵGM method is handling better. For example, in the TREC9 results, 10% of the system-topic scores are zero, and a further 46% are below 0.1, making a total of 56% low scores, shown in the first row of Table 2. For TREC2001 the corresponding rates were 4% and 45%, totalling 49%; and for TREC8 the rates were 3% and 33%, totalling 36%. That is, in the TREC2001 and TREC8 experiments there were fewer low AP values in the score matrix, AM-AP suffers less vulnerability, and there is thus less scope for ϵGM-AP to be superior.

Metric	$\% = 0$	$\% \le 0.1$
AP	10	56
P@10	37	37
nDCG	10	23
RBP0.95	17	52
SP	0	4

Table 2: Proportion of low scores among the TREC9 system-topic combinations when assessed using different effectiveness metrics.

5.3 Effectiveness measures

Figure 5 and Figure 6 show what happens in the TREC9 environment when AP is replaced by P@10, RBP0.95 (see Moffat & Zobel (2009) for a definition of this metric), and nDCG (see Järvelin & Kekäläinen (2002)) as the underlying similarity measure. The score density plot in Figure 6 suggests that ϵGM-nDCG should be stable as ϵ is changed, because of the low density of nDCG near-zero scores, and that is what is observed in Figure 5. On the other hand, RBP0.95 has a relatively high density of near-zero scores, and ϵGM-RBP0.95 is accordingly sensitive to ϵ, with more consistent system rankings being generated with $\epsilon = 0.1$ than when ϵ tends to ∞ and the arithmetic mean is being used. Note also the comparatively poor performance of P@10 – it is relatively immune to the choice of ϵGM or AM aggregation, but nor is it a terribly good basis for ordering systems.

Webber et al. (2008a) recently introduced a *standardized* version of AP that we denote here as SP. The critical difference between AP and SP is that a set of t topic means and standard deviations are computed across the st per-system per-topic scores, and each of the st scores is then converted into a z score with regard to that topic's statistics:

$$e'_{s,t} = \frac{e_{s,t} - \mathit{mean}_t}{\mathit{sd}_t} ,$$

where $e'_{s,t}$ is the z-score corresponding to the average precision effectiveness $e_{s,t}$. Because z-scores are centered on zero, and can be negative as well as positive, Webber et al. further transformed the $e'_{s,t}$ values through the use of the cumulative normal probability distribution. The result is a set of $e''_{s,t}$ scores that lie strictly between zero and one, and which, for each topic, has a mean of 0.5 and a uniform standard deviation. The last row of Table 2 shows the rather unique properties of the set of effectiveness scores that result from this two-stage standardization of AP scores to generate SP scores.

Given the intention of the standardization process, it is unsurprising (Figure 7) that there is no change in sys-

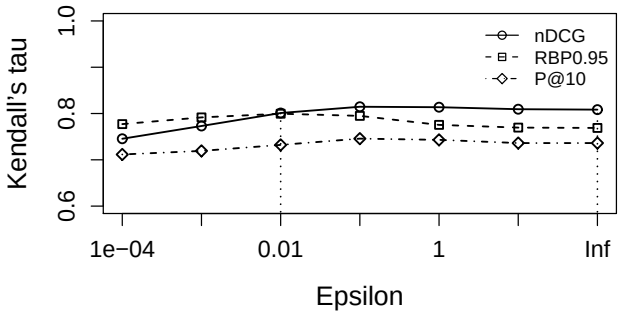

Figure 5: Average system ordering correlations when ϵGM-P@10, ϵGM-RBP0.95, and ϵGM-nDCG are used to score the TREC9 systems. When ϵ approaches infinity the methods converge to AM-P@10, AM-RBP0.95, and AM-nDCG respectively.

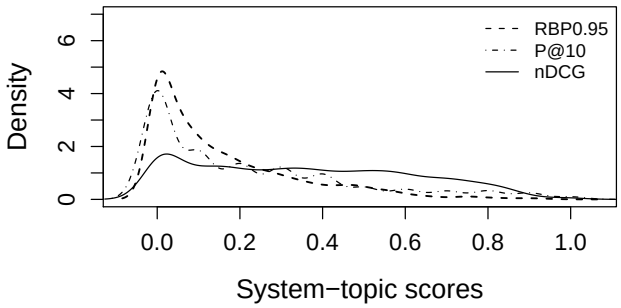

Figure 6: Density distribution of TREC9 system-topic scores using RBP0.95, P@10, and nDCG ($t = 50$ topics and $s = 105$ systems).

tem ranking consistency as ϵ is changed; and also unsurprising (Figure 8, Table 2) that the fraction of SP scores that are near zero is small. That is, the condition that led to geometric mean average precision being proposed as an alternative to arithmetic mean average precision – the presence of important low-scoring topics in amongst the high-scoring ones – is removed by the standardization process, and there is no benefit in shifting to ϵGM aggregation. Instead, AM aggregation, with the elimination of the need to select ϵ, is the preferred approach, because standardization means that all topics are already contributing equally to any assessment in regard to differences in system effectiveness, and no further benefit arises from emphasizing low scores.

Note that standardization is only possible when a set of retrieval systems are being mutually compared, and topic means and standard deviations can be calculated; or when pooled relevance judgements based on a set of previous systems are available, together with the runs that led to those judgements.

Figure 9 draws together the observations we have made. A total of fifteen collection and effectiveness metric couplings are shown, with the percentage of low effectiveness scores for that coupling plotted horizontally, and the extent of the superiority (or inferiority) of ϵGM as an aggregation method relative to AM plotted vertically. The behavior of the SP metric does not depend on a specific aggregation technique to perform well, and for all three collections used is insensitive to the score averaging mechanism. In the case of the other four metrics the trend is as we have observed already: if there are many low scores, ϵGM gives more consistent system orderings than does AM.

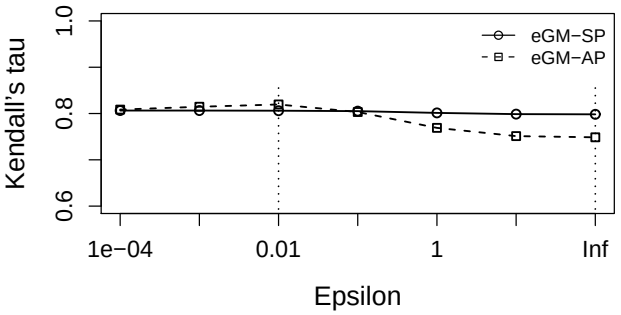

Figure 7: Average system ordering correlations when ϵGM-SP (based on standardized average precision) is used as the evaluation metric, compared to the ϵGM-AP combination. The methodology and data set used in this experiment are identical to those used in Figure 1.

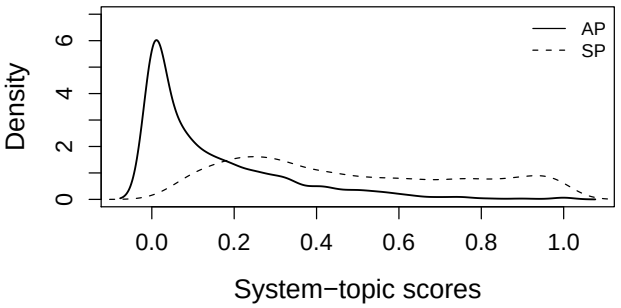

Figure 8: Density distribution of AP and SP scores for the TREC9 Web Track data ($t = 50$ topics and $s = 105$ systems).

5.4 Query "hardness"

All of the results presented thus far have been based on the average performance of the aggregation methods over large numbers of random splittings of the topics in the query set. As an alternative, we also constructed two particular topic subsets based on the nominal topic difficulty scores defined by Equation 2, to evaluate the extent to which the various aggregation mechanisms were affected by topic difficulty. In undertaking this part of the experimentation, we additionally sought to determine whether the additive/subtractive version of the adjusted geometric mean gave rise to consistent rankings, or whether the `trec_eval` ϵ-thresholding version was more consistent.

The two query partitionings were constructed to illustrate extreme behavior. In the first partitioning, the Hard/Easy split, Equation 2 was used to assign a difficulty rating to each of the $t = 50$ TREC9 topics, and then the highest-scoring 25 were taken into one set, and the lowest 25 topics were taken into the other. The second Middle/Rest partitioning again ranked the topics by difficulty, but then extracted the mid-scoring (most internally consistent) group into one subset, and left the remaining 12 hardest topics and 13 easiest ones in the other set. System orderings based on aggregate score over AP and SP for each pair of sets were then computed, and compared using Kendall's τ. The results of these experiments are shown in Table 3. The columns headed " Random" relate to the previous methodology of taking the average τ value over 10,000 random splittings of the 50 topics, and represent the numeric values of some of the points already plotted in the various graphs.

There are a number of trends that can be distilled from Table 3. Looking at the AP halves of the three subtables it is clear that the additive version of ϵGM is preferable to the $\epsilon\text{GM}_{\texttt{trec\_eval}}$ version – in eight of the nine cases, ϵGM detects more similarity between the system orderings,

Aggregation method	AP			SP		
	Random	Hard/Easy	Middle/Rest	Random	Hard/Easy	Middle/Rest
AM	0.748	0.694	0.722	0.798	0.704	0.820
ϵGM, $\epsilon = 0.01$	0.819	0.779	0.826	0.805	0.712	0.824
$\epsilon\text{GM}_{\texttt{trec\_eval}}$, $\epsilon = 0.00001$	0.798	0.818	0.800	0.806	0.802	0.826

(a) TREC9 data set, $t = 50$ and $s = 105$.

Aggregation method	AP			SP		
	Random	Hard/Easy	Middle/Rest	Random	Hard/Easy	Middle/Rest
AM	0.693	0.584	0.611	0.741	0.622	0.798
ϵGM, $\epsilon = 0.01$	0.729	0.650	0.674	0.740	0.622	0.798
$\epsilon\text{GM}_{\texttt{trec\_eval}}$, $\epsilon = 0.00001$	0.682	0.639	0.567	0.737	0.657	0.816

(b) TREC2001 data set, $t = 50$ and $s = 97$.

Aggregation method	AP			SP		
	Random	Hard/Easy	Middle/Rest	Random	Hard/Easy	Middle/Rest
AM	0.773	0.719	0.753	0.781	0.739	0.790
ϵGM, $\epsilon = 0.01$	0.755	0.732	0.760	0.769	0.714	0.785
$\epsilon\text{GM}_{\texttt{trec\_eval}}$, $\epsilon = 0.00001$	0.681	0.672	0.676	0.768	0.680	0.785

(c) TREC8 data set, $t = 50$ and $s = 129$.

Table 3: System ranking correlation coefficients using Kendall's τ, for two different effectiveness metrics, three different score aggregation methods, three different collections, and three different ways of splitting the topics into two halves. The three subtables are ordered by decreasing percentage of system-topic AP scores that are below 0.1. Note that the Hard/Easy and Middle/Rest splits are both one-off arrangements in each of the three data sets.

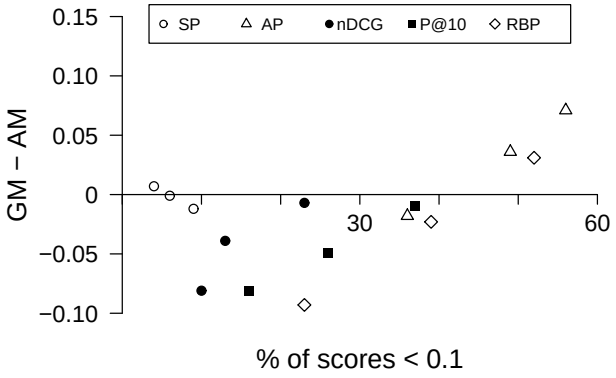

Figure 9: Comparing the ϵGM and AM score aggregation methods across three collections and across five effectiveness metrics. The horizontal axis shows the percentage of low effectiveness scores generated by that collection/metric combination; the vertical axis plots the difference in the Kendall's τ score obtained using $\epsilon\text{GM} - \text{AM}$ and $\epsilon = 0.01$.

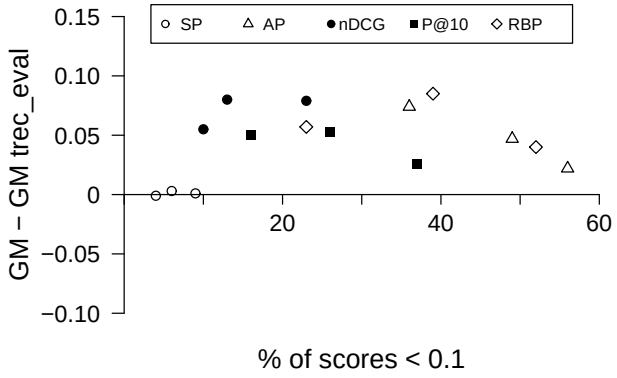

Figure 10: Comparing the ϵGM aggregation method and the thresholded GM variant used in the `trec_eval` program, with other experimental details as for Figure 9. The ϵGM approach is more self-consistent on all three collections, and for four of the five effectiveness metrics used.

including in all three of the Random evaluations.

Second, looking at the top row of each subtable, the AM-SP combination of effectiveness metric and aggregation method is uniformly better than the AM-AP combination that has dominated IR reporting for more than a decade. The third effect to be noted is that the Hard/Easy pairing, with just two exceptions in connection with the $\epsilon\text{GM}_{\texttt{trec\_eval}}$ aggregation, is more likely to lead to different system orderings than is the Middle/Rest topics split. And fourth, it also appears that the Middle/Rest is no less likely to generate different system orderings than a Random split, and hence there is no sense in which it provides a problematic arrangement for the aggregation metrics to handle. This is a slightly surprising outcome, since the Rest group contains topics of widely varying difficulty, at least in terms of Equation 2.

We also explored the use of Equation 1 as a topic difficulty rating, and in results not included here, obtained correlation patterns similar to those shown in Table 3.

Finally in this section, to reinforce our contention that `trec_eval`'s thresholding $\epsilon\text{GM}_{\texttt{trec\_eval}}$ aggregation method is less reliable than the additive/subtractive ϵGM version, Figure 10 repeats the "differences between the average Kendall's τ score" experiment of Figure 9. The three points representing SP on the three different collections are again unaffected by the aggregation method used. But in all of the other twelve cases, ϵGM generates more consistent system rankings than does $\epsilon\text{GM}_{\texttt{trec\_eval}}$. We can see no basis for persisting with the thresholding version of GM-AP provided by `trec_eval`, and suggest that other authors be similarly vary of using the scores it generates.

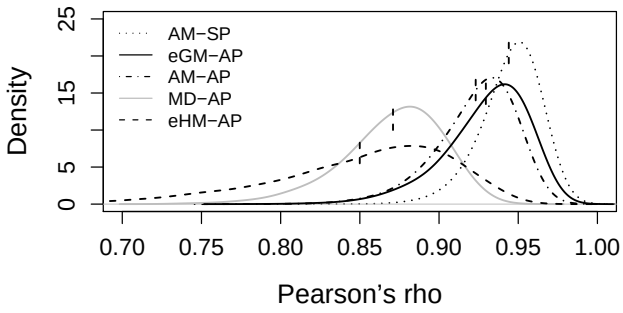

Figure 11: Density distribution of $10{,}000$ Pearson's ρ values for AM-AP, GM-AP ($\epsilon = 0.01$), HM-AP ($\epsilon = 0.01$), MD-AP, and AM-SP, based on $10{,}000$ random splittings of the TREC9 data set ($t = 50$ and $s = 105$). The experimental methodology was as for Figure 2.

5.5 Correlation coefficient methods

We have made extensive use of Kendall's τ in our analysis, starting from the presumption outlined in Section 1 that whether or not significance of any particular pairwise relationships had been established, it was likely that overall system scores would be used to derive system rankings, and thus, that study of score aggregation mechanisms was of merit. Use of Kendall's τ allowed the "closeness" of pairs of system rankings to then be quantified.

Kendall's τ takes into account the system ordering that is generated, but not the scores that led to that ordering, meaning that when there are clusters of near-similar scores, modest changes in the scores can lead to more dramatic changes in the correlation coefficient. An alternative metric that is based on scores rather than rankings is the Pearson product-moment coefficient. To verify that the relationships between rankings that have been noted above are not specific to the use of Kendall's τ, we repeated the experiments that led to Figure 2, using Pearson's ρ to compare pairs of lists of "system, score" pairs. The results are shown in Figure 11, again using the TREC9 resource.

Broadly speaking, Figure 11 shows the same trends as had been identified using Kendall's τ. The best method for obtaining a per-system score is the SP metric coupled with the AM averaging technique. Earlier in this paper we noted that because AM relied on addition, it could technically only be applied to values that were on the same scale, a restriction that did not apply to the geometric mean. The superior performance of AM-SP compared to AM-AP can be interpreted as a verification of this observation, since the process of standardizing the AP scores (subtracting the mean, and then dividing by the standard deviation for that topic) renders them into unitless values on a common scale. On the other hand, the pre-standardization AP values have different scales (in the sense of millimeters versus inches), because what is good performance on one topic might be substandard performance on another. In this sense, the process of standardization makes it "right" to then compute the arithmetic mean as the gross statistic for a system. And, even though the fully standardized scores are constrained to the $(0, 1)$ interval, the fact that equality is not possible at either end of the scale means that no matter how good (or bad) a system is on a particular topic, it is possible – at least in theory – for a different system to be better (or worse). That is, there are no bookends in SP that force systems to be considered to be "equal" on very easy or very hard topics.

Note also in Figure 11 the good performance of ϵGM – by ϵ-adjusting the scores, and computing a geometric mean, small effectiveness values are allowed to contribute to the final outcome. However use of the harmonic mean is not appropriate, and the HM-AP method is inferior to the baseline AM-AP approach. Nor – predictably – is the median an especially good score aggregation technique.

6 Conclusion

Our exploration of AM-AP and ϵGM-AP has confirmed that for the TREC9 Web data the ϵ-adjusted geometric mean is a more appropriate score aggregation mechanism than is the arithmetic mean. This appears to be a consequence of the large number of low AP scores (more than 50% are less than 0.1) across the TREC9 systems and topics. Experiments conducted with other TREC data resources confirm that when a collection has a majority of system-topic AP scores that are zero or close to zero, the ϵ-adjusted geometric mean is a more appropriate score aggregation method than is the arithmetic mean. On the other hand, when there are only a minority of low system-topic scores, the arithmetic mean is resilient, and tends to perform well, where "performs well" is in the sense of system orderings derived from one subset of the topics being similar to the system orderings derived from a different set.

We also experimented with other effectiveness metrics, including P@10, nDCG, RBP0.95, and SP, and observed the same overall outcome – when a large fraction of small effectiveness scores are generated by the metric, it is better to use the ϵGM aggregation approach.

On the other hand, experiments using the standardized precision (SP) metric showed that it anticipates the benefits brought about through the use of the geometric mean, and that the best aggregation rule for it was the standard arithmetic mean. The SP metric has the useful attribute of transforming the effectiveness scores so that, for every topic in the set, the mean score for that topic is 0.5, and the standard deviation is also fixed. Standardizing thus converts all system-topic scores to the same "units", and allows averaging as a logically correct operation.

To broaden the scope of our investigation we also plan to explore both further aggregation mechanisms, such as the MEDRANK approach of Fagin et al. (2003); and also other rank correlation approaches, including the τ_{AP} top-weighted approach of Yilmaz et al. (2008). Top-weighting of the rank correlation scoring mechanism is important if we are more interested in fidelity near the top of each ranking than in (say) the bottom half of the two system orderings being compared.

We conclude by reiterating that our experiments – in which we regard a metric and aggregation technique to be "good" if they yield similar overall system rankings from different topic sets – are founded in practice rather than in theory, and reflect common evaluation custom rather than an underlying principle. We also stress that to be plausible, experimentation should be accompanied by significance testing, and because significance tests are carried out over sets of values, the details of the aggregation technique used to obtain representative gross scores are of somewhat parenthetical interest if significance cannot be asserted. Nevertheless, and even given these caveats, we have found that ϵGM has a role to play when there are many small score values to be handled, and that AM-SP is the combined metric and aggregation technique that is most strongly self-consistent in terms of score-induced system rankings.

Acknowledgment: This work was supported by the Australian Research Council, by the Government of Malaysia, and by the University of Malaya. We thank the referees for their helpful comments.

References

Al-Maskari, A., Sanderson, M., Clough, P. & Airio, E. (2008), The good and the bad system: Does the test collection predict users' effectiveness?, *in* 'Proc. 31st Ann. Int. ACM SIGIR Conf. on Research and Development in Information Retrieval', Singapore, pp. 59–66.

Buckley, C. (2004*a*), Topic prediction based on comparative retrieval rankings, *in* 'Proc. 27th Ann. Int. ACM SIGIR Conf. on Research and Development in Information Retrieval', Sheffield, England, pp. 506–507.

Buckley, C. (2004*b*), Why current IR engines fail, *in* 'Proc. 27th Ann. Int. ACM SIGIR Conf. on Research and Development in Information Retrieval', Sheffield, England, pp. 584–585.

Cormack, G. V. & Lynam, T. R. (2007), Validity and power of t-test for comparing MAP and GMAP, *in* 'Proc. 30th Ann. Int. ACM SIGIR Conf. on Research and Development in Information Retrieval', Amsterdam, The Netherlands, pp. 753–754.

Fagin, R., Kumar, R. & Sivakumar, D. (2003), Efficient similarity search and classification via rank aggregation, *in* 'Proc. SIGMOD Int. Conf. on Management of Data', San Diego, CA, pp. 301–312.

Järvelin, K. & Kekäläinen, J. (2002), 'Cumulated gain-based evaluation of IR techniques', *ACM Transactions on Information Systems* **20**(4), 422–446.

Kendall, M. & Gibbons, J. D. (1990), *Rank Correlation Methods*, Oxford University Press, New York.

Mandl, T., Womser-Hacker, C., Nunzio, G. D. & Ferro, N. (2008), How robust are multilingual information retrieval systems?, *in* 'Proc. 23rd Ann. ACM Symp. on Applied Computing', Fortaleza, Cear, Brazil, pp. 16–20.

Mizzaro, S. (2008), The good, the bad, the difficult, and the easy: Something wrong with information retrieval evaluation?, *in* 'Proc. 30th European Conf. on Information Retrieval', Glasgow, Scotland, pp. 642–646.

Mizzaro, S. & Robertson, S. (2007), HITS hits TREC: Exploring IR evaluation results with network analysis, *in* 'Proc. 30th Ann. Int. ACM SIGIR Conf. on Research and Development in Information Retrieval', Amsterdam, The Netherlands, pp. 479–486.

Moffat, A. & Zobel, J. (2009), 'Rank-biased precision for measurement of retrieval effectiveness', *ACM Transactions on Information Systems*. To appear.

O'Brien, M. & Keane, M. T. (2007), Modeling user behavior using a search-engine, *in* 'Proc. 12th Int. Conf. on Intelligent User Interfaces', Honolulu, Hawaii, USA, pp. 357–360.

Robertson, S. (2006), On GMAP: And other transformations, *in* 'Proc. 15th ACM Int. Conf. on Information and Knowledge Management', Virginia, USA, pp. 78–83.

Robertson, S. (2008), A new interpretation of average precision, *in* 'Proc. 31st Ann. Int. ACM SIGIR Conf. on Research and Development in Information Retrieval', Singapore, pp. 689–690.

Sanderson, M. & Zobel, J. (2005), Information retrieval system evaluation: effort, sensitivity, and reliability, *in* 'Proc. 28th Ann. Int. ACM SIGIR Conf. on Research and Development in Information Retrieval', Salvador, Brazil, pp. 162–169.

Voorhees, E. M. (2003), Overview of the TREC 2003 Robust Retrieval Track, *in* 'Proc. 12th Text REtrieval Conference (TREC 2003)', Gaithersburg, Maryland.

Voorhees, E. M. (2005), Overview of the TREC 2005 Robust Retrieval Track, *in* 'Proc. 14th Text REtrieval Conference (TREC 2005)', Gaithersburg, Maryland.

Webber, W., Moffat, A. & Zobel, J. (2008a), Score standardization for inter-collection comparison of retrieval, *in* 'Proc. 31st Ann. Int. ACM SIGIR Conf. on Research and Development in Information Retrieval', Singapore, pp. 51–58.

Webber, W., Moffat, A., Zobel, J. & Sakai, T. (2008b), Precision-at-ten considered redundant, *in* 'Proc. 31st Ann. Int. ACM SIGIR Conf. on Research and Development in Information Retrieval', Singapore, pp. 695–696.

Yilmaz, E., Aslam, J. A. & Robertson, S. (2008), A new rank correlation coefficient for information retrieval, *in* 'Proc. 31st Ann. Int. ACM SIGIR Conf. on Research and Development in Information Retrieval', Singapore, pp. 587–594.

Zobel, J. (1998), How reliable are the results of large-scale information retrieval experiments?, *in* 'Proc. 21st Ann. Int. ACM SIGIR Conf. on Research and Development in Information Retrieval', Melbourne, Australia, pp. 307–314.