System Scoring Using Partial Prior Information

Sri Devi Ravana Laurence A. F. Park Alistair Moffat
Department of Computer Science and Software Engineering
The University of Melbourne, Australia
{sravana,lapark,alistair}@csse.unimelb.edu.au

ABSTRACT

We introduce smoothing of retrieval effectiveness scores, which balances results from prior incomplete query sets against limited additional complete information, in order to obtain more refined system orderings than would be possible on the new queries alone.

Categories and Subject Descriptors: H.3.4 [Information Storage and Retrieval]: Systems and software – *performance evaluation*

General Terms: Measurement, Performance, Experimentation

Keywords: Score aggregation, system ordering, smoothing, partial information

1. PROBLEM STATEMENT

When selecting an information retrieval system, the three factors that influence the decision are the storage occupied by the system and its files, the speed of index construction and querying, and the effectiveness of the retrieval results.

To evaluate effectiveness, and allow system comparisons to be made, it is usual to use a test collection, comprising a set of documents, a set of queries deemed to be somehow representative of what the retrieval service will deal with in practice, and a set of relevance judgments that indicate which documents are relevant to which query. If each of the test systems is used to index the document collection and execute each of the queries, and scored using an effectiveness metric, an overall average score for each system can be computed, and used to order the systems. Typically, the more queries that are available as part of the test collection, the more stable the overall system ranking is likely to be, providing a direct tension between experimental cost, measured in terms of the number of relevance judgements to be carried out; and experimental stability, measured as the likelihood of having generated the "correct" system ordering.

In this paper we describe a method for blending partial prior system scores into extended experiments, so as to increase experimental stability, without increasing experimental cost. The scenario we consider is this: we suppose that two different sets of retrieval systems have already been ordered using different associated sets of queries, and that we now wish to prepare an overall system ordering that combines the two. What process should be adopted to merge the two separate orderings? And can the prior different-queries system scores be used to advantage, to reduce the cost of creating the overall ordering?

More precisely, we suppose that a set of retrieval systems S_x has been evaluated on a set of topics Q_a, and that a disjoint set of retrieval systems S_y has been evaluated on a disjoint set of topics Q_b, in the arrangement shown in the shaded region in Figure 1. Each of the system subsets is evaluated over the same document collection, using any suitable effectiveness measure, such as average precision, to compute mean average precision (MAP) scores from which a system ordering can be derived. But, because the two query sets Q_a and Q_b are disjoint, it is not clear that the effectiveness scores for S_x (on Q_a) can be directly compared with those for S_y (on Q_b).

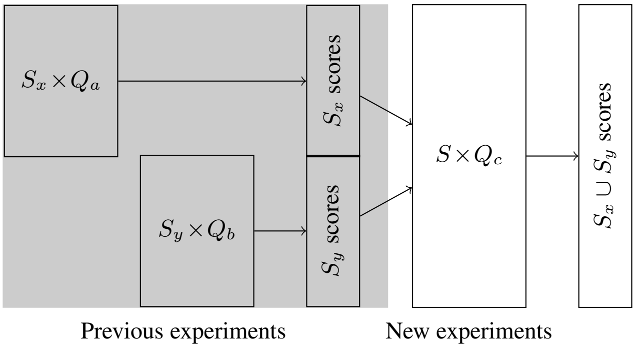

Figure 1: The structure of the experiments undertaken. A set of systems S_x has been executed against query set Q_a to get overall effectiveness scores; another set of systems S_y has been scored using queries Q_b; and now the objective is to compute effectiveness scores for the systems $S_x \cup S_y$ using some number of additional queries Q_c. In the experiments, each of Q_a, Q_b and Q_c contain a sample of 25 distinct topics from a pool containing 249 topics. The query set Q_c is evaluated over all 110 systems, while sets Q_a and Q_b are evaluated over systems S_x and S_y, each of size 55.

2. SMOOTHING SYSTEM SCORES

One obvious way of obtaining overall effectiveness scores and a merged system ordering for $S = S_x \cup S_y$ would be to evaluate S_x against Q_b, to evaluate S_y against Q_a, and thus obtain a ranking of S based on $Q_a \cup Q_b$. But suppose that this is not possible – perhaps because the topic sets or relevance judgements used for subsets Q_a and Q_b are deemed to be of commercial value and have not been made public – and that all that is known are the two sets of system effectiveness scores. If Q_a and Q_b are not available for further testing, then additional queries need to be evaluated against the full set of systems, shown as the box Q_c in Figure 1. Our question then becomes: should the experiment $S \times Q_c$ be regarded as standalone? Or can the system scores calculated for $S_x \times Q_a$ and $S_y \times Q_b$ be used to inform the outcomes of $S \times Q_c$?

SIGIR'09, July 19–23, 2009, Boston, Massachusetts, USA.
ACM 978-1-60558-483-6/09/07.

Smoothing is a process that allows approximate statistics to be computed from insufficient data [1]. To allow blending of effectiveness scores between prior knowledge and new observations, a smoothing term is introduced into the effectiveness metric. This allows the new observations to be tempered by other information, even if that other information is not provided with the same coverage of systems as the new. For example, we propose that *smoothed average precision* for a system s on query q in the context of prior information for a query set Q be computed as:

$$\widetilde{\mathrm{AP}}_{s,q,Q} = \alpha \mathrm{AP}_{s,q} + (1-\alpha)\mathrm{MAP}_{s,Q} \qquad (1)$$

where $\mathrm{MAP}_{s,Q}$ is the mean average precision of system s on query set Q; $\mathrm{AP}_{s,q}$ is the average precision of system s on an additional query q; and $\alpha \in [0,1]$ is a constant.

Once $\widetilde{\mathrm{AP}}_{s,q,Q}$ has been computed for each $q \in Q_c$, the mean smoothed average precision for Q_c in the context of Q can be calculated, and used in place of mean average precision. Those adjusted scores can then be used to obtain a system ordering in which the system behavior on the new queries Q_c is moderated by the prior scores obtained by the systems on Q. Note that the query set Q_c is same for all systems being compared, so that the score comparisons are fair; but that the smoothing query set Q does not have this restriction, and might differ on a system-by-system basis. Note also that smoothing can be applied to any effectiveness metric, and in our experiments we have used both smoothed average precision and smoothed standardized average precision [3].

3. EXPERIMENT

To examine the effect of smoothing on the accuracy of the system ordering, we used the 249 topics prepared for the 2004 TREC Robust track [2], and the 110 systems that submitted runs against them. To perform each experiment, we randomly sampled three sets of 25 mutually exclusive topics to form Q_a, Q_b and Q_c, and randomly split the 110 systems S into 55 systems for S_x and 55 systems for S_y. Scores were then computed for $S \times Q_c$, for $S_x \times Q_a$, and for $S_y \times Q_b$, as shown in Figure 1. We also computed, as a baseline, the experiment $S \times (Q_a \cup Q_c)$, to measure how stable the system orderings would be if 50 topics could be used against the full set of systems.

The various system orderings resulting from each experiment (baseline, plus $\alpha \in \{0, 0.5, 0.8, 1\}$) were then compared against the "whole population" system ranking obtained using all of the 249 queries, and Kendall's tau rank correlation scores computed. This experiment was then repeated 10,000 times, with two different effectiveness metrics used to generate the scores.

Figure 2 shows the τ distribution arising when AP is used as the effectiveness metric. The inclusion of the two sets of background scores means that a system ordering is generated that on average is closer to the presumed ground truth than the system ordering generated by the background queries only (when $\alpha = 0$), or by the set Q_c only (when $\alpha = 1$). In this experiment, the best value of α was approximately 0.8, with the prior scores given one quarter of the weight of the all-systems scores on the Q_c query set. The system ranking that results is still not as accurate as can be obtained if 50 queries are used against all of the systems (the baseline in the graph), but nor is it as expensive to compute.

Figure 3 shows the same experiment, but with standardized average precision used as the underlying effectiveness metric, with the standardizing factors computed according to the subset of queries and topics being used in each case. The same overall pattern of behavior can be observed, providing further evidence of the usefulness of smoothing. Now the best value of α is 0.5.

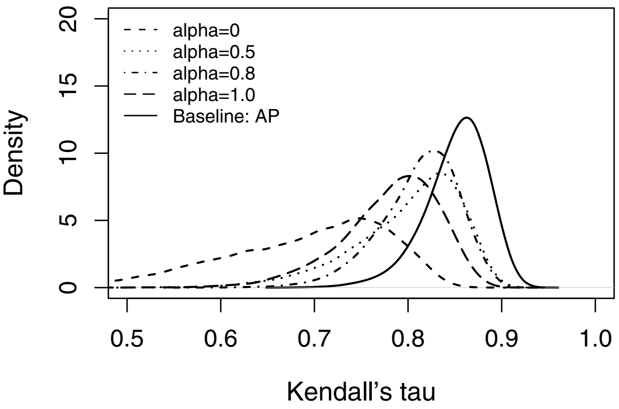

Figure 2: The distribution of τ values for $\widetilde{\mathrm{MAP}}$ (mean smoothed average precision) using $\alpha = 0$, 0.5, 0.8 and 1.

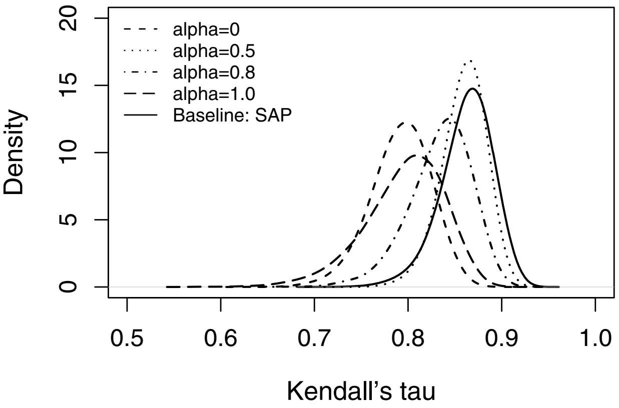

Figure 3: The distribution of τ values for $\widetilde{\mathrm{MSAP}}$ (mean smoothed standardized average precision) using $\alpha = 0$, 0.5, 0.8 and 1.

4. CONCLUSION

We have introduced smoothing of effectiveness scores, using information obtained from prior queries to temper the system scores assigned during subsequent evaluations. We have shown that blending previous system scores, even though those scores were not obtained on a consistent topic set, boosts the quality of followup system evaluations. This simple process means that all-systems evaluations can be carried out as a follow-on to preliminary small-scale experiments.

Acknowledgment: This work was supported by the Australian Research Council, by the Government of Malaysia, and by the University of Malaya.

5. REFERENCES

[1] S. F. Chen and J. Goodman. An empirical study of smoothing techniques for language modeling. In *ACL '96: Proc. 34th Ann. Meet. Assoc. Computational Linguistics*, pages 310–318, Morristown, NJ, 1996. Association for Computational Linguistics.

[2] E. M. Voorhees. The TREC robust retrieval track. *SIGIR Forum*, 39(1):11–20, 2005.

[3] W. Webber, A. Moffat, and J. Zobel. Score standardization for inter-collection comparison of retrieval systems. In *SIGIR '08: Proc. 31st Ann. Int. ACM SIGIR Conf. Research and Development in Information Retrieval*, pages 51–58, New York, NY, 2008. ACM.