Has Adhoc Retrieval Improved Since 1994?

Timothy G. Armstrong, Alistair Moffat, William Webber, Justin Zobel
Computer Science and Software Engineering
The University of Melbourne
Victoria 3010, Australia
{tgar,alistair,wew,jz}@csse.unimelb.edu.au

ABSTRACT

Evaluation forums such as TREC allow systematic measurement and comparison of information retrieval techniques. The goal is consistent improvement, based on reliable comparison of the effectiveness of different approaches and systems. In this paper we report experiments to determine whether this goal has been achieved. We ran five publicly available search systems, in a total of seventeen different configurations, against nine TREC adhoc-style collections, spanning 1994 to 2005. These runsets were then used as a benchmark for reassessing the relative effectiveness of the original TREC runs for those collections. Surprisingly, there appears to have been no overall improvement in effectiveness for either median or top-end TREC submissions, even after allowing for several possible confounds. We therefore question whether the effectiveness of adhoc information retrieval has improved over the past decade and a half.

Categories and Subject Descriptors

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

General Terms

Retrieval experiment, evaluation, system measurement.

1. INTRODUCTION

Public information retrieval (IR) evaluation efforts such as TREC and CLEF use pooled resources to develop shared materials for system evaluation; see Voorhees and Harman [2005] for details. Since TREC is a yearly effort, drawing participation from leading IR research groups, it is natural to ask whether (and by how much) retrieval systems have improved over time. However, comparing results between different TREC years is problematic. First, query difficulty is highly variable, so changes in the mix of queries between collections leads to variation in collection difficulty. Second, even within the same track designation, such as the Adhoc track, the nature of the task can change from year to year.

Variability in topic difficulty can be addressed with score standardization [Webber et al., 2008]. The observed scores of a set of systems for a topic are used to estimate the difficulty and variability of the topic, and effectiveness scores for that topic are then standardized using these factors. Change in the nature of the task, on the other hand, is not fully addressed by standardization, and poses an open challenge.

In this paper, we investigate trends in the effectiveness of retrieval systems submitted to nine Adhoc and Robust tracks of TREC from 1994 to 2005. Several publicly available search engines are used, with varying parameters, across all test collections. These form a standardizing set, and are a reference point the original TREC runs can be compared to. Our starting hypothesis was that we would observe an upward trend in effectiveness, possibly plateauing in later TRECs. However, the results show no such trend. This raises the question of whether adhoc retrieval on newswire-style data has actually improved since 1994.

2. EXPERIMENTAL METHOD

The TREC tracks examined are Adhoc in TRECs 3 to 8 (1994 to 1999) and Robust in TRECs 2003 (new topics only), 2004 (new topics only), and 2005. We selected five publicly available search engines: Apache Lucene 2.4.0; Indri (bundled with the Lemur 4.8.0 toolkit); Zettair 0.9.2; Terrier 2.2; and mg 1.2.1.[1] These five systems were run against all of the test collections. For each system, one out-of-the-box run was made. Another 12 parameterized runs were also performed (3 for Zettair, 4 for Terrier, and 5 for Indri). This reference set of systems was then scored against the nine different collections, to derive standardization factors for each topic in each collection. The metric used is average precision (AP). All runs used title plus description fields as queries (description only for TREC 4, which lacked titles).

A limitation of this approach is that the set of reference systems was not as diverse as the population of original systems submitted to TREC, and the AP standard deviation for each search topic was markedly smaller than for submitted TREC runs. This gave rise to a large proportion of standardized topic scores close to 0.0 or 1.0, reducing the fidelity of standardized scores. To compensate for this, two virtual (or *background*) systems were added to the standardizing set, one scoring 0 for every topic, the other scoring 1.

Using these standardizing factors, mean standardized AP scores were calculated for all of the automatic systems participating in the original TREC experiments. A standardized score of 0.5 for a system means that it has average performance relative to the augmented set of 19 standardizing systems. See Webber et al. [2008] for details of the standardization process. The distributions of the standardized system scores for different TRECs were then compared to see if any clear trend in performance emerged.

Our hypothesis was one of gradual improvement in effectiveness over time. More concretely, based on Table 13.1 on page 311 of Voorhees and Harman [2005], we guessed that performance might increase by a factor of around 1.5, reflecting the gains accruing as

[1] Software available from `lucene.apache.org`, `www.lemurproject.org/indri`, `www.seg.rmit.edu.au/zettair`, `ir.dcs.gla.ac.uk/terrier`, and `www.cs.mu.oz.au/mg`.

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

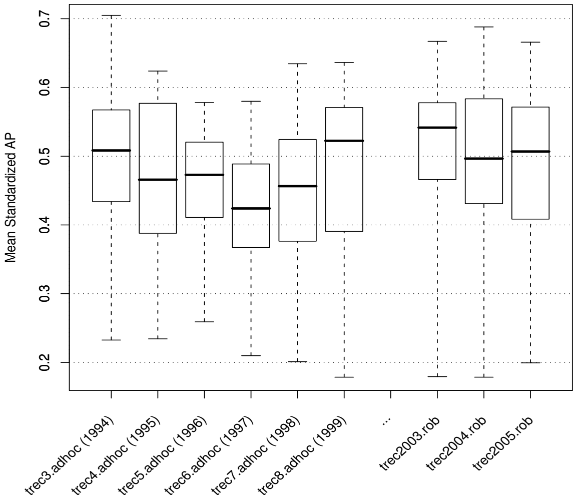

Figure 1: Mean standardized AP scores of runsets submitted to 9 TREC events, excluding manual systems, with standardization factors established by a pool of 17 current public systems and their variants, plus 2 background systems. The central line in the box is the median score; the top and bottom of the boxes are the quartiles.

	T3	T5	T6	T8	T03	T05
indri_ootb_dirichlet	.494	.482	.592	.450	.484	.442
lucene_ootb	.393	.465	.576	.430	.490	.405
mg_cosine	.281	.318	.390	.266	.200	.302
terrier_DFRee	.568	.532	.622	.582	.578	.500
zettair_ootb_dirichlet	.292	.441	.501	.415	.446	.364

Table 1: Scores for out-of-the-box IR systems, standardized by the set of automatic runs submitted to the corresponding TREC events.

typical systems moved from older approaches, such as cosine similarity with simple weighting, to probabilistic and language models, and as techniques such as query expansion were adopted. Figure 1 shows the results: our hypothesis is far from confirmed. The earliest set of systems, those from TREC 3, are as a group superior to those of the subsequent four years, and are competitive with the Robust runs of a full decade later. Indeed, the best system from TREC 3 – one of the first generation of BM25 runs from City University, London – remains, when standardized, one of the best systems in the entire 12-year dataset.

A possible confound is the changing nature of query formulation from topics over time. Earlier TRECs had no title-only runs, which in general (though not universally) demonstrate retrieval performance inferior to that of longer queries, because of the additional information from other topic fields that is omitted. We tested the effect of excluding title-only runs from our analysis. The results, in Figure 2, show a marginal improvement in median and quartile scores from TREC 8 onwards, but still no clear upwards trend. Title-only runs (not shown) are only available from TREC 6 onwards, but here too no clear trend of improvement was observed, and the mid-range TREC 6 systems were competitive with mid-range runs from later years. There did appear to be some improvement in the performance of the top runs, although the sample size of title-only runs, on average 21 runs per TREC, or 5 runs in the top quartile, was too small to draw any firm conclusions.

There are many other aspects that go towards determining the quality of a set of TREC systems. Participants may put more effort into tuning their systems one year than in another. Perhaps, as the popularity of TREC grew, the median quality of the participating groups fell. And, of course, there is a certain degree of randomness involved. But even so, it is striking that, as a group, the participating systems in the premiere public IR evaluation forum appear to demonstrate no consistent improvement over time.

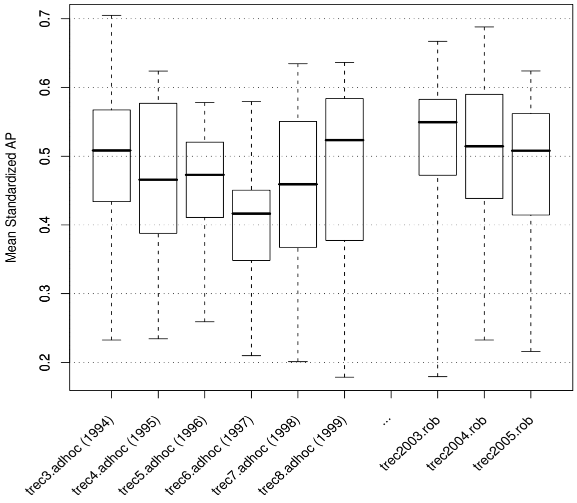

Figure 2: Mean standardized AP scores for runsets submitted to TREC, excluding manual systems and runsets using only titles.

We also note that the current publicly available IR systems have not captured the effectiveness achievements observed in the better historical TREC runs. Table 1 shows the scores achieved by the systems in their out-of-the-box configuration, standardized relative to the corresponding TREC systems for that year. Only one system scored consistently above 0.5, meaning that the other publicly available systems would have been at best average performers in each TREC event – not merely from the TRECs as a whole, but even fifteen years ago in 1994! The comparison is possibly unfair, in that TREC teams put considerable effort into tuning their systems for each test collection, whereas the out-of-the-box systems were not tuned. Providing exact scores for the reference systems with non-default settings is inappropriate, since we cannot claim to have tuned them to achieve optimal performance. Better performance was achieved by turning on features such as query expansion. Even so, none of the reference runs came close to the best TREC systems in performance.

3. CONCLUSION

We found no evidence that the TREC Adhoc and Robust track systems improved overall from 1994 to 2005. There are many factors that could account for the apparent stagnation of adhoc retrieval, and a more detailed investigation may uncover factors we have not considered. Nevertheless, the question remains: have adhoc retrieval techniques improved since 1994? The evidence of the TREC experiments suggests, possibly not. The challenge now is to either find the problem in our methodology, or face the possibility that the gains in performance for over a decade have been illusory.

4. ACKNOWLEDGEMENTS

This work was supported by the Australian Research Council, and the Australian Government through NICTA.

References

E. M. Voorhees and D. K. Harman, editors. *TREC: Experiment and Evaluation in Information Retrieval*. Addison-Wesley, 2005.

W. Webber, A. Moffat, and J. Zobel. Score standardization for inter-collection comparison of retrieval systems. *SIGIR 2008*, pages 51–58.