Relative Significance is Insufficient: Baselines Matter Too

Timothy G. Armstrong, Justin Zobel, William Webber, Alistair Moffat

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

ABSTRACT

We have tabulated retrieval effectiveness claims from a large number of information retrieval research papers from 1998–2008, a period that has seen many innovations. The results of our analysis are not encouraging. Over this period, although a great many papers claimed significant effectiveness improvements, there has been no overall gain in absolute retrieval effectiveness on TREC ad hoc collections. A decade of development has not, it appears, led to better systems.

To promote verifiable improvement, reporting practices that allow rigorous comparison with prior results are needed. We propose several measures: ongoing longitudinal surveys; better reporting of baselines and use of standard systems; and use of resources such as our `evaluatIR.org`, an accessible database of test results.

1. INTRODUCTION

A core goal of information retrieval (IR) research is to make ongoing improvements in retrieval system effectiveness. A tenet of our community is that – through incremental improvement, and innovations such as language models and query expansion – we have gradually improved the effectiveness of search systems. To verify claimed improvements, we create standard test collections, in particular through the TREC mechanism; and we carry out "before" and "after" trials, measuring performance using a standard metric such as mean average precision (MAP). We also use the literature to argue the details of test collection creation and of effectiveness measures, but are confident that their systematic adoption has let us measure progress in the field.

However, a careful tabulation of the last decade of IR literature reveals a picture that for ad-hoc retrieval is far from encouraging. The reported effectiveness results show no pattern of improvement in MAP at all, and even in 2008 many new results that were validated via experiments using old collections were below the median results of a decade ago. Furthermore, these "improved" results are often worse than those available from the publicly available Terrier system. It seems that over a decade or more, authors have published and referees have approved work that, taken collectively, has done little to advance the effectiveness of IR systems.

We see this problem as a broad failure of experimental method. There are straightforward mechanisms that could lead to better outcomes, but adopting them will require determination on the part of the community, as at face value they would mean that many current papers would not be publishable.

SIGIR Workshop on the Future of IR Evaluation, July 23, 2009, Boston
.

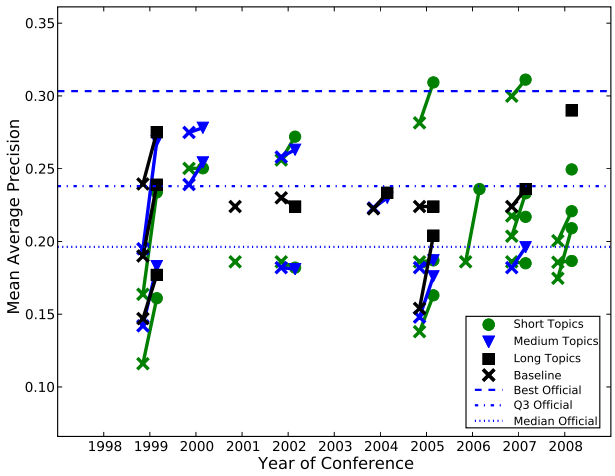

Figure 1: Published MAP scores for the TREC 7 Ad-Hoc collection. The connections show before-after pairs.

2. METHODOLOGY

All papers published at the annual ACM SIGIR conference for the period 1998–2008, and at the ACM CIKM conference for 2004–2008, were scrutinized for experimental effectiveness results. A large proportion of new IR techniques are first presented in SIGIR, so it is where we expect to find results that are indicative of the overall state of IR research. In recent years the CIKM conference has also become a significant forum for IR research.

Results were tabulated for papers that presented effectiveness scores for ad-hoc style retrieval on TREC collections, meaning TREC Ad-Hoc, Robust, Web, and Terabyte collections, and subsets thereof. Note was made of all MAP and P@10 effectiveness scores, as these are the most commonly reported metrics and the only ones used regularly enough in the period surveyed to permit a longitudinal analysis. Careful attention was paid to the distinction between "baseline" and "improved" (or "before" and "after") values. The analysis identified 87 SIGIR papers and 21 CIKM papers. Of these, 90 were focused on retrieval effectiveness; 8 on efficiency; 5 on distributed retrieval; and 5 reported scores but did not make clear claims. The set of papers studied included four that had authors in common with this abstract.

Results for a representative test collection and measure (TREC 7, using MAP), are shown in Figure 1. The trend visible in this plot is typical of what we found for all of the Ad-Hoc retrieval tasks, including the Robust track, and the Web tracks in TREC 9 and TREC 2001. There is no clear upward or downward trend in retrieval effectiveness, and since 1998 the vast majority of scores

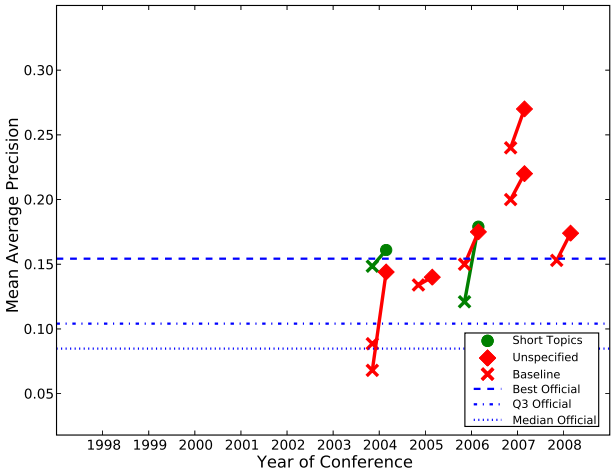

Figure 2: Published MAP scores for the TREC 2003 web track topic distillation task in SIGIR and CIKM Proceedings.

fluctuated in the range of the upper 50% of official TREC run scores. Baselines show a similar trend: the relationship between baseline score and the claimed score in each paper is stronger than any incremental performance improvements over time.

With the exception of the TREC 6, TREC 7 and TREC Robust 2004 and 2005 scores reported in two papers [Liu et al., 2005, Zhang et al., 2007], and a TREC 4 score reported in a 1998 paper [Mitra et al., 1998], we found no ad-hoc retrieval results that exceeded the scores of the best corresponding automatic TREC run. It might be argued that the maximum TREC scores are an unstable and unfair baseline, and that because of per-topic variation in system performance we would expect some outlier systems in a big pool such as the TREC competitions, purely by chance. However, given the time that has elapsed and the number of publications claiming significant (and sometimes substantial) effectiveness improvements, it is surprising that the original best systems are so rarely bettered – especially given the fact that the original runs were the only ones conducted without the benefit of hindsight. As a contrast to the ad-hoc retrieval tasks, Figure 2 shows that there have been ongoing performance improvements for the web topic distillation task of TREC 2003.

Another finding of our analysis was the large number of variant test collections used, despite the survey's restriction to 11 base collections. In 108 publications, 83 different test collections were used, with variants derived by subsetting or combining topics and corpuses from different base collections. There was also little use of standard retrieval systems, even though public domain systems are competitive with published results, and are natural baseline candidates. For instance, Terrier achieves a MAP of 0.248 on the TREC 7 Ad-Hoc collection[1], beating all but four results published since Terrier's 2005 release.

3. PROPOSALS

Future IR evaluations will need to consider the issues raised by our analysis, including the lack of gains overall, the apparent readiness of reviewers to accept papers that have results that are demonstrably weak, and the lack of transparency in many retrieval experiments. It is our view that even significant improvements on a poor baseline should not in themselves merit publication, as such results do not prove that the method being tested would be effective when added to a more competitive baseline. Yet many papers report experimental results using non-standard test-collections, make poor baseline choices, do not report best prior results, and do not provide sufficient experimental detail that would allow their claims to be independently reproduced.

Having an expectation of thorough and consistent reporting of past results would go some way to addressing these concerns, but in our view more is required. We have created a resource for researchers that can bring together all relevant effectiveness results in a way that permits easy comparisons and benchmarking, namely `evaluatIR.org` [Armstrong et al., 2009]. We see several uses for the system: as a resource for analysis of a researcher's own runs against a large database of existing results; as a repository for use by readers and reviewers of papers who wish to evaluate published claims; and as a database that allows the IR community to perform longitudinal and other comparative analyses. Use of this resource is, however, a challenging step: few new methods appear to be competitive with established benchmark systems, and the papers describing them would thus be at risk of summary rejection.

As a related step, we should expect researchers to use multiple test collections, and, more significantly, multiple retrieval systems, to demonstrate that new techniques provide verifiable improvements in combination with a range of configurations.

4. CONCLUSION

Our longitudinal survey of published IR results in SIGIR and CIKM proceedings from 1998–2008 has revealed that ad-hoc retrieval does not appear to have measurably improved. There are many possible explanations for this apparent stagnation, but it is troubling that it appears to have gone largely unremarked within the IR community. It is also paradoxical that the stream of incremental "significant" effectiveness improvements in the literature has not resulted in any apparent cumulative improvements.

Whatever the future direction of IR evaluation, there are fundamental issues with reporting practices that must be addressed. Our evaluation suggests that current methods for measuring improvement are not adequate, and that unless we adopt rigorous strategies for identifying which techniques in the field are of genuine value, we risk remaining on a treadmill of inconclusive experimentation.

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

References

T. G. Armstrong, A. Moffat, W. Webber, and J. Zobel. EvaluatIR: An online tool for evaluating and comparing IR systems. In *Proc. 32nd Ann. Int. ACM SIGIR Conf. on Research and Development in Information Retrieval*, Boston, USA, 2009. Demo.

S. Liu, C. Yu, and W. Meng. Word sense disambiguation in queries. In *Proc. 14th ACM Int. Conf. on Information and Knowledge Management*, pages 525–532, Bremen, Germany, 2005.

M. Mitra, A. Singhal, and C. Buckley. Improving automatic query expansion. In *Proc. 21st Ann. Int. ACM SIGIR Conf. on Research and Development in Information Retrieval*, pages 206–214, Melbourne, Australia, 1998.

W. Zhang, S. Liu, C. Yu, C. Sun, F. Liu, and W. Meng. Recognition and classification of noun phrases in queries for effective retrieval. In *Proc. 16th ACM Conf. on Information and Knowledge Management*, pages 711–720, Lisbon, Portugal, 2007.

[1]From `ir.dcs.gla.ac.uk/terrier`, specifically Terrier 2.2 with BM25 similarity ($b = 0.3$) and query expansion (Bose-Einstein 1 term weighting model with 3 documents and 10 terms) using Title+Description queries.