EvaluatIR: An Online Tool for Evaluating and Comparing IR Systems

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

Categories and Subject Descriptors

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

General Terms

Retrieval experiment, evaluation, system measurement.

Extended Abstract

A fundamental goal of information retrieval research is to develop new retrieval techniques, and to demonstrate that they attain improved effectiveness compared to their predecessors. To quantitatively compare IR techniques, the community has developed a range of standard test collections, in particular the TREC collections; see Voorhees and Harman [2005].

Researchers use these collections as experimental test-beds, and use the observed improvements as evidence of the significance of their research contribution. Most commonly, a baseline system is chosen and improvements relative to this are measured and then presented as evidence of superiority. However, these baselines are frequently inappropriate, and there is often little consistency between researchers or research groups as to how effectiveness experiments are carried out and then reported. Ideally, the current best published results would be used as a baseline, but such practice is rare; and – a further confound on good practice – researchers usually only publish summary metrics, which cannot be used to establish statistical significance when used in subsequent comparisons. The original TREC runs are available for detailed analysis, but are rarely referred to when new methods are proposed.

Instead, authors make use of off-the-shelf software, or of variants of their own software, but neither of these approaches is particularly compelling. Any claims based on comparison to such baselines must be treated with scepticism, and researchers can easily (either inadvertently or deliberately) publish non-competitive "improvements" simply by comparing to an even poorer baseline. For example, in some papers the developers of query expansion techniques compare to unexpanded baselines; whether the methods improve on other expansion techniques is not demonstrated. More broadly, it is often the case that a method that improves on a poor baseline is in effect doing no more than compensating for a defect, and the method cannot improve a system that is already effective.

These issues mean that a reader or referee cannot easily establish whether published results demonstrate a genuine advance in effectiveness, and the enormous labor invested in developing test collections is greatly undermined.

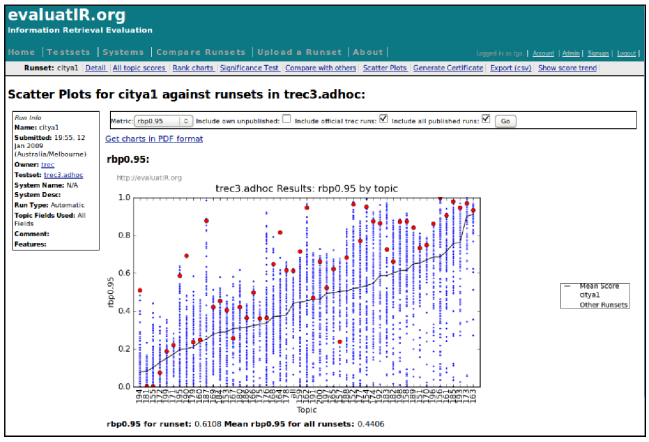

Figure 1: The per-topic RBP scores for a runset on `evaluatIR.org` are compared graphically with others for the same collection.

We have developed a mechanism for authenticating tests undertaken on a standard corpus. Our website `evaluatIR.org` provides an independent, permanent measure of effectiveness that can be referred to by both authors and subsequent readers. Researchers seeking a comparison upload runs via the browser-based interface, and the website returns a link to a page with performance results and statistical comparisons to baselines, using measures such as MAP, nDCG, and RBP, and techniques such as longitudinal standardization. By comparing against standard baselines and up-to-date runs submitted by others, researchers can determine whether their methods provide a true improvement over earlier work, and readers and referees can more easily assess claimed results.

A permanent URL will be provided for each submitted run that a researcher has asked to be "published", so that reviewers and readers can explore the details of the reported results, and use them as a baseline in future work. We envision that as a community we would develop a culture of expecting that published results be made available for scrutiny in this way, and that, with widespread use of `evaluatIR.org`, research in IR will be placed on a stronger, more verifiable footing.

Acknowledgements

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

References

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

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