Document Ranking for Effectiveness-Efficiency Tradeoffs

Vo Ngoc Anh *Alistair Moffat*

Department of Computer Science and Software Engineering
The University of Melbourne
Victoria 3010, Australia

{*vo,alistair*}*@csse.unimelb.edu.au*

Aim: A large-scale document ranking system should be both effective and efficient. Here we summarize the main features of a prototype system we built for that purpose. The success of the system is demonstrated through its relative performance for the efficiency task of the TREC 2005 Terabyte Track.

Method: Building a document ranking system involves two key decisions: choosing a retrieval model, and choosing a suitable index representation. The former determines the *effectiveness* of the system, the latter the *efficiency*; and each of them affects the other.

The impact-based document ranking mechanism described by Anh and Moffat [2] was chosen for our system because of its balance between effectiveness and efficiency. In terms of effectiveness, it is highly competitive, although still inferior to advanced language modelling implementations. On the other hand, in terms of efficiency the mechanism is excellent, as it ranks documents using a small number of calculations, all on integer numbers. To further facilitate query efficiency, we compress the index using the `slide-8` coding scheme [1], which allows an excellent balance between compression ratio and decoding speed.

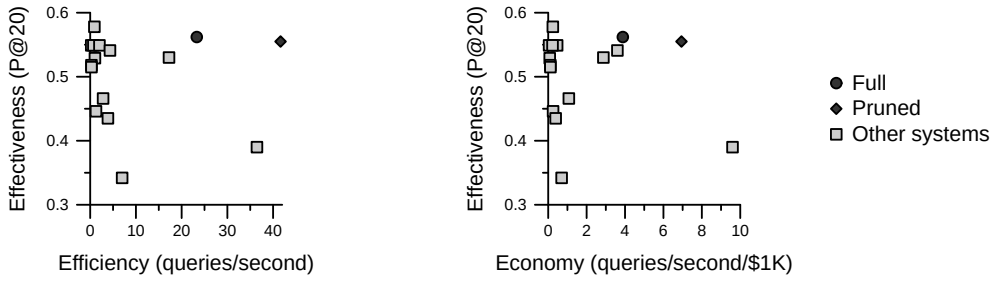

Figure 1: Relative performance in the 2005 TREC Terabyte Track (using the 426 GB `GOV2` collection), with effectiveness compared to efficiency (left-hand graph) and economy (right-hand graph), the latter calculated as throughput normalized by estimated system cost ($US). Efficiency is measured over 50,000 real-life queries; effectiveness over a subset of 50 queries.

Results: Figure 1 shows the relative performance of our system in the efficiency task of the TREC 2005 Terabyte Track. The graphs were compiled from data covering the 16 runs submitted by the 8 groups with the best scores according to the metric P@20 [3]. Our two runs are labelled `Full` and `Pruned`. The former refers to the full processing of all pointers in all inverted lists, the latter to a run in which low-impact pointers were ignored.

Conclusion: This year's involvement in the TREC Terabyte Track showed that our system provides a good balance between effectiveness and efficiency.

References

[1] V. N. Anh and A. Moffat. Improved word-aligned binary compression for text indexing. Submitted, 2005.

[2] V. N. Anh and A. Moffat. Simplified similarity scoring using term ranks. In G. Marchionini, A. Moffat, J. Tait, R. Baeza-Yates and N. Ziviani (editors), *Proc. 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 226–233, Salvador, Brazil, August 2005. ACM Press, New York.

[3] C. L. A. Clarke and F. Scholer. The TREC 2005 Terabyte Track. In *The Fourthteenth Text REtrieval Conference (TREC 2005) Notebook*, Gaithersburg, MD, November 2005. National Institute of Standards and Technology. Available at http://trec.nist.gov/act_part/t14_notebook/t14.notebook.html.

Proceedings of the 10th Australasian Document Computing Symposium, Sydney, Australia, December 12, 2005.