# Document Priors for Query Prediction

Steven Garcia Nicholas Lester Justin Zobel
School of Computer Science and Information Technology
RMIT University, GPO Box 2476V, Melbourne 3001, Australia
{garcias,nml,jz}@cs.rmit.edu.au

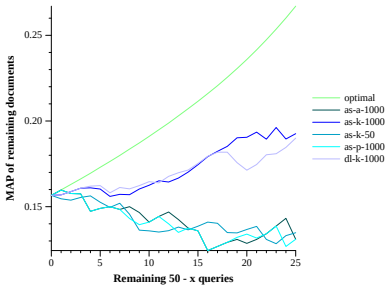
**Aim:** To predict the difficulty of a query issued to an information retrieval system. A query is considered difficult when the effectiveness of the search results are poor.

**Methods:** It has been shown that the likelihood of document access from a collection is non-uniform [2]. As such, for each document in a collection, a probability can be obtained that indicates the likelihood of seeing that document in any given result set. We propose several approaches to query difficulty prediction that take advantage of the non-uniform likelihood of document access by a search system.

Given a set of document probabilities, an absolute ordering of documents from most to least likely to be retrieved indicates a default ranking for documents in the collection. For a generic query, we expect the documents in the result set to be ranked in much the same order as the absolute ordering based on the document prior probabilities. Therefore, a query that produces a result set with documents that do not significantly differ in order to the prior based absolute ordering, is considered to be a difficult query. Conversely, a query that produces a result set that significantly differs in order to the absolute ordering is considered to have high discriminatory power, and therefore is considered a simple query to resolve. We propose five query prediction measures based on document priors, discussion of each technique is presented elsewhere [1].

In recent years, query difficulty prediction has been incorporated into TREC as a part of the Robust track [3]. One metric to measure the quality of a set of predictions is the area under the curve metric that measures the quality of the prediction by measuring the difference in average precision between the worst 25 predicted topics of a run, to the worst 25 performing topics. We use this metric on the TREC 2005 Robust topics and the Aquaint collection.

**Results:** The figure illustrates the performance of our five techniques using the area under the curve approach. Each line shows the mean average precision for the remaining TREC queries after removing the worst $x$ queries. The best possible prediction for this run is shown with the optimal curve. The area between the optimal curve and the other curves shows the gap between that approach and the optimal prediction.




**Conclusions:** We explore a novel approach to query difficulty prediction and propose five metrics to determine the query difficulty based on document priors. Two of the five techniques show promise as predictors. We plan to further explore difficulty prediction using document priors in combination with other query prediction techniques, with the hope of further improving query difficulty prediction.

## References

[1] Y. Bernstein, B. Billerbeck, S. Garcia, N. Lester, F. Scholer, J. Zobel and W. Webber. Rmit university at trec 2005: Terabyte and robust track. In E. M. Voorhees and L. P. Buckland (editors), *Proc. Text Retrieval Conf. (TREC)*, Gaithersburg, MD, November 2005. National Institute of Standards and Technology. Proceedings to appear.

[2] S. Garcia, H. E. Williams and A. Cannane. Access-ordered indexes. In V. Estivill-Castro (editor), *Proceedings of the 27th Conference on Australasian Computer Science*, Volume 26, pages 7–14, Dunedin, New Zealand, January 2004. Australian Computer Society.

[3] E. M. Voorhees. Overview of the TREC 2004 robust track. In E. M. Voorhees and L. P. Buckland (editors), *Proc. Text Retrieval Conf. (TREC)*, Gaithersburg, MD, November 2004. National Institute of Standards and Technology Special Publication 500-261.

**Proceedings of the 10th Australasian Document Computing Symposium, Sydney, Australia, December 12, 2005.** **Copyright for this abstract remains with the authors.**