Low-Cost and Robust Evaluation of Information Retrieval Systems

Benjamin A. Carterette[*]
Center for Intelligent Information Retrieval
Department of Computer Science
University of Massachusetts Amherst
carteret@cis.udel.edu

Abstract

Research in Information Retrieval has progressed against a background of rapidly increasing corpus size and heterogeneity, with every advance in technology quickly followed by a desire to organize and search more unstructured, more heterogeneous, and even bigger corpora. But as retrieval problems get larger and more complicated, evaluating the ranking performance of a retrieval engine gets harder: evaluation requires human judgments of the relevance of documents to queries, and for very large corpora the cost of acquiring these judgments may be insurmountable. This cost limits the types of problems researchers can study as well as the data they can be studied on.

We present methods for understanding performance differences between retrieval engines in the presence of missing and noisy relevance judgments. The work introduces a model of the cost of experimentation that incorporates the cost of human judgments as well as the cost of drawing incorrect conclusions about differences between engines in both the training and testing phases of engine development. Through adopting a view of evaluation that is more concerned with distributions over performance differences rather than estimates of absolute performance, the expected cost can be minimized so as to reliably differentiate between engines with less than 1% of the human effort that has been used in past experiments.

Contributions. The primary contributions of this work are the elements of our models of experimental design and cost. These are:

1. Algorithms for acquiring relevance judgments to rank systems by common evaluation measures such as precision, recall, NDCG, and average precision.
2. An understanding of the space of hypotheses about rankings of systems through the idea of “confidence” and distributions of measures over the space of possible judgments.
3. Models for estimating the probabilities of relevance of unjudged documents, using available judgments as training data.

From these elements we can show through both theoretical and empirical analysis that:

4. The optimal experimental design for information retrieval consists of several hundred queries for which assessors make a few dozen judgments each.

[*] Author’s current affiliation: Department of Computer and Information Sciences, University of Delaware