Towards Good Evaluation of Individual Topics

Chris Buckley
Sabir Research, Inc.

Test collection evaluation in information retrieval has necessarily focused on comparing systems over reasonably large sets of topics and averaging results—there are too many system-topic interactions to rely on just a couple of topics. This dependency on large numbers of topics has allowed us to sweep several non-flattering truths under the rug; the most important one being that our standard test collections really do a poor job at evaluating system performance on individual topics. Some of the reasons and important consequences of this are examined, and suggestions for improving individual topic evaluation are presented.

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