Low Cost Evaluation in Information Retrieval

Ben Carterette
University of Delaware, carteret@cis.udel.edu

Evangelos Kanoulas
University of Sheffield, e.kanoulas@sheffield.ac.uk

Emine Yilmaz
Microsoft Research Cambridge, eminey@microsoft.com

Abstract:

Search corpora are growing larger and larger: over the last 10 years, the IR research community has moved from the several hundred thousand documents on the TREC disks to the tens of millions of U.S. government web pages of GOV2 to the one billion general-interest web pages in the new ClueWeb09 collection. But traditional means of acquiring relevance judgments and evaluating – e.g. pooling documents to calculate average precision – do not seem to scale well to these new large collections. They require substantially more cost in human assessments for the same reliability in evaluation; if the additional cost goes over the assessing budget, errors in evaluation are inevitable.

Some alternatives to pooling that support low-cost and reliable evaluation have recently been proposed. A number of them have already been used in TREC and other evaluation forums (TREC Million Query, Legal, Chemical, Web, Relevance Feedback Tracks, CLEF Patent IR, INEX). Evaluation via implicit user feedback (e.g. clicks) and crowdsourcing have also recently gained attention in the community. Thus it is important that the methodologies, the analysis they support, and their strengths and weaknesses are well-understood by the IR community. Furthermore, these approaches can support small research groups attempting to start investigating new tasks on new corpora with relatively low cost. Even groups that do not participate in TREC, CLEF, or other evaluation conferences can benefit from understanding how these methods work, how to use them, and what they mean as they build test collections for tasks they are interested in.

The goal of this tutorial is to provide attendees with a comprehensive overview of techniques to perform low cost (in terms of judgment effort) evaluation. A number of topics will be covered, including alternatives to pooling, evaluation measures robust to incomplete judgments, evaluating with no relevance judgments, statistical inference of evaluation metrics, inference of relevance judgments, query selection, techniques to test the reliability of the evaluation and reusability of the constructed collections.

The tutorial should be of interest to a wide range of attendees. Those new to the field will come away with a solid understanding of how low cost evaluation methods can be applied to construct inexpensive test collections and evaluate new IR technology, while those with intermediate knowledge will gain deeper insights and further understand the risks and gains of low cost evaluation. Attendees should have a basic knowledge of the traditional evaluation framework (Cranfield) and metrics (such as average precision and nDCG), along with some basic knowledge on probability theory and statistics. More advanced concepts will be explained during the tutorial.

SIGIR'10, July 19–23, 2010, Geneva, Switzerland.
ACM 978-1-60558-896-4/10/07.

ACM Categories & Descriptors:

H.3.4 Information Storage and Retrieval; Performance evaluation (efficiency and effectiveness)

General Terms:

Experimentation, Measurement

Keywords:

information retrieval, evaluation, test collections

Bios

Ben Carterette is an Assistant Professor of Computer and Information Sciences at the University of Delaware, where he teaches Information Retrieval, Databases, and Artificial Intelligence. He has published extensively on constructing and using test collections for low cost. He has co-organized two SIGIR workshops on test collections that go beyond binary independent relevance judgments, and co-coordinated the TREC Million Query track from 2007–2009 as well as the TREC Session track in 2010. He is also co-organizing a SIGIR workshop on constructing the next generation of information retrieval test collections.

Evangelos Kanoulas is a postdoctoral researcher (Marie Curie fellow) in the Department of Information Studies at the University of Sheffield. He received his Ph.D. from Northeastern University. His main research interest is evaluation methods for information retrieval. He has published papers in SIGIR and CIKM. Further, Evangelos was actively involved in coordinating the Million Query Track in TREC 2007 – 2009 and he is one of the co-coordinators of the TREC 2010 Session Track.

Emine Yilmaz is a postdoctoral researcher in the Information Retrieval and Analysis Group at Microsoft Research Cambridge. She obtained her Ph.D. from Northeastern University. Most of her current work involves evaluation of retrieval systems, the effect of evaluation metrics on learning to rank problems and modeling user behavior. Her main interests are information retrieval and applications of information theory, statistics and machine learning. She has previously published research papers at major information retrieval venues, including SIGIR, CIKM and served as one of the organizers of the ICTIR Conference in 2009. She is also one of the organizers of a SIGIR workshop on crowdsourcing.