Abstract

Experimental evaluation plays a critical role in driving progress in information retrieval (IR) today. Careful evaluation is necessary for advancing the state-of-the-art; yet, many published papers present work that was ill-evaluated. Furthermore, many papers are not accepted for publication due to non-satisfactory evaluation. Therefore, there is a strong need to educate students, researchers, and practitioners about the proper way to carry out IR experiments. This tutorial focuses on the methodologies that should be used to perform empirical evaluation, specifically, for ad hoc retrieval, of the highest scientific standards.

CVs of the presenters

Donald Metzler is a Senior Software Engineer at Google. Prior to that he was a Senior Research Scientist at Yahoo! and a Research Assistant Professor of Computer Science at the University of Southern California (USC). He obtained his Ph.D. from the University of Massachusetts. As an active member of the information retrieval, Web search, and natural language processing research communities, he has served on the senior program committees of SIGIR, CIKM, and WWW, and is a member of the Information Retrieval Journal editorial board. He has published over 40 research papers, has 2 patents granted, 14 patents pending, and is the author of "A Feature-Centric View of Information Retrieval" and co-author of "Search Engines: Information Retrieval in Practice".

Oren Kurland is a Senior Lecturer in the Faculty of Industrial Engineering and Management at the Technion --- Israel Institute of Technology. He obtained his Ph.D. in Computer Science from Cornell University. The information retrieval research group that Oren leads at the Technion focuses on developing formal models for information retrieval. Oren published more than 30 peer-reviewed papers in IR conferences and journals. He served as a senior program committee member for the SIGIR and CIKM conferences. He is also a member of the editorial board of the Information Retrieval Journal. Oren was awarded the IBM's, Google's and Yahoo's faculty research awards.

Summary

Intended audience: Graduate students, researchers from other disciplines, and industrial practitioners who are interested in experimentally validating the effectiveness of search engines in a rigorous, scientifically sound manner.
Experimental evaluation plays a critical role in driving progress in information retrieval (IR) today. There are few alternative ways of exploring, in depth, the empirical merits (or lack thereof) of newly devised search techniques. Careful evaluation is necessary for advancing the state-of-the-art; yet, many published papers present work that was ill-evaluated. Indeed, this phenomenon has garnered attention from the community recently, after the publication of a study by Armstrong et al. (2009) that suggested ad hoc search quality has not meaningfully advanced since 1984. The authors noted that the root of the problem was generally lax evaluation methodologies (e.g., weak baselines, etc.). Therefore, there is a strong need to educate students, researchers, and practitioners about the proper way to carry out IR experiments.

The primary goals of this tutorial are as follows:

1. Highlight the importance of experimental evaluation in IR. This will be accomplished by providing a brief explanation of the importance of experimental evaluations as from the early days of IR, and the ramifications of poorly executed experimental studies.

2. Provide attendees with an in-depth overview of the fundamental IR evaluation paradigm. The tutorial will explain all of the key ``tools'' (e.g., test collections, baselines, statistical significance testing, result analysis, etc.) that make up an ``experimental toolbox''. We will cover a number of case studies and provide specific examples of how the various ``tools'' can be used to evaluate new ad hoc search techniques.

3. More broadly, we hope that this tutorial will be an important first step towards developing a culture of strong experimental evaluations within the IR community. We will make our slides publicly available to help fill the gap in knowledge left by the fact the topic is largely ignored within IR courses and textbooks.

It is very important to note that this is not a tutorial about devising evaluation measures or building test collections. Those issues are orthogonal to the topics covered by this tutorial.