CHEERS: CHeap & Engineered Evaluation of Retrieval Systems

Kevin Roitero
University of Udine
Udine, Italy
roitero.kevin@spes.uniud.it

ABSTRACT

In test collection based evaluation of retrieval effectiveness, many research investigated different directions for an economical and a semi-automatic evaluation of retrieval systems. Although several methods have been proposed and experimentally evaluated, their accuracy seems still limited. In this paper we present our proposal for a more engineered approach to information retrieval evaluation.

CCS CONCEPTS

• **Information systems → Test collections**;

KEYWORDS

test collections, TREC, economical evaluation

ACM Reference Format:
Kevin Roitero. 2018. CHEERS: CHeap & Engineered Evaluation of Retrieval Systems. In *SIGIR '18: The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval, July 8–12, 2018, Ann Arbor, MI, USA.* ACM, New York, NY, USA, 1 page. https://doi.org/10.1145/3209978.3210229

1 INTRODUCTION AND BACKGROUND

A common approach to measure Information Retrieval (IR) effectiveness is to use test collections, where participants run their own systems over a set of topics, and then systems are ranked according to an evaluation metric. The whole evaluation process is rather expensive, in terms of both human time and money. Analyzing the literature, we notice that a large amount of work has been carried out, but there is still much to do; many works study how to reduce the evaluation cost, in different ways: to consider few topics [3, 5, 10, 12, 15, 17, 18, 20], to Evaluate the Effectiveness without expressing any Relevance Judgment (EEwRJ) [2, 9, 13, 14, 16, 19], to rely on Crowd-Sourced (CS) judgments [1, 4, 6, 7], to consider topic-system relationships [8, 11], etc.

2 RESEARCH METHODOLOGY

We propose to develop the research on IR evaluation, focusing on different dimensions: number of topics and systems, and pool depth. Concerning the former approach, we make use of a novel software to run experiments on larger datasets featuring many topics and systems; we optimize and study the generality of topic/system subsets, trying to define a practical approach to find subsets of a few good topics/best performing systems, by considering features that characterize such sets; we consider also to reduce the pool depth and the number of documents considered in the evaluation. This set of experiments will address as well the reproducibility and scalability of the test collections approach. A more general and novel approach, that we call Learning to Evaluate (LeToE), which features a seamless integration with the previous experiments, is based on the idea of using Machine Learning to evaluate IR systems in a TREC-like setting. The aim is to estimate system effectiveness, topic ease, and the metric values at a certain pool depth. We make use of SVMs, decision/boosted trees, (deep) neural networks, and other algorithms to exploit relationships between document ranked lists, system effectiveness, and topic ease. We consider also transfer learning between different collections/settings. Concerning the features, we consider artificial metric values that can be computed using EEwRJ and CS methods, real values up to a pool depth, for a small subset of topics, and ranked lists/systems/topic features.

SIGIR '18, July 8–12, 2018, Ann Arbor, MI, USA

ACM ISBN 978-1-4503-5657-2/18/07.
https://doi.org/10.1145/3209978.3210229

REFERENCES

[1] Omar Alonso and Stefano Mizzaro. 2009. Can we get rid of TREC assessors? Using Mechanical Turk for relevance assessment. In *Proc. of SIGIR 2009.*
[2] Javed A. Aslam and Robert Savell. 2003. On the Effectiveness of Evaluating Retrieval Systems in the Absence of Relevance Judgments. In *Proc. of SIGIR 2003.*
[3] Andrea Berto, Stefano Mizzaro, and Stephen Robertson. 2013. On using fewer topics in information retrieval evaluations. In *Proc. of ICTIR 2013.*
[4] A. Checco, K. Roitero, E. Maddalena, S. Mizzaro, and G. Demartini. 2017. Let's Agree to Disagree: Fixing Agreement Measures for Crowdsourcing. In *HCOMP2017.*
[5] John Guiver, Stefano Mizzaro, and Stephen Robertson. 2009. A few good topics: Experiments in topic set reduction for retrieval evaluation. *TOIS* (2009).
[6] Mehdi Hosseini, Ingemar J. Cox, Natasa Milic-Frayling, Gabriella Kazai, and Vishwa Vinay. 2012. On Aggregating Labels from Multiple Crowd Workers to Infer Relevance of Documents. In *Proc. of ECIR 2012.*
[7] Eddy Maddalena, Kevin Roitero, Gianluca Demartini, and Stefano Mizzaro. 2017. Considering Assessor Agreement in IR Evaluation. In *Proc. of ICTIR 2017.*
[8] Stefano Mizzaro and Stephen Robertson. 2007. Hits Hits TREC: Exploring IR Evaluation Results with Network Analysis. In *Proc. of SIGIR 2007.*
[9] Rabia Nuray and Fazli Can. 2006. Automatic ranking of information retrieval systems using data fusion. *Inf. Proc. & Management* 42, 3 (May 2006), 595–614.
[10] Stephen Robertson. 2011. On the Contributions of Topics to System Evaluation. In *Proc. of ECIR 2011.*
[11] Kevin Roitero, Eddy Maddalena, and Stefano Mizzaro. 2017. Do Easy Topics Predict Effectiveness Better Than Difficult Topics?. In *Proc. of ECIR 2017.*
[12] Tetsuya Sakai. 2016. Topic set size design. *IRJ* 19, 3 (2016), 256–283.
[13] Tetsuya Sakai and Chin-Yew Lin. 2010. Ranking Retrieval Systems without Relevance Assessments, Revisited. In *Proc. of EVIA 2010.*
[14] Ian Soboroff, Charles Nicholas, and Patrick Cahan. 2001. Ranking Retrieval Systems Without Relevance Judgments. In *Proc. of SIGIR 2001.*
[15] Karen Sparck Jones and Cornelis Joost van Rijsbergen. 1976. Information retrieval test collections. *Journal of documentation* 32, 1 (1976), 59–75.
[16] Anselm Spoerri. 2007. Using the structure of overlap between search results to rank retrieval systems without relevance judgments. *Inf. Proc. & Mgmt* (2007).
[17] Ellen M Voorhees. 2001. Evaluation by highly relevant documents. In *Proc. of SIGIR 2001.*
[18] William Webber, Alistair Moffat, and Justin Zobel. 2008. Statistical power in retrieval experimentation. In *Proc. of CIKM 2008.*
[19] S. Wu and F. Crestani. 2003. Methods for Ranking Information Retrieval Systems Without Relevance Judgments. In *Proc of Symposium on Applied Computing.*
[20] Justin Zobel. 1998. How reliable are the results of large-scale information retrieval experiments?. In *Proc. of SIGIR 1998.*