Inf Retrieval J (2016) 19:225–229
DOI 10.1007/s10791-016-9281-7

# Information retrieval evaluation using test collections

**Falk Scholer[1] · Diane Kelly[2] · Ben Carterette[3]**

Published online: 10 June 2016
© Springer Science+Business Media New York 2016

## 1 Introduction

Test collections are perhaps the most widely used tool for evaluating the effectiveness of information retrieval (IR) technologies. Test collections consist of a set of topics or information need descriptions, a set of information objects to be searched, and relevance judgments indicating which objects are relevant for which topics. Based on pioneering work carried out by Cyril Cleverdon and colleagues at Cranfield University in the 1960s (Cleverdon 1997), the popularity of test collections in IR evaluation has flourished in large part thanks to campaigns such as the Text Retrieval Conference[1] (TREC), the Cross-Language Evaluation Forum[2] (CLEF), the NII Testbeds and Community for Information Access Research project[3] (NTCIR), the Initiative for the Evaluation of XML Retrieval[4]

[1] http://trec.nist.gov.

[2] http://www.clef-initiative.eu.

[3] http://research.nii.ac.jp/ntcir.

[4] http://inex.mmci.uni-saarland.de.

✉ Diane Kelly
dianek@email.unc.edu

Falk Scholer
falk.scholer@rmit.edu.au

Ben Carterette
carteret@cis.udel.edu

[1] School of Computer Science and Information Technology, RMIT University, Melbourne, Australia

[2] School of Information and Library Science, University of North Carolina at Chapel Hill, Chapel Hill, NC, USA

[3] Department of Computer and Information Sciences, University of Delaware, Newark, DE, USA

(INEX), and the Forum for Information Retrieval Evaluation[5] (FIRE). In particular, TREC, which has ran since 1992, has generated and made available a number of test collections and enabled hundreds of groups from all over the world to participate in the development of next generation retrieval technologies (Voorhees and Harman 2005). These evaluation initiatives have also stimulated a substantial amount of research about evaluation methodology, including investigations of tasks and measures, topic set size design, the pooling process, assessor variability, the elicitation of relevance judgments, and the use of statistical significance testing.

Test collections have played a vital role in providing a basis for the measurement and comparison of the effectiveness of different information retrieval algorithms and techniques. However, test collections also present a number of issues, from being expensive and complex to construct, to instantiating a particular abstraction of the retrieval process. In his essay on the history of IR evaluation, Robertson (2008) attributes TREC with stimulating a "series of substantial advances in information retrieval techniques" (p. 452). He further notes several concerns about TREC, and more generally, the emphasis on the laboratory experiment and the use of test collections, which can create bias in what is studied, how it is studied and what is considered as a valid contribution to IR research. This is in part because laboratory experiments (and test collections) are abstractions where choices have been made about which real-world aspects should (and could) be represented and studied, and which need to be abstracted away to make the experiment tidier and more tractable. Such abstractions are necessary, but the widespread use of test collections and focus on results often means that researchers tend to forget about these abstractions. Consequently, the assumptions built-into the approach are reified through reuse and sometimes become mistaken for reality.

Questioning aspects of the IR evaluation paradigm can be uncomfortable and paralyzing; however, it is necessary if we want our evaluation methodologies to evolve and be more democratic and supportive of innovation and creativity. Thus, the purpose of this special issue of *Information Retrieval Journal* is to highlight research that questions and explores different aspects of test collection-based evaluation in information retrieval research.

## 2 Overview of papers

We received ten submissions for the special issue, of which four were accepted. Accepted papers address a number of issues including reproducibility and representativeness of test collections, topic set size design, methods for predicting relevance based on assessor disagreement and measures of test collection reliability and their robustness to statistical assumptions. A summary of each paper is presented below.

In *The Strange Case of Reproducibility versus Representativeness in Contextual Suggestion Test Collections*, Samar et al. (2016) examine issues of reproducibility and representativeness in the context of the TREC Contextual Suggestion Track, which aims to make recommendations to users according to their geographic location. *Reproducibility* is the extent to which the results of an evaluation can be regenerated under similar, or the same, conditions. Reproducibility is a key characteristic of empirical research and is considered desirable by most researchers, especially those conducting evaluations. One strength of test collection-based evaluations is that they are reproducible so long as the collection is static and the researchers have done an adequate job describing their systems,

[5] http://fire.irsi.res.in.

settings and research designs. Samar et al. (2016) consider *representativeness* as the extent to which the techniques developed by teams participating in the TREC Contextual Suggestion Track would work in commercial web search services.

In the TREC Contextual Suggestion Track, teams can recommend items from the open Web or from a collection of Web pages that have been crawled to facilitate evaluation, the ClueWeb12 collection. A consequence of using the open Web for retrieval is that results are not necessarily reproducible since the Web is always changing. However, results might be more representative of what actually happens during real-time search situations, thus introducing the reproducibility and representativeness dichotomy. Samar et al's (2016) work is in part based on the observation that systems using the open Web performed better than systems using ClueWeb12. Samar et al. investigate differences in relevance assessments given to documents from the open Web and ClueWeb12, including overlapping documents, and find that documents from the open Web are generally assigned higher relevance scores by assessors. They further identify a sample of ClueWeb12 documents that can potentially enhance the representativeness of techniques developed by researchers using the collection and propose a method that can be used to identify additional documents from ClueWeb12 in the future that allow for the development of more representative retrieval techniques.

In *Topic Set Size Design*, Sakai (2016) investigates the issue of how many topics should be selected for test collection-based evaluation. Previous analysis suggested that a minimum of 50 topics should be used to obtain stable effectiveness estimates, and this number has been used for many years as a heuristic to guide test collection construction, even when the context of the test collection (including the type of search task being evaluated, and the range of effectiveness metrics being used) varied. This work demonstrates how the required number of topics for a test collection can be analytically estimated based on formal sample size design techniques.

Three approaches based on the $t$ test, ANOVA, and confidence intervals are presented and compared. Since the main cost component in the construction of a new test collection is the creation of human relevance judgments, which is itself a function of the number of topics, a key benefit arising from this new approach is that it can assist researchers in estimating how many topics are required so that a test collection has adequate power for the identification of statistically significant differences between retrieval algorithms. A second important benefit of the formal size design approach is that it provides a robust framework for investigating the balance between the number of topics that are used in a test collection versus the pool depth to which relevance judgments are made. The analysis also demonstrates that different effectiveness metrics lead to substantially different within-system variances, and therefore topic set size estimates, highlighting the importance of choosing metrics that are appropriate for the search task being analyzed.

Demeester et al. (2016) address the issue of the generalizability of relevance assessments and how this impacts the reliability of retrieval results in their paper, *Predicting Relevance based on Assessor Disagreement: Analysis and Practical Applications for Search Evaluation*. When test collections are built, it is common for one, or a small number of people to make relevance assessments of the information objects that have been returned for topics. These assessments form the gold standard assessments on which all other evaluation measures are computed, such as mean average precision and discounted cumulated gain. It is well known that such assessments are not always generalizable; that is, different people often make different assessments of the same information objects for the same topic. However, this is accepted as one of the limitations of test collection-based evaluation.

Demeester et al. (2016) introduce the *Predicted Relevance Model (PRM)*, which predicts the relevance of a result for a random user, based on an observed assessment and knowledge of the average disagreement between assessors. The basic idea is that a greater degree of disagreement leads to a more uncertain prediction of relevance. One nice aspect of this model is that it recognizes for some topics, there is consensus amongst assessors, but for other topics, it is more difficult to predict what relevance value a person will assign to an information object. Another nice aspect of this model is that it allows for binary judgments to be transformed into graded relevance judgments, which addresses another shortcoming of many standard test collections and measures: the use of binary relevance. The model is evaluated using two test collections, which contain different types of search queries (e.g., informational and navigational) and different types of assessors.

In the final contribution to this special issue, *Measures of Test Collection Reliability and their Robustness to Statistical Assumptions*, Urbano (2016) compares different measures of test collection reliability with a focus on topic set size. These include ad-hoc measures such as Kendall's tau, average precision, and absolute and relative sensitivity, which do not consider the magnitude of the differences or variability, as well as measures that focus on the analysis of variance such as the F-statistic, the generalizability coefficient and the dependability index. Urbano's (2016) analysis is particularly concerned with the statistical assumptions behind the measures and the extent to which these are violated when using them to measure test collection reliability. To obtain the population performance parameters so that statistical assumptions could be tested, stochastic simulations are generated from four TREC collections. Two cases are then considered when examining the robustness of test collection reliability measures with respect to statistical assumptions: (1) where an IR researcher has a test collection with a certain number of topics (sample) and wants to estimate its reliability and (2) where an IR researcher has access to a test collection with a certain number of sample topics and wants to estimate the reliability of a new collection with a different set of sample topics from the same population of topics as the initial collection.

Urbano's (2016) results show that ad-hoc measures of test collection reliability tend to underestimate the reliability of collections, especially when the number of topics is small, and that measures from generalizability theory provide better estimates, although they also underestimate reliability for small collections. Urbano further found that all the test collection reliability measures were robust to the normality and homoscedasticity assumptions and showed slight instability with respect to the assumption regarding uncorrelated measures. Urbano (2016) provides all of the data and code used in this study online, which reflects a growing trend in IR research towards experimental transparency and reproducibility of results, furthering the rigor of IR experiments and researchers' abilities to scrutinize methods and results.

## 3 Summary

The use of standard test collections has helped distinguish the field of information retrieval within computer science as one with strong experimental rigor (Voorhees 2007). In IR, the tradition has been to not only build innovative technologies, but also demonstrate their value by comparing them to known baselines using agreed upon experimental tools and procedures. Robertson (2008) concludes his essay on the history of IR evaluation by observing that although we might be tempted to conclude that all the basic methodological

work in IR experimentation has already been done, there is still much left to do. This is especially true considering new technologies and retrieval contexts are constantly emerging. The papers in this special issue address a breadth of factors that impact the validity and reliability of information retrieval evaluation using test collections. We believe these papers reflect some of the major challenges facing test collection-based evaluation and present thoughtful proposals for moving forward.

**Acknowledgments** We thank all the authors who submitted papers and allowed us to review their work for this special issue. We also thank all those who reviewed papers for this special issue.

## References

Cleverdon, C. (1997). The Cranfield tests on index language devices. In K. SpärckJones & P. Willett (Eds.), *Readings in information retrieval*. San Francisco: Morgan Kaufmann.

Demeester, D., Aly, R., Hiemstra, D., Nguyen, D., & Develder, C. (2016). Predicting relevance based on assessor disagreement: Analysis and practical applications for search evaluation. *Information Retrieval Journal*. doi:10.1007/s10791-015-9275-x.

Robertson, S. (2008). On the history of evaluation in IR. *Journal of Information Science, 34*(4), 439–456.

Sakai, T. (2016). Topic set size design. *Information Retrieval Journal*. doi:10.1007/s10791-015-9273-z.

Samar, T., Bellogin, A., & de Vries, A. (2016). The strange case of reproducibility vs. representativeness in contextual suggestion test collections. *Information Retrieval Journal*. doi:10.1007/s10791-015-9276-9.

Urbano, J. (2016). Measures of test collection reliability and their robustness to statistical assumptions. *Information Retrieval Journal*. doi:10.1007/s10791-015-9274-y.

Voorhees, E. M. (2007). TREC: Continuing information retrieval's tradition of experimentation. *Communications of the ACM, 50*(11), 51–54.

Voorhees, E. M., & Harman, D. (2005). *TREC: Experiment and evaluation in information retrieval*. Cambridge, MA: MIT Press.