Introduction to the Special Issue on Reproducibility in Information Retrieval: Evaluation Campaigns, Collections, and Analyses

NICOLA FERRO, University of Padua, Italy
NORBERT FUHR, University of Duisburg-Essen, Germany
ANDREAS RAUBER, Technical University of Wien, Austria

CCS Concepts: • **Information systems → Evaluation of retrieval results**;

Additional Key Words and Phrases: reproducibility

ACM Reference Format:
Nicola Ferro, Norbert Fuhr, and Andreas Rauber. 2018. Introduction to the Special Issue on Reproducibility in Information Retrieval: Evaluation Campaigns, Collections, and Analyses. *ACM J. Data Inform. Quality* 10, 3, Article 1 (July 2018), 4 pages. https://doi.org/0000001.0000001

1 INTRODUCTION

Information Retrieval (IR) is a discipline that has been strongly rooted in experimentation since its inception. Experimental evaluation has always been a strong driver for IR research and innovation, and these activities have been shaped by large scale evaluation campaigns such as *Text REtrieval Conference (TREC)*[1] in the US, *Conference and Labs of the Evaluation Forum (CLEF)*[2] in Europe, *NII Testbeds and Community for Information access Research (NTCIR)*[3] in Japan and Asia, and *Forum for Information Retrieval Evaluation (FIRE)*[4] in India.

IR systems are getting increasingly complex. They need to cross language and media barriers; they span from unstructured, via semi-structured, to highly structured data; and they are faced with diverse. complex and frequently underspecified (ambiguously specified) information needs, search tasks, and societal challenges. As a consequence, evaluation and experimentation, which has remained a fundamental element, has in turn become increasingly sophisticated and challenging.

Replicability and *reproducibility* of the experimental results are becoming a primary concern in many areas of science [8, 12] and, in particular, in computer science as also witnessed by the recent ACM policy on *Artifact Review and Badging*[5].

Also the IR research community is increasingly focused on issues concerned with the replicability and reproducibility of the experimental results [1, 4, 5, 9, 11, 13]. We now commonly find questions

[1] https://trec.nist.gov/
[2] http://www.clef-initiative.eu/
[3] http://research.nii.ac.jp/ntcir/index-en.html
[4] http://fire.irsi.res.in/
[5] https://www.acm.org/publications/policies/artifact-review-badging

Authors' addresses: Nicola Ferro, University of Padua, Italy, ferro@dei.unipd.it; Norbert Fuhr, University of Duisburg-Essen, Germany, norbert.fuhr@uni-due.de; Andreas Rauber, Technical University of Wien, Austria, rauber@ifs.tuwien.ac.at.

1936-1955/2018/7-ART1 $15.00
https://doi.org/0000001.0000001

about the extent of reproducibility of the reported experiments in the review forms of all the major IR conferences, such as SIGIR, CHIIR, ICTIR and ECIR, as well as journals, such as ACM TOIS. We also witness the raise of new activities aimed at verifying the reproducibility of results: for example, the "Reproducibility Track" at ECIR since 2015 hosts papers which replicate, reproduce and/or generalize previous research results while CLEF/NTCIR/TREC REproducibility[6] (CENTRE) is a new joint evaluation activity, started in 2018, to assess and quantify the extent of replicability and reproducibility of our experimental results [7].

Nevertheless, it has been repeatedly shown that best TREC systems still outperform off-the-shelf open source systems [1–3, 10, 11]. This is due to many different factors, such as using default configuration instead of tuning on a specific collection, or lack of the specific and advanced components and resources adopted by the best systems. It has been also shown that additivity is an issue, since adding a component on top of a weak or strong base does not produce the same level of gain [3, 10]. This poses a serious challenge when off-the-shelf open source systems are used as stepping stone to test a new component on top of them, because the gain might appear bigger starting from a weak baseline.

Moreover, as it also emerged from a recent survey within the SIGIR community [6], while there is a very positive attitude towards reproducibility and it is considered very important from a scientific point of view, there are many obstacles to it such as the effort required to put it into practice, the lack of rewards for achieving it, the possible barriers for new and inexperienced groups, and, last but not least, the (somehow optimistic) researcher's perception that their own research is already reproducible.

Overall, the above considerations stress the need and urgency for a systematic approach to reproducibility in IR. Indeed, repeatability, reproducibility, and generalizability of experiments and results cannot be taken for granted. We need to emphasize these aspects as key requirements if we wish to continue to reliably and durably advance research and technology in the field. In turn, we need to actively pursue them as a core part of our experimental methodology and practice.

In this special issue of JDIQ, we aspire to provide an overview of innovative research at the intersection of information retrieval and data quality, from theory to practice, with a focus on challenges, solutions, and experiences in reproducibility of IR experimental results.

The special issue is split into two parts, each one containing 4 papers. The first part concerns evaluation campaigns, experimental collections, the way they are built, and the methodology we adopt to analyse the experimental results from the perspective of the challenges posed by reproducibility. The second part deals with tools and infrastructures to ease the reproducibility of experiments in IR.

Several of the articles included in this part of the special issue refer in one form or another to the organisation of evaluation campaigns, investigating possible improvements of the methods currently applied.

Moffat et al. in "Estimating Measurement Uncertainty for Information Retrieval Effectiveness Metrics" deals with the reliability of the pooling methodology and with the problem of the approximation introduced by not judged documents, showing how to estimate the measure uncertainty introduced by these factors in order to improve the reproducibility of experimental results.

Roitero et al. investigate in "Reproduce. Generalize. Extend. On Information Retrieval Evaluation without Relevance Judgments" various methods for reducing the effort for relevance assessments, as this is the most 'expensive' step in evaluation campaigns. They reproduce previous work in this area, analyze the effect of various parameters on the quality of this method and also generalize it to semi-automatic evaluation.

[6] http://www.centre-eval.org/

The paper "Reproduce and Improve: An Evolutionary Approach to Select a Few Good Topics for Information Retrieval Evaluation" by Roteiro et al. tackles another core issue in building experimental collections that is the selection of the topics used to evaluate systems. Usually a large number of topics is used and this then requires a great amount of effort to create the ground-truth. Roteiro et al. show how to select a smaller subset of topics in a more efficient way than before, opening the way for a wider adoption of this approach, and they reproduce and generalize previous findings in this area of research.

Finally, most evaluation campaigns are restricted to either system-oriented evaluations lacking any user-system interaction, or perform lab experiments with a limited number of users who solve predefined tasks. In contrast, living labs allow researchers for experimentation with real users of a live website. "OpenSearch: Lessons Learned from an Online Evaluation Campaign" by Jagerman et al. presents such platform along with the experiment results obtained.

Reviewers

This special issue would have not been possible without the help and careful work of many colleagues who helped to select and improve the papers presented here.

Ismail Altingövde, Middle East Technical University, Turkey

Ben Carterette, University of Delaware and Spotify, USA

Franca Debole, ISTI CNR Pisa, Italy

Julio Gonzalo, UNED Madrid, Spain

Donna Harman, National Institute of Standards and Technology (NIST), USA

Claudia Hauff, Technical University of Delft, The Netherlands

Yubin Kim , Carnegie Mellon University, USA

Udo Kruschwitz, University of Essex, UK

Aldo Lipani, Technical University of Vienna, Austria

Xiaolu Lu, RMIT University, Australia

Maria Maistro, University of Padua, Italy

Jian-Yun Nie , Université de Montréal, Canada

Heiko Schuldt, University of Basel, Switzerland

Ellen Voorhees, National Institute of Standards and Technology (NIST), USA

Arjen P. de Vries, Radboud University, The Netherlands

Guido Zuccon, Queensland University of Technology, Australia

REFERENCES

[1] J. Arguello, M. Crane, F. Diaz, J. Lin, and A. Trotman. 2015. Report on the SIGIR 2015 Workshop on Reproducibility, Inexplicability, and Generalizability of Results (RIGOR). *SIGIR Forum* 49, 2 (December 2015), 107–116.

[2] T. G. Armstrong, A. Moffat, W. Webber, and J. Zobel. 2009. Has Adhoc Retrieval Improved Since 1994?. In *Proc. 32nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2009)*, J. Allan, J. A. Aslam, M. Sanderson, C. Zhai, and J. Zobel (Eds.). ACM Press, New York, USA, 692–693.

[3] T. G. Armstrong, A. Moffat, W. Webber, and J. Zobel. 2009. Improvements That Don't Add Up: Ad-Hoc Retrieval Results Since 1998. In *Proc. 18th International Conference on Information and Knowledge Management (CIKM 2009)*, D. W.-L. Cheung, I.-Y. Song, W. W. Chu, X. Hu, and J. J. Lin (Eds.). ACM Press, New York, USA, 601–610.

[4] N. Ferro. 2017. Reproducibility Challenges in Information Retrieval Evaluation. *ACM Journal of Data and Information Quality (JDIQ)* 8, 2 (February 2017), 8:1–8:4.

[5] N. Ferro, N. Fuhr, K. Järvelin, N. Kando, M. Lippold, and J. Zobel. 2016. Increasing Reproducibility in IR: Findings from the Dagstuhl Seminar on "Reproducibility of Data-Oriented Experiments in e-Science". *SIGIR Forum* 50, 1 (June 2016), 68–82.

[6] N. Ferro and D. Kelly. 2018. SIGIR Initiative to Implement ACM Artifact Review and Badging. *SIGIR Forum* 52, 1 (June 2018), 4–10.

[7] N. Ferro, M. Maistro, T. Sakai, and I. Soboroff. 2018. Overview of CENTRE@CLEF 2018: a First Tale in the Systematic Reproducibility Realm. In *Experimental IR Meets Multilinguality, Multimodality, and Interaction. Proceedings of the Nineth International Conference of the CLEF Association (CLEF 2018)*, P. Bellot, C. Trabelsi, J. Mothe, F. Murtagh, J.-Y. Nie, L. Soulier, E. SanJuan, L. Cappellato, and N. Ferro (Eds.). Lecture Notes in Computer Science (LNCS) 11018, Springer, Heidelberg, Germany, 225–232.

[8] J. Freire, N. Fuhr, and A. Rauber (Eds.). 2016. *Report from Dagstuhl Seminar 16041: Reproducibility of Data-Oriented Experiments in e-Science*. Schloss Dagstuhl–Leibniz-Zentrum für Informatik, Germany.

[9] N. Fuhr. 2017. Some Common Mistakes In IR Evaluation, And How They Can Be Avoided. *SIGIR Forum* 51, 3 (December 2017), 32–41.

[10] S. Kharazmi, F. Scholer, D. Vallet, and M. Sanderson. 2016. Examining Additivity and Weak Baselines. *ACM Transactions on Information Systems (TOIS)* 34, 4 (June 2016), 23:1–23:18.

[11] J. Lin, M. Crane, A. Trotman, J. Callan, I. Chattopadhyaya, J. Foley, G. Ingersoll, C. Macdonald, and S. Vigna. 2016. Toward Reproducible Baselines: The Open-Source IR Reproducibility Challenge. In *Advances in Information Retrieval. Proc. 38th European Conference on IR Research (ECIR 2016)*, N. Ferro, F. Crestani, M.-F. Moens, J. Mothe, F. Silvestri, G. M. Di Nunzio, C. Hauff, and G. Silvello (Eds.). Lecture Notes in Computer Science (LNCS) 9626, Springer, Heidelberg, Germany, 357–368.

[12] M. R. Munafò, BB. A. Nosek, D. V. M. Bishop, K. S. Button, C. D. Chambers, N. Percie du Sert, U. Simonsohn, E.-J. Wagenmakers, J. J. Ware, and J. P. A. Ioannidis. 2017. A manifesto for reproducible science. *Nature Human Behaviour* 1 (January 2017), 0021:1–0021:9.

[13] J. Zobel, W. Webber, M. Sanderson, and A. Moffat. 2011. Principles for Robust Evaluation Infrastructure. In *Proc. Workshop on Data infrastructurEs for Supporting Information Retrieval Evaluation (DESIRE 2011)*, M. Agosti, N. Ferro, and C. Thanos (Eds.). ACM Press, New York, USA, 3–6.

Received 16 July 2018