Reproducible Online Search Experiments

Timo Breuer[1,2](✉)

[1] TH Köln (University of Applied Sciences), Cologne, Germany
timo.breuer@th-koeln.de
[2] Universität Duisburg-Essen, Duisburg, Germany

Abstract. In the empirical sciences, the evidence is commonly manifested by experimental results. However, very often, these findings are not reproducible, hindering scientific progress. Innovations in the field of information retrieval (IR) are mainly driven by experimental results as well. While there are several attempts to assure the reproducibility of offline experiments with standardized test collections, reproducible outcomes of online experiments remain an open issue. This research project will be concerned with the reproducibility of online experiments, including real-world user feedback. In contrast to previous living lab attempts by the IR community, this project has a stronger focus on making IR systems and corresponding results reproducible. The project aims to provide insights concerning key components that affect reproducibility in online search experiments. Outcomes help to improve the design of reproducible IR online experiments in the future.

Keywords: Reproducibility · Online evaluation · Living lab

1 Motivation

Reproducible findings are fundamental for scientific progress and validity. In 2016, a Nature survey [2] revealed that lack of reproducibility nearly affects all scientific disciplines and can be considered as a general concern. Non-reproducible results limit the trustworthiness of publications and hinder progress. Besides investigating various reasons for non-reproducibility, the study showed that scientists mostly agree upon the importance of the problem that became known as *reproducibility crisis* during the last years. Especially in the field of information retrieval (IR), new findings are manifested by empirical studies and experiments. Innovations are assumed to be valid if their results are superior compared to those of previous findings. Despite this intuitive but rather naive assumption, achieving reproducibility in the field of IR is a many-faceted problem. For instance, the meta-evaluation by Armstrong et al. [1] reveal the illusory progress of ad-hoc retrieval performance over an entire decade, caused by comparisons to weak baselines. Ten years later, Yang et al. [16] report similar results as part of their meta-evaluation. The lacking upwards trend in retrieval

J. M. Jose et al. (Eds.): ECIR 2020, LNCS 12036, pp. 597–601, 2020.
https://doi.org/10.1007/978-3-030-45442-5_77

performance can be traced back to non-reproducible findings. If baselines of previous results are not or only laboriously reproducible, the community does not use them adequately.

We see a gap between reproducibility efforts for offline evaluations on the one side and online retrieval experiments trying to include real-world user interactions on the other side. While several initiatives are trying to establish reproducible IR research for offline evaluations on standard test collections, there is little research effort concerning the reproducibility of online experiments. This dissertation project will be concerned with the reproducibility of online experiments in the field of information retrieval.

2 Related Work

Progress in information retrieval revolves around the evaluation of experimental results. This research project will focus specifically on two aspects of evaluation in IR - reproducible experiments and the living lab paradigm. This section gives a brief overview of the two evaluation branches.

As mentioned in the previous section, meta-evaluations of IR systems revealed limited progress over the years [1,16]. During the last years, the IR community tried to tackle this problem with several attempts concerned with reproducibility. These can be broadly categorized into attempts on a conceptual level and initiatives in the form of workshops, infrastructures, and frameworks. Conceptually, Ferro and Kelly elaborate an implementation for the field of information retrieval [10] of the ACM Artifact and Review Badging[1]. The PRIMAD model [8] offers orientation which components of an IR experiment may affect reproducibility or have to be considered when trying to reproduce the corresponding experiment. The Evaluation-as-a-Service (EaaS) paradigm [13] reverses the conventional evaluation approach of a shared task like it is applied at the TREC conference. Instead of letting participants submit the results (runs) only, the complete retrieval system is submitted in a form such that it can be rerun independently by others to produce the results. Workshops deal with the reproducibility either re- or proactively. For example, the CENTRE workshop [9] challenges participants to reconstruct IR systems and their results, whereas The Open-Source IR Replicability Challenge (OSIRRC) [7] motivated participants to package their retrieval systems and corresponding software dependencies in advance to prepare them for appropriate reuse.

Compared to offline ad-hoc retrieval, online search experiments are affected by non-deterministic variables including user behavior, updated data collections, modifications of web interfaces, or traffic dependencies [11]. Balog et al. introduced the first living lab campaign in 2014 [3]. The infrastructure found application in several workshops and intiviates at the CLEF and TREC conferences from 2015 to 2017 [14]. Despite these elegant solutions for implementing living lab infrastructures, the aspect of reproducibility remained neglected, e.g., there was no specification of how the experiments could be archived for later use [12].

[1] https://www.acm.org/publications/policies/artifact-review-badging.

On the other hand, research efforts towards reproducible IR experiments have a strong focus on ad-hoc retrieval experiments and do not include any insights beyond offline environments at the time of writing.

3 Preliminary Work and Research Proposal

Preliminary Work. We participated in the CENTRE@CLEF2019 workshop dedicated to the replicability, reproducibility, and generalizability of ad-hoc retrieval experiments [5]. The workshop's organizers challenge the participants to reconstruct results of previous submissions to the CLEF, NTCIR, and TREC conferences. CENTRE defines replicability and reproducibility by using the same or another test collection of the original setup, respectively. The results of our experimental setups showed that we can replicate the outcomes fairly well, whereas reproduced outcomes are significantly lower. Having the reimplementation of an ad-hoc retrieval system at hand, we decided to contribute it to the OSIRRC@SIGIR2019 workshop [7]. All contributions resulted in an image library of Docker images to which we contributed the IRC-CENTRE2019 image [4]. Additionally, we introduced STELLA - a new interpretation of the living lab paradigm - at the OSIRRC workshop [6]. We propose to transfer the idea of encapsulating retrieval systems with Docker containers to the online search scenario. In order to underline the feasibility and benefits of this proposal, we aligned components of the STELLA framework to the PRIMAD model.

Based on this preliminary work, we investigate the reproducibility of retrieval systems with the main focus on online environments. In the following, we present the research questions of this project.

RQ1 - How is the ACM terminology of repeatability, replicability, and reproducibility applied to online search experiments? While the ACM definitions can be implemented for offline ad-hoc experiments, an analogy for the online case is less obvious. Did previous online search experiments consider reproducibility? If so, is it possible to go a step further and align them with the PRIMAD model?

RQ2 - How can simulations of search sessions, based on user logs, help to identify key components of reproducible retrieval performance? To what extent affect the identified components of an online experiment the reproducibility? Compared to offline experiments, the user of the search engine is a key component in the online case. User logs comprise implicit feedback such that they can be used to model the user component of an experiment. What influence do the user and other session-related components have on the reproducibility?

RQ3 - What requirements must a living lab infrastructure meet in order to guarantee reproducible online search experiments? By identifying key components that affect the reproducibility of online search experiments, we gain insights about the requirements for reproducible online search experiments. What kind of practical steps have to be considered when implementing a framework for reproducible retrieval experiments in production environments?

4 Methodology and Experiments

Addressing *RQ1*, we want to conduct a literature survey and evaluate how previous living lab approaches and online experiments paid attention to the topic of reproducibility. Since there exist different terminologies, we use the ACM definitions of repeatability, replicability, and reproducibility as a starting point. As a result, we do not only want to give an overview of how existing literature paid attention to these concepts, but also provide an ontology that is inspired by the PRIMAD model [8]. While the ACM terminology is defined by the two experimental components of the research team and setup, the PRIMAD model conceptualizes the experiment on a more granular level. More specifically, it pays attention to the platform, research goal, implementation, method, actor, and data. This point of view is mainly data-focused and applies well to the offline ad-hoc experiment. However, it could be extended such that it also considers the actual user of a retrieval system.

Regarding *RQ2*, we primarily focus on vertical search experiments. As a result, we want to provide insights concerning key components that affect the reproducibility of online search experiments. Beforehand, the reusability of user logs is particularly interesting, since reusable test collections are fundamental to offline retrieval experiments. Tan et al. [15] examine the reusability of user judgments that contributed to a relevance pool by performing a *leave-one-out* analysis. As a starting point, we propose to repeat this study with the user logs of another search engine. Assuming we have retrieved a fair amount of interaction logs that deliver relevance feedback in the form of clicks and other interactions [11], we systematically assess the influence of specific components. For instance, we can simulate sessions with different durations, tasks, or users. By comparing a diverse set of different session constellations and corresponding outcomes, we identify significant influences. Are specific components more important than others or even crucial for successful reproduction? Furthermore, it is of interest to relate to previous offline reproducibility efforts. Consider two rankers A and B, that are compared by the conventional offline ad-hoc experiment. The retrieval effectiveness of A outperforms that of B, which is denoted as $A \succ B$ and is confirmed to be reproducible. Under which circumstances and to which extent can $A \succ B$ be reproduced in an online environment? Which components affect the reproducibility?

Having identified major influences and key components, we address *RQ3* by deriving requirements that have to be met by an adequate living lab infrastructure. On a functional level, technical components of the infrastructure have to be included. Quality requirements play an essential role, as well. Since experimental systems will be deployed in production environments, a certain degree of quality has to be guaranteed. Subpar retrieval performance and latencies caused by long query processing may affect user behavior, and at the worst, damage the reputation of the sites. Furthermore, we have to consider general conditions like the ethical and juridical aspects of data logging. On an organizational level, it has to be specified, which prerequisites an embedded search engine provider has to fulfill.

References

1. Armstrong, T.G., Moffat, A., Webber, W., Zobel, J.: Improvements that don't add up: ad-hoc retrieval results since 1998. In: Proceedings of CIKM, pp. 601–610 (2009)
2. Baker, M.: 1,500 scientists lift the lid on reproducibility. Nature **533**, 452–454 (2016)
3. Balog, K., Kelly, L., Schuth, A.: Head first: living labs for ad-hoc search evaluation. In: Proceedings of CIKM, pp. 1815–1818 (2014)
4. Breuer, T., Schaer, P.: Dockerizing automatic routing runs for the open-source IR replicability challenge (osirrc 2019). In: Proceedings of the Open-Source IR Replicability Challenge (OSIRRC) @ SIGIR (2019)
5. Breuer, T., Schaer, P.: Replicability and reproducibility of automatic routing runs. In: Working Notes of CLEF. CEUR Workshop Proceedings (2019)
6. Breuer, T., Schaer, P., Tavalkolpoursaleh, N., Schaible, J., Wolff, B., Müller, B.: STELLA: towards a framework for the reproducibility of online search experiments. In: Proceedings of the Open-Source IR Replicability Challenge (OSIRRC) @ SIGIR (2019)
7. Clancy, R., Ferro, N., Hauff, C., Lin, J., Sakai, T., Wu, Z.Z.: The SIGIR 2019 open-source ir replicability challenge (OSIRRC 2019). In: Proceedings of SIGIR, pp. 1432–1434 (2019)
8. Ferro, N., Fuhr, N., Järvelin, K., Kando, N., Lippold, M., Zobel, J.: Increasing reproducibility in IR: findings from the dagstuhl seminar on "Reproducibility of Data-Oriented Experiments in e-Science". SIGIR Forum **50**, 68–82 (2016)
9. Ferro, N., Fuhr, N., Maistro, M., Sakai, T., Soboroff, I.: CENTRE@CLEF2019: overview of the replicability and reproducibility tasks. In: Working Notes of CLEF (2019)
10. Ferro, N., Kelly, D.: SIGIR initiative to implement ACM artifact review and badging. SIGIR Forum **52**, 4–10 (2018)
11. Hofmann, K., Li, L., Radlinski, F.: Online evaluation for information retrieval. Found. Trends Inf. Retrieval **10**, 1–117 (2016)
12. Hopfgartner, F., et al.: Continuous evaluation of large-scale information access systems: a case for living labs. In: Information Retrieval Evaluation in a Changing World - Lessons Learned from 20 Years of CLEF, pp. 511–543 (2019)
13. Hopfgartner, F., et al.: Evaluation-as-a-service for the computational sciences: overview and outlook. J. Data Inf. Qual. **10**, 15:1–15:32 (2018)
14. Jagerman, R., Balog, K., de Rijke, M.: OpenSearch: lessons learned from an online evaluation campaign. J. Data Inf. Qual. **1**, 13:1–13:15 (2018)
15. Tan, L., Baruah, G., Lin, J.: On the reusability of "Living Labs" test collections: a case study of real-time summarization. In: Proceedings of SIGIR, pp. 793–796 (2017)
16. Yang, W., Lu, K., Yang, P., Lin, J.: Critically examining the "Neural Hype": weak baselines and the additivity of effectiveness gains from neural ranking models. In: Proceedings of SIGIR, pp. 1129–1132 (2019)