The SIGIR 2019 open-source IR replicability challenge (OSIRRC 2019)

Clancy, Ryan; Ferro, Nicola; Hauff, Claudia; Lin, Jimmy; Sakai, Tetsuya; Wu, Ze Zhong

DOI
10.1145/3331184.3331647

Publication date
2019

Document Version
Final published version

Published in
SIGIR 2019 - Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval

The SIGIR 2019 Open-Source IR Replicability Challenge (OSIRRC 2019)

Ryan Clancy,[1] Nicola Ferro,[2] Claudia Hauff,[3] Jimmy Lin,[1] Tetsuya Sakai,[4] Ze Zhong Wu[1]
[1] University of Waterloo [2] University of Padua [3] TU Delft [4] Waseda University

1 INTRODUCTION

The importance of repeatability, replicability, and reproducibility is broadly recognized in the computational sciences, both in supporting desirable scientific methodology as well as sustaining empirical progress. In order to precisely articulate the goals of this workshop, it is first necessary to establish common terminology. We use the above terms in the same manner as recent ACM guidelines pertaining to artifact review and badging:[1]

- *Repeatability* (same team, same experimental setup): a researcher can reliably repeat her own computation.
- *Replicability* (different team, same experimental setup): an independent group can obtain the same result using the authors' own artifacts.
- *Reproducibility* (different team, different experimental setup): an independent group can obtain the same result using artifacts which they develop completely independently.

This workshop tackles the replicability challenge for *ad hoc* document retrieval, with three explicit goals:

(1) Develop a common Docker interface specification to support images that capture systems performing *ad hoc* retrieval experiments on standard test collections. The solution that we have developed is known as "the jig".
(2) Build a curated library of Docker images that work with the jig to capture a diversity of systems and retrieval models.
(3) Explore the possibility of broadening our efforts to include additional tasks, evaluation methodologies, and benchmark initiatives.

Trivially, by supporting replicability, our proposed solution enables repeatability as well (which, as a recent case study has shown [13], is not as easy as one might imagine). It is *not* our goal to directly address reproducibility, although we do see our efforts as an important stepping stone.

We hope that the fruits of this workshop can fuel empirical progress in *ad hoc* retrieval by providing competitive baselines that are easily replicable. The "prototypical" research paper of this mold proposes an innovation and demonstrates its value by comparing against one or more baselines. The often-cited meta-analysis of Armstrong et al. [2] from a decade ago showed that researchers compare against weak baselines, and a recent study by Yang et al. [12] revealed that, a decade later, the situation has not improved much—researchers are *still* comparing against weak baselines. Lin [8] discussed social aspects of why this persists, but there are genuine technical barriers as well. The growing complexity of modern retrieval techniques, especially neural models that are sensitive to hyperparameters and other minor aspects of the training regime, poses challenges for researchers who wish to demonstrate that their proposed innovation improves upon a particular method. In contrast to NLP, for instance, where state-of-the-art results are often copied directly from published papers or public leaderboards, in IR greater emphasis is placed on in-depth comparisons between existing and proposed approaches, thus requiring access to actual result runs. Solutions that address replicability would greatly simplify such comparisons.

2 BACKGROUND

There has been much discussion about reproducibility in the sciences, with most scientists agreeing that the situation can be characterized as a crisis [3]. We lack the space to provide a comprehensive review of relevant literature in the medical, natural, and behavioral sciences. Even within the computational sciences to which at least a large portion of IR belongs, there have been many studies and proposed solutions. Here, we focus on summarizing the immediate predecessor of this workshop.

Our workshop was conceived as the next iteration of the Open-Source IR Reproducibility Challenge (OSIRRC), organized as part of the SIGIR 2015 Workshop on Reproducibility, Inexplicability, and Generalizability of Results (RIGOR) [1]. This event in turn traces its roots back to a series of workshops focused on open-source IR systems, which is widely understood as an important component of reproducibility. The Open-Source IR Reproducibility Challenge[2] brought together developers of open-source search engines to provide replicable baselines of their systems in a common environment on Amazon EC2. The product is a repository that contains all code necessary to generate *ad hoc* retrieval baselines, such that with a single script, anyone with a copy of the collection can replicate the submitted runs. Developers from seven different systems contributed to the evaluation, which was conducted on the GOV2 collection. The details of their experience are captured in an ECIR 2016 paper [9].

In OSIRRC 2019, we aim to address two shortcomings with the previous exercise as a concrete step in moving the field forward. From the technical perspective, the RIGOR 2015 participants developed scripts in a shared VM environment, and while this was sufficient to support cross-system comparisons at the time, the scripts were not sufficiently constrained, and the entire setup suffered from portability and isolation issues. Thus, it would have

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

[2] Note that the exercise is more accurately characterized as replicability and not reproducibility; the event predated ACM's standardization of terminology.

SIGIR '19, July 21–25, 2019, Paris, France

ACM ISBN 978-1-4503-6172-9/19/07.
https://doi.org/10.1145/3331184.3331647

been difficult for others to reuse the infrastructure to replicate the results—in other words, the replicability experiments themselves were difficult to replicate. We believe that Docker, which is a popular standard for containerization, offers a potential solution to these technical challenges.

Another limitation of the previous exercise was its focus on "bag of words" baselines, and while some participants did submit systems that exploited richer models (e.g., term dependence models and pseudo-relevance feedback), there was insufficient diversity in the retrieval models that were examined. Primarily due to these two issues, the exercise has received less follow-up and uptake than the organizers had hoped.

3 DOCKER AND "THE JIG"

From a technical perspective, our efforts are built around Docker, a widely-adopted Linux-centric technology for delivering software in lightweight packages called containers. The Docker Engine hosts one or more of these containers on physical machines and manages their lifecycle. One key feature of Docker is that all containers run on a single operating system kernel; isolation is handled by Linux kernel features such as cgroups and kernel namespaces. This makes containers far more lightweight than virtual machines, and hence easier to manipulate. Containers are created from *images*, which are typically built by importing base images (for example, capturing a specific software distribution) and then overlaying custom code. The images themselves can be manipulated, combined, and modified as first-class citizens in a broad ecosystem. For example, a group can overlay several existing images from public sources, add in its own code, and in turn publish the resulting image to be further used by others.

As defined by the Merriam-Webster dictionary, a jig is "a device used to maintain mechanically the correct positional relationship between a piece of work and the tool or between parts of work during assembly". The central activity of this workshop revolves around the co-design and co-implementation of a jig and Docker images that work with the jig for *ad hoc* retrieval. Of course, in our context, the relationship is computational instead of mechanical.

Shortly after the acceptance of the workshop proposal, we issued a call for participants who were interested in contributing Docker images to our effort; the jig was designed with the input of these participants. In other words, the jig and the images co-evolved with feedback from members of the community. The code of the jig is open source and available on GitHub.[3]

Our central idea is that each image would expose a number of "hooks" that correspond to a point in the prototypical lifecycle of an *ad hoc* retrieval experiment: for example, indexing a collection, running a batch of queries, etc. In our current specification, each hook corresponds to a script in the image that has a specific name and resides at a fixed location. Each script can invoke its own interpreter: common implementations include bash and Python. These scripts then tie into code that captures whatever retrieval model a particular researcher wishes to encapsulate in the image—for example, a search engine implemented in Java or C++.

Note that by design the current jig does not make any demands about the transparency of a particular image. For example, the search hook can run an executable whose source code is not publicly available. Such an image, while demonstrating replicability, would not allow other researchers to inspect the inner workings of a particular retrieval method. While such images are not forbidden in our design, they are obviously less desirable than images based on open code. In practice, however, we anticipate that most images will be based on open-source code.

The jig is responsible for triggering the hooks in each image in a particular sequence according to a predefined lifecycle model, e.g., first index the collection, then run a batch of queries, finally evaluate the results. We have further built tooling that applies the jig to multiple images, aggregates results from each, and performs various analyses.

One technical design choice that we have grappled with is how to get data "into" and "out of" a container. To be more concrete, for *ad hoc* retrieval the container needs access to the document collection and also the topics. The jig also needs to be able to obtain the run files generated by the image for evaluation. Generically, there are three options for feeding data to an image: first, the data can be part of the image itself; second, the data can be fetched from a remote location by the image (e.g., via `curl`, `wget`, or some other network transfer mechanism); third, the jig could mount an external data directory that the container has access to. The first two approaches are problematic for our use case: images need to be shareable, or resources need to be placed at a publicly-accessible location online. This is not permissible for document collections where researchers are required to sign license agreements before using. Furthermore, both approaches do not allow the possibility of testing on blind held-out data.

We ultimately opted for the third approach: the jig mounts a (read-only) data directory that makes the document collection available at a known location, as part of the contract between the jig and the image (and similarly for topics). A separate directory that is writable serves as the mechanism for the jig to gather output runs from the image for evaluation. This method makes it possible for images to be tested on blind held-out documents and topics, as long as the formats have been agreed to in advance.

Note that we have been intentionally vague in our description of the jig because it is a work in progress and constantly evolving as we gather more image contributions. The above description provides a broad overview that is likely to remain accurate, although specific details will inevitably evolve over time. Our plan is to version the jig as one would any other piece of software, and periodically declare stable versions of the specification for deployment. We invite interested readers to consult our code repository for the latest updates.

4 FUTURE VISION AND ONGOING WORK

Our efforts complement other concurrent activities in the community. SIGIR has established a task force to implement ACM's policy on artifact review and badging [5], and our efforts can be viewed as a technical feasibility study.

This workshop also complements the recent CENTRE[4] evaluation tasks jointly run at CLEF, NTCIR, and TREC [6, 10]. One of

[3] https://github.com/osirrc/jig

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

the goals of CENTRE is to define appropriate measures to determine whether and to what extent replicability and reproducibility have been achieved, while our efforts focus on how these properties can be demonstrated technically. Thus, the jig can provide the means to achieve CENTRE goals. Given fortuitous alignment in schedules, this collaboration has already begun: participants of CENTRE@CLEF2019 [4] have explicitly been encouraged to participate in our workshop. We also have ongoing discussions with organizers of tracks in TREC 2019 to adopt the jig (or some variant thereof) as one mechanism for submitting results (or more accurately, delivering the code by which results can be generated). One attractive property of submissions based on Docker images is the possibility of evaluating on blind held-out data.

We have proposed and prototyped a technical solution to the replicability challenge specifically for the SIGIR community, but the changes we envision will not occur without a corresponding cultural shift. Sustained, cumulative empirical progress will only be made if researchers use our tools in their evaluations, and this will only be possible if images for the comparison conditions are available. This means that the community needs to adopt the norm of associating research papers with source code for replicating results in those papers. However, as Voorhees et al. [11] reported recently, having a link to a repository in a paper is far from sufficient. The jig provides the tools to "wrap" *ad hoc* retrieval experiments in a standard way, but these tools are useless without broad adoption. The incentive structures of academic publishing need to adapt to encourage such behavior, but unfortunately this is beyond the scope of our workshop.

Although there remain technical details to iron out, we believe that the jig with proper extensions can accommodate a range of batch retrieval tasks. One important future direction is to build extensions that would enable tasks beyond batch retrieval, for example, to support interactive retrieval (with real or simulated user input) and evaluation on private and other sensitive data. Moreover, our effort represents a first systematic attempt to embody the Evaluation-as-a-Service paradigm [7] via Docker containers. We believe that there are many possible paths forward building on the ideas presented here.

Finally, we view our efforts as a stepping stone toward reproducibility, and beyond that, generalizability. While these two important desiderata are not explicit goals of our workshop, we note that the jig itself can provide the technical vehicle for delivering reproducibility and generalizability. In this workshop, we are assuming that the authors of a particular retrieval method contribute the image. However, there is nothing that would prevent researchers from reproducing another team's results, that is then captured in a Docker image conforming to our specifications. This would demonstrate reproducibility as well as replicability of those reproducibility efforts. The jig also supports mechanisms for evaluations on document collections and information needs beyond those that an image was originally designed for. This aligns with intuitive notions of what it means for a technique to be generalizable.

Overall, we believe that our efforts have moved the field of information retrieval forward both in terms of supporting "good science" as well as sustained, cumulative empirical progress. We look forward to responses from the community that will help further advance these worthy goals!

REFERENCES

[1] Jaime Arguello, Matt Crane, Fernando Diaz, Jimmy Lin, and Andrew Trotman. 2015. Report on the SIGIR 2015 Workshop on Reproducibility, Inexplicability, and Generalizability of Results (RIGOR). *SIGIR Forum* 49, 2 (2015), 107–116.

[2] Timothy G. Armstrong, Alistair Moffat, William Webber, and Justin Zobel. 2009. Improvements That Don't Add Up: Ad-Hoc Retrieval Results Since 1998. In *Proceedings of the 18th International Conference on Information and Knowledge Management (CIKM 2009)*. Hong Kong, China, 601–610.

[3] Monya Baker. 2016. Is There a Reproducibility Crisis? *Nature* 533 (2016), 452–454.

[4] Nicola Ferro, Norbert Fuhr, Maria Maistro, Tetsuya Sakai, and Ian Soboroff. 2019. Overview of CENTRE@CLEF 2019. In *Proceedings of the Tenth International Conference of the CLEF Association (CLEF 2019)*.

[5] Nicola Ferro and Diane Kelly. 2018. SIGIR Initiative to Implement ACM Artifact Review and Badging. *SIGIR Forum* 52, 1 (2018), 4–10.

[6] Nicola Ferro, Maria Maistro, Tetsuya Sakai, and Ian Soboroff. 2018. Overview of CENTRE@CLEF 2018: A First Tale in the Systematic Reproducibility Realm. In *Proceedings of the Ninth International Conference of the CLEF Association (CLEF 2018)*. Avignon, France, 239–246.

[7] Frank Hopfgartner, Allan Hanbury, Henning Müller, Ivan Eggel, Krisztian Balog, Torbben Brodt, Gordon V. Cormack, Jimmy Lin, Jayashree Kalpathy-Cramer, Noriko Kando, Makoto P. Kato, Anastasia Krithara, Tim Gollub, Martin Potthast, Evelyne Viegas, and Simon Mercer. 2018. Evaluation-as-a-Service for the Computational Sciences: Overview and Outlook. *ACM Journal of Data and Information Quality (JDIQ)* 10, 4 (November 2018), 15:1–15:32.

[8] Jimmy Lin. 2018. The Neural Hype and Comparisons Against Weak Baselines. *SIGIR Forum* 52, 2 (2018), 40–51.

[9] Jimmy Lin, Matt Crane, Andrew Trotman, Jamie Callan, Ishan Chattopadhyaya, John Foley, Grant Ingersoll, Craig Macdonald, and Sebastiano Vigna. 2016. Toward Reproducible Baselines: The Open-Source IR Reproducibility Challenge. In *Proceedings of the 38th European Conference on Information Retrieval (ECIR 2016)*. Padua, Italy, 408–420.

[10] Tetsuya Sakai, Nicola Ferro, Ian Soboroff, Zhaohao Zeng, Peng Xiao, and Maria Maistro. 2019. Overview of the NTCIR-14 CENTRE Task. In *Proceedings of the 14th NTCIR Conference on Evaluation of Information Access Technologies*. Tokyo, Japan.

[11] Ellen M. Voorhees, Shahzad Rajput, and Ian Soboroff. 2016. Promoting Repeatability Through Open Runs. In *Proceedings of the 7th International Workshop on Evaluating Information Access (EVIA 2016)*. Tokyo, Japan, 17–20.

[12] Wei Yang, Kuang Lu, Peilin Yang, and Jimmy Lin. 2019. Critically Examining the "Neural Hype": Weak Baselines and the Additivity of Effectiveness Gains from Neural Ranking Models. In *Proceedings of the 42nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2019)*. Paris, France.

[13] Ruifan Yu, Yuhao Xie, and Jimmy Lin. 2018. H_2oloo at TREC 2018: Cross-Collection Relevance Transfer for the Common Core Track. In *Proceedings of the Twenty-Seventh Text REtrieval Conference (TREC 2018)*. Gaithersburg, Maryland.