Reproducibility Challenges in Information Retrieval Evaluation

NICOLA FERRO, University of Padua

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

Additional Key Words and Phrases: reproducibility

ACM Reference Format:
Nicola Ferro, 2016.Reproducibility Challenges in Information Retrieval Evaluation *ACM J. Data Inform. Quality* XX, YY, Article ZZ (November 2016), 4 pages.
DOI: 0000001.0000001

1. INTRODUCTION

Information Retrieval (IR) is concerned with *ranking* information resources with respect to user information needs, delivering a wide range of key applications for industry and society, like Web search engines [Croft et al. 2009], intellectual property and patent search [Lupu and Hanbury 2013], and many others.

Performances of IR systems are determined not only by their efficiency but also and most importantly by their *effectiveness*, i.e. their ability to retrieve and better rank relevant information resources while at the same time suppressing the retrieval of not relevant ones. Due to the many sources of uncertainty, as for example vague user information needs, unstructured information sources, or subjective notion of relevance, *experimental evaluation* is the only mean to assess the performances of IR systems from the effectiveness point of view. Experimental evaluation relies on the Cranfield paradigm which makes use of *experimental collections*, consisting of documents, sampled from a real domain of interest; topics, representing real user information needs in that domain; and, relevance judgements, determining which documents are relevant to which topics [Harman 2011].

To share the effort and optimize the use of resources, experimental evaluation is usually carried out in publicly open and large-scale evaluation campaigns at international level, like the *Text REtrieval Conference (TREC)*[1] in the United States [Harman and Voorhees 2005], the *Conference and Labs of the Evaluation Forum (CLEF)*[2] in Europe [Ferro 2014], the *NII Testbeds and Community for Information access Research (NTCIR)*[3] in Japan and Asia, and the *Forum for Information Retrieval Evaluation (FIRE)*[4] in India. These initiatives produce, every year, huge amounts of scientific data which are extremely valuable since they are the foundations for all the subsequent scientific production and system development [Ferro et al. 2011].

[1] http://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/

Author's addresses: Nicola Ferro, Department of Information Engineering, University of Padua, Italy; email: ferro@dei.unipd.it.

DOI: 0000001.0000001

Reproducibility is becoming a primary concern in many areas of science [Freire et al. 2016] and, in particular, in computer science as also witnessed by the recent *Association for Computing Machinery (ACM)* policy on result and artifact review and badging[5]. An increasing attention is paid to reproducibility also in IR [Ferro et al. 2016b; Zobel et al. 2011] where discussion is ongoing on: infrastructures for managing experimental data [Agosti et al. 2012a; Agosti et al. 2012b]; use of private data in evaluation [Callan and Moffat 2012]; evaluation as a service [Hanbury et al. 2015]; reproducible baselines [Arguello et al. 2015; Di Buccio et al. 2015]; and, considering it as part of the review process of major conferences and in dedicated tracks, such as the new ECIR Reproducibility Track [Fuhr et al. 2015; Ferro et al. 2016a].

2. CHALLENGES

Reproducing IR experiments is extremely challenging, even when they are very well-documented [Ferro and Silvello 2015]. There are three main different areas that are of major concern for reproducibility: system runs, experimental collections, and meta-evaluation studies.

The most common concern for reproducibility are *system runs*, i.e. the outputs of the execution of an IR system, since they are what typically researchers and developers want to compare their new ideas against. Even if you use the same datasets and even if you rely on shared open source software, there are often many hidden parameters and tunings which hamper the reproducibility of algorithms and techniques. The situation is even more challenging when you also rely on user-interaction data. Approaches like Evaluation-as-a-Service [Hanbury et al. 2015], based on open interfaces and virtual machines [Potthast et al. 2014], or Open Runs [Voorhees et al. 2016], i.e. system runs backed by a software repository that captures the code to recreate the run, are now starting to explore how to face these issues.

Experimental collections are the core of evaluation and they are used for many years, often for purposes different from those that led to their creation. Nevertheless, they are not yet a primary focus for reproducibility, even if they should be, given their central role in experimentation. Indeed, it is important to understand their limitations and their generalizability as well as to reproduce the process that led to their creation. This is not always trivial since, for example, documents may be ephemeral data such as tweets [Amigó et al. 2012], topics may be sampled from real system logs [Allan et al. 2008], relevance judgements are made by (disagreeing) humans [Voorhees 2000] and, more and more often, using crowdsourcing [Alonso and Mizzaro 2012].

Finally, IR has a strong tradition in assessing its own evaluation methodologies, such as robustness of the experimental collections [Zobel 1998], reliability of the adopted evaluation measures [Sakai 2006] and many others [Sanderson 2010]. All these *meta-evaluation experiments* strongly rely on existing experimental collections and gathered systems runs and their reproducibility should be a key concern, since they probe our own experimental methods.

3. TOWARDS SOLUTIONS

Even if IR has a long tradition in ensuring that the due scientific rigor is guaranteed in producing experimental data, it has not a similar tradition in managing and taking care of such valuable data [Agosti et al. 2007; Dussin and Ferro 2009]. This represents a serious obstacle to facing the above mentioned challenges. For example, there is a lack of commonly agreed formats for modeling and describing the experimental data as well as almost no metadata (descriptive, administrative, copyright, etc.) for annotating and enriching them. The semantics of the data themselves is often not explicit and it

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

is demanded to the scripts typically used for processing them, which are often not well documented, rely on rigid assumptions on the data format or even on side effects in processing the data. Finally, IR lacks a commonly agreed mechanism for citing and linking data to the papers describing them [Silvello and Ferro 2016].

All these issues may be addressed by adapting solutions developed in other fields with similar problems. However, the biggest issue is the community itself, which would need to evolve its own experimental methodologies to take into account reproducibility and the actions needed to guarantee it. This calls for an orchestrated effort and a cultural change which are the most compelling challenges towards a proper management and curation of the experimental data.

We have moved the first steps in this direction by proposing an initial model of the entities involved in IR evaluation [Silvello et al. 2016], based on semantic Web and *Linked Open Data (LOD)* technologies, and by making (a subset of) the CLEF experimental data accessible through a running prototype. Moreover, we have shown how this model can be effectively leveraged for enriching and curating the experimental data themselves by automatically connecting them with expertise topics and expert profiles.

Another example in this direction is the ontology of IR concepts proposed by [Lipani et al. 2014] to move towards nanopublications, i.e. the possibility to publish statements about data and experiments, together with references that establish the authorship and provenance of the statements in a machine-readable format.

The goal of these first seeds is to demonstrate how it is possible to address the lack of advanced experimental data management in the IR field and to, hopefully, act as a catalyst for shared and commonly agreed approaches that will enable the community to tackle the above mentioned challenges.

References

M. Agosti, E. Di Buccio, N. Ferro, I. Masiero, S. Peruzzo, and G. Silvello. 2012a. DIRECTions: Design and Specification of an IR Evaluation Infrastructure. In *CLEF 2012*. LNCS 7488, 88–99.

M. Agosti, G. M. Di Nunzio, and N. Ferro. 2007. Scientific Data of an Evaluation Campaign: Do We Properly Deal With Them?. In *CLEF 2006*. LNCS 4730, 11–20.

M. Agosti, N. Ferro, and C. Thanos. 2012b. DESIRE 2011 Workshop on Data infrastructurEs for Supporting Information Retrieval Evaluation. *SIGIR Forum* 46, 1, 51–55.

J. Allan, B. A. Carterette, J. A. Aslam, V. Pavlu, B. Dachev, and E. Kanoulas. 2008. Million Query Track 2007 Overview. In *TREC 2007*. NIST SP 500-274.

O. Alonso and S. Mizzaro. 2012. Using Crowdsourcing for TREC Relevance Assessment. *IPM* 48, 6, 1053–1066.

E. Amigó, A. Corujo, J. Gonzalo, E. Meij, and M. de Rijke. 2012. Overview of RepLab 2012: Evaluating Online Reputation Management Systems. In *CLEF 2012 Working Notes*. http://ceur-ws.org/Vol-1178/.

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, 107–116.

J. Callan and A. Moffat. 2012. Panel on Use of Proprietary Data. *SIGIR Forum* 46, 2, 10–18.

W. B. Croft, D. Metzler, and T. Strohman. 2009. *Search Engines: Information Retrieval in Practice*. Pearson.

E. Di Buccio, G. M. Di Nunzio, N. Ferro, D. K. Harman, M. Maistro, and G. Silvello. 2015. Unfolding Off-the-shelf IR Systems for Reproducibility. In *Proc. SIGIR Workshop on Reproducibility, Inexplicability, and Generalizability of Results (RIGOR 2015)*.

M. Dussin and N. Ferro. 2009. Managing the Knowledge Creation Process of Large-Scale Evaluation Campaigns. In *ECDL 2009*. LNCS 5714, 63–74.

N. Ferro. 2014. CLEF 15th Birthday: Past, Present, and Future. *SIGIR Forum* 48, 2, 31–55.

N. Ferro, F. Crestani, M.-F. Moens, J. Mothe, F. Silvestri, G. M. Di Nunzio, C. Hauff, and G. Silvello (Eds.). 2016a. *Proc. 38th European Conference on IR Research (ECIR 2016)*. LNCS 9626.

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

N. Ferro, A. Hanbury, H. Müller, and G. Santucci. 2011. Harnessing the Scientific Data Produced by the Experimental Evaluation of Search Engines and Information Access Systems. In *ICCS 2011*. Procedia Computer Science 4, 740–749.

N. Ferro and G. Silvello. 2015. Rank-Biased Precision Reloaded: Reproducibility and Generalization, In *ECIR 2015*. LNCS 9022, 768–780.

J. Freire, N. Fuhr, and A. Rauber (Eds.). 2016. *Report from Dagstuhl Seminar 16041: Reproducibility of Data-Oriented Experiments in e-Science*. Dagstuhl Reports 6, 1.

N. Fuhr, A. Rauber, G. Kazai, and A. Hanbury (Eds.). 2015. *Proc. 37th European Conference on IR Research (ECIR 2015)*. LNCS 9022.

A. Hanbury, H. Müller, K. Balog, T. Brodt, G. V. Cormack, I. Eggel, T. Gollub, F. Hopfgartner, J. Kalpathy-Cramer, N. Kando, A. Krithara, J. Lin, S. Mercer, and M. Potthast. 2015. Evaluation-as-a-Service: Overview and Outlook. *CoRR* abs/1512.07454.

D. K. Harman. 2011. *Information Retrieval Evaluation*. Morgan & Claypool Publishers.

D. K. Harman and E. M. Voorhees (Eds.). 2005. *TREC. Experiment and Evaluation in Information Retrieval*. MIT Press.

A. Lipani, F. Piroi, L. Andersson, and A. Hanbury. 2014. An Information Retrieval Ontology for Information Retrieval Nanopublications. In *CLEF 2014*. LNCS 8685, 44–49.

M. Lupu and A. Hanbury. 2013. Patent Retrieval. *FnTIR* 7, 1 (2013), 1–97.

M. Potthast, T. Gollub, F. Rangel Pardo, P. Rosso, E. Stamatatos, and B. Stein. 2014. Improving the Reproducibility of PAN's Shared Tasks: Plagiarism Detection, Author Identification, and Author Profiling. In *CLEF 2014*. LNCS 8685, 268–299.

T. Sakai. 2006. Evaluating Evaluation Metrics based on the Bootstrap. In *SIGIR 2006*. 525–532.

M. Sanderson. 2010. Test Collection Based Evaluation of Information Retrieval Systems. *FnTIR* 4, 4 (2010), 247–375.

G. Silvello, G. Bordea, N. Ferro, P. Buitelaar, and T. Bogers. 2016. Semantic Representation and Enrichment of Information Retrieval Experimental Data. *IJDL* (in print).

G. Silvello and N. Ferro. 2016. "Data Citation is Coming". Introduction to the Special Issue on Data Citation. *Bulletin of IEEE Technical Committee on Digital Libraries (IEEE-TCDL)* 12, 1, 1–5.

E. M. Voorhees. 2000. Variations in relevance judgments and the measurement of retrieval effectiveness. *IPM* 36, 5, 697–716.

E. M. Voorhees, S. Rajput, and I. Soboroff. 2016. Promoting Repeatability Through Open Runs. In *Proc. 7th International Workshop on Evaluating Information Access (EVIA 2016)*. 17–20.

J. Zobel. 1998. How Reliable are the Results of Large-Scale Information Retrieval Experiments. In *SIGIR 1998*. 307–314.

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)*. ACM Press, 3–6.

Received 15 September 2016; revised 4 November 2016; accepted 7 November 2016