The Information Retrieval Experiment Platform

Extended Abstract

Maik Fröbe[1], Jan Heinrich Reimer[2], Sean MacAvaney[3], Niklas Deckers[4], Janek Bevendorff[5], Benno Stein[6], Matthias Hagen[7] and Martin Potthast[8]

[1]Friedrich-Schiller-Universität Jena
[3]Leipzig University and ScaDS.AI
[4]Bauhaus-Universität Weimar

Abstract

In this extended abstract,[1] we present the Information Retrieval Experiment Platform (TIREx) that integrates ir_datasets, ir_measures, PyTerrier, and TIRA for standardized, reproducible, collaborative, scalable, and blinded experiments in IR. Information retrieval experiments face potential problems concerning (1) internal validity, (2) external validity, (3) leakage by large pre-trained models, and (4) a high barrier of entry to built on top of state-of-the-art approaches. TIREx aims to support IR experiments to mitigate those issues. We will focus our talk on collaborations enabled by TIREx that lower the barrier to entry for students to shared tasks in IR.

1. Introduction

Research and development in information retrieval (IR) has been predominantly experimental. In its early days in the 1960s, the IR community saw the need to develop and validate experimental procedures, giving rise to the Cranfield paradigm [2], which became the de facto standard for shared tasks hosted at TREC [3] and many spin-off evaluations. Organizers of typical IR-shared tasks provide a task definition, a document collection, and topics. Participants implement retrieval approaches for the task, and run them on the topics. They then submit the resulting document rankings ("runs") to the organizers. The task organizers pool all submitted runs, and evaluate them [4], to produce a reusable set of relevance assessments. Finally, participants share a written description of their runs (a "notebook" paper) to disseminate their methodology and findings. This division of labor allowed the community to scale up collaborative experiments. With many research labs working independently on the same task, the community descends on a "wisdom of the crowd", while ensuring a rigorous comparative evaluation.

[1] Condensed version of a resource paper at SIGIR 2023 [1]

LWDA 2023: Lernen, Wissen, Daten, Analysen October 09–11, 2023, Marburg, Germany

maik.froebe@uni-jena.de (M. Fröbe); heinrich.reimer@uni-jena.de (J. Reimer); Sean.MacAvaney@glasgow.ac.uk (S. MacAvaney); niklas.deckers@uni-leipzig.de (N. Deckers); janek.bevendorff@uni-weimar.de (J. Bevendorff); benno.stein@uni-weimar.de (B. Stein); heinrich.reimer@uni-jena.de (M. Hagen); martin.potthast@uni-leipzig.de (M. Potthast)

0000-0002-1003-981X (M. Fröbe); 0000-0003-1992-8696 (J. Reimer); 0000-0002-8914-2659 (S. MacAvaney); 0000-0001-6803-1223 (N. Deckers); 0000-0002-3797-0559 (J. Bevendorff); 0000-0001-9033-2217 (B. Stein); 0000-0003-1992-8696 (M. Hagen); 0000-0003-2451-0665 (M. Potthast)

© 2023 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

CEUR Workshop Proceedings (CEUR-WS.org)

Despite their lasting success, this way of organizing shared tasks also has shortcomings. First, as with many other disciplines in computer science and beyond, approaches described in a given notebook underlying a given run submission might not be reproducible. There are well-documented cases where reproductions failed, despite putting much effort into it, even for approaches with diligently archived code repositories [5, 6]. Second, run submissions require that participants have access to the test data, which has severe implications [7], such as informing (biasing) the research hypothesis or approach, unless researchers make a point of not looking at the test topics during development. Third, it cannot be ruled out that not a single one of future large language models has been trained, by mistake or deliberately, on publicly available test data to maximize its ability, or that a usage warning that states not to use the data for training would go unnoticed. In any case, the current best practices for shared tasks do not enforce "blinded experimentation" with sufficient rigor, compared to other empirical disciplines.

We develop the IR Experiment Platform [1] to address those problems. Its key features include full integration of open source tools for working with IR data (`ir_datasets` [8]), for executing retrieval pipelines (PyTerrier [9]), and for evaluating IR systems (`ir_measures` [10]) with TIRA [11], a continuous integration platform for reproducible shared tasks. The IR Experiment Platform aims to promote the standardization of IR experiments and to enable the submission of working software rather than runs.

2. New Perspectives on Cooperations and Teaching Initiatives

The main focus of the presentation at the FGIR workshop will be on how we can lower the barrier of entry to shared tasks and on how TIREx can promote new types of cooperation among participants of shared tasks. We specifically want to address teaching initiatives where students participate in shared tasks as part of their coursework. One main difficulty in this setting is that organizers become part of the debug cycle. To mitigate this problem, we aim to reduce the gap between the development environment and the submission environment. For Docker submissions, we often observed that participants dockerized their software only as an afterthought (the instructions to set up the development environment did not match the setup in the Docker image; submitted images often had missing libraries or wrong versions), and we currently try to address this by promoting procedures where participants directly develop in the docker image that they will submit (e.g., via dev-containers). Furthermore, TIREx allows to execute pipelines where the output of one pipeline component can be used as input to subsequent components. These components can be software submissions or manual data uploads. Cooperations enabled by this are, for instance, if one team creates user query variants that can be used as additional input in pipelines by other teams. We also envision similar cooperations for standard components of retrieval pipelines. For instance, the team behind the query performance prediction toolkit qpptk [12] currently dockerizes their framework so that it can be used as a pipeline component in TIREx, which enables other teams to directly use query performance prediction without the need to learn a new framework. We are currently in the process of advertising this type of cooperation and try to get more such retrieval components into TIREx (e.g., the Splade [13] and REL [14] teams agreed to submit their systems to TIREx). TIREx supports caching of results, so each component is executed only once, aiming at GreenIR [15].

Acknowledgments

This work has been partially supported by the OpenWebSearch.eu project (funded by the EU; GA 101070014).

References

[1] M. Fröbe, J. Reimer, S. MacAvaney, N. Deckers, S. Reich, J. Bevendorff, B. Stein, M. Hagen, M. Potthast, The Information Retrieval Experiment Platform, in: H. Chen, W. E. Duh, H. Huang, M. P. Kato, J. Mothe, B. Poblete (Eds.), 46th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2023), ACM, 2023, pp. 2826–2836. URL: https://dl.acm.org/doi/10.1145/3539618.3591888. doi:10.1145/3539618.3591888.

[2] C. Cleverdon, The Cranfield tests on index language devices, in: ASLIB Proceedings, MCB UP Ltd. (Reprinted in Readings in Information Retrieval, Karen Sparck-Jones and Peter Willett, editors, Morgan Kaufmann, 1997), 1967, pp. 173–192.

[3] E. M. Voorhees, The evolution of cranfield, in: N. Ferro, C. Peters (Eds.), Information Retrieval Evaluation in a Changing World - Lessons Learned from 20 Years of CLEF, volume 41 of *The Information Retrieval Series*, Springer, 2019, pp. 45–69.

[4] E. M. Voorhees, The philosophy of information retrieval evaluation, in: C. Peters, M. Braschler, J. Gonzalo, M. Kluck (Eds.), Evaluation of Cross-Language Information Retrieval Systems, Second Workshop of the Cross-Language Evaluation Forum, CLEF 2001, Darmstadt, Germany, September 3-4, 2001, Revised Papers, volume 2406 of *Lecture Notes in Computer Science*, Springer, 2001, pp. 355–370.

[5] J. Arguello, F. Diaz, J. Lin, A. Trotman, SIGIR 2015 workshop on reproducibility, inexplicability, and generalizability of results (RIGOR), in: R. Baeza-Yates, M. Lalmas, A. Moffat, B. A. Ribeiro-Neto (Eds.), Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval, Santiago, Chile, August 9-13, 2015, ACM, 2015, pp. 1147–1148.

[6] J. Lin, Q. Zhang, Reproducibility is a process, not an achievement: The replicability of IR reproducibility experiments, in: J. M. Jose, E. Yilmaz, J. Magalhães, P. Castells, N. Ferro, M. J. Silva, F. Martins (Eds.), Advances in Information Retrieval - 42nd European Conference on IR Research, ECIR 2020, Lisbon, Portugal, April 14-17, 2020, Proceedings, Part II, volume 12036 of *Lecture Notes in Computer Science*, Springer, 2020, pp. 43–49.

[7] N. Fuhr, Proof by experimentation?: towards better IR research, SIGIR Forum 54 (2020) 2:1–2:4.

[8] S. MacAvaney, A. Yates, S. Feldman, D. Downey, A. Cohan, N. Goharian, Simplified data wrangling with ir_datasets, in: F. Diaz, C. Shah, T. Suel, P. Castells, R. Jones, T. Sakai (Eds.), SIGIR '21: The 44th International ACM SIGIR Conference on Research and Development in Information Retrieval, Virtual Event, Canada, July 11-15, 2021, ACM, 2021, pp. 2429–2436.

[9] C. Macdonald, N. Tonellotto, S. MacAvaney, I. Ounis, Pyterrier: Declarative experimentation in python from BM25 to dense retrieval, in: G. Demartini, G. Zuccon, J. S. Culpepper, Z. Huang, H. Tong (Eds.), CIKM '21: The 30th ACM International Conference on Informa-

tion and Knowledge Management, Virtual Event, Queensland, Australia, November 1 - 5, 2021, ACM, 2021, pp. 4526–4533.

[10] S. MacAvaney, C. Macdonald, I. Ounis, Streamlining evaluation with ir-measures, in: M. Hagen, S. Verberne, C. Macdonald, C. Seifert, K. Balog, K. Nørvåg, V. Setty (Eds.), Advances in Information Retrieval - 44th European Conference on IR Research, ECIR 2022, Stavanger, Norway, April 10-14, 2022, Proceedings, Part II, volume 13186 of *Lecture Notes in Computer Science*, Springer, 2022, pp. 305–310.

[11] M. Fröbe, M. Wiegmann, N. Kolyada, B. Grahm, T. Elstner, F. Loebe, M. Hagen, B. Stein, M. Potthast, Continuous Integration for Reproducible Shared Tasks with TIRA.io, in: Advances in Information Retrieval. 45th European Conference on IR Research (ECIR 2023), Lecture Notes in Computer Science, Springer, Berlin Heidelberg New York, 2023.

[12] G. Faggioli, O. Zendel, J. S. Culpepper, N. Ferro, F. Scholer, An enhanced evaluation framework for query performance prediction, in: D. Hiemstra, M. Moens, J. Mothe, R. Perego, M. Potthast, F. Sebastiani (Eds.), Advances in Information Retrieval - 43rd European Conference on IR Research, ECIR 2021, Virtual Event, March 28 - April 1, 2021, Proceedings, Part I, volume 12656 of *Lecture Notes in Computer Science*, Springer, 2021, pp. 115–129. URL: https://doi.org/10.1007/978-3-030-72113-8_8. doi:10.1007/978-3-030-72113-8_8.

[13] T. Formal, C. Lassance, B. Piwowarski, S. Clinchant, Splade v2: Sparse lexical and expansion model for information retrieval, CoRR abs/2109.10086 (2021). arXiv:2109.10086.

[14] J. M. van Hulst, F. Hasibi, K. Dercksen, K. Balog, A. P. de Vries, REL: an entity linker standing on the shoulders of giants, in: J. X. Huang, Y. Chang, X. Cheng, J. Kamps, V. Murdock, J. Wen, Y. Liu (Eds.), Proceedings of the 43rd International ACM SIGIR conference on research and development in Information Retrieval, SIGIR 2020, Virtual Event, China, July 25-30, 2020, ACM, 2020, pp. 2197–2200. URL: https://doi.org/10.1145/3397271.3401416. doi:10.1145/3397271.3401416.

[15] H. Scells, S. Zhuang, G. Zuccon, Reduce, reuse, recycle: Green information retrieval research, in: E. Amigó, P. Castells, J. Gonzalo, B. Carterette, J. S. Culpepper, G. Kazai (Eds.), SIGIR '22: The 45th International ACM SIGIR Conference on Research and Development in Information Retrieval, Madrid, Spain, July 11 - 15, 2022, ACM, 2022, pp. 2825–2837. URL: https://doi.org/10.1145/3477495.3531766. doi:10.1145/3477495.3531766.