The First International Workshop on Open Web Search (WOWS)*

Sheikh Mastura Farzana[1], Maik Fröbe[2], Michael Granitzer[3], Gijs Hendriksen[4], Djoerd Hiemstra[4], Martin Potthast[5], and Saber Zerhoudi[3]

[1] German Aerospace Center (DLR), sheikh.farzana@dlr.de
[2] Friedrich-Schiller-Universität Jena, maik.froebe@uni-jena.de
[3] University of Passau, michael.granitzer@uni-passau.de
[4] Radboud University, djoerd.hiemstra@ru.nl
[5] Leipzig University and ScaDS.AI, martin.potthast@uni-leipzig.de

Abstract We organize the first international Workshop on Open Web Search (WOWS) at ECIR 2024 with two calls for contributions. The first call targets scientific contributions on cooperative search engine development. This includes cooperative crawling of the web and cooperative deployment and evaluation of search engines. We specifically highlight the potential of enabling public and commercial organizations to use an indexed web crawl as a resource to create innovative search engines tailored to specific user groups, instead of relying on one search engine provider. The second call aims at gaining practical experience with joint, cooperative evaluation of search engine prototypes and their components using the Information Retrieval Experiment Platform (TIREx).

1 Introduction

Web search is a critical technology for the digital economy, dominated by a few gatekeepers [4]. These gatekeepers take care of and thus control every aspect of search: they crawl the web, index it, develop search engines, and provide search-based applications. Since these gatekeepers also control the market for search engine advertisements that are displayed alongside search results, they have incentives to sacrifice search result quality by promoting best-paying results over the most relevant results. The gatekeepers have every opportunity to dictate their will to their users: They even control the web browsers and operating systems that users need to access the web [15]. Additionally, these companies also have a profound influence on academic research, both indirectly and directly through funding of research facilities, researchers, and sponsorship of conferences [1].[6]

With these tech giants of web search controlling every aspect of search, web publishers need to optimize their content for the search engines instead of the search engines optimizing their results for the content. This has led to a closed ecosystem of search engines and risks compromising quality for publishers. Many search engine users are aware of this fact [11].

*https://opensearchfoundation.org/wows2024/

[6]Some authors were previously funded by grants awarded by these gatekeepers.

The first International Workshop on Open Web Search (WOWS) promotes and discusses ideas and approaches to opening up today's closed search ecosystem. We are particularly interested in approaches that allow organizations to provide search engines cooperatively, e.g., by specializing in one task and sharing their expertise, tools, and resources with others for mutual benefit. We are also interested in ideas and approaches that allow organizations to collaborate on these tasks and open standards for search and open source retrieval.

Beyond search, the web has demonstrated its critical role as a resource, e.g., for training large language models, for analyzing human behavioral data, or for studying the web as a structure itself. However, tapping into such a resource requires a large infrastructure, appropriate technical capabilities, and much data cleaning and pre-processing. While some large organizations can provide these resources, small research groups or young startups either lack the skills or would have to devote much time and resources outside their core area, which is usually beyond their budget. Since the open source landscape of web data processing is surprisingly small, we are interested in approaches where organizations share their data processing infrastructures and treat their data as open and accessible.

From a practical point of view, the workshop will also focus on the scientific evaluation of search engines and their components, using test collections from shared tasks. We will encourage participants to submit and evaluate new retrieval approaches or components of retrieval pipelines using TIRA/TIREx [7, 8]. Modern web search engines use complex pipelines with many components (e.g., query/document understanding), so different organizations can bring different components to open search ecosystems. Therefore, we ask participants to implement retrieval pipelines and/or components. Using TIREx, these components can be systematically combined and run on many test collections, making their results publicly available, and comparing their effectiveness. If the outputs of standard components are available for multiple test collections, researchers can reuse them without having to rerun the software itself, promoting GreenIR [18] along the entire retrieval pipeline [9]. Therefore, the goals of the workshop are:

1. Sharing novel concepts, algorithms, and ideas for cooperatively building web search engines to open the web as a resource for researchers and innovators.
2. Gaining hands-on experience with joint, collaborative evaluation of search engines and/or their components with TIRA/TIREx.

2 Workshop Program

The workshop's intended audience includes researchers and practitioners in open web search and has a call for research contributions and a call for software. We see the Workshop on Open Web Search as a continuation of workshops on the topic of open source information retrieval that began with the first International Workshop on Open Source Web Information Retrieval in 2005, the second International Workshop on Open Source Information Retrieval in 2006 [22], the Workshop on Open Source Information Retrieval in 2012 [19], the Lucene for information access and retrieval research (LIARR) workshop in 2017 [3], and the

workshops on replicability, including the Reproducibility, Inexplicability, and Generalizability of Results (RIGOR) workshop in 2016 [2] and the Open-Source Information Retrieval Replicability Challenge (OSIRRC) in 2019 [5].

2.1 Call for Research Contributions

We seek research contributions that address elements of a traditional web search pipeline and also incorporate recent developments such as (open source) large language models to interface with retrieval systems. Specifically, our focus is on contributions that address the importance of an open web search pipeline and the creation of an open web index as a basis for the development of search applications for specific purposes and communities. Interesting topics include, but are not limited to (1) crawling for an open web index (e.g., fast DNS resolution, distribution of crawl jobs, etc.). (2) pre-processing and enrichment (e.g., entity extraction/linking, geo-parsing, spam classification, etc.). (3) indexing and search architectures (e.g., with tools like Terrier [14], OldDog [17], GeeseDB [10], Spinque [6], CIFF [12], Anserini [21], PISA [16], JASSv2 [20]). (4) search interfaces and paradigms (e.g., novel search paradigms like conversational search, temporal argument search, or geospatial search with aspects like trust, privacy, bias, and ethics). Overall, we invite contributions in the form of research papers and resource papers along the following non-exclusive list of topics: standards for search and interoperability; deployment of search engines; collaborative crawling; open infrastructures for evaluation; open source search engines; open source replicability; ethical and legal aspects of web search; alternatives for query logs and click logs; vertical search engines; search engines for low-resource languages; web master control (robots.txt and more); energy efficiency of web search; and large scale web data pre-processing pipelines.

2.2 Call for Evaluation Components

Evaluation plays a critical role for every (web) search engine. Robust and insightful evaluation is essential to guide the development efforts of cooperative open web search engines, especially when multiple decoupled organizations with potentially incompatible goals contribute to the ecosystem. However, evaluations of general-purpose web search engines are difficult due to the size of the web and the immense technical effort required to implement a practical web search engine. Therefore, we request the submission of software to TIRA/TIREx [7, 8] that implements search engine components that the community can declaratively assemble to evaluate retrieval pipelines composed of submitted components.

This call leverages the open infrastructure for evaluation provided by TIRA/TIREx that enables the construction of experimental retrieval pipelines that can be evaluated on many information retrieval test collections without having to own these collections or even invest in preprocessing or indexing them (e.g., by submitting re-rankers). Each step of a retrieval pipeline in TIREx is a Docker image, and submitted Docker images can be used in declarative PyTerrier [14] pipelines after the shared task has finished [7]. All software submissions

to TIREx are immutable to allow caching of a component's output. Typical IR test collections are static, so that many steps in retrieval pipelines need to be executed only once in a lifetime. Thus, we will make all outputs of the submitted software publicly available where the dataset licenses allow this so that the community can build complex retrieval pipelines without re-executing many of its components. At the same time, TIREx provides a provenance "chain of trust" to users who reuse the output of a software component, as it documents which Docker image produced which output. Because TIRA runs each submitted Docker image in a sandbox not under the participant's control, reproducibility and replicability are improved while simplifying reusing components.

The outcomes of the shared task are the dockerized components submitted by the participants and the outputs produced by those docker images on standard IR benchmarks in TIRA/TIREx. The focus of our call for individual (potentially fine-grained) components of retrieval pipelines is in contrast to previous efforts that collected complete retrieval systems [2, 5], as small components (or their outputs) can be more easily reused in a wide range of IR experiments than complete systems, especially in fast-changing research environments. All software submissions in TIREx are implemented against the interface provided by ir_datasets [13], which provides abstraction and standardization over typical data structures of IR experiments such as information needs and documents.

Our call for software asks for submissions that fall into one or more of the following classes of retrieval components: (1) query processing (e.g., query performance prediction, query reformulation, entity detection, query segmentation, query expansion, etc.), (2) document processing (e.g., document expansion, document reduction, spam classification, web genre classification, topic modeling, etc.), and (3) query–document processing (e.g., re-ranking, rank fusion, etc.).

3 Conclusion

In summary, the ECIR 2024 Workshop on Open Web Search (WOWS) addresses some of the most pressing concerns related to web search such as transparency, availability, and accessibility of web resources. We expect that WOWS will foster collaboration and innovation in building open search ecosystems. WOWS will explore new approaches that enable organizations to collaborate on various aspects of search engines, promote open standards, and challenge the current closed ecosystem of search. In addition, WOWS aims to facilitate data sharing and open access to web resources. The practical focus of the workshop is to promote scientific evaluation through shared tasks, collaboration in building search pipelines and retrieval test collections, with the vision that different organizations contribute to building a sustainable open search framework.

Acknowledgments

This work has received funding from the European Union's Horizon Europe research and innovation programme under grant agreement No 101070014 (OpenWebSearch.EU, https://doi.org/10.3030/101070014).

Bibliography

[1] Abdalla, M., Abdalla, M.: The grey hoodie project: Big tobacco, big tech, and the threat on academic integrity. In: Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society, pp. 287–297 (2021)

[2] Arguello, J., Crane, M., Diaz, F., Lin, J., Trotman, A.: Report on the SIGIR 2015 workshop on reproducibility, inexplicability, and generalizability of results (RIGOR). SIGIR Forum **49**(2), 107–116 (2016)

[3] Azzopardi, L., Crane, M., Fang, H., Ingersoll, G., Lin, J., Moshfeghi, Y., Scells, H., Yang, P., Zuccon, G.: The Lucene for information access and retrieval research (LIARR) workshop at SIGIR 2017. In: Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 1429–1430 (2017)

[4] Bahrke, J., Podesta, A., Regnier, T., Simonini, S.: Digital markets act: Commission designates six gatekeepers. European Commission Press Release (2023), https://ec.europa.eu/commission/presscorner/detail/en/IP_23_4328

[5] Clancy, R., Ferro, N., Hauff, C., Lin, J., Sakai, T., Wu, Z.Z.: The SIGIR 2019 open-source IR replicability challenge (OSIRRC 2019). In: Piwowarski, B., Chevalier, M., Gaussier, É., Maarek, Y., Nie, J., Scholer, F. (eds.) Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 1432–1434 (2019)

[6] Cornacchia, R.: Graph Ranking – part 1. Spinque (2021), https://spinque.com/blog/graph-ranking-part-1/

[7] Fröbe, M., Reimer, J.H., MacAvaney, S., Deckers, N., Reich, S., Bevendorff, J., Stein, B., Hagen, M., Potthast, M.: The information retrieval experiment platform. In: Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval (2023)

[8] Fröbe, M., Wiegmann, M., Kolyada, N., Grahm, B., Elstner, T., Loebe, F., Hagen, M., Stein, B., Potthast, M.: 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 (2023)

[9] Granitzer, M., Voigt, S., et al.: Impact and development of an open web index for open web search. Journal of the Association for Information Science and Technology (2023)

[10] Kamphuis, C., de Vries, A.P.: GeeseDB: A Python graph engine for exploration and search. In: Proceedings of the 2nd International Conference on Design of Experimental Search and Information REtrieval Systems (DESIRES) (2021)

[11] Lewandowski, D., Schultheiß, S.: Public awareness and attitudes towards search engine optimization. Behaviour & Information Technology **42**, 1025–1044 (2022)

[12] Lin, J., Mackenzie, J., Kamphuis, C., Macdonald, C., Mallia, A., Siedlaczek, M., Trotman, A., de Vries, A.: Supporting interoperability between open-source search engines with the common index file format. In: Proceedings

of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 2149–2152 (2020)
[13] MacAvaney, S., Yates, A., Feldman, S., Downey, D., Cohan, A., Goharian, N.: Simplified data wrangling with ir_datasets. In: SIGIR (2021)
[14] Macdonald, C., Tonellotto, N., MacAvaney, S., Ounis, I.: Pyterrier: Declarative experimentation in Python from BM25 to dense retrieval. In: Proceedings of the 30th ACM International Conference on Information & Knowledge Management (CIKM), pp. 4526–4533 (2021)
[15] Mager, A., Norocel, C., Rogers, R. (eds.): The State of Google Critique and Intervention, Sage Journals (2023)
[16] Mallia, A., Siedlaczek, M., Mackenzie, J., Suel, T.: PISA: performant indexes and search for academia. In: Proceedings of the Open-Source IR Replicability Challenge co-located with 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval, OSIRRC@SIGIR 2019, Paris, France, July 25, 2019., pp. 50–56 (2019), URL http://ceur-ws.org/Vol-2409/docker08.pdf
[17] Mühleisen, H., Samar, T., Lin, J., De Vries, A.: Old dogs are great at new tricks: Column stores for information retrieval prototyping. In: Proceedings of the 37th International ACM SIGIR conference on Research & Development in Information Ietrieval, pp. 863–866 (2014)
[18] Scells, H., Zhuang, S., Zuccon, G.: Reduce, reuse, recycle: Green information retrieval research. In: Amigó, E., Castells, P., Gonzalo, J., Carterette, B., Culpepper, J.S., Kazai, G. (eds.) SIGIR '22: The 45th International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 2825–2837 (2022)
[19] Trotman, A., Clarke, C.L., Ounis, I., Culpepper, S., Cartright, M.A., Geva, S.: Open source information retrieval: A report on the SIGIR 2012 workshop. SIGIR Forum **46**(2), 95–101 (December 2012)
[20] Trotman, A., Lilly, K.: Jassjr: The minimalistic BM25 search engine for teaching and learning information retrieval. In: Huang, J.X., Chang, Y., Cheng, X., Kamps, J., Murdock, V., Wen, J., Liu, Y. (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, pp. 2185–2188, ACM (2020), https://doi.org/10.1145/3397271.3401413, URL https://doi.org/10.1145/3397271.3401413
[21] Yang, P., Fang, H., Lin, J.: Anserini: Enabling the use of lucene for information retrieval research. In: Kando, N., Sakai, T., Joho, H., Li, H., de Vries, A.P., White, R.W. (eds.) Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval, Shinjuku, Tokyo, Japan, August 7-11, 2017, pp. 1253–1256, ACM (2017), https://doi.org/10.1145/3077136.3080721, URL https://doi.org/10.1145/3077136.3080721
[22] Yee, W.G., Beigbeder, M., Buntine, W.: SIGIR06 workshop report: Open source information retrieval systems (OSIR06). ACM SIGIR Forum **40**(2), 61–65 (2006)