Proposal for the 2nd International Workshop on Open Web Search (WOWS) at ECIR 2025

Sheikh Mastura Farzana[1], Maik Fröbe[2], Michael Granitzer[3], Gijs Hendriksen[4], Djoerd Hiemstra[4], Martin Potthast[5], Arjen P. de Vries[4], 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, firstname.lastname@ru.nl
[5] Leipzig University martin.potthast@uni-leipzig.de

Abstract We propose the second International Workshop on Open Web Search (WOWS) at ECIR 2025 with two calls for contributions: the first call aims at scientific contributions to building search engines cooperatively, including crawling, deployment, evaluation, and the use of the web as a resource by researchers and innovators. The second call introduces a shared task named WOWS-Eval that aims at gaining practical experience with joint, cooperative evaluation of search engines by focusing to enrich the Open Web Index (OWI) with relevance judgments cooperatively transferred from existing TREC-style test collections.

1 Introduction

Web search is a crucial technology for the digital economy that is dominated by a few gatekeepers like Google and Microsoft. These gatekeepers control every aspect of search: they crawl the web, they build the index, they build the search engines and provide the search applications. More problematically, these gatekeepers also control the advertisements that are shown with the search results, giving them incentives to sacrifice quality by showing the best paying results over the most relevant results. These companies have every opportunity to impose their will on their users: including controlling the web browsers and operating systems that users need in order to access the web [1]. These gatekeepers also have a profound influence on the academic research agendas by funding research institutions, sponsoring conferences, and funding researchers directly [2].

Because the gatekeepers of web search control every aspect of search, web publishers optimize their content to the search engines instead of the search engines optimizing results to the content. This has resulted in a closed ecosystem of search engines and the risk of publishers sacrificing quality. Many search engine users are aware of this but do not always have access to better alternatives [20].

The goal of this workshop is to continue the previous edition of the WOWS workshop held at ECIR 2024[6] [11, 12] that had a scientific and a practical hands-on part.The scientific focus of WOWS 2024, which we aim to continue, is the

[6]https://opensearchfoundation.org/wows2024/

exploration of ideas and approaches for opening up the closed search ecosystem we have today. We are particularly interested in approaches that allow organizations to collaboratively build search engines, for instance by allowing organizations to specialize on one task (crawling, indexing, search, web search applications) while sharing their results with others. Furthermore, we are interested in ideas and approaches that allow organizations to cooperate on each individual task, for example, in open standards for search and open source information retrieval systems. Beyond search, the web has demonstrated its critical role as a resource, for several applications such as training large language models, analysing human behavioural data, etc. However, tapping into such a resource requires suitable infrastructure, corresponding technical skills and large efforts in terms of data cleaning and pre-processing that small research groups or young startups might lack. We are interested in approaches where organizations share and combine their data processing infrastructures and treat their data openly.

From a practical perspective, we are interested in conducting a shared task named WOWS-Eval that transfers relevance judgments from TREC-style test collections to the Open Web Index (OWI) [15, 17]. The previous edition of WOWS already included a hands-on shared task that aimed to collaboratively develop and evaluate components of retrieval pipelines, where 11 teams participated (notably, 4 teams through a dedicated one-week student hackathon at TU Dresden [9]). To continue this hands-on part of the workshop, we aim to collaboratively enable the evaluation of retrieval systems and their components on the Open Web Index for diverse use-cases by transferring relevance judgments. To maximize the re-usability of the transferred relevance judgments for the scientific community, we will use the ClueWeb22 [25] as target corpus and we aim to enrich it with relevance judgments from the ClueWeb09, ClueWeb12, MS MARCO, etc. The ClueWeb22 is a broad corpus that contains subsets that are available free of charge,[7] thereby ensuring that transferred relevance judgments benefit the community even if the Open Web Index discontinues. We will host two subtasks, the first for candidate retrieval that aims to collect a set of documents that might be relevant for a topic in the ClueWeb22, and the second subtask aiming at relevance judgments that assess if a document is relevant for a given topic. Especially large language models are currently receiving much attention for conducting relevance judgments, e.g., as part of the LLMJudge shared task [27][8] at the LLM4Eval workshop [26], and other approaches that use large language models for annotating data [3, 10, 28, 30]. We will ensure that the data format for WOWS-Eval is a superset of the previous LLMJudge task so that approaches submitted to LLMJudge can also be evaluated and will use TIRA/TIREx [13, 14] for software and run submissions.

To sum up the above, the workshop aims are as follows:

1. Exchange novel concepts, algorithms and ideas for building web search engines cooperatively, e.g., crawling, deployment, and evaluation, and for tapping the web as a resource for researchers and innovators; and

[7] http://lemurproject.org/clueweb22/obtain.php
[8] https://llm4eval.github.io/challenge/

Table 1. Approximate outline of the proposed Open Web Search Workshop.

Time	Event
09:00–10:00	Keynote
10:30–12:00	Research Contribution Talks
13:00–14:30	WOWS-Eval Talks on Transferring Relevance Judgments
15:00–16:30	Breakout Groups
16:30–17:00	Reports of the Breakout Groups + Planning SIGIR Forum Paper

2. Enabling the evaluation of retrieval systems on the Open Web Index by gaining practical experience by cooperatively transferring relevance judgments.

The workshop will be organised by researchers participating in the currently ongoing project OpenWebSearch.eu, which aims to build a fully open web index, associated infrastructures, algorithms along the whole retrieval pipeline as well as open machine learning and knowledge representation models [16].

In the remainder, we highlight our workshop program with two calls for contributions (Section 2) and will discuss related events that inspired us (Section 3).

2 Workshop Program

The goal of this workshop is to encourage and discuss ideas and approaches to open the closed search ecosystem. We are particularly interested in approaches that allow organizations to provide search engines cooperatively, for instance by organizations that specialize on one task while sharing their results with others.

Table 1 shows the outline of the proposed full-day workshop. The workshop starts with a keynote, followed by two 90-minute sessions with talks for accepted research contributions (Section 2.1) and the WOWS-Eval relevance label transfer contributions (Section 2.2). Similar to WOWS 2024 [11], we plan to conclude the workshop with a 2 hour session where breakout groups discuss challenges and possible next steps for successful open web search, with the goal of writing a SIGIR Forum paper with organizers and participants as co-authors that summarizes the challenges and possible solutions discussed in the break-out groups.

2.1 Call for Research Contributions

We aim for research contributions that address elements of a traditional web search pipeline and also include recent developments like (open source) large language models as an interface to retrieval systems, and that demonstrate the need of an open web search pipeline and the creation of an open web index, i.e., where the web index itself becomes open data and can be reused as needed.

Crawling for an Open Web Index: Web Crawling describes the process of navigating the graph structure of the web for discovering and fetching web-data. Web crawling is the predominant method for web search engines to build-up their index. It usually involves technical challenges (e.g., fast DNS resolution,

distributing crawl jobs), algorithmic challenges (e.g., link traversal order, duplicate detection) and policy considerations (e.g., crawling etiquette). Web crawling takes up significant resources on both the crawled servers and the crawler itself. In order to ease the process of crawling and to give webmasters control over the crawling process, several standards like sitemaps and robots.txt exist [19]. However, with the advanced usage of web content for training AI systems, more explicitly the new trend of training Generative AI algorithms, these standards have been shown to not be expressible enough for allowing webmasters and content owners to express the intended use of their content. Furthermore, for more green computing, questions arise for making crawling processes more efficient. Consequently, for building an open web index we aim for contributions on more efficient crawling strategies, collaboration and coordination of independent crawling and standards for a higher expressiveness for legal and ethical limitations for content usage.

Preprocessing and Enrichment Web search and other large-scale web data analytics rely on processing archives of web pages stored in a standardized and efficient format. Since its introduction in 2008, the IIPC's Web ARCive (WARC) format[9] has become the standard format for this purpose. As a list of individually compressed HTTP records, it allows for constant-time random access to all kinds of web data. Processing large samples of web crawls (on the order of hundreds of thousands of WARC files), however, still poses a challenge, particularly in terms of throughput performance and also resilience against unexpected, erroneous, or malicious input data. Furthermore, organising preprocessed content on scale requires enrichment of content through entity extraction, entity linking as well as on more advanced concepts for evaluating information quality or potentially hateful content. Questions when building an open web index hereby involve not only the extraction quality, but also efficiency and scalability. It also involves questions on how to leverage novel algorithms from a research lab towards web-scale and how researchers can benefit from analysing web-scale data, like for example analysing the distribution of content bias in the web.

Indexing and Search Architectures: A web index is the core of any traditional web search engine and kept a closed secret by web search engine providers. From our point of view, this hinders the development of new search architectures that combine central and shared /federated indices. Recent research has shown that it is possible to define search engines declaratively over pre-defined indices. Examples are Terrier [23], OldDog [24], and GeeseDB [18], and the commercial offering by Spinque [8]. Pre-defined indices may be provided using the common index file format (CIFF) that is supported by the engines mentioned above, as well as Anserini (which uses Lucene), PISA and JASSv2 [21]. The key insight here is that you could create a "search engine mash-up" by downloading an index and a search engine specification, and run this specification on your

[9]ISO 28500:2017; https://iipc.github.io/warc-specifications/

own hardware (possibly still deployed in the cloud of course). Along this topic we aim to discuss contributions towards the indexing process itself (e.g. partitioning schemes, indexing algorithms, index distribution and storage) as well as potential new web search engine architectures, including, but not limited to declarative search engine, hybrid federated search engines, etc.

Search Interfaces and Paradigms: An open web index, i.e. an web index available for download and open use comparable to open software, would significantly boost the research, development and evaluation of new search interfaces and search paradigms at scale and over realistic, up-to-date data. This could include search paradigms like conversational search, temporal argument search or geospatial search. Also, cross cutting concerns like trust, privacy, bias and ethics can be conceptualised and analysed on a different kind of scale and with a different kind of scope. Consequently, we are looking for contributions that present conceptual, methodological or analytical insights into these topics and also on search verticals benefiting from an open web index.

2.2 WOWS-Eval Call for Relevance Label Transfer Contributions

Evaluation plays a critical role for any (web) search engine. A solid and insightful evaluation is vital to guide the development efforts of cooperative open web search engines, especially if multiple decoupled organizations with potentially incompatible goals contribute different components of the ecosystem. However, generic purpose evaluations of web search engines are very challenging due to the size of the web and the immense engineering efforts required to implement a practical web search engine. Consequently, we aim to enrich the Open Web Index with relevance judgments to enable the direct evaluation of retrieval approaches. Therefore, we want to start to evaluate if automatic relevance transfer from existing relevance judgments to the ClueWeb22 is possible, as this would allow to re-use the high judgment efforts used to create the original judgments while the retrieval systems would still process the Open Web Index. We aim to maintain a mono-repository that contains the code for all submissions.[10] We ask for both run submissions (e.g., approaches that leverage external APIs like ChatGPT) and software submissions as Docker images to TIRA/TIREx [13, 14] and aim to make the resulting transferred relevance judgments (manually spot checked) publicly available, preferrably within ir_datasets [22] via a follow-up publication to an IR conference with all collaborators as co-authors.

Conceptually, we aim to transfer relevance judgments in two subtasks, (1) candidate retrieval, and (2) candidate judgment.

Candidate Retrieval. Given a topic including title, description, narrative, and previously relevant documents, the candidate retrieval aims to pool relevant documents. We make all data of a topic available and encourage participants to use

[10] https://github.com/OpenWebSearch/wows-code/tree/main/conf25

as much information as possible to create diverse judgment pools. Documents retrieved in task 1 are subsequently assessed in task 2. To simplify participation, we have indexed the ClueWeb22 dataset into ChatNoir[11] [6] so that participants can retrieve via an REST API. Potential submissions might use documents previously labeled as relevant as query, can use those documents as relevance feedback, or use classical retrieval approaches on the title, description, or variations.

Candidate Judgment. Given a topic (title, description, narrative, and previously relevant documents) and a candidate document, the candidate judgment task aims to predict the relevance label of the document. Potential solutions could involve the similarity of the to-be-judged documents to previously relevant documents or, similar to the LLM4Eval scenario, an large language model for the relevance prediction. We will provide interested participants access to open-source LLMs via REST APIs and will make the predictions of submitted approaches publicly available during the shared task to allow strong ensembles.

3 WOWS as a Follow-up of Related Workshops

As WOWS 2024, also the proposed 2025 edition is inspired by workshops on open source information retrieval that started in 2005 with the first International Workshop on Open Source Web Information Retrieval; the second International Workshop on Open Source Information Retrieval in 2006 [31]; the 2012 Workshop on Open Source Information Retrieval [29]; the Lucene for information access and retrieval research (LIARR) workshop 2017 [5], the Reproducibility, Inexplicability, and Generalizability of Results (RIGOR) [4]; and the the Open-Source Information Retrieval Replicability Challenge (OSIRRC) [7]. Our second call for WOWS-Eval is inspired by the LLMJudge shared task [27].

4 Conclusion

The ECIR 2025 Workshop on Open Web Search (WOWS) addresses some 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 closed search ecosystem. 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 of retrieval pipelines on the Open Web Index through shared tasks, with the vision that different organizations contribute to 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).

[11] https://chatnoir.web.webis.de/

Bibliography

[1] The State of Google Critique and Intervention, Sage Journals (2023)

[2] 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)

[3] Alizadeh, M., Kubli, M., Samei, Z., Dehghani, S., Bermeo, J.D., Korobeynikova, M., Gilardi, F.: Open-source large language models outperform crowd workers and approach chatgpt in text-annotation tasks. CoRR **abs/2307.02179** (2023)

[4] 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)

[5] 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 SIGIR 2017, pp. 1429–1430, Association for Computing Machinery (2017), ISBN 9781450350228

[6] Bevendorff, J., Stein, B., Hagen, M., Potthast, M.: Elastic chatnoir: Search engine for the clueweb and the common crawl. In: Pasi, G., Piwowarski, B., Azzopardi, L., Hanbury, A. (eds.) Proceedings of ECIR 2018, LNCS, vol. 10772, pp. 820–824, Springer (2018)

[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 2019, pp. 1432–1434, ACM (2019)

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

[9] Erben, L., Hampel, M., Kuns, M., Melisch, V., Natzschka, P., Pertsch, W., Razouk, L., Stolle, R., Thoss, R.T., Trinh, T.G., Gonsior, J., Reusch, A.: Assembling four open web search components: TU dresden at WOWS 2024. In: Proceedings WOWS 2024, CEUR Workshop Proceedings, vol. 3689, pp. 73–93, CEUR-WS.org (2024), URL https://ceur-ws.org/Vol-3689/WOWS_2024_paper_8.pdf

[10] Faggioli, G., Dietz, L., Clarke, C.L.A., Demartini, G., Hagen, M., Hauff, C., Kando, N., Kanoulas, E., Potthast, M., Stein, B., Wachsmuth, H.: Perspectives on large language models for relevance judgment. In: Proceedings of ICTIR 2023, pp. 39–50, ACM (2023)

[11] Farzana, S.M., Fröbe, M., Granitzer, M., Hendriksen, G., Hiemstra, D., Potthast, M., de Vries, A.P., Zerhoudi, S.: Report on the 1st international workshop on open web search (WOWS 2024) at ECIR 2024. SIGIR Forum **58**(1), 1–13 (2024)

[12] Farzana, S.M., Fröbe, M., Granitzer, M., Hendriksen, G., Hiemstra, D., Potthast, M., Zerhoudi, S.: The first international workshop on open web search (WOWS). In: Proceedings of ECIR 2024, LNCS, vol. 14612, pp. 426–431, Springer (2024)

[13] 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 SIGIR 2023 (2023)

[14] 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), LNCS, Springer (2023)

[15] Granitzer, M., Voigt, S., Fathima, N.A., Golasowski, M., Guetl, C., Hecking, T., Hendriksen, G., Hiemstra, D., Martinovic, J., Mitrovic, J., Mlakar, I., Moiras, S.,

Nussbaumer, A., Öster, P., Potthast, M., Srdic, M.S., Megi, S., Slaninová, K., Stein, B., de Vries, A.P., Vondrák, V., Wagner, A., Zerhoudi, S.: Impact and development of an open web index for open web search. J. Assoc. Inf. Sci. Technol. **75**(5), 512–520 (2024)
[16] 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)
[17] Hendriksen, G., Dinzinger, M., Farzana, S.M., Fathima, N.A., Fröbe, M., Schmidt, S., Zerhoudi, S., Granitzer, M., Hagen, M., Hiemstra, D., Potthast, M., Stein, B.: The open web index - crawling and indexing the web for public use. In: Proceedings of ECIR 2024, LNCS, vol. 14612, pp. 130–143, Springer (2024)
[18] Kamphuis, C., de Vries, A.P.: GeeseDB: A Python graph engine for exploration and search (2021)
[19] Koster, M., Illyes, G., Zeller, H., Sassman, L.: Rfc 9309 robots exclusion protocol (2022)
[20] Lewandowski, D., Schultheiß, S.: Public awareness and attitudes towards search engine optimization. Behaviour & Information Technology **42**, 1025–1044 (2022)
[21] 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 SIGIR 2020, pp. 2149–2152 (2020)
[22] MacAvaney, S., Yates, A., Feldman, S., Downey, D., Cohan, A., Goharian, N.: Simplified data wrangling with ir_datasets. In: Proceedings of SIGIR 2021, pp. 2429–2436, ACM (2021)
[23] Macdonald, C., Tonellotto, N., MacAvaney, S., Ounis, I.: Pyterrier: Declarative experimentation in Python from BM25 to dense retrieval. In: Proceedings of CIKM 2021, pp. 4526–4533 (2021)
[24] 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 SIGIR 2014, pp. 863–866 (2014)
[25] Overwijk, A., Xiong, C., Callan, J.: Clueweb22: 10 billion web documents with rich information. In: Proceedings of SIGIR 2022, pp. 3360–3362, ACM (2022)
[26] Rahmani, H.A., Siro, C., Aliannejadi, M., Craswell, N., Clarke, C.L.A., Faggioli, G., Mitra, B., Thomas, P., Yilmaz, E.: Llm4eval: Large language model for evaluation in IR. In: Proceedings of SIGIR 2024, pp. 3040–3043, ACM (2024)
[27] Rahmani, H.A., Siro, C., Aliannejadi, M., Craswell, N., Clarke, C.L.A., Faggioli, G., Mitra, B., Thomas, P., Yilmaz, E.: Report on the 1st workshop on large language model for evaluation in information retrieval (llm4eval 2024) at SIGIR 2024. CoRR **abs/2408.05388** (2024)
[28] Thomas, P., Spielman, S., Craswell, N., Mitra, B.: Large language models can accurately predict searcher preferences. In: Proceedings SIGIR 2024, pp. 1930–1940, ACM (2024)
[29] 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 (dec 2012), ISSN 0163-5840
[30] Upadhyay, S., Pradeep, R., Thakur, N., Craswell, N., Lin, J.: UMBRELA: umbrela is the (open-source reproduction of the) bing relevance assessor. CoRR **abs/2406.06519** (2024)
[31] Yee, W.G., Beigbeder, M., Buntine, W.: SIGIR06 workshop report: Open source information retrieval systems (OSIR06). ACM SIGIR Forum **40**(2), 61–65 (2006)