SIGIR 2024 Workshop on Simulations for Information Access (Sim4IA 2024)

Philipp Schaer
TH Köln
Cologne, Germany
pschaer@th-koeln.de

Christin Katharina Kreutz
TH Köln
Cologne, Germany
TH Mittelhessen
Gießen, Germany
ckreutz@acm.org

Krisztian Balog
University of Stavanger
Stavanger, Norway
krisztian.balog@uis.no

Timo Breuer
TH Köln
Cologne, Germany
timobreuer@acm.org

Norbert Fuhr
University of Duisburg-Essen
Duisburg, Germany
norbert.fuhr@uni-due.de

ABSTRACT

Simulations in various forms have been used to evaluate information access systems, like search engines, recommender systems, or conversational agents. In the form of the Cranfield paradigm, a simulation setup is well-known in the IR community, but user simulations have recently gained interest. While user simulations help to reduce the complexity of evaluation experiments and help with reproducibility, they can also contribute to a better understanding of users. Building on recent developments in methods and toolkits, the Sim4IA workshop aims to bring together researchers and practitioners to form an interactive and engaging forum for discussions on the future perspectives of the field. An additional aim is to plan an upcoming TREC/CLEF campaign.

CCS CONCEPTS

• **Information systems → Task models**; **Information retrieval**.

KEYWORDS

Information Access; Simulation; User Models; Evaluation

ACM Reference Format:
Philipp Schaer, Christin Katharina Kreutz, Krisztian Balog, Timo Breuer, and Norbert Fuhr. 2024. SIGIR 2024 Workshop on Simulations for Information Access (Sim4IA 2024). In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '24), July 14–18, 2024, Washington, DC, USA.* ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/3626772.3657991

1 MOTIVATION

Evaluating the effectiveness of information access systems (like search engines, recommender systems, or conversational agents) is a timeless and complex scientific task. The common understanding of evaluation is closely coupled to the Cranfield paradigm, the dominating evaluation method in the field to deal with the inherent complexity in the information retrieval context. Cranfield studies are a special form of simulating the search process where implicit and explicit assumptions about the information system and its users are made. These assumptions are helpful for system-oriented experiments as they allow researchers to reduce the complexity of comparing systems. Despite its long history and roots within the community, Cranfield is not uncriticized [12]. The underlying assumptions lead to (over-)simplifications and potentially unrealistic search evaluations. Therefore other evaluation methods, including interactive retrieval settings [14], living labs [10], or (user) simulation studies [5] were proposed and discussed in venues like CHIIR, ECIR, or SIGIR and some of them also tested in shared tasks at TREC or CLEF. While some labs focused on living labs or interactive evaluations (like iCLEF [9], OpenSearch [13], or LiLaS [20]), no shared tasks in TREC/CLEF did primarily focus on user simulations. Only very few labs were concerned with user interactions and their simulation. One example is NewsREEL [15] where so-called replay data was recorded in the online evaluation of the lab to simulate user interactions. Recently, the TREC Interactive Knowledge Assistance Track (iKAT) [1] illustrated the need for evaluating interactive information access systems and the limits of current approaches. Some submissions to iKAT included simulated user feedback in their interactive IA systems, while the lab itself did not employ such an evaluation strategy. These alternative evaluation endeavors aim to enable a more realistic and holistic information access evaluation by including richer user models or more complex representations of the search processes (like sessions).

In addition to evaluating IA systems with simulations, they can also contribute to a better understanding of users. Formalizing a user model for simulation delivers explicit hypotheses on user behavior, which can produce insights into the validity of assumptions on users [5].

One recent example for a re-started interest in the topic of simulation is the Sim4IR workshop that was held at SIGIR 2021, building on past efforts to create "a forum for researchers and practitioners to promote methodology and development of more widespread use of simulation for IR evaluation." Around 80 participants demonstrated

SIGIR '24, July 14–18, 2024, Washington, DC, USA

ACM ISBN 979-8-4007-0431-4/24/07
https://doi.org/10.1145/3626772.3657991

the interest in the topic and the engagement from the community. The main conclusions from the workshop's discussion were "that simulation has the potential to offer solutions to the limitations of existing evaluation methodologies, but there is more research needed toward developing realistic user simulators; and the development and sharing of simulators, in the form of toolkits and online services, is critical for successful uptake" [3].

In this current workshop[1] we dwell deeper into the field by focusing on the user simulation part. This form of simulation recently gained popularity due to available toolkits like SIMIIR 2.0[2] [22] or tutorials on this exact topic (CIKM 2023 [4]). Nevertheless, a specific venue to present and discuss new and experimental approaches is missing. No track or workshop on simulations was present at ECIR or SIGIR 2022/2023. However, a reasonable number of relevant papers on user simulations were accepted, and even a study on simulating user queries won the best paper award at ECIR 2022 [19]. Additionally, since the introduction of generative AI methods into the field, the possibility of integrating LLMs to simulate users has opened up a new chapter.

2 WORKSHOP OVERVIEW

2.1 Goals

To understand how and whether the evaluation of information access technology can truly benefit from simulating user interactions, not only tools and frameworks are critical, but a multidisciplinary discussion and mutual understanding among the broad and sometimes conflicting perspectives is necessary. Simulations have to be aligned to the real-world settings of users and their complex information needs, contexts, and requirements. This workshop should serve as a forum to bring together researchers and practitioners. Additionally, this workshop aims to provide a much-needed forum for the community to discuss the emerging challenges when applying (user) simulations to evaluate information access systems in simulation-based shared tasks. Our goals are to:

- Kick off a series of workshops to generate an open conversation about possible future scenarios, applications, and methods to include simulations in the evaluation of IA systems;
- Provide a forum at SIGIR to discuss the pressing and emerging issues the IR community faces, and how simulations can help to overcome these;
- Produce a SIGIR Forum report detailing the initial outcomes of this first workshop;
- Develop and advertise the idea to organize a TREC/CLEF campaign that includes simulations as a core element;
- Continue this workshop series at subsequent IR conferences to foster further and wider perspectives.

2.2 Topics of Interest

While we would start with a broad definition and a wide range of relevant topics around simulations and how to use these in evaluations of IA systems, a specific focus in this first iteration of the workshop should lie on user simulations. In general, the fields of synthetic queries, user variation, knowledge modeling and conversational information access continue to attract a great deal of interest [1, 6–8, 11, 16–18, 21].

Time	Agenda
9:00–9:30	Welcome
9:30–10:00	Keynote 1
10:00–10:30	Lightning talks
11:00–12:00	Panel discussion
12:00–12:30	Lightning talks
13:30–14:00	Keynote 2
14:00–15:30	Breakout group discussions
16:00–17:00	Reports of the group discussions and closing

Table 1: Tentative schedule for the workshop.

Without limitation, the following topics address the ongoing research interest in the context of user simulation:

- Methods for simulating user actions/behavior
  - Synthetic queries (and their relations to real queries)
  - Simulating browsing and clicking behavior
  - Simulating stopping behavior
  - Simulating rich search interactions
  - Interactions with conversational agents
- Modeling information needs, knowledge, and cognitive states
  - Connecting information needs and the information hole (ASK)
  - Modeling the connection between knowledge and information need
  - Knowledge states beyond a vocabulary
- Modeling contexts and personas
  - Users' contexts—what to model?
  - Personas and simulation
  - First-time usage vs. learning processes
- Building user simulators
  - Toolkits and resources
  - NLP methods and LLMs
- Validation of user simulators
  - Requirements for conversational simulators

2.3 Format and Structure

We anticipate a *highly interactive* and *engaging*, *full-day* workshop to foster collaborations and bring the community together. We are keen to hear different views and opinions on the current and future directions for simulations in information access systems. Therefore, two keynotes, a panel discussion with selected participants, lightning talks, and guided breakout sessions are planned. Table 1 gives a tentative schedule. To enable interaction with a broader set of participants, we offer a limited hybrid participation in addition to onsite attendance via Zoom.

The technical setup for the hybrid participation would consist of a laptop, with a good room microphone but without any speakers. This way we can transfer the onsite discussion into the virtual environment, and enable a back-channel by allowing online attendees to pose questions via chat. An organizer will then moderate and repeat the questions/comments to the onsite audience. We as the

[1] https://sim4ia.org/sigir2024/
[2] https://github.com/padre-lab-eu/simiir-2

organizers will bring our own equipment (laptop and microphone) to be independent from local settings.

Welcome. We welcome the onsite and online participants and set the stage for a day of interaction.

Keynotes. Our two keynotes will be given by Leif Azzopardi[3] (University of Strathclyde) and Gabriella Pasi[4] (University of Milano-Bicocca).

Lightning talks. We allocate short slots (3-5 minute talks) for accepted opinion pieces and encore talks from a selection of different viewpoints for onsite presentation.

Panel discussion. We will moderate a panel with invited panelists from a diverse set of backgrounds with different levels of seniority championing different subfields and views. Our two keynote speakers will be joined by two more panelists.

Breakout group discussions. All attendees of the workshop will be split into smaller groups, each focusing on different topics or questions from the keynotes, panel or lightning talks to foster exchange and multi-faceted discussions. One task for the group work will be the conceptualization of a TREC/CLEF campaign. Other topics could include the specification of personas for evaluations which could become a guideline to refer back to, the extension of pre-existing user models with new components and their incorporation into evaluation measures or the definition and formalization of the connection between knowledge and cognitive states. Breakout groups' discussions will be moderated by some previously generated questions. Ideas, opinions and outcomes of the groups are captured in shared Google docs. Online attendees will be split into online breakout groups. An organizer will be joining these virtual breakouts towards the end of the group discussion to compose a summary of findings to present to the audience in the reporting session.

Reports of the group discussions and closing. Joint discussion and merging of findings or opinions from the breakout group discussions by in-person attendees and organizers for the online attendees. Final notes and discussion of future directions for the community.

2.4 Contribution Type

We welcome contributions of original perspectives or already published work to support exchange as lightning talks (extended abstracts) given by in-person attendees. A lightning talk might present an unusual perspective, rough sketch, fresh idea, research vision or understudied problem. It can explore controversial topics or discuss exploratory solutions but generally sparks a discussion at the workshop. Works that are already published are highly welcome in this format. For a lightning talk, an extended abstract of up to two pages should be submitted, which will be editorially reviewed. Presented talks will be featured in a SIGIR Forum publication for which presenters will be invited as co-authors.

Our selection process of lightning talks will be conducted via EasyChair. Each submission will be reviewed by at least two organizers. The pieces will be evaluated based on criteria such as their novelty, polarization power, significance, visionariness, and clarity.

2.5 Expected Outcomes

Our goal is to strengthen the community, connect researchers in different academic stages, and foster a common understanding of fundamental concepts in the topic of the workshop. As one of the topics for the group discussion is the conceptualization of a TREC/CLEF campaign, this endeavor will be followed up on by the organizers and other interested participants. A write-up report will be submitted to the SIGIR Forum to retain and share the ideas and outcomes of the panel, lightning talks, and breakout group discussions.

3 ORGANIZERS

The following organziers will contribute to this workshop.

Philipp Schaer holds a focus professorship for data science (previously information retrieval) at the Institute of Information Science (TH Köln – University of Applied Sciences, Cologne, Germany). He published at IR-related venues like ECIR, SIGIR, and JCDL on topics like evaluation, digital libraries, and reproducibility. Recently, he was co-program chair of ACM/IEEE JCDL 2022, co-organizer of the CLEF LiLAS lab.

Christin Kreutz is a postdoctoral researcher at TH Köln – University of Applied Sciences, Cologne, Germany. She works on digital libraries and user aspects in IA systems. She recently acquired a two-year funding project to work on users' perspectives on artificially LLM-created and personalized convincing arguments.

Krisztian Balog is a full professor at the University of Stavanger and a staff research scientist at Google. His research focuses on the development and application of information retrieval, natural language processing, and machine learning techniques for intelligent information access, with a current emphasis on conversational systems and user simulation. Balog has co-organized numerous workshops at SIGIR and CIKM (including Sim4IR at SIGIR'21) as well as large-scale benchmarking efforts at TREC and CLEF, and has recently co-authored an FnTIR monograph on user simulation.

Timo Breuer is a postdoctoral researcher interested in reproducible IR evaluations and user-oriented living lab experiments. He sees user simulations as a key element to bridge the gap between these two disciplines. After completing his Ph.D. studies, he was a guest researcher at the Retrieval Group at NIST.

Norbert Fuhr is a senior professor for computer science at the University of Duisburg-Essen, Germany. His past research dealt with topics such as probabilistic retrieval models, the integration of IR and databases, retrieval in distributed digital libraries and XML documents, and user friendly retrieval interfaces. His current research interests are models for interactive retrieval, user-oriented retrieval methods (especially in medicine and online shops), and reproducibility of IR experiments.

4 RELATED WORKSHOPS

Two workshops directly related to IR simulations have been held so far, SIGINT [2] in 2010 and Sim4IR [3] in 2021. SIGINT focused on simulating interactions mainly for interactive IR. The potential of simulations as an evaluation methodology beyond static test collection experiments was explored. Sim4IR further explored the capabilities of simulation-based evaluations. Amongst other topics, it was discussed how accurate evaluations need to reflect human

[3] https://www.strath.ac.uk/staff/azzopardileifdr/

[4] https://ikr3.disco.unimib.it/people/gabriella-pasi/

behavior, how simulated and human judges relate, and how simulations as an evaluation methodology interplay with established methods.

Both workshops view simulations as a potentially invaluable experimental methodology what fosters the need to further progress the field. With steadily progressing research on simulations, it is time to update the research agenda and exchange views to strengthen a common ground for the state of the art. In particular, to reflect the new requirements for simulations raised by emerging search paradigms and take concrete steps towards establishing a simulation-based benchmarking campaign.

ACKNOWLEDGMENTS

This works is partially funded by Deutsch Forschungsgemeinschaft (DFG) under grant number 509543643 (https://gepris.dfg.de/gepris/projekt/509543643?language=en).

REFERENCES

[1] Mohammad Aliannejadi, Zahra Abbasiantaeb, Shubham Chatterjee, Jeffery Dalton, and Leif Azzopardi. 2024. TREC iKAT 2023: The Interactive Knowledge Assistance Track Overview. arXiv:2401.01330 [cs.IR]

[2] Leif Azzopardi, Kalervo Järvelin, Jaap Kamps, and Mark D. Smucker. 2010. Report on the SIGIR 2010 workshop on the simulation of interaction. *SIGIR Forum* 44, 2 (2010), 35–47. https://doi.org/10.1145/1924475.1924484

[3] Krisztian Balog, David Maxwell, Paul Thomas, and Shuo Zhang. 2021. Report on the 1st simulation for information retrieval workshop (Sim4IR 2021) at SIGIR 2021. *SIGIR Forum* 55, 2 (2021), 10:1–10:16. https://doi.org/10.1145/3527546.3527559

[4] Krisztian Balog and ChengXiang Zhai. 2023. Tutorial on User Simulation for Evaluating Information Access Systems. In *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management, CIKM 2023, Birmingham, United Kingdom, October 21-25, 2023*, Ingo Frommholz, Frank Hopfgartner, Mark Lee, Michael Oakes, Mounia Lalmas, Min Zhang, and Rodrygo L. T. Santos (Eds.). ACM, 5200–5203. https://doi.org/10.1145/3583780.3615296

[5] Krisztian Balog and ChengXiang Zhai. 2023. User Simulation for Evaluating Information Access Systems. arXiv:2306.08550 [cs.HC]

[6] Timo Breuer, Norbert Fuhr, and Philipp Schaer. 2022. Validating Simulations of User Query Variants. In *Advances in Information Retrieval - 44th European Conference on IR Research, ECIR 2022, Stavanger, Norway, April 10-14, 2022, Proceedings, Part I (Lecture Notes in Computer Science, Vol. 13185)*, Matthias Hagen, Suzan Verberne, Craig Macdonald, Christin Seifert, Krisztian Balog, Kjetil Nørvåg, and Vinay Setty (Eds.). Springer, 80–94. https://doi.org/10.1007/978-3-030-99736-6_6

[7] Björn Engelmann, Timo Breuer, and Philipp Schaer. 2023. Simulating Users in Interactive Web Table Retrieval. In *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management, CIKM 2023, Birmingham, United Kingdom, October 21-25, 2023*, Ingo Frommholz, Frank Hopfgartner, Mark Lee, Michael Oakes, Mounia Lalmas, Min Zhang, and Rodrygo L. T. Santos (Eds.). ACM, 3875–3879. https://doi.org/10.1145/3583780.3615187

[8] Lukas Gienapp, Harrisen Scells, Niklas Deckers, Janek Bevendorff, Shuai Wang, Johannes Kiesel, Shahbaz Syed, Maik Fröbe, Guido Zuccon, Benno Stein, Matthias Hagen, and Martin Potthast. 2023. Evaluating Generative Ad Hoc Information Retrieval. arXiv:2311.04694 [cs.IR]

[9] Julio Gonzalo, Víctor Peinado, Paul D. Clough, and Jussi Karlgren. 2009. Overview of iCLEF 2009: Exploring Search Behaviour in a Multilingual Folksonomy Environment. In *Multilingual Information Access Evaluation II. Multimedia Experiments - 10th Workshop of the Cross-Language Evaluation Forum, CLEF 2009, Corfu, Greece, September 30 - October 2, 2009, Revised Selected Papers (Lecture Notes in Computer Science, Vol. 6242)*, Carol Peters, Barbara Caputo, Julio Gonzalo, Gareth J. F. Jones, Jayashree Kalpathy-Cramer, Henning Müller, and Theodora Tsikrika (Eds.). Springer, 13–20. https://doi.org/10.1007/978-3-642-15751-6_2

[10] Frank Hopfgartner, Krisztian Balog, Andreas Lommatzsch, Liadh Kelly, Benjamin Kille, Anne Schuth, and Martha Larson. 2019. Continuous Evaluation of Large-Scale Information Access Systems: A Case for Living Labs. In *Information Retrieval Evaluation in a Changing World*, Nicola Ferro and Carol Peters (Eds.). Vol. 41. Springer International Publishing, Cham, 511–543. https://doi.org/10.1007/978-3-030-22948-1_21 Series Title: The Information Retrieval Series.

[11] Zhiyuan Hu, Yue Feng, Anh Tuan Luu, Bryan Hooi, and Aldo Lipani. 2023. Unlocking the Potential of User Feedback: Leveraging Large Language Model as User Simulators to Enhance Dialogue System. In *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management (CIKM '23)*. ACM. https://doi.org/10.1145/3583780.3615220

[12] Peter Ingwersen and Kalervo Järvelin. 2005. *The turn - integration of information seeking and retrieval in context.* Springer, Dordrecht.

[13] Rolf Jagerman, Krisztian Balog, Philipp Schaer, Johann Schaible, Narges Tavakolpoursaleh, and Maarten de Rijke. 2017. Overview of TREC OpenSearch 2017. In *Proceedings of The Twenty-Sixth Text REtrieval Conference, TREC 2017, Gaithersburg, Maryland, USA, November 15-17, 2017 (NIST Special Publication, Vol. 500-324)*, Ellen M. Voorhees and Angela Ellis (Eds.). National Institute of Standards and Technology (NIST). https://trec.nist.gov/pubs/trec26/papers/Overview-O.pdf

[14] Diane Kelly. 2009. Methods for Evaluating Interactive Information Retrieval Systems with Users. *Foundations and Trends in Information Retrieval* 3, 1—2 (2009), 1–224. https://doi.org/10.1561/1500000012

[15] Benjamin Kille, Andreas Lommatzsch, Gebrekirstos G. Gebremeskel, Frank Hopfgartner, Martha A. Larson, Jonas Seiler, Davide Malagoli, András Serény, Torben Brodt, and Arjen P. de Vries. 2016. Overview of NewsREEL'16: Multi-dimensional Evaluation of Real-Time Stream-Recommendation Algorithms. In *Experimental IR Meets Multilinguality, Multimodality, and Interaction - 7th International Conference of the CLEF Association, CLEF 2016, Évora, Portugal, September 5-8, 2016, Proceedings (Lecture Notes in Computer Science, Vol. 9822)*, Norbert Fuhr, Paulo Quaresma, Teresa Gonçalves, Birger Larsen, Krisztian Balog, Craig Macdonald, Linda Cappellato, and Nicola Ferro (Eds.). Springer, 311–331. https://doi.org/10.1007/978-3-319-44564-9_27

[16] Binsheng Liu, Xiaolu Lu, and J. Shane Culpepper. 2021. Strong Natural Language Query Generation. *Inf. Retr.* 24, 4–5 (oct 2021), 322–346. https://doi.org/10.1007/s10791-021-09395-3

[17] Bulou Liu, Yueyue Wu, Fan Zhang, Yiqun Liu, Zhihong Wang, Chenliang Li, Min Zhang, and Shaoping Ma. 2022. Query Generation and Buffer Mechanism: Towards a better conversational agent for legal case retrieval. *Information Processing & Management* 59, 5 (2022), 103051. https://doi.org/10.1016/j.ipm.2022.103051

[18] Jiqun Liu and Ran Yu. 2021. State-Aware Meta-Evaluation of Evaluation Metrics in Interactive Information Retrieval. In *Proceedings of the 30th ACM International Conference on Information & Knowledge Management* (Virtual Event, Queensland, Australia) *(CIKM '21)*. Association for Computing Machinery, New York, NY, USA, 3258–3262. https://doi.org/10.1145/3459637.3482190

[19] Gustavo Penha, Arthur Câmara, and Claudia Hauff. 2022. Evaluating the Robustness of Retrieval Pipelines with Query Variation Generators. In *Advances in Information Retrieval - 44th European Conference on IR Research, ECIR 2022, Stavanger, Norway, April 10-14, 2022, Proceedings, Part I (Lecture Notes in Computer Science, Vol. 13185)*, Matthias Hagen, Suzan Verberne, Craig Macdonald, Christin Seifert, Krisztian Balog, Kjetil Nørvåg, and Vinay Setty (Eds.). Springer, 397–412. https://doi.org/10.1007/978-3-030-99736-6_27

[20] Philipp Schaer, Timo Breuer, Leyla Jael Castro, Benjamin Wolff, Johann Schaible, and Narges Tavakolpoursaleh. 2021. Overview of LiLAS 2021 - Living Labs for Academic Search (Extended Overview). In *Proceedings of the Working Notes of CLEF 2021 - Conference and Labs of the Evaluation Forum, Bucharest, Romania, September 21st - to - 24th, 2021 (CEUR Workshop Proceedings, Vol. 2936)*, Guglielmo Faggioli, Nicola Ferro, Alexis Joly, Maria Maistro, and Florina Piroi (Eds.). CEUR-WS.org, 1668–1699. https://ceur-ws.org/Vol-2936/paper-143.pdf

[21] Shuai Wang, Harrisen Scells, Bevan Koopman, and Guido Zuccon. 2023. Can ChatGPT Write a Good Boolean Query for Systematic Review Literature Search? arXiv:2302.03495 [cs.IR]

[22] Saber Zerhoudi, Sebastian Günther, Kim Plassmeier, Timo Borst, Christin Seifert, Matthias Hagen, and Michael Granitzer. 2022. The SimIIR 2.0 Framework: User Types, Markov Model-Based Interaction Simulation, and Advanced Query Generation. In *Proceedings of the 31st ACM International Conference on Information & Knowledge Management* (Atlanta, GA, USA) *(CIKM '22)*. Association for Computing Machinery, New York, NY, USA, 4661–4666. https://doi.org/10.1145/3511808.3557711