Theory and Toolkits for User Simulation in the Era of Generative AI: User Modeling, Synthetic Data Generation, and System Evaluation

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

Nolwenn Bernard
University of Stavanger
Stavanger, Norway
nolwenn.m.bernard@uis.no

Saber Zerhoudi
University of Passau
Passau, Germany
saber.zerhoudi@uni-passau.de

ChengXiang Zhai
University of Illinois at Urbana-Champaign
Urbana, Illinois, USA
czhai@illinois.edu

Abstract

Interactive AI systems, including search engines, recommender systems, conversational agents, and generative AI applications, are increasingly central to user experiences. However, rigorously evaluating their performance, training them effectively with interaction data, and modeling user behavior for personalization remain significant challenges, often difficult to address reproducibly and at scale. User simulation, which employs intelligent agents to mimic human interaction patterns, offers a powerful and versatile methodology to tackle these interconnected issues. This half-day tutorial provides a comprehensive overview of modern user simulation techniques for interactive AI systems. We will explore the theoretical foundations and practical applications of simulation for system evaluation, algorithm training, and user modeling, emphasizing the crucial connections between these uses. The tutorial covers key simulation methodologies, with a particular focus on recent advancements leveraging large language models, discussing both the opportunities they present and the open challenges they entail. Crucially, we will also provide practical guidance, highlighting relevant toolkits, libraries, and datasets available to researchers and practitioners.

CCS Concepts

• **Information systems** → **Evaluation of retrieval results**; **Users and interactive retrieval**.

Keywords

User simulation, evaluation, interactive information access

ACM Reference Format:
Krisztian Balog, Nolwenn Bernard, Saber Zerhoudi, and ChengXiang Zhai. 2025. Theory and Toolkits for User Simulation in the Era of Generative AI: User Modeling, Synthetic Data Generation, and System Evaluation. In *Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '25), July 13–18, 2025, Padua, Italy.* ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/3726302.3731697

SIGIR '25, Padua, Italy

ACM ISBN 979-8-4007-1592-1/2025/07
https://doi.org/10.1145/3726302.3731697

1 Motivation

Interactive systems, including search engines, recommender systems, conversational agents, and increasingly, generative AI applications, are ubiquitous. However, evaluating their effectiveness, particularly in supporting users through complex tasks, remains a significant scientific challenge, compounded by user variability. Traditional evaluation methodologies like reusable test collections [50], user studies [28], and online A/B testing [23] face limitations, especially in comparing multiple *interactive* systems reproducibly and cost-effectively. Advancements in generative AI introduce further complexities. Effective personalization demands precise user modeling, training robust interactive algorithms requires extensive interaction data, and evaluating these dynamic systems reproducibly at scale is difficult. User simulation, which employs intelligent agents to mimic user interactions, emerges as a powerful approach to address these interconnected challenges [10].

Beyond enabling repeatable and low-cost *evaluation*, simulation provides controlled environments crucial for *training* interactive AI systems and developing nuanced *user models*. These distinct uses are deeply connected: simulated interactions used for evaluation can inform user model development, which in turn can improve the simulators used for training and further evaluation. Simulation thus offers a versatile methodology to gain insights, augment traditional evaluations, and facilitate the responsible development of interactive AI. However, the broader adoption of user simulation techniques is currently hindered by factors such as the scarcity of accessible resources and standardized tools [9], alongside a degree of skepticism concerning the validity of simulation outcomes compared to real user interactions [15]. This tutorial synthesizes relevant research, addressing both (1) the underlying theoretical foundations, with a special focus on elucidating the connections between leveraging user simulation for evaluation, training, and user modeling, and (2) practical application, featuring discussion and guidance on specific toolkits and libraries, aiming to lower the barrier to entry and demonstrate the value of simulation.

2 Objectives

This tutorial aims to equip participants with a solid understanding of the goals, underlying principles, and diverse applications of user simulation within interactive AI, spanning system evaluation, model training, and user modeling. We will cover essential theoretical background, specifically highlighting the connections between these different simulation uses.

The tutorial will provide an overview of key simulation methodologies, paying particular attention to the newest generation of approaches that leverage large language models (LLMs). Furthermore, we will discuss practical resources available to researchers and practitioners, including relevant datasets and software toolkits, to facilitate the adoption of these techniques.

3 Relevance

The evaluation of interactive information access systems like search engines, recommender systems, and conversational assistants remains a core question for the SIGIR community, critical for both industrial practitioners assessing product utility and academic researchers benchmarking algorithmic development. The rapid emergence of Generative AI and LLM-powered interactive systems introduces new layers of complexity and associated evaluation challenges. Furthermore, the relevance of user simulation extends significantly beyond evaluation; it offers methods for generating synthetic data at scale for training data-hungry models and developing sophisticated user models essential for personalization—both central themes for SIGIR. This tutorial aims to provide the necessary synthesis and practical grounding to leverage simulation effectively for these interconnected tasks.

3.1 Previous Editions

An initial version of this tutorial was given at CIKM '23 and SIGIR-AP '23, with a broad coverage of the evaluation of information access systems using simulation in general. Variants of the tutorial have also been given at AAAI '24 and WWW '24, catering specifically to the AI and Web communities, respectively.

In contrast, the present tutorial significantly broadens the scope by explicitly incorporating the use of simulation not just for evaluation, but equally for training interactive systems and enabling advanced user modeling, emphasizing the connections between these applications. Simultaneously, this edition aims to be more practical, featuring dedicated discussion and guidance on specific toolkits, libraries, and datasets to lower the barrier to entry. Reflecting the rapid developments in the field, we place a much stronger emphasis on leveraging LLMs within simulation frameworks, exploring both the novel possibilities they unlock and the challenges associated with their use.

4 Target Audience and Prerequisites

This introductory tutorial primarily targets graduate students, academic researchers, and industry practitioners working on information access or, more broadly, interactive AI systems. We expect our tutorial to be broadly appealing to participants of SIGIR.

5 Format and Schedule

We aim for a **half-day tutorial**, i.e., 3 hours in total, excluding breaks. The contents are organized into the following sections.

- **Background, Motivation, and Context**
  We introduce the landscape of modern interactive AI systems, encompassing search engines, recommender systems, conversational agents, and the rapidly evolving domain of Generative AI. We position simulation as a versatile methodology, grounded in user understanding, capable of addressing the interconnected challenges of evaluation, training, and modeling for today's complex AI, contrasting its role with traditional methods and highlighting its increasing relevance. We briefly discuss historical roots and explorations across various communities: early work investigated its potential in information retrieval [20, 55], while in dialogue systems, simulation has been employed for tasks like policy learning [52]. Fundamentally, user simulation involves creating operational user models, drawing inspiration from decades of research into search intent [16], information seeking behavior [22, 39], cognitive modeling [13, 25], and even economic IR models [4].
- **Foundations of User Simulation: Behavior Models, Formalisms, and Metrics**
  This section establishes the theoretical groundwork for user simulation. We begin with a high-level overview of influential conceptual models of human information-seeking behavior, categorized as *cognitive models* [13, 25, 26], *process models* [30, 35], and *strategic models* [12, 44, 45], which provide guidance for modeling user decisions and processes. We then introduce Markov Decision Processes (MDPs) as a general mathematical framework for formalizing sequential decision-making, widely used for constructing user simulators. Building on this foundation, we demonstrate how simulation principles underpin even traditional IR evaluation, interpreting metrics like Precision, Recall, and NDCG@k as representing implicit, naive user models [27, 63]. Finally, we briefly discuss metrics explicitly derived from simulating user behavior based on assumptions about tasks, interaction patterns, perceived rewards [7, 18, 40–42, 54, 59].
- **Simulating Interactions**
  This section covers the fundamental approaches to modeling user interactions for simulation. We start with models describing interaction workflows (specifying actions, responses, transitions) [11, 38]. We then delve into modeling specific user actions common in search, such as query formulation [3, 17, 32], result scanning [21], clicks [19], perceived effort [54], and stopping decisions [36]. We extend this to conversational contexts, discussing conceptualizations based on intents [6, 46] and dialogue structure [57, 62]. We cover different simulator architectures (modular vs. end-to-end [2, 56]) and established techniques from dialogue systems like agenda-based simulation [51] and sequence-to-sequence models [29, 33], alongside approaches specific to conversational information access [1, 24, 34, 43, 49, 53, 58, 60].
- **Simulation Toolkits and Frameworks**
  Building on the techniques discussed, this section surveys available open-source toolkits, libraries, and frameworks designed to

facilitate the practical implementation of user simulators. We will cover resources relevant for both traditional search/recommendation interaction and conversational systems [1, 2, 5, 8, 24, 33, 37, 43, 47–49, 53, 58, 60, 61]. For each key resource, we will discuss its architecture, scope (models supported, tasks targeted), ease of use, extensibility, and associated datasets or dependencies, aiming to provide attendees with practical starting points for implementing simulators.

- **Simulator Quality: Validation Principles and Methods**
  This section addresses the crucial topic of validation of simulators. We define key requirements for effective simulators, such as validity, fidelity, and interpretability. We then introduce and discuss common validation methodologies used to assess these properties. These include quantitative comparisons against real user interaction logs (e.g., matching distributions of actions, sequence similarity analysis [31]), sensitivity analysis of simulator parameters [14], task-based evaluation (assessing if downstream conclusions hold), and qualitative assessments. We also touch upon the inherent challenges in validating complex simulated behavior [52, 62].
- **Resources for Validation: Benchmarks and Protocols**
  Complementing the discussion of validation methods, this section focuses on the practical resources available for carrying out simulator validation. We will identify and discuss standard benchmark datasets that contain real user interaction logs (e.g., from TREC tracks, public search/session logs, dialogue corpora) suitable for comparative validation studies. We will cover established validation protocols, common metrics used, and recent work on platforms dedicated to the evaluation of simulation approaches.
- **Conclusion and Future Challenges**
  We conclude by highlighting open issues and providing several potential research directions. We discuss how simulation technologies can help foster collaboration between academia and industry. We also argue that some of the major challenges that remain require research from multiple subject areas, including information science, information retrieval, recommender systems, machine learning, natural language processing, knowledge representation, human-computer interaction, and psychology, making user simulation a truly interdisciplinary area for research.

6 Materials

The theory part is based on a recent survey that has been published in Foundations and Trends in Information Retrieval [9]. The https://usersim.ai site features an annotated bibliography of recent literature along with a collection of toolkits and will host the slides.

7 Presenters

Krisztian Balog is a full professor at the University of Stavanger and a staff research scientist at Google DeepMind. His general research interests lie in the use and development of information retrieval, natural language processing, and machine learning techniques for intelligent information access tasks. His current research concerns novel evaluation methodologies, conversational information access, user modeling, transparency, and explainability. Balog regularly serves on the senior programme committee of SIGIR, WSDM, WWW, CIKM, and ECIR. He previously served as general co-chair of ICTIR'20 and ECIR'22, program committee co-chair of ICTIR'19 (full papers), CIKM'21 (short papers), and SIGIR'24 (resource and reproducibility), Associate Editor of ACM Transactions on Information Systems, and coordinator of IR benchmarking efforts at TREC and CLEF. Balog is the recipient of the 2018 Karen Spärck Jones Award. He has previously given tutorials at WWW'13, SIGIR'13, WSDM'14, ECIR'16, SIGIR'19, CIKM'23, AAAI'24, and WWW'24.

Nolwenn Bernard is a final-year PhD student at the University of Stavanger, Norway. Her research specifically focuses on the use of user simulation for the development and evaluation of conversational information access systems. Part of her work involves the development of resources to make user simulation more accessible to the community and support future research in this area. She has published papers at SIGIR, ICTIR, WSDM, and CUI.

Saber Zerhoudi is a final-year PhD student at the University of Passau, Germany. His research focuses on simulating and evaluating user search behavior with interactive information retrieval systems, with applications extending to digital libraries contexts. He has published his work at conferences including SIGIR, CIKM, ECIR, CHIIR, JCDL, and SIGIR-AP. He is one of the main authors behind the SimIIRv2 and SimIIRv3 toolkits.

ChengXiang Zhai is a Donald Biggar Willett Professor in Engineering of Department of Computer Science at the University of Illinois at Urbana-Champaign. His research interests include intelligent information retrieval, text mining, natural language processing, machine learning, and their applications. He serves as a Senior Associate Editor of ACM Transactions on Intelligent Systems and Technology and previously served as Associate Editors of ACM TOIS, ACM TKDD, and Elsevier's IPM, and Program Co-Chair of NAACL-HLT'07, SIGIR'09, and WWW'15. He is an ACM Fellow and a member of the ACM SIGIR Academy. He received the ACM SIGIR Gerard Salton Award and ACM SIGIR Test of Time Award (three times). He has previously given tutorials at HLT-NAACL'04, SIGIR'05, SIGIR'06, HLT-NAACL'07, ICTIR'13, SIGIR'14, KDD'17, SIGIR'17, SIGIR'18, SIGIR'20, SIGIR'21, CIKM'23, AAAI'24, SIGIR-AP'24, and WWW'24.

Acknowledgments

This research was supported by the Norwegian Research Center for AI Innovation, NorwAI (Research Council of Norway, nr. 309834).

References

[1] Zahra Abbasiantaeb, Yifei Yuan, Evangelos Kanoulas, and Mohammad Aliannejadi. 2024. Let the LLMs Talk: Simulating Human-to-Human Conversational QA via Zero-Shot LLM-to-LLM Interactions. In *Proceedings of the 17th ACM International Conference on Web Search and Data Mining (WSDM '24)*. 8–17.
[2] Jafar Afzali, Aleksander Mark Drzewiecki, Krisztian Balog, and Shuo Zhang. 2023. UserSimCRS: A User Simulation Toolkit for Evaluating Conversational Recommender Systems. In *Proc. of WSDM '23*. 1160–1163.
[3] Leif Azzopardi. 2009. Query Side Evaluation: An Empirical Analysis of Effectiveness and Effort. In *Proc. of SIGIR '09*. 556–563.
[4] Leif Azzopardi. 2011. The Economics in Interactive Information Retrieval. In *Proc. of SIGIR '11*. 15–24.
[5] Leif Azzopardi, Timo Breuer, Björn Engelmann, Christin Kreutz, Sean MacAvaney, David Maxwell, Andrew Parry, Adam Roegiest, Xi Wang, and Saber Zerhoudi.

2024. SimIIR 3: A Framework for the Simulation of Interactive and Conversational Information Retrieval. In *Proceedings of the 2024 Annual International ACM SIGIR Conference on Research and Development in Information Retrieval in the Asia Pacific Region (SIGIR-AP 2024)*. 197–202.

[6] Leif Azzopardi, Mateusz Dubiel, Martin Halvey, and Jeffery Dalton. 2018. Conceptualizing Agent-Human Interactions During the Conversational Search Process. In *Proc. of CAIR '18 workshop*.

[7] Leif Azzopardi, Joel Mackenzie, and Alistair Moffat. 2021. ERR is Not C/W/L: Exploring the Relationship Between Expected Reciprocal Rank and Other Metrics. In *Proc. of ICTIR '21*. 231–237.

[8] Leif Azzopardi, Paul Thomas, and Alistair Moffat. 2019. Cwl_eval: An Evaluation Tool for Information Retrieval. In *Proc. of SIGIR '19*. 1321–1324.

[9] Krisztian Balog and ChengXiang Zhai. 2024. User Simulation for Evaluating Information Access Systems. *Foundations and Trends in Information Retrieval* 18, 1-2 (2024), 1–261. https://doi.org/10.1561/1500000098

[10] Krisztian Balog and ChengXiang Zhai. 2025. User Simulation in the Era of Generative AI: User Modeling, Synthetic Data Generation, and System Evaluation. arXiv:2501.04410 [cs.AI]

[11] Feza Baskaya, Heikki Keskustalo, and Kalervo Järvelin. 2013. Modeling Behavioral Factors in Interactive Information Retrieval. In *Proc. of CIKM '13*. 2297–2302.

[12] Marcia J. Bates. 1989. The Design of Browsing and Berrypicking Techniques for the Online Search Interface. *Online Review* 13 (1989), 407–424. Issue 5.

[13] Nicholas J Belkin, Robert N Oddy, and Helen M Brooks. 1982. ASK for Information Retrieval: Part I. Background and Theory. *J. Doc.* (1982).

[14] Timo Breuer, Norbert Fuhr, and Philipp Schaer. 2022. Validating Simulations of User Query Variants. In *Proc. of ECIR '22*. 80–94.

[15] Timo Breuer, Christin Katharina Kreutz, Norbert Fuhr, Krisztian Balog, Philipp Schaer, Nolwenn Bernard, Ingo Frommholz, Marcel Gohsen, Kaixin Ji, Gareth J. F. Jones, Jüri Keller, Jiqun Liu, Martin Mladenov, Gabriella Pasi, Johanne Trippas, Xi Wang, Saber Zerhoudi, and ChengXiang Zhai. 2024. Report on the Workshop on Simulations for Information Access (Sim4IA 2024) at SIGIR 2024. arXiv:2409.18024 [cs.IR]

[16] Andrei Broder. 2002. A Taxonomy of Web Search. *SIGIR Forum* 36, 2 (2002), 3–10.

[17] Ben Carterette, Ashraf Bah, and Mustafa Zengin. 2015. Dynamic Test Collections for Retrieval Evaluation. In *Proc. of ICTIR '15*. 91–100.

[18] Olivier Chapelle, Donald Metlzer, Ya Zhang, and Pierre Grinspan. 2009. Expected Reciprocal Rank for Graded Relevance. In *Proc. of CIKM '09*. 621–630.

[19] Aleksandr Chuklin, Ilya Markov, and Maarten de Rijke. 2015. *Click Models for Web Search*. Morgan & Claypool.

[20] Michael D. Cooper. 1973. A Simulation Model of an Information Retrieval System. *Information Storage and Retrieval* 9, 1 (1973), 13–32.

[21] Nick Craswell, Onno Zoeter, Michael Taylor, and Bill Ramsey. 2008. An Experimental Comparison of Click Position-Bias Models. In *Proc. of WSDM '08*. 87–94.

[22] David Ellis. 1989. A Behavioural Approach to Information Retrieval System Design. *J. Doc.* 45 (1989), 171–212. Issue 3.

[23] Katja Hofmann, Lihong Li, and Filip Radlinski. 2016. Online Evaluation for Information Retrieval. *Found. Trends Inf. Retr.* 10, 1 (2016), 1–117.

[24] Chen Huang, Peixin Qin, Yang Deng, Wenqiang Lei, Jiancheng Lv, and Tat-Seng Chua. 2024. Concept – An Evaluation Protocol on Conversation Recommender Systems with System-centric and User-centric Factors. arXiv:2404.03304 [cs.CL]

[25] Peter Ingwersen. 1996. Cognitive Perspectives of Information Retrieval Interaction: Elements of a Cognitive IR Theory. *J. Doc.* 52 (1996), 3–50. Issue 1.

[26] Peter Ingwersen and Kalervo Järvelin. 2005. *The Turn: Integration of Information Seeking and Retrieval in Context*. Springer.

[27] Shubhra (Santu) K Karmaker, Parikshit Sondhi, and ChengXiang Zhai. 2020. Empirical Analysis of Impact of Query-Specific Customization of NDCG: A Case-Study with Learning-to-Rank Methods. In *Proc. of CIKM '20*. 3281–3284.

[28] Diane Kelly. 2009. Methods for Evaluating Interactive Information Retrieval Systems with Users. *Found. Trends Inf. Retr.* 3, 1–2 (2009), 1–224.

[29] Florian Kreyssig, Iñigo Casanueva, Paweł Budzianowski, and Milica Gašić. 2018. Neural User Simulation for Corpus-based Policy Optimisation of Spoken Dialogue Systems. In *Proc. of SIGDIAL '18*. 60–69.

[30] Carol C. Kuhlthau. 1991. Inside the Search Process: Information Seeking from the User's Perspective. *J. Am. Soc. Inf. Sci. Technol.* 42, 5 (1991), 361–371.

[31] Sahiti Labhishetty and Chengxiang Zhai. 2021. An Exploration of Tester-Based Evaluation of User Simulators for Comparing Interactive Retrieval Systems. In *Proc. of SIGIR '21*. 1598–1602.

[32] Sahiti Labhishetty and ChengXiang Zhai. 2022. PRE: A Precision-Recall-Effort Optimization Framework for Query Simulation. In *Proc. of ICTIR '22*. 51–60.

[33] Hsien-chin Lin, Christian Geishauser, Shutong Feng, Nurul Lubis, Carel van Niekerk, Michael Heck, and Milica Gasic. 2022. GenTUS: Simulating User Behaviour and Language in Task-oriented Dialogues with Generative Transformers. In *Proc. of SIGDIAL '22*. 270–282.

[34] Aldo Lipani, Ben Carterette, and Emine Yilmaz. 2021. How Am I Doing?: Evaluating Conversational Search Systems Offline. *ACM Trans. Inf. Syst.* 39, 4, Article 51 (2021), 22 pages.

[35] Gary Marchionini. 1995. *Information Seeking in Electronic Environments*. Cambridge University Press.

[36] David Maxwell. 2019. *Modelling Search and Stopping in Interactive Information Retrieval*. Ph.D. Dissertation. University of Glasgow.

[37] David Maxwell and Leif Azzopardi. 2016. Simulating Interactive Information Retrieval: SimIIR: A Framework for the Simulation of Interaction. In *Proc. of SIGIR '16*. 1141–1144.

[38] David Maxwell and Leif Azzopardi. 2018. Information Scent, Searching and Stopping: Modelling SERP Level Stopping Behaviour. In *Proc. of ECIR '18*. 210–222.

[39] Lokman I. Meho and Helen R. Tibbo. 2003. Modeling the Information-Seeking Behavior of Social Scientists: Ellis's Study Revisited. *J. Am. Soc. Inf. Sci. Technol.* 54, 6 (2003), 570–587.

[40] Alistair Moffat, Peter Bailey, Falk Scholer, and Paul Thomas. 2017. Incorporating User Expectations and Behavior into the Measurement of Search Effectiveness. *ACM Trans. Inf. Syst.* 35, 3, Article 24 (2017).

[41] Alistair Moffat, Joel Mackenzie, Paul Thomas, and Leif Azzopardi. 2022. A Flexible Framework for Offline Effectiveness Metrics. In *Proc. of SIGIR '22*. 578–587.

[42] Alistair Moffat and Justin Zobel. 2008. Rank-biased Precision for Measurement of Retrieval Effectiveness. *ACM Trans. Inf. Syst.* 27, 1 (2008), 1–27.

[43] Paul Owoicho, Ivan Sekulic, Mohammad Alianejadi, Jeffery Dalton, and Fabio Crestani. 2023. Exploiting Simulated User Feedback for Conversational Search: Ranking, Rewriting, and Beyond. In *Proc. of SIGIR '23*.

[44] Peter Pirolli. 2007. *Information Foraging Theory: Adaptive Interaction with Information*. Oxford University Press.

[45] Peter Pirolli and Stuart Card. 1999. Information Foraging. *Psychological Review* 106 (1999), 643–675. Issue 4.

[46] Filip Radlinski and Nick Craswell. 2017. A Theoretical Framework for Conversational Search. In *Proc. of CHIIR '17*. 117–126.

[47] Zerhoudi Saber and Granitzer Michael. 2024. Generative Agents Navigating Digital Libraries. In *Proc. of ICADL '24'*. 171–188.

[48] Zerhoudi Saber and Granitzer Michael. 2025. SearchLab: Exploring Conversational and Traditional Search Interfaces in Information Retrieval. In *Proc, of CHIIR '25*.

[49] Alexandre Salle, Shervin Malmasi, Oleg Rokhlenko, and Eugene Agichtein. 2021. Studying the Effectiveness of Conversational Search Refinement Through User Simulation. In *Proc. of ECIR '21*. 587–602.

[50] Mark Sanderson. 2010. Test Collection Based Evaluation of Information Retrieval Systems. *Found. Trends Inf. Retr.* 4, 4 (2010), 247–375.

[51] Jost Schatzmann, Blaise Thomson, Karl Weilhammer, Hui Ye, and Steve Young. 2007. Agenda-Based User Simulation for Bootstrapping a POMDP Dialogue System. In *Proc. of NAACL-HLT '07*. 149–152.

[52] Jost Schatzmann, Karl Weilhammer, Matt Stuttle, and Steve Young. 2006. A Survey of Statistical User Simulation Techniques for Reinforcement-Learning of Dialogue Management Strategies. *Knowl. Eng. Rev.* 21, 2 (2006), 97–126.

[53] Ivan Sekulić, Mohammad Aliannejadi, and Fabio Crestani. 2022. Evaluating Mixed-Initiative Conversational Search Systems via User Simulation. In *Proc. of WSDM '22*. 888–896.

[54] Mark D. Smucker and Charles L.A. Clarke. 2012. Time-Based Calibration of Effectiveness Measures. In *Proc. of SIGIR '12*. 95–104.

[55] Karen Spärck Jones. 1979. Search Term Relevance Weighting given Little Relevance Information. *J. Doc.* 35, 1 (1979), 30–48.

[56] Bo-Hsiang Tseng, Yinpei Dai, Florian Kreyssig, and Bill Byrne. 2021. Transferable Dialogue Systems and User Simulators. In *Proc. of ACL '21*. 152–166.

[57] Svitlana Vakulenko, Kate Revoredo, Claudio Di Ciccio, and Maarten de Rijke. 2019. QRFA: A Data-Driven Model of Information-Seeking Dialogues. In *Proc. of ECIR '19*. 541–557.

[58] Xiaolei Wang, Xinyu Tang, Xin Zhao, Jingyuan Wang, and Ji-Rong Wen. 2023. Rethinking the Evaluation for Conversational Recommendation in the Era of Large Language Models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP '23)*. 10052–10065.

[59] Emine Yilmaz, Milad Shokouhi, Nick Craswell, and Stephen Robertson. 2010. Expected Browsing Utility for Web Search Evaluation. In *Proc. of CIKM '10*. 1561–1564.

[60] Se-eun Yoon, Zhankui He, Jessica Echterhoff, and Julian McAuley. 2024. Evaluating Large Language Models as Generative User Simulators for Conversational Recommendation. In *Proc. of NAACL '24*. 1490–1504.

[61] 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 *Proc. of CIKM '22*. 4661–4666.

[62] Shuo Zhang and Krisztian Balog. 2020. Evaluating Conversational Recommender Systems via User Simulation. In *Proc. of KDD '20*. 1512–1520.

[63] Yinan Zhang, Xueqing Liu, and ChengXiang Zhai. 2017. Information Retrieval Evaluation as Search Simulation: A General Formal Framework for IR Evaluation. In *Proc. of ICTIR '17*. 193–200.