Tutorial on User Simulation for Evaluating Information Access Systems on the Web

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

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

ABSTRACT

Users routinely interact with the Web via information access systems such as search engines and recommender systems. How to accurately evaluate such interactive systems with reproducible experiments is an important, yet difficult challenge. To address this challenge, user simulation has emerged as a promising solution. This half-day tutorial focuses on providing a thorough introduction to user simulation techniques designed specifically for evaluating information access systems on the Web. We systematically review major research progress, covering both general frameworks for designing user simulators, and specific models and algorithms for simulating user interactions with search engines, recommender systems, and conversational assistants. We also highlight some important future research directions.

CCS CONCEPTS

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

KEYWORDS

User simulation; Evaluation; Interactive information access

ACM Reference Format:
Krisztian Balog and ChengXiang Zhai. 2024. Tutorial on User Simulation for Evaluating Information Access Systems on the Web. In *Companion Proceedings of the ACM Web Conference 2024 (WWW '24 Companion), May 13–17, 2024, Singapore, Singapore.* ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/3589335.3641243

1 TOPIC AND RELEVANCE

Although Web information access systems, such as search engines, recommender systems, and conversational assistants, are used by millions on a daily basis, how to appropriately evaluate those systems remains an open scientific challenge. For example, the weak correlation of online and offline evaluation results makes it hard to choose the best algorithm to deploy in a production environment, while inaccurate evaluation of algorithms would result in misleading conclusions and thus hinder progress in research.

WWW '24 Companion, May 13–17, 2024, Singapore, Singapore

ACM ISBN 979-8-4007-0172-6/24/05…$15.00
https://doi.org/10.1145/3589335.3641243

The emergence of large language models (LLMs) such as ChatGPT makes information access on the Web increasingly more interactive and conversational. It is especially challenging to evaluate an *interactive* system's overall effectiveness in helping a user finish a task via interactive support, because the utility of such a system can only be assessed by a user interacting with the system. Moreover, the fact that users vary significantly in terms of their behaviour and preferences makes it very difficult to perform system evaluation with reproducible experiments.

There are three widely-used evaluation methodologies for information access systems: reusable test collections [65], user studies [37], and online evaluation [31]. However, none of these methodologies can be used to compare multiple interactive information access systems (in terms of their overall effectiveness in supporting users) using reproducible experiments; the test collection-based approach is static in nature, while there is an inherent lack of reproducibility when real users are involved. User simulation has the potential to enable repeatable and reproducible evaluations at low cost, without using invaluable user time (human assessor time or online experimentation bandwidth). Further, simulation can augment traditional evaluation methodologies by providing insights into how system performance changes under different conditions and user behaviour. Relevant research work, however, is scattered across multiple research communities, including information retrieval, recommender systems, dialogue systems, and user modeling. This tutorial aims to synthesize this extensive body of research into a coherent framework with a focus on applications of user simulation to evaluate Web information access systems.

2 CONTENT AND ORGANIZATION

The contents are organized into the following sections.

- **Background: Evaluation of Web Information Access Systems** *[20 min]*
  We first describe the spectrum of Web information access tasks. Next, we briefly discuss the goals of evaluation and general methodologies of evaluation (reusable test collections, user studies, and online evaluation). We then highlight the challenges involved in evaluating Web information access systems and how user simulation can help address those challenges.
- **Overview of User Simulation** *[15 min]*
  This part provides a brief historical account on the use of simulation techniques, and highlight how various research communities focused on different but complementary areas of evaluation and user simulation. This includes early work on simulation in information retrieval [21, 29, 71] and studies in interactive information retrieval pointing out discrepancies between interactive

and non-interactive evaluation results [30, 69, 74]. In dialogue systems research, simulation-based techniques have been used for dialogue policy learning [67, 78], and to a limited extent also for evaluation [26]. User simulation can be regarded as developing a complete and operational user model, which makes work on search tasks and intent [16, 48], information seeking models [25, 54], cognitive models of users [13, 32], and economic IR models [3, 4] highly relevant to us.

- **Simulation-based Evaluation Frameworks** *[25 min]*
  We make the key observation that traditional evaluation measures used in IR and Web search may be viewed as naive user simulators, and discuss how to interpret Precision, Recall, and NDCG@k from an user simulation perspective [36, 82]. Next, we discuss metrics based on explicit models of user behavior, based on (1) the assumed user task, (2) the assumed user behavior when interacting with results, (3) the measurement of the reward a user would receive from examining a result, and (4) the measurement of the effort a user would need to make in order to receive the reward. Specifically, we cover the RBP [57], ERR [19], EBU [77], and the time-biased gain [70] measures, as well as the more general frameworks of C/W/L [7, 55], C/W/L/A [56], and the model-based framework by Carterette [17]. Finally, we present a general simulation-based evaluation framework [82] and the Interface Card Model [83], which can be used to evaluate an interactive information access system with a computationally generated dynamic browsing interface using user simulation.
- **User Simulation and Human Decision-making** *[15 min]*
  In this part, we provide a high-level overview of research on conceptual models that can provide theoretical guidance for modeling processes and decisions from an individual's perspective. We cover models of search behavior within three main categories: (1) *cognitive models*, focusing on the cognitive processes underlying the information-seeking activity [13, 32, 33], (2) *process models*, representing the different stages and activities during the search process [39, 47], and (3) *strategic models*, describing tactics that users employ when searching for information [12, 60, 61]. Then, we discuss how to model decision-making processes mathematically using Markov decision processes (MDP). The MDP framework provides a general formal framework for constructing user simulators, which we will use to discuss specific user simulation techniques in the next two sections.
- **Simulating Interactions with Search and Recommender Systems** *[45 min]*
  We start by presenting models that describe interaction workflows, that is, specify the space of user actions and system responses, and possible transitions between them [11, 52, 53]. Then, we discuss specific user actions: query formulation [2, 5, 10, 18, 35, 41], scanning behavior [22, 23, 27], clicks [20, 27, 34, 84], effort involved in processing documents [18, 50, 70, 85], and stopping [49, 52, 53, 58, 76]. The connection between the extensive work on click modeling for Web search and user simulation is discussed. We also provide an overview of toolkits and resources [8, 51, 79] and discuss approaches to validating simulators [15, 40, 42].
- **Simulating Interactions with Conversational Assistants** *[30 min]*
  We begin with a conceptualization of conversational information access in terms of intents [6, 63, 72] and dialogue structure [14, 46, 62, 62, 75, 80, 81], and discuss two fundamentally different simulator architectures: modular [1] and end-to-end systems [43, 73]. There is a solid body of work within dialogue systems research on simulating user decisions to build on, including the widely used agenda-based simulation [66] and more recent sequence-to-sequence models [24, 28, 38, 43, 44]. This is followed by the discussion of simulation approaches developed specifically for conversational information access [45, 59, 64, 68, 80, 81]. We review toolkits and resources [1, 14, 59, 64, 68], followed by a discussion on how simulators themselves can be evaluated [67, 80].
- **Conclusion and Future Challenges** *[15 min]*
  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.
- **Discussion** *[15 min]*
  We dedicate the last bit of the tutorial to open-ended discussion and feedback from participants.

3 AUDIENCE

Since the question of how to accurately evaluate a search engine, a recommender system, or a conversational assistant is important to both practitioners who would like to assess the utility of their product systems and researchers who would like to know whether their new algorithms are truly more effective than the existing ones, we expect our tutorial to be broadly appealing to many participants of the Web Conference, including undergraduate and graduate students, academic and industry researchers, practitioners from the industry, and government policy/decision makers. As the tutorial is mostly self-contained with only minimum pre-required background knowledge, it is expected to be accessible to most attendants of the Web Conference.

Participants of the tutorial can expect to learn what user simulation is, why it is important to use it for evaluation, how existing user simulation techniques can already be useful for evaluating interactive Web information access systems, how to develop new user simulators, and how to use user simulation broadly to evaluate assistive AI systems. They can also expect to learn about associated challenges and where additional research is still needed.

4 PREVIOUS EDITIONS

An initial version of this tutorial was given at the 32nd ACM International Conference on Information and Knowledge Management (CIKM '23) in October, 2023 in Birmingham, UK, with a broad coverage of the evaluation of information access systems using simulation in general. The same initial version of the tutorial was

also given at the 1st International ACM SIGIR Conference on Information Retrieval in the Asia Pacific (SIGIR-AP '23), mainly for the Asian audience in November, 2023 in Beijing, China. The present tutorial, given at the Web Conference, is based on this initial version, undergoing significant customization to cater specifically to the conference's audience. This includes an expanded focus on relevant topics, such as the background of Web information access systems, the connection between Web click modeling and user simulation, and the application of simulation techniques in e-commerce.

A variant of the tutorial has also been given at the 38th Annual AAAI Conference on Artificial Intelligence (AAAI '24) in February in Vancouver, Canada. That edition adopts a broader perspective of user simulation for evaluating an interactive AI system and focuses more on simulation algorithms and techniques that are well connected with various sub-fields of AI, such as machine learning and agent-based systems. In contrast, the present tutorial emphasizes more on applications of user simulation for evaluating Web information access systems.

5 TUTORIAL MATERIALS

The tutorial is based on a survey that is currently under review at Foundations and Trends in Information Retrieval; a preprint is available at [9]. There is a companion website to the survey, https://usersim.ai, which also hosts the slides for the tutorial.

6 PRESENTERS

Krisztian Balog is a full professor at the University of Stavanger and a staff research scientist at Google. 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, and conversational and explainable search and recommendation methods. 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, and AAAI'24.

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, and AAAI'24.

REFERENCES

[1] 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.

[2] Leif Azzopardi. 2009. Query Side Evaluation: An Empirical Analysis of Effectiveness and Effort. In *Proc. of SIGIR '09*. 556–563.

[3] Leif Azzopardi. 2011. The Economics in Interactive Information Retrieval. In *Proc. of SIGIR '11*. 15–24.

[4] Leif Azzopardi. 2014. Modelling Interaction with Economic Models of Search. In *Proc. of SIGIR '14*. 3–12.

[5] Leif Azzopardi, Maarten de Rijke, and Krisztian Balog. 2007. Building Simulated Queries for Known-Item Topics: An Analysis Using Six European Languages. In *Proc. of SIGIR '07*. 455–462.

[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. 2023. User Simulation for Evaluating Information Access Systems. arXiv:2306.08550 [cs.HC]

[10] Feza Baskaya, Heikki Keskustalo, and Kalervo Järvelin. 2012. Time Drives Interaction: Simulating Sessions in Diverse Searching Environments. In *Proc. of SIGIR '12*. 105–114.

[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] Nolwenn Bernard and Krisztian Balog. 2023. MG-ShopDial: A Multi-Goal Conversational Dataset for e-Commerce. In *Proc. of SIGIR '23*.

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

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

[17] Ben Carterette. 2011. System Effectiveness, User Models, and User Utility: A Conceptual Framework for Investigation. In *Proc. of SIGIR '11*. 903–912.

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

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

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

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

[22] 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.

[23] Georges E. Dupret and Benjamin Piwowarski. 2008. A User Browsing Model to Predict Search Engine Click Data from Past Observations. In *Proc. of SIGIR '08*. 331–338.

[24] Layla El Asri, Jing He, and Kaheer Suleman. 2016. A Sequence-to-Sequence Model for User Simulation in Spoken Dialogue Systems. In *Proc. of Interspeech '16*. 1151–1155.

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

[26] David Griol, Javier Carbó, and José M. Molina. 2013. An Automatic Dialog Simulation Technique to Develop and Evaluate Interactive Conversational Agents. *Applied Artificial Intelligence* 27, 9 (2013), 759–780.

[27] Fan Guo, Chao Liu, Anitha Kannan, Tom Minka, Michael Taylor, Yi-Min Wang, and Christos Faloutsos. 2009. Click Chain Model in Web Search. In *Proc. of WWW '09*. 11–20.

[28] Izzeddin Gür, Dilek Hakkani-Tür, Gokhan Tür, and Pararth Shah. 2018. User Modeling for Task Oriented Dialogues. In *Proc. of SLT '18 workshop*. 900–906.

[29] Donna Harman. 1992. Relevance Feedback Revisited. In *Proc. of SIGIR '92*. 1–10.

[30] William Hersh, Andrew Turpin, Susan Price, Benjamin Chan, Dale Kramer, Lynetta Sacherek, and Daniel Olson. 2000. Do Batch and User Evaluations Give the Same Results?. In *Proc. of SIGIR '00*. 17–24.

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

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

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

[34] Thorsten Joachims, Laura Granka, Bing Pan, Helene Hembrooke, Filip Radlinski, and Geri Gay. 2007. Evaluating the Accuracy of Implicit Feedback from Clicks and Query Reformulations in Web Search. *ACM Trans. Inf. Syst.* 25, 2 (2007).

[35] Chris Jordan, Carolyn Watters, and Qigang Gao. 2006. Using Controlled Query Generation to Evaluate Blind Relevance Feedback Algorithms. In *Proc. of JCDL '06*. 286–295.

[36] 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.

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

[38] 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.

[39] 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.

[40] 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.

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

[42] Sahiti Labhishetty and ChengXiang Zhai. 2022. RATE: A Reliability-Aware Tester-Based Evaluation Framework of User Simulators. In *Proc. of ECIR '22*. 336–350.

[43] 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.

[44] Hsien-chin Lin, Nurul Lubis, Songbo Hu, Carel van Niekerk, Christian Geishauser, Michael Heck, Shutong Feng, and Milica Gasic. 2021. Domain-independent User Simulation with Transformers for Task-oriented Dialogue Systems. In *Proc. of SIGDIAL '21*. 445–456.

[45] 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.

[46] Shengnan Lyu, Arpit Rana, Scott Sanner, and Mohamed Reda Bouadjenek. 2021. A Workflow Analysis of Context-Driven Conversational Recommendation. In *Proc. of WWW '21*. 866–877.

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

[48] Gary Marchionini. 2006. Exploratory Search: From Finding to Understanding. *Commun. ACM* 49, 4 (2006), 41–46.

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

[50] David Maxwell and Leif Azzopardi. 2016. Agents, Simulated Users and Humans: An Analysis of Performance and Behaviour. In *Proc. of CIKM '16*. 731–740.

[51] 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.

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

[53] David Maxwell, Leif Azzopardi, Kalervo Järvelin, and Heikki Keskustalo. 2015. Searching and Stopping: An Analysis of Stopping Rules and Strategies. In *Proc. of CIKM '15*. 313–322.

[54] 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.

[55] 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).

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

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

[58] Kathryn Ritgerod Nickles. 1995. *Judgment-based and reasoning-based stopping rules in decision making under uncertainty*. Ph.D. Dissertation. University of Minnesota.

[59] 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*.

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

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

[62] Chen Qu, Liu Yang, W. Bruce Croft, Johanne R. Trippas, Yongfeng Zhang, and Minghui Qiu. 2018. Analyzing and Characterizing User Intent in Information-Seeking Conversations. In *Proc. of SIGIR '18*. 989–992.

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

[64] 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.

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

[66] 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.

[67] 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.

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

[69] Catherine L. Smith and Paul B. Kantor. 2008. User Adaptation: Good Results from Poor Systems. In *Proc. of SIGIR '08*. 147–154.

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

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

[72] Johanne R. Trippas, Damiano Spina, Lawrence Cavedon, Hideo Joho, and Mark Sanderson. 2018. Informing the Design of Spoken Conversational Search: Perspective Paper. In *Proc. of CHIIR '18*. 32–41.

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

[74] Andrew H. Turpin and William Hersh. 2001. Why Batch and User Evaluations Do Not Give the Same Results. In *Proc. of SIGIR '01*. 225–231.

[75] 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.

[76] J. Frank Yates. 1990. *Judgment and decision making*. Prentice Hall Englewood Cliffs, N.J.

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

[78] Steve Young, Milica Gašić, Simon Keizer, FranÁois Mairesse, Jost Schatzmann, Blaise Thomson, and Kai Yu. 2010. The Hidden Information State Model: A Practical Framework for POMDP-based Spoken Dialogue Management. *Computer Speech & Language* 24, 2 (2010), 150–174.

[79] 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.

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

[81] Shuo Zhang, Mu-Chun Wang, and Krisztian Balog. 2022. Analyzing and Simulating User Utterance Reformulation in Conversational Recommender Systems. In *Proc. of SIGIR '22*. 133–143.

[82] 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.

[83] Yinan Zhang and Chengxiang Zhai. 2015. Information Retrieval as Card Playing: A Formal Model for Optimizing Interactive Retrieval Interface. In *Proc. of SIGIR '15*. 685–694.

[84] Guorui Zhou, Xiaoqiang Zhu, Chenru Song, Ying Fan, Han Zhu, Xiao Ma, Yanghui Yan, Junqi Jin, Han Li, and Kun Gai. 2018. Deep Interest Network for Click-Through Rate Prediction. In *Proc. of KDD '18*. 1059–1068.

[85] Guido Zuccon. 2016. Understandability Biased Evaluation for Information Retrieval. In *Proc. of ECIR '16*. 280–292.