User Simulation for Evaluating Information Access Systems

Other titles in Foundations and Trends® in Information Retrieval

Multi-hop Question Answering
Vaibhav Mavi, Anubhav Jangra and Adam Jatowt
ISBN: 978-1-63828-374-4

Conversational Information Seeking
Hamed Zamani, Johanne R. Trippas, Jeff Dalton and Filip Radlinski
ISBN: 978-1-63828-200-6

Perspectives of Neurodiverse Participants in Interactive Information Retrieval
Laurianne Sitbon, Gerd Berget and Margot Brereton
ISBN: 978-1-63828-202-0

Efficient and Effective Tree-based and Neural Learning to Rank
Sebastian Bruch, Claudio Lucchese and Franco Maria Nardini
ISBN: 978-1-63828-198-6

User Simulation for Evaluating Information Access Systems

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

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

the essence of knowledge
Boston — Delft

Foundations and Trends® in Information Retrieval

Published, sold and distributed by:
now Publishers Inc.
PO Box 1024
Hanover, MA 02339
United States
Tel. +1-781-985-4510
www.nowpublishers.com
sales@nowpublishers.com

Outside North America:
now Publishers Inc.
PO Box 179
2600 AD Delft
The Netherlands
Tel. +31-6-51115274

The preferred citation for this publication is

K. Balog and C. Zhai. *User Simulation for Evaluating Information Access Systems*. Foundations and Trends® in Information Retrieval, vol. 18, no. 1-2, pp. 1–261, 2024.

ISBN: 978-1-63828-379-9

Full text available at: http://dx.doi.org/10.1561/1500000098

Foundations and Trends® in Information Retrieval
Volume 18, Issue 1-2, 2024

Editorial Scope

Foundations and Trends® in Information Retrieval publishes survey and tutorial articles in the following topics:

- Applications of IR
- Architectures for IR
- Collaborative filtering and recommender systems
- Cross-lingual and multilingual IR
- Distributed IR and federated search
- Evaluation issues and test collections for IR
- Formal models and language models for IR
- IR on mobile platforms
- Indexing and retrieval of structured documents
- Information categorization and clustering
- Information extraction
- Information filtering and routing
- Metasearch, rank aggregation and data fusion
- Natural language processing for IR
- Performance issues for IR systems, including algorithms, data structures, optimization techniques, and scalability
- Question answering
- Summarization of single documents, multiple documents, and corpora
- Text mining
- Topic detection and tracking
- Usability, interactivity, and visualization issues in IR
- User modelling and user studies for IR
- Web search

Information for Librarians

Foundations and Trends® in Information Retrieval, 2024, Volume 18, 5 issues. ISSN paper version 1554-0669. ISSN online version 1554-0677. Also available as a combined paper and online subscription.

Contents

User Simulation for Evaluating Information Access Systems

Krisztian Balog[1] and ChengXiang Zhai[2]

[1] *University of Stavanger, Norway; krisztian.balog@uis.no*
[2] *University of Illinois at Urbana-Champaign, USA; czhai@illinois.edu*

ABSTRACT

Information access systems, such as search engines, recommender systems, and conversational assistants, have become integral to our daily lives as they help us satisfy our information needs. However, evaluating the effectiveness of these systems presents a long-standing and complex scientific challenge. This challenge is rooted in the difficulty of assessing a system's overall effectiveness in assisting users to complete tasks through interactive support, and further exacerbated by the substantial variation in user behaviour and preferences. To address this challenge, user simulation emerges as a promising solution.

This monograph focuses on providing a thorough understanding of user simulation techniques designed specifically for evaluation purposes. We begin with a background of information access system evaluation and explore the diverse applications of user simulation. Subsequently, we systematically review the major research progress in user simulation, covering both general frameworks for designing user simulators, utilizing user simulation for evaluation, and specific models and algorithms for simulating user interactions with search engines, recommender systems, and conversational

---

Krisztian Balog and ChengXiang Zhai (2024), "User Simulation for Evaluating Information Access Systems", Foundations and Trends® in Information Retrieval: Vol. 18, No. 1-2, pp 1–261. DOI: 10.1561/1500000098.

assistants. Realizing that user simulation is an interdisciplinary research topic, whenever possible, we attempt to establish connections with related fields, including machine learning, dialogue systems, user modeling, and economics. We end the monograph with a broad discussion of important future research directions, many of which extend beyond the evaluation of information access systems and are expected to have broader impact on how to evaluate interactive intelligent systems in general.

1

Introduction

Information access systems, such as search engines, recommender systems, and conversational assistants, have become increasingly intelligent in understanding users' intents, supporting their tasks, and interacting with them using natural language dialogue, thanks to recent progress in research in artificial intelligence (AI), especially in machine learning and natural language processing. These information access systems are used by millions on a daily basis to perform a wide range of tasks where humans need help to find information relevant to a task. In general, the interactions with these systems involve a user entering information needs or preferences (by typing queries, rating items, or asking natural language questions) and interacting with information objects (by clicking, typing, or speaking) that are presented by the system on some device (e.g., desktop, laptop, tablet, smart phone, or smart speaker) in some modality or combination of modalities (e.g,. text, rich snippets, voice). With intelligent home devices becoming available, information access systems may also be used to power a wide range of intelligent agent systems, such as Google Home or Alexa, which can go beyond supporting information access to also support other user tasks (e.g., controlling home appliances, making appointments, or placing orders).

Although information access systems have already become useful products for people, how to appropriately evaluate those systems remains an open scientific challenge. It is especially challenging to evaluate a system's overall effectiveness in helping a user finish a task via interactive support. The fact that users vary significantly in terms of their behaviour and preferences makes evaluation even more difficult. As a promising strategy for evaluating information access systems using reproducible experiments, user simulation has attracted much attention recently. In this monograph, we review the recent progress in this area with a focus on user simulation for evaluating information access systems.

In the rest of this section, we first describe the spectrum of information access tasks. Next, we briefly discuss the goals of evaluation and general methodologies of evaluation. We then highlight the challenges involved in evaluating information access systems and how user simulation can help address those challenges. Finally, we describe the aims and scope of this monograph.

1.1 Information Access Tasks

Information access refers to the ability to identify, retrieve, and use information effectively.[1] Access to the right information at the right time plays an important role in everyone's life and is vital to business operations. At a high-level, information access can happen in two modes (Zhai and Massung, 2016): (1) *pull mode*, where the user takes the initiative and uses a search engine to find needed information ("pull" relevant information to the user), and (2) *push mode*, where the system takes the initiative and recommends relevant information to a user ("push" relevant information to the user). The two modes can be naturally mixed in conversational AI systems, which are increasingly common due to the emergence of large language models (LLMs) (McTear and Ashurkina, 2024).

Search engines and recommender systems are the two most common widely used applications for information access. The two modes of information access are complementary and often supported simultane-

[1] https://www.encyclopedia.com/computing/news-wires-white-papers-and-books/information-access

ously using a single system. For example, a search engine can not only support querying (pull mode) but also recommend related information to the user (push mode). Similarly, a recommender system may also recommend information in the form of a ranked list to enable a user to further interact with the recommended information and potentially enter a query to further explore the information space. Indeed, search and recommendation have been suggested as "two sides of the same coin" in Belkin and Croft (1992). As such, it is not surprising that search engines and recommender systems share many common technical challenges (e.g., modeling a user's information need and preferences, matching an information item with a user's interest, ranking items accurately, learning from a user's feedback information, evaluating a ranked list to assess its utility to a user), and tend to benefit from using similar techniques, including user simulation techniques. For this reason, we intend to cover the topic of user simulation in the broad context of information access with the understanding that most discussions are relevant to both search engines and recommender systems, even though the actual research work that we discuss may have been done for either just search engines or recommender systems.

Recently, conversational assistants have attracted much attention (McTear, 2021). They generally support mixed-initiative interaction via natural language to facilitate both search and recommendation in the same information access session (Zamani *et al.*, 2023). Compared with traditional search engines and recommender systems, where the actions a user could potentially take are well specified by the user interface of the system, conversational assistants have more open-ended functions in the sense that a user can potentially ask questions, provide clarifications, and explore related topics using unrestricted natural language, thus adding complexity to user simulation. We note that conversational assistants, casually referred to as "conversational AI," can cater to diverse user goals, including, e.g., social chatting. However, our primary focus in this monograph is on systems that are designed to support information access, i.e., tasks where there is an underlying information need and the system returns information objects (which may be documents, entities, answers, utterances, etc.). This focus increases the commonality between conversational assistants, search engines, and recommender systems in terms of user simulation.

1.2 Evaluation Methodologies

There are three widely-used evaluation methodologies for information access systems: reusable test collections (Sanderson, 2010), user studies (Kelly, 2009), and online evaluation (Hofmann *et al.*, 2016).

Reusable test collections (a.k.a. *offline evaluation*) facilitate large-scale automatic evaluation and have been invaluable for comparing different algorithms and improving the state of the art. They ensure repeatability and enable comparison between different approaches and study of effectiveness of individual components within complex methods. However, they are static and are severely limited in their ability to capture many aspects of users and interactions adequately. They measure system performance on an abstraction of a given process (e.g., search or recommendation), where the user is also abstracted away. Specifically, they are based on simplified models of the information access process and user behaviour. Examples include the assumption that a system always presents a ranked list of results to a user or that a user can always recognize whether a document is relevant in a search result list. This has so far been the standard evaluation methodology for making relative comparisons between two systems in a repeatable and reproducible manner. However, it is generally not so useful for the purpose of evaluating the actual utility of a system due to the significant deviation between the evaluation environment and the real-world application and the very limited inclusion of users. It is also generally hard or impossible to develop test collections for evaluating an *interactive* system, a limitation that can be addressed by user simulation.

User studies provide the highest fidelity in terms of capturing real users' interactions with an actual system in a controlled setting. However, experiments that involve real users are costly to run. Further, if multiple rounds of experimentation are performed, new users need to be involved in order to avoid misinterpretation of the findings due to fatigue and learning effects. Also, experiments that involve an actual service have a bandwidth limit, which is set by the amount of users and their activity using that service. User studies are useful for assessing the actual utility of a system, but they suffer from several limitations. First, the result is generally not reproducible (even the same user would not behave in the same way when repeating an experiment due to learning effects). Thus, they have limited value for making relative comparisons between systems, especially when new systems—to be developed in the future—need to be included in the comparison. Second, the cost of recruitment is often high, and it is a challenge to recruit enough people from the right population. Major industry labs often have to invest significantly in recruiting users and conducting user studies, which smaller companies typically cannot afford. In academia, only a few examples of user studies of reasonable scale involve participants beyond university students (Brennan *et al.*, 2014).

Online evaluation (a.k.a. log-based studies) is based on the idea of observing real users of a fully operational system and assessing the system's performance by analyzing the recorded user behaviour. For instance, A/B tests can be used to evaluate different versions of a system. Online evaluation is widely used by companies that deploy real-world applications or services, and is regarded as the most direct and reliable measurement of quality and user experience. Like user studies, online evaluation enables measuring the actual utility of a system and comparing systems with real users, but at a much larger scale in terms of the number of users used for evaluation. Unlike user studies, however, it generally does not provide control over the users, making it harder to interpret the results. As in the case of user studies, online evaluation suffers from being not reproducible and thus cannot be "reused" to compare different systems or analyze the effectiveness of various components. Another limitation of online evaluation is that the user interactions to be evaluated are limited to natural interactions with the system, thus it cannot accommodate counterfactual evaluation, i.e., where the potential outcomes of a hypothetical system (e.g., one with a new algorithm) and being compared to those of an existing system. This hypothetical system cannot be used by real users, making online evaluation infeasible. Additionally, there is a risk of leaving a negative impression on users about a production system's performance if a system to be evaluated online turns out to perform poorly.

1.3 Challenges in Evaluating Information Access Systems and Simulation-Based Evaluation

In general, all the three methodologies discussed earlier can be applied to, and indeed have been regularly used for, evaluating information access systems. However, none of those methodologies can be used to compare multiple interactive information access systems (in terms of their overall effectiveness in supporting users) using reproducible experiments due to the static nature of the test collection-based approach and the lack of reproducibility when real users are involved. These limitations can be addressed by using user simulation: simulated users can be controlled and thus enable reproducible experiments.

Note that the existing test collection-based evaluation methodology (Sanderson, 2010) can be viewed as a simple form of user simulation. This means we are already utilizing user simulation, albeit implicitly, without explicitly articulating what kind of users are being simulated. A key advantage of evaluation based on user simulation is to make assumptions about simulated users and their behaviour more explicit, while modeling a broader range of user actions than current measures consider (see Section 4.3 for more discussion on this).

We want to highlight the importance of evaluating the *overall effectiveness* of a system (not just various components of a system), especially in the case of information access systems. This is because users are likely to benefit the most from intelligent assistance in case of complex tasks. Commonly, complex tasks are decomposed into a series of smaller and simpler components. The decomposition process generally involves close collaboration between a user and a system in an interactive way, in which a user would iteratively direct the system to perform specific component functions and the results from multiple steps can be synthesized to generate solutions to a complex problem. Most component-level tasks can be abstracted, studied and addressed in isolation, and evaluated using the reusable test collection methodology with reproducible experiments. While this is clearly necessary to allow for systematic progress to be made, evaluation of individual components alone is insufficient. It is arguably more important to view them as components of a larger information access system and study how to evaluate the *whole* system from a user's perspective. Indeed, the ultimate goal of evaluation is to measure how well the user is aided in achieving their end goal. It is vitally important to do this evaluation correctly; if not done appropriately, it would mislead us to draw wrong conclusions or deploy an inferior application system that negatively impacts the user experience.

1.4 User Simulation

What do we exactly mean by user simulation? Informally, user simulation is to have an intelligent agent simulate how a user interacts with a system. The agent can be built based on models/algorithms/rules and any knowledge we have about the user (their behaviour, knowledge, etc.). The agent can also have parameters that can be varied in a meaningful way to simulate variations of users. Once a user simulator is constructed, it can then be used to interact with any system that needs to be evaluated. In turn, we can measure the system's utility based on the observed behaviour of the agent while interacting with the system. Simulation thus has the potential to enable repeatable and reproducible evaluations at a low cost, without using invaluable user time (human assessor time or online experimentation bandwidth). Further, simulation can augment traditional evaluation methodologies by offering possibilities to gain insights into how system performance changes under different conditions and user behaviour.

1.4.1 Problem Definition

User simulation is the process of modeling a user's behaviour and decision-making patterns within an interactive system, specifically designed to mimic and predict how a user will act in various interaction contexts or scenarios related to completing a task. To effectively simulate a user's behaviour within an interactive system, configuration variables that influence this behaviour must be defined:

- Task (T): A user's behaviour varies according to nature of the user's task. Tasks vary in complexity, and different tasks require different types and levels of interaction, decision-making processes, and completion strategies.

- System (S): A user's behaviour depends on the system they interact with. This includes the system's functionality, user interface, and overall usability and support for task goals. It is the system that dictates the types of possible actions, denoted as $\mathcal{A}$, that a user can perform at any given point during their interactions.

- User information (U): Different users may behave differently when completing the same task using the same system. Simulations must account for variations in individual user characteristics such as age, technical proficiency, preferences, and cognitive styles.

With these variables defined, the task of user simulation can be stated as the following computational problem:

> Given the variables T, S, and U, the goal is to create an agent that can simulate every action that user U may take when attempting to complete task T using system S.

This problem involves developing a computational model that can dynamically generate user actions, reflecting the behavioural patterns and decision-making processes of a user, based on a specific task context. Formally, we define the computational model as $\pi : \mathcal{S} \rightarrow A$, where $\mathcal{S} = (T, U, S, H)$ represents the current state, encompassing information about the task T, system S, user U, as well as the history of previous interactions H (including the actions taken by the user, the responses provided by the system, and any other relevant events that have occurred during the interaction), and $A \in \mathcal{A}$ is the action taken by the (simulated) user. The choice of computational model (e.g., rule-based, probabilistic, or machine-learned algorithm) is influenced by the nature of the task, system, and user information.

1.4.2 Scope

User simulation encompasses a wide spectrum, ranging from predicting single actions to modeling complex behaviour across multiple tasks.

In our formulation, this scope is primarily determined by how the task information (T) is defined. At one end of the spectrum, T might represent a very specific interaction context, such as predicting whether a user would click on a particular search result snippet. Here, the focus is on simulating a single, isolated action. Moving along the spectrum, T could encompass a sequence of actions within a given context, such as reformulating search queries within a search session, requiring the model to consider dependencies between actions. Further expanding the scope, T might represent an entire task, such as finding information on a particular topic or completing a purchase, where the simulation would involve multiple sequences of interactions. Finally, at the broadest level, T could encompass a user's general preferences and behaviour across various tasks, necessitating models that capture long-term patterns and adapt to different contexts. Thus, by varying the granularity and breadth of T, our formulation allows for user simulations in a wide range of application scenarios at different levels of complexity. Table 1.1 lists specific examples of user simulation for various information access tasks.

1.4.3 Approaches

Approaches to user simulation can be classified based on the specific types of actions that they attempt to simulate. For example, some approaches may simulate how a user generates a query while others might simulate how a user responds to a search result list (e.g., simulating when a user might click on or skip a result). Approaches simulating different types of actions can also be combined to simulate a whole session of actions of a user. As will be elaborated later in the monograph, most existing work tends to focus on simulating each type of action separately with significantly less work on simulating a whole user session.

The problem of simulating a user's action can often be framed as a classification problem when there is a relatively small set of actions to choose from; for example, simulation of a user's clicking action may be framed as a binary classification problem, where the algorithm would predict whether the simulated user would click on a result or not after examining a snippet. When there are potentially infinitely many actions

Table 1.1: Examples of user simulation, ranging from single actions to more complex behaviours.

Task (T)	**System** (S)	**User information** (S)	**Actions** ($\mathcal{A}$)
Rating a product to express satisfaction	E-commerce website with product pages and rating features	User's purchase history, browsing behaviour, and demographic information	Browsing, Rating
Refining a search query to find specific information	Search engine with a search box, query suggestions, and navigable search result lists	User's initial query, search history, and click behaviour	Reformulating, Clicking
Collecting as many relevant information items as possible	Search engine with a query box and navigable search result lists	University researcher conducting a comprehensive literature review on a topic	Querying, Clicking
Finding a movie to watch	Recommender system with slates of items	Previous watch history	Clicking, Watching
Seeking assistance with a technical issue	Conversational assistant with natural language chat interface	User's description of the problem, technical expertise, and previous interactions	Prompting

to choose from (e.g., when formulating a query, any valid query would be potentially an option), in practice, we often make assumptions to restrict the number of actions to be considered when simulating those actions (e.g., limit the length of a query to be considered).

With the problem framed as a classification problem, different approaches generally vary in how they perform the classification (equivalently prediction) task. At a high level, we can distinguish two broad approaches: model-based and data-driven.

- ***Model-based*** approaches may be based on rules designed with knowledge about how users behave or on interpretable probabilistic models that can more flexibly capture uncertainties using interpretable parameters. The parameters of such models may be set heuristically or empirically derived from observed user data. By varying those parameters, different types of users can be simulated.

- ***Data-driven*** *(or machine-learned)* approaches emphasize maximizing accuracy of fitting any observed real user data, without necessarily imposing interpretability. Almost all such approaches are based on supervised machine learning, notably using deep neural networks which can learn effective, but non-interpretable representations from the data for predictive modeling.

These two families of approaches may also be combined, e.g., by utilizing model-based techniques to compute effective features for data-driven approaches or employing machine-learned models in specific components of model-based approaches. However, interpretability is desirable when building user simulators for evaluation to ensure that evaluation results are meaningful and to allow for the testing of verifiable hypotheses. Hence, this monograph primarily focuses on interpretable model-based approaches.

1.4.4 Uses of Simulation

In general, user simulation has many uses, including at least the following:

- Performing large-scale automatic evaluation of interactive systems (i.e., without the involvement of real users).
- Gaining insight into user behaviour to inform the design of systems and evaluation measures.
- Analyzing system performance under various conditions and user behaviours (answering *what-if* questions, such as "What is the influence of X on Y?").
- Augmenting data with human feedback and generating synthetic data with the purpose of training machine learning models and addressing data scarcity or privacy concerns. More broadly, user

simulation can facilitate machine learning approaches that require human input (interactive learning, reinforcement learning, or human-in-the-loop systems).

We note that all these uses require similar techniques, but our focus is on evaluation.

1.4.5 Requirements and Desiderata

When utilizing user simulation for system evaluation, it is critical that simulators provide reliable and insightful assessments. Two essential properties that ensure this are *validity* and *interpretability*.

- **Validity**: Simulated users must exhibit behaviours that align with empirical observations of real user behaviour in similar contexts. This includes both high-level strategies (e.g., information seeking patterns) and low-level actions (e.g., clicking behaviour). Without validity, the insights gained from simulation cannot be trusted.
- **Interpretability**: While not strictly a requirement, interpretability is a highly desirable property. Interpretability means that the simulated behaviour can be understood and adjusted through controllable parameters. This allows researchers to (1) understand why the simulator produced certain behaviours and (2) investigate how changes in specific parameters influence the behaviour of users. In general, as user behaviour and preferences vary significantly across users, interpretability is needed to facilitate interpretation of the evaluation results generated by user simulation, i.e., to understand what kind of real world users can be expected to produce similar results.

However, while striving for high validity is important, simulation does not need to be perfect in order to be useful. For example, although the relevance judgments in almost all the test collections for information retrieval evaluation are incomplete, the conclusions about relative performance of different retrieval systems tend not be affected much by the approximation made in a test collection (Voorhees, 2000). In fact, creating a "perfect" user simulator, i.e., one that flawlessly replicates human behaviour across all possible tasks and contexts, is likely an AI-complete problem, on par with achieving Artificial General Intelligence (AGI, cf. Section 9.4). When evaluating systems, we are often interested in a *relative comparison* between them with regards to some measure of utility, which is a weaker requirement than quantifying the *actual utility* of technology (in terms of some measurable impact, such as enhanced productivity or user satisfaction). Nevertheless, the practical utility of user simulation for relative comparisons lies in its *sensitivity*: the better an evaluation can distinguish between systems, the more practically useful it is.

While both validity and interpretability are desirable, there often exists a trade-off between the two. Data-driven (machine-learned) simulators, trained on large datasets of real user behaviour, can often achieve high predictive accuracy, capturing complex patterns and nuances in user actions. However, this predictive power comes at the cost of reduced interpretability. The internal workings of these models, often involving complex neural networks or ensemble methods, can be opaque and difficult to understand. This makes it challenging to pinpoint the specific reasons behind a simulated user's behaviour or to adjust the model's parameters in a controlled manner.

Beyond the essential requirements of validity and interpretability, there are several other desirable properties that can enhance the realism of user simulation.

- **Cognitive plausibility**: The decision-making processes underlying simulated user behaviour should be grounded in theories or models of human cognition, ensuring that the simulated actions are not arbitrary or random.

- **Variation**: While reflecting general user behaviour patterns, simulated users should also exhibit variability and occasional outliers, "not replicating average behaviour completely" (Bignold *et al.*, 2021). That is, simulation should reflect the unpredictable nature of real human interactions.

- **Adaptability**: Simulated users should be able to learn from their interactions with the system, update their expectations about the system and adjust their behaviour accordingly (Balog, 2021).

By incorporating these desirable properties, user simulators can achieve a higher level of realism and sophistication, enabling more accurate predictions of user behaviour and more insightful evaluations of interactive systems.

1.5 Aims and Organization

With the emergence of various information access systems exhibiting increasing complexity, there is a critical need for sound and scalable means of automatic evaluation. Simulation has the potential to offer a solution here. It has attracted attention from multiple angles and much progress has been made in the last decade. Relevant research work, however, has been scattered in multiple research communities, including information retrieval, recommender systems, dialogue systems, and user modeling. This monograph aims to synthesize that research into a coherent framework. Given the substantial amount of work performed within the context of information access systems, this is where our main focus will lie.

Specifically, our main objective is to discuss how simulation may be employed to undertake evaluation of information access systems in order to (1) estimate how well they will perform under various circumstances, and (2) analyze how performance changes under different conditions and user behaviours. However, we attempt to make our discussions as generic as possible, such that those working on other types of interactive systems, or applications of assistive AI, would also find it useful. Specifically, whenever possible, we would attempt to lay out general conceptual frameworks, discuss general challenges, and extract general ideas from specific research work, which we hope to be broadly useful to more readers as well as provide a stable meaningful high-level structure where future research work can be naturally incorporated and discussed. Our emphasis on the generality of discussion, however, means that our treatment of any specific research work is inevitably brief. Detailed information can be conveniently found in the numerous research papers cited throughout this monograph. When selecting specific work to elaborate, we have also chosen to focus more on representative work that is useful for illustrating major ideas instead of having an even coverage of all the work. Due to the interdisciplinary nature of the topic and quick growth of research, the references cited in our monograph are inevitably incomplete. Considering this limitation and anticipating the rapid progress of research in this area in the future, we have created a website (https://usersim.ai/) for an envisioned broad interdisciplinary research community on user simulation. This platform aims to foster collaboration among researchers from diverse fields, enabling them to collectively maintain a repository of up-to-date and relevant references over time.

The main intended audience of this monograph includes both researchers, who wish to further the research and development of simulation-based evaluation methods, as well as industry practitioners, who are interested in employing these techniques in operational settings. Considering the fact that user simulation is broadly connected with multiple fields (see a more detailed discussion of this in Section 9), we also hope this monograph will be broadly useful to an audience beyond those interested in using user simulation for evaluation.

The rest of this monograph is organized as follows. We start in Section 2 by providing a background on the development of simulation techniques within different research communities. Following this, Section 3 gives an overview of how user simulation has been employed in the past. Section 4 introduces a conceptual framework for generally modeling interactions between a user and a system and evaluating any interactive system using user simulation. Next, in Section 5, we discuss decision-making and cognitive processes of users, followed by the mathematical framework we will employ in subsequent sections to model these. We present simulation techniques for search engines and recommender systems in Section 6 and for conversational assistants in Section 7. Section 8 focuses on applying user simulation in practice, covering issues related to configuring, validating, and building simulators. Section 9 broadens the perspective on user simulation as an interdisciplinary research area intersecting with various fields beyond computer science, ultimately contributing to the progress towards AGI. Finally, Section 10 concludes the monograph by highlighting open issues and possible future research directions.

References

Abbasiantaeb, Z., Y. Yuan, E. Kanoulas, and M. 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. DOI: 10.1145/3616855.3635856.

Abolghasemi, A., S. Verberne, A. Askari, and L. Azzopardi. (2023). "Retrievability Bias Estimation Using Synthetically Generated Queries". In: *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management. CIKM '23*. 3712–3716. DOI: 10.1145/3583780.3615221.

Afzali, J., A. M. Drzewiecki, K. Balog, and S. Zhang. (2023). "UserSimCRS: A User Simulation Toolkit for Evaluating Conversational Recommender Systems". In: *Proceedings of the Sixteenth ACM International Conference on Web Search and Data Mining. WSDM '23*. 1160–1163. DOI: 10.1145/3539597.3573029.

Aher, G. V., R. I. Arriaga, and A. T. Kalai. (2023). "Using Large Language Models to Simulate Multiple Humans and Replicate Human Subject Studies". In: *International Conference on Machine Learning. ICML '23*. 337–371.

Aliannejadi, M., L. Azzopardi, H. Zamani, E. Kanoulas, P. Thomas, and N. Craswell. (2021). "Analysing Mixed Initiatives and Search Strategies during Conversational Search". In: *Proceedings of the 30th ACM International Conference on Information and Knowledge Management. CIKM '21*. 16–26. DOI: 10.1145/3459637.3482231.

Aliannejadi, M., J. Kiseleva, A. Chuklin, J. Dalton, and M. Burtsev. (2020). "ConvAI3: Generating Clarifying Questions for Open-Domain Dialogue Systems (ClariQ)". arXiv: 2009.11352 [cs.CL].

Aliannejadi, M., H. Zamani, F. Crestani, and W. B. Croft. (2019). "Asking Clarifying Questions in Open-Domain Information-Seeking Conversations". In: *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval. SIGIR '19*. 475–484. DOI: 10.1145/3331184.3331265.

Anand, A., L. Cavedon, H. Joho, M. Sanderson, and B. Stein. (2020). "Conversational Search (Dagstuhl Seminar 19461)". *Dagstuhl Reports*. 9(11): 34–83. Ed. by A. Anand, L. Cavedon, H. Joho, M. Sanderson, and B. Stein. DOI: 10.4230/DagRep.9.11.34.

Arguello, J., W.-C. Wu, D. Kelly, and A. Edwards. (2012). "Task Complexity, Vertical Display and User Interaction in Aggregated Search". In: *Proceedings of the 35th International ACM SIGIR Conference on Research and Development in Information Retrieval. SIGIR '12*. 435–444. DOI: 10.1145/2348283.2348343.

Arora, S. and P. Doshi. (2021). "A Survey of Inverse Reinforcement Learning: Challenges, Methods and Progress". *Artificial Intelligence*. 297: 103500. DOI: 10.1016/j.artint.2021.103500.

Arvola, P., J. Kekäläinen, and M. Junkkari. (2010). "Expected Reading Effort in Focused Retrieval Evaluation". *Information Retrieval*. 13(5): 460–484. DOI: 10.1007/s10791-010-9133-9.

Aula, A., P. Majaranta, and K.-J. Räihä. (2005). "Eye-Tracking Reveals the Personal Styles for Search Result Evaluation". In: *Proceedings of the 2005 IFIP TC13 International Conference on Human-Computer Interaction. INTERACT '05*. 1058–1061. DOI: 10.1007/11555261_104.

Azzopardi, L. (2009). “Query Side Evaluation: An Empirical Analysis of Effectiveness and Effort”. In: *Proceedings of the 32nd International ACM SIGIR Conference on Research and Development in Information Retrieval. SIGIR ’09*. 556–563. DOI: 10.1145/1571941.1572037.

Azzopardi, L. (2011). “The Economics in Interactive Information Retrieval”. In: *Proceedings of the 34th International ACM SIGIR Conference on Research and Development in Information Retrieval. SIGIR ’11*. 15–24. DOI: 10.1145/2009916.2009923.

Azzopardi, L. (2014). “Modelling Interaction with Economic Models of Search”. In: *Proceedings of the 37th International ACM SIGIR Conference on Research and Development in Information Retrieval. SIGIR ’14*. 3–12. DOI: 10.1145/2600428.2609574.

Azzopardi, L., M. Dubiel, M. Halvey, and J. Dalton. (2018a). “Conceptualizing Agent-Human Interactions During the Conversational Search Process”. In: *Proceedings of the 2nd International Workshop on Conversational Approaches to Information Retrieval. CAIR ’18*.

Azzopardi, L., K. Järvelin, J. Kamps, and M. D. Smucker. (2011). “Report on the SIGIR 2010 Workshop on the Simulation of Interaction”. *SIGIR Forum*. 44(2): 35–47. DOI: 10.1145/1924475.1924484.

Azzopardi, L., J. Mackenzie, and A. Moffat. (2021). “ERR is Not C/W/L: Exploring the Relationship Between Expected Reciprocal Rank and Other Metrics”. In: *Proceedings of the 2021 ACM SIGIR International Conference on Theory of Information Retrieval. ICTIR ’21*. 231–237. DOI: 10.1145/3471158.3472239.

Azzopardi, L. and M. de Rijke. (2006). “Automatic Construction of Known-Item Finding Test Beds”. In: *Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. SIGIR ’06*. 603–604. DOI: 10.1145/1148170.1148276.

Azzopardi, L., M. de Rijke, and K. Balog. (2007). “Building Simulated Queries for Known-Item Topics: An Analysis Using Six European Languages”. In: *Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. SIGIR ’07*. 455–462. DOI: 10.1145/1277741.1277820.

Azzopardi, L., P. Thomas, and N. Craswell. (2018b). "Measuring the Utility of Search Engine Result Pages: An Information Foraging Based Measure". In: *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval. SIGIR '18*. 605–614. DOI: 10.1145/3209978.3210027.

Azzopardi, L., P. Thomas, and A. Moffat. (2019). "Cwl_eval: An Evaluation Tool for Information Retrieval". In: *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval. SIGIR '19*. 1321–1324. DOI: 10.1145/3331184.3331398.

Azzopardi, L., R. W. White, P. Thomas, and N. Craswell. (2020). "Data-Driven Evaluation Metrics for Heterogeneous Search Engine Result Pages". In: *Proceedings of the 2020 Conference on Human Information Interaction and Retrieval. CHIIR '20*. 213–222. DOI: 10.1145/3343413.3377959.

Azzopardi, L. and G. Zuccon. (2015). "An Analysis of Theories of Search and Search Behavior". In: *Proceedings of the 2015 International Conference on The Theory of Information Retrieval. ICTIR '15*. 81–90. DOI: 10.1145/2808194.2809447.

Azzopardi, L. and G. Zuccon. (2016). "An Analysis of the Cost and Benefit of Search Interactions". In: *Proceedings of the 2016 ACM International Conference on the Theory of Information Retrieval. ICTIR '16*. 59–68. DOI: 10.1145/2970398.2970412.

Balog, K. (2021). "Conversational AI from an Information Retrieval Perspective: Remaining Challenges and a Case for User Simulation". In: *Proceedings of the 2nd International Conference on Design of Experimental Search and Information REtrieval Systems. DESIRES '21*. 80–90.

Balog, K. and T. Kenter. (2019). "Personal Knowledge Graphs: A Research Agenda". In: *Proceedings of the 2019 ACM SIGIR International Conference on Theory of Information Retrieval. ICTIR '19*. 217–220. DOI: 10.1145/3341981.3344241.

Balog, K., D. Maxwell, P. Thomas, and S. Zhang. (2022). "Report on the 1st Simulation for Information Retrieval Workshop (Sim4IR 2021) at SIGIR 2021". *SIGIR Forum*. 55(2): 1–16. DOI: 10.1145/3527546.3527559.

Baskaya, F. (2014). "Simulating Search Sessions in Interactive Information Retrieval Evaluation". *PhD thesis*. University of Tampere.

Baskaya, F., H. Keskustalo, and K. Järvelin. (2011). "Simulating Simple and Fallible Relevance Feedback". In: *Proceedings of the 33rd European Conference on IR Research. ECIR '11*. 593–604. DOI: 10.1007/978-3-642-20161-5_59.

Baskaya, F., H. Keskustalo, and K. Järvelin. (2012). "Time Drives Interaction: Simulating Sessions in Diverse Searching Environments". In: *Proceedings of the 35th International ACM SIGIR Conference on Research and Development in Information Retrieval. SIGIR '12*. 105–114. DOI: 10.1145/2348283.2348301.

Baskaya, F., H. Keskustalo, and K. Järvelin. (2013). "Modeling Behavioral Factors in Interactive Information Retrieval". In: *Proceedings of the 22nd ACM International Conference on Information and Knowledge Management. CIKM '13*. 2297–2302. DOI: 10.1145/2505515.2505660.

Bates, M. J. (1989). "The Design of Browsing and Berrypicking Techniques for the Online Search Interface". *Online Review*. 13(5): 407–424. DOI: 10.1108/eb024320.

Belkin, N. J. and W. B. Croft. (1992). "Information Filtering and Information Retrieval: Two Sides of the Same Coin?" *Communications of the ACM*. 35(12): 29–38. DOI: 10.1145/138859.138861.

Belkin, N. J., R. N. Oddy, and H. M. Brooks. (1982). "ASK for Information Retrieval: Part I. Background and Theory". *Journal of Documentation*. DOI: 10.1108/eb026722.

Bellogin, A., P. Castells, and I. Cantador. (2011). "Precision-oriented Evaluation of Recommender Systems: An Algorithmic Comparison". In: *Proceedings of the Fifth ACM Conference on Recommender Systems. RecSys '11*. 333–336. DOI: 10.1145/2043932.2043996.

Bendada, W., G. Salha, and T. Bontempelli. (2020). "Carousel Personalization in Music Streaming Apps with Contextual Bandits". In: *Proceedings of the 14th ACM Conference on Recommender Systems. RecSys '20*. 420–425. DOI: 10.1145/3383313.3412217.

Bernard, N. and K. Balog. (2023). "MG-ShopDial: A Multi-Goal Conversational Dataset for e-Commerce". In: *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval. SIGIR '23*. 2775–2785. DOI: 10.1145/3539618.3591883.

Bernard, N. and K. Balog. (2024). "Identifying Breakdowns in Conversational Recommender Systems using User Simulation". In: *Proceedings of the 6th International Conference on Conversational User Interfaces. CUI '24*. DOI: 10.1145/3640794.3665539.

Berryman, J. M. (2006). "What Defines "Enough" Information? How Policy Workers Make Judgements and Decisions during Information Seeking: Preliminary Results from an Exploratory Study". *Information Research*. 11(4).

Bi, K., Q. Ai, and W. B. Croft. (2021). "Asking Clarifying Questions Based on Negative Feedback in Conversational Search". In: *Proceedings of the 2021 ACM SIGIR International Conference on Theory of Information Retrieval. ICTIR '21*. 157–166. DOI: 10.1145/3471158.3472232.

Bignold, A., F. Cruz, R. Dazeley, P. Vamplew, and C. Foale. (2021). "An Evaluation Methodology for Interactive Reinforcement Learning with Simulated Users". *Biomimetics*. 6(1): 13. DOI: 10.3390/biomimetics6010013.

Birchler, U., M. Butler, and Routledge. (2007). *Information Economics*. Routledge.

Blunt, C. R. (1965). "An Information Retrieval System Model". *Tech. rep.* HRB-SINGER Inc.

Bollen, D., M. Graus, and M. C. Willemsen. (2012). "Remembering the Stars? Effect of Time on Preference Retrieval from Memory". In: *Proceedings of the Sixth ACM Conference on Recommender Systems. RecSys '12*. 217–220. DOI: 10.1145/2365952.2365998.

Borgman, C. L. (1996). "Why are Online Catalogs Still Hard to Use?" *Journal of the American Society for Information Science*. 47(7): 493–503. DOI: 10.1002/(SICI)1097-4571(199607)47:7<493::AID-ASI3>3.0.CO;2-P.

Bota, H., K. Zhou, and J. M. Jose. (2016). "Playing Your Cards Right: The Effect of Entity Cards on Search Behaviour and Workload". In: *Proceedings of the 2016 ACM on Conference on Human Information Interaction and Retrieval. CHIIR '16*. 131–140. DOI: 10.1145/2854946.2854967.

Bountouridis, D., J. Harambam, M. Makhortykh, M. Marrero, N. Tintarev, and C. Hauff. (2019). "SIREN: A Simulation Framework for Understanding the Effects of Recommender Systems in Online News Environments". In: *Proceedings of the Conference on Fairness, Accountability, and Transparency. FAT* '19*. 150–159. DOI: 10.1145/3287560.3287583.

Boyce, B. (1982). "Beyond Topicality: A Two Stage View of Relevance and the Retrieval Process". *Information Processing & Management*. 18(3): 105–109. DOI: 10.1016/0306-4573(82)90033-4.

Brennan, K., D. Kelly, and J. Arguello. (2014). "The Effect of Cognitive Abilities on Information Search for Tasks of Varying Levels of Complexity". In: *Proceedings of the 5th Information Interaction in Context Symposium. IIiX '14*. 165–174. DOI: 10.1145/2637002.2637022.

Breuer, T., N. Fuhr, and P. Schaer. (2022). "Validating Simulations of User Query Variants". In: *Proceedings of the 44th European Conference on IR Research. ECIR '22*. 80–94. DOI: 10.1007/978-3-030-99736-6_6.

Broder, A. (2002). "A Taxonomy of Web Search". *SIGIR Forum*. 36(2): 3–10. DOI: 10.1145/792550.792552.

Brown, T., B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh, D. Ziegler, J. Wu, C. Winter, C. Hesse, M. Chen, E. Sigler, M. Litwin, S. Gray, B. Chess, J. Clark, C. Berner, S. McCandlish, A. Radford, I. Sutskever, and D. Amodei. (2020). "Language Models are Few-Shot Learners". In: *Advances in Neural Information Processing Systems. NeurIPS '20*. 1877–1901.

Browne, G. J., M. G. Pitts, and J. C. Wetherbe. (2005). "Stopping Rule Use During Web-Based Search". In: *Proceedings of the 38th Annual Hawaii International Conference on System Sciences*. Vol. 9. 271b. DOI: 10.1109/HICSS.2005.556.

Browne, G. J., M. G. Pitts, and J. C. Wetherbe. (2007). "Cognitive Stopping Rules for Terminating Information Search in Online Tasks". *MIS Quarterly*. 31(1): 89–104. DOI: 10.2307/25148782.

Buckley, C. and J. A. Walz. (1999). "The TREC-8 Query Track". In: *Proceedings of The Eighth Text REtrieval Conference. TREC '99*.

Budzianowski, P., T.-H. Wen, B.-H. Tseng, I. Casanueva, S. Ultes, O. Ramadan, and M. Gašić. (2018). "MultiWOZ - A Large-Scale Multi-Domain Wizard-of-Oz Dataset for Task-Oriented Dialogue Modelling". In: *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing. EMNLP '18*. 5016–5026. DOI: 10.18653/v1/D18-1547.

Bunt, H., V. Petukhova, E. Gilmartin, C. Pelachaud, A. Fang, S. Keizer, and L. Prévot. (2020). "The ISO Standard for Dialogue Act Annotation, Second Edition". In: *Proceedings of the Twelfth Language Resources and Evaluation Conference. LREC '20*. 549–558.

Bunt, H., V. Petukhova, D. Traum, and J. Alexandersson. (2017). "Dialogue Act Annotation with the ISO 24617-2 Standard". In: *Multimodal Interaction with W3C Standards: Toward Natural User Interfaces to Everything*. Springer International Publishing. 109–135. DOI: 10.1007/978-3-319-42816-1_6.

Burnell, R., W. Schellaert, J. Burden, T. D. Ullman, F. Martinez-Plumed, J. B. Tenenbaum, D. Rutar, L. G. Cheke, J. Sohl-Dickstein, M. Mitchell, *et al.* (2023). "Rethink Reporting of Evaluation Results in AI". *Science*. 380(6641): 136–138. DOI: 10.1126/science.adf6369.

Câmara, A., D. Maxwell, and C. Hauff. (2022). "Searching, Learning, and Subtopic Ordering: A Simulation-Based Analysis". In: *Proceedings of the 44th European Conference on IR Research. ECIR '22*. 142–156. DOI: 10.1007/978-3-030-99736-6_10.

Card, S. K., P. Pirolli, M. Van Der Wege, J. B. Morrison, R. W. Reeder, P. K. Schraedley, and J. Boshart. (2001). "Information Scent as a Driver of Web Behavior Graphs: Results of a Protocol Analysis Method for Web Usability". In: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems. CHI '01*. 498–505. DOI: 10.1145/365024.365331.

Carterette, B. (2011). "System Effectiveness, User Models, and User Utility: A Conceptual Framework for Investigation". In: *Proceedings of the 34th International ACM SIGIR Conference on Research and Development in Information Retrieval. SIGIR '11*. 903–912. DOI: 10.1145/2009916.2010037.

Carterette, B., A. Bah, and M. Zengin. (2015). "Dynamic Test Collections for Retrieval Evaluation". In: *Proceedings of the 2015 International Conference on The Theory of Information Retrieval. ICTIR '15*. 91–100. DOI: 10.1145/2808194.2809470.

Carterette, B., P. Clough, M. Hall, E. Kanoulas, and M. Sanderson. (2016). "Evaluating Retrieval over Sessions: The TREC Session Track 2011-2014". In: *Proceedings of the 39th International ACM SIGIR Conference on Research and Development in Information Retrieval. SIGIR '16*. 685–688. DOI: 10.1145/2911451.2914675.

Carterette, B., E. Kanoulas, and E. Yilmaz. (2011). "Simulating Simple User Behavior for System Effectiveness Evaluation". In: *Proceedings of the 20th ACM International Conference on Information and Knowledge Management. CIKM '11*. 611–620. DOI: 10.1145/2063576.2063668.

Chaney, A. J. B. (2021). "Recommendation System Simulations: A Discussion of Two Key Challenges". arXiv: 2109.02475 [cs.IR].

Chapelle, O., D. Metlzer, Y. Zhang, and P. Grinspan. (2009). "Expected Reciprocal Rank for Graded Relevance". In: *Proceedings of the 18th ACM Conference on Information and Knowledge Management. CIKM '09*. 621–630. DOI: 10.1145/1645953.1646033.

Charnov, E. L. (1976). "Optimal Foraging, the Marginal Value Theorem". *Theoretical Population Biology*. 9(2): 129–136. DOI: 10.1016/0040-5809(76)90040-X.

Chen, D., W. Chen, H. Wang, Z. Chen, and Q. Yang. (2012). "Beyond Ten Blue Links: Enabling User Click Modeling in Federated Web Search". In: *Proceedings of the Fifth ACM International Conference on Web Search and Data Mining. WSDM '12*. 463–472. DOI: 10.1145/2124295.2124351.

Chen, H., X. Liu, D. Yin, and J. Tang. (2017). "A Survey on Dialogue Systems: Recent Advances and New Frontiers". *ACM SIGKDD Explorations Newsletter*. 19(2): 25–35. DOI: 10.1145/3166054.3166058.

Chen, L. and P. Pu. (2012). "Critiquing-based Recommenders: Survey and Emerging Trends". *User Modeling and User-Adapted Interaction*. 22: 125–150. DOI: 10.1007/s11257-011-9108-6.

Chen, X., L. Yao, J. McAuley, G. Zhou, and X. Wang. (2021). "A Survey of Deep Reinforcement Learning in Recommender Systems: A Systematic Review and Future Directions". arXiv: 2109.03540 [cs.IR].

Cheng, H.-T., L. Koc, J. Harmsen, T. Shaked, T. Chandra, H. Aradhye, G. Anderson, G. Corrado, W. Chai, M. Ispir, *et al.* (2016). "Wide & Deep Learning for Recommender Systems". In: *Proceedings of the 1st Workshop on Deep Learning for Recommender Systems*. 7–10. DOI: 10.1145/2988450.2988454.

Cheng, Q., L. Li, G. Quan, F. Gao, X. Mou, and X. Qiu. (2022). "Is MultiWOZ a Solved Task? An Interactive TOD Evaluation Framework with User Simulator". In: *Findings of the Association for Computational Linguistics: EMNLP 2022. EMNLP '22*. 1248–1259. DOI: 10.18653/v1/2022.findings-emnlp.90.

Chi, E. H., A. Rosien, G. Supattanasiri, A. Williams, C. Royer, C. Chow, E. Robles, B. Dalal, J. Chen, and S. Cousins. (2003). "The Bloodhound Project: Automating Discovery of Web Usability Issues using the InfoScent™ Simulator". In: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems. CHI '03*. 505–512. DOI: 10.1145/642611.642699.

Chuklin, A., I. Markov, and M. de Rijke. (2015). *Click Models for Web Search*. Morgan & Claypool. DOI: 10.2200/S00654ED1V01Y201507ICR043.

Chuklin, A. and P. Serdyukov. (2012). "Good Abandonments in Factoid Queries". In: *Proceedings of the 21st International Conference on World Wide Web. WWW '12 Companion*. 483–484. DOI: 10.1145/2187980.2188088.

Clarke, C. L. A., N. Craswell, I. Soboroff, and E. M. Voorhees. (2011). "Overview of the TREC 2011 Web Track". In: *Proceedings of The Twentieth Text REtrieval Conference. TREC '11*.

Clarke, C. L., M. Kolla, G. V. Cormack, O. Vechtomova, A. Ashkan, S. Büttcher, and I. MacKinnon. (2008). "Novelty and Diversity in Information Retrieval Evaluation". In: *Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. SIGIR '08*. 659–666. DOI: 10.1145/1390334.1390446.

Cleverdon, C. and M. Kean. (1968). "Factors Determining the Performance of Indexing Systems". Aslib Cranfield Research Project, Cranfield, England.

Coleman, M. and T. L. Liau. (1975). "A Computer Readability Formula Designed for Machine Scoring". *Journal of Applied Psychology*. 60(2): 283–284. DOI: 10.1037/h0076540.

Collins-Thompson, K., P. Hansen, and C. Hauff. (2017). "Search as Learning (Dagstuhl seminar 17092)".

Cooper, M. D. (1973a). "A Simulation Model of an Information Retrieval System". *Information Storage and Retrieval*. 9(1): 13–32. DOI: 10.1016/0020-0271(73)90004-1.

Cooper, W. S. (1968). "Expected Search Length: A Single Measure of Retrieval Effectiveness based on the Weak Ordering Action of Retrieval Systems". *American Documentation*. 19(1): 30–41. DOI: 10.1002/asi.5090190108.

Cooper, W. S. (1973b). "On Selecting a Measure of Retrieval Effectiveness Part II. Implementation of the Philosophy". *Journal of the American Society for Information Science*. 24(6): 413–424. DOI: 10.1002/asi.4630240603.

Cosijn, E. and P. Ingwersen. (2000). "Dimensions of Relevance". *Information Processing & Management*. 36(4): 533–550. DOI: 10.1016/S0306-4573(99)00072-2.

Craswell, N., O. Zoeter, M. Taylor, and B. Ramsey. (2008). "An Experimental Comparison of Click Position-Bias Models". In: *Proceedings of the 2008 International Conference on Web Search and Data Mining. WSDM '08*. 87–94. DOI: 10.1145/1341531.1341545.

Cremonesi, P., Y. Koren, and R. Turrin. (2010). "Performance of Recommender Algorithms on Top-N Recommendation Tasks". In: *Proceedings of the Fourth ACM Conference on Recommender Systems. RecSys '10*. 39–46. DOI: 10.1145/1864708.1864721.

Crook, P. and A. Marin. (2017). "Sequence to Sequence Modeling for User Simulation in Dialog Systems". In: *Proceedings of Interspeech 2017*. 1706–1710. DOI: 10.21437/Interspeech.2017-161.

Culpepper, J. S., F. Diaz, and M. D. Smucker. (2018). "Research Frontiers in Information Retrieval: Report from the Third Strategic Workshop on Information Retrieval in Lorne (SWIRL 2018)". *SIGIR Forum*. 52(1): 34–90. DOI: 10.1145/3274784.3274788.

Dalton, J., C. Xiong, V. Kumar, and J. Callan. (2020). "CAsT-19: A Dataset for Conversational Information Seeking". In: *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval. SIGIR '20*. 1985–1988. DOI: 10.1145/3397271.3401206.

Davidson, S., S. Romeo, R. Shu, J. Gung, A. Gupta, S. Mansour, and Y. Zhang. (2023). "User Simulation with Large Language Models for Evaluating Task-Oriented Dialogue". arXiv: 2309.13233 [cs.CL].

De Mey, M. (1977). "The Cognitive Viewpoint: Its Development and its Scope". *Communication & Cognition*. 10(2): 7–23.

Deldjoo, Y., J. R. Trippas, and H. Zamani. (2021). "Towards Multi-Modal Conversational Information Seeking". In: *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval. SIGIR '21*. 1577–1587. DOI: 10.1145/3404835.3462806.

Deriu, J., A. Rodrigo, A. Otegi, G. Echegoyen, S. Rosset, E. Agirre, and M. Cieliebak. (2021). "Survey on Evaluation Methods for Dialogue Systems". *Artificial Intelligence Review*. 54(1): 755–810. DOI: 10.1007/s10462-020-09866-x.

Devlin, J., M.-W. Chang, K. Lee, and K. Toutanova. (2019). “BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding”. In: *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers). NAACL ’19*. 4171–4186. DOI: 10.18653/v1/N19-1423.

Diaz, F. and J. Arguello. (2009). “Adaptation of Offline Vertical Selection Predictions in the Presence of User Feedback”. In: *Proceedings of the 32nd International ACM SIGIR Conference on Research and Development in Information Retrieval. SIGIR ’09*. 323–330. DOI: 10.1145/1571941.1571998.

Diaz, F., R. White, G. Buscher, and D. Liebling. (2013). “Robust Models of Mouse Movement on Dynamic Web Search Results Pages”. In: *Proceedings of the 22nd ACM International Conference on Information and Knowledge Management. CIKM ’13*. 1451–1460. DOI: 10.1145/2505515.2505717.

Diriye, A., R. White, G. Buscher, and S. Dumais. (2012). “Leaving so Soon? Understanding and Predicting Web Search Abandonment Rationales”. In: *Proceedings of the 21st ACM International Conference on Information and Knowledge Management. CIKM ’12*. 1025–1034. DOI: 10.1145/2396761.2398399.

Dodge, J., A. Gane, X. Zhang, A. Bordes, S. Chopra, A. Miller, A. Szlam, and J. Weston. (2016). “Evaluating Prerequisite Qualities for Learning End-to-End Dialog Systems”. In: *4th International Conference on Learning Representations. ICLR ’16*.

Dorça, F. (2015). “Implementation and use of Simulated Students for Test and Validation of new Adaptive Educational Systems: a Practical Insight”. *International Journal of Artificial Intelligence in Education*. 25: 319–345. DOI: 10.1007/s40593-015-0037-0.

Dumais, S., E. Cutrell, J. Cadiz, G. Jancke, R. Sarin, and D. C. Robbins. (2003). “Stuff I’ve Seen: A System for Personal Information Retrieval and Re-use”. In: *Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Informaion Retrieval. SIGIR ’03*. 72–79. DOI: 10.1145/860435.860451.

Dumais, S. T., G. Buscher, and E. Cutrell. (2010). "Individual Differences in Gaze Patterns for Web Search". In: *Proceedings of the Third Symposium on Information Interaction in Context. IIiX '10*. 185–194. DOI: 10.1145/1840784.1840812.

Dupret, G. E. and B. Piwowarski. (2008). "A User Browsing Model to Predict Search Engine Click Data from Past Observations". In: *Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. SIGIR '08*. 331–338. DOI: 10.1145/1390334.1390392.

Dzyabura, D. and A. Tuzhilin. (2013). "Not by Search Alone: How Recommendations Complement Search Results". In: *Proceedings of the 7th ACM Conference on Recommender Systems. RecSys '13*. 371–374. DOI: 10.1145/2507157.2507231.

Eckert, W., E. Levin, and R. Pieraccini. (1997). "User Modeling for Spoken Dialogue System Evaluation". In: *1997 IEEE Workshop on Automatic Speech Recognition and Understanding Proceedings*. 80–87. DOI: 10.1109/ASRU.1997.658991.

Eickhoff, C., J. Teevan, R. White, and S. Dumais. (2014). "Lessons from the Journey: A Query Log Analysis of within-Session Learning". In: *Proceedings of the 7th ACM International Conference on Web Search and Data Mining. WSDM '14*. 223–232. DOI: 10.1145/2556195.2556217.

Ekstrand, M. D., A. Chaney, P. Castells, R. Burke, D. Rohde, and M. Slokom. (2021). "SimuRec: Workshop on Synthetic Data and Simulation Methods for Recommender Systems Research". In: *Proceedings of the 15th ACM Conference on Recommender Systems. RecSys '21*. 803–805. DOI: 10.1145/3460231.3470938.

El Asri, L., J. He, and K. Suleman. (2016). "A Sequence-to-Sequence Model for User Simulation in Spoken Dialogue Systems". In: *Proceedings of Interspeech 2016*. 1151–1155. DOI: 10.21437/Interspeech.2016-1175.

Ellis, D. (1989). "A Behavioural Approach to Information Retrieval System Design". *Journal of Documentation*. 45(3): 171–212. DOI: 10.1108/eb026843.

Faggioli, G., L. Dietz, C. L. A. Clarke, G. Demartini, M. Hagen, C. Hauff, N. Kando, E. Kanoulas, M. Potthast, B. Stein, and H. Wachsmuth. (2023). “Perspectives on Large Language Models for Relevance Judgment”. In: *Proceedings of the 2023 ACM SIGIR International Conference on Theory of Information Retrieval. ICTIR ’23*. 39–50. DOI: 10.1145/3578337.3605136.

Felicioni, N., M. Ferrari Dacrema, and P. Cremonesi. (2021). “A Methodology for the Offline Evaluation of Recommender Systems in a User Interface with Multiple Carousels”. In: *Adjunct Proceedings of the 29th ACM Conference on User Modeling, Adaptation and Personalization. UMAP ’21*. 10–15. DOI: 10.1145/3450614.3461680.

Ferrari Dacrema, M., N. Felicioni, and P. Cremonesi. (2022). “Offline Evaluation of Recommender Systems in a User Interface With Multiple Carousels”. *Frontiers in Big Data*. 5. DOI: 10.3389/fdata.2022.910030.

Ferraro, A., X. Serra, and C. Bauer. (2021). “Break the Loop: Gender Imbalance in Music Recommenders”. In: *Proceedings of the 2021 Conference on Human Information Interaction and Retrieval. CHIIR ’21*. 249–254. DOI: 10.1145/3406522.3446033.

Ferro, N. and C. Peters, eds. (2019). *Information Retrieval Evaluation in a Changing World Lessons Learned from 20 Years of CLEF*. Springer.

Fleder, D. M. and K. Hosanagar. (2007). “Recommender Systems and Their Impact on Sales Diversity”. In: *Proceedings of the 8th ACM Conference on Electronic Commerce. EC ’07*. 192–199. DOI: 10.1145/1250910.1250939.

Fleder, D. M. and K. Hosanagar. (2009). “Blockbuster Culture’s Next Rise or Fall: The Impact of Recommender Systems on Sales Diversity”. *Management Science*. 55(5): 697–712. DOI: 10.1287/mnsc.1080.0974.

Friedman, L., S. Ahuja, D. Allen, Z. Tan, H. Sidahmed, C. Long, J. Xie, G. Schubiner, A. Patel, H. Lara, B. Chu, Z. Chen, and M. Tiwari. (2023). “Leveraging Large Language Models in Conversational Recommender Systems”. arXiv: 2305.07961 [cs.IR].

Fu, W.-T. and P. Pirolli. (2007). "SNIF-ACT: A Cognitive Model of User Navigation on the World Wide Web". *Human-Computer Interaction*. 22(4): 355–412.

Fuhr, N. (2008). "A Probability Ranking Principle for Interactive Information Retrieval". *Information Retrieval*. 11(3): 251–265. DOI: 10.1007/s10791-008-9045-0.

Gao, C., X. Lan, N. Li, Y. Yuan, J. Ding, Z. Zhou, F. Xu, and Y. Li. (2023). "Large Language Models Empowered Agent-based Modeling and Simulation: A Survey and Perspectives". arXiv: 2312.11970 [cs.AI].

Gao, C., F. Xu, X. Chen, X. Wang, X. He, and Y. Li. (2024). "Simulating Human Society with Large Language Model Agents: City, Social Media, and Economic System". In: *Companion Proceedings of the ACM on Web Conference 2024*. 1290–1293.

Gao, C., W. Lei, X. He, M. de Rijke, and T.-S. Chua. (2021). "Advances and Challenges in Conversational Recommender Systems: A Survey". *AI Open*. 2: 100–126. DOI: 10.1016/j.aiopen.2021.06.002.

Gao, J., M. Galley, and L. Li. (2019). "Neural Approaches to Conversational AI". *Foundations and Trends in Information Retrieval*. 13(2–3): 127–298. DOI: 10.1561/1500000074.

Gemmis, M. de, P. Lops, C. Musto, F. Narducci, and G. Semeraro. (2015). "Semantics-Aware Content-Based Recommender Systems". In: *Recommender Systems Handbook*. Ed. by F. Ricci, L. Rokach, and B. Shapira. 2nd ed. Springer US. 119–159. DOI: 10.1007/978-1-4899-7637-6_4.

Gettys, C. F. and S. Fisher. (1979). "Hypothesis Plausibility and Hypothesis Generation". *Organizational Behavior and Human Performance*. 24(1): 93–110. DOI: 10.1016/0030-5073(79)90018-7.

Ghanem, N., S. Leitner, and D. Jannach. (2022). "Balancing Consumer and Business Value of Recommender Systems: A Simulation-based Analysis". *Electronic Commerce Research and Applications*. 55: 101195. DOI: 10.1016/j.elerap.2022.101195.

Ghosh, S., M. Rath, and C. Shah. (2018). "Searching as Learning: Exploring Search Behavior and Learning Outcomes in Learning-related Tasks". In: *Proceedings of the 2018 Conference on Human Information Interaction & Retrieval. CHIIR '18*. 22–31. DOI: 10.1145/3176349.3176386.

Gordon, M. D. (1990). "Evaluating the Effectiveness of Information Retrieval Systems Using Simulated Queries". *Journal of the American Society for Information Science*. 41(5): 313–323. DOI: 10.1002/(SICI)1097-4571(199007)41:5<313::AID-ASI1>3.0.CO;2-G.

Griffiths, J.-M. (1976). "The Computer Simulation of Information Retrieval Systems". *PhD thesis*. University of London (University College).

Griol, D., J. Carbó, and J. M. Molina. (2013). "An Automatic Dialog Simulation Technique to Develop and Evaluate Interactive Conversational Agents". *Applied Artificial Intelligence*. 27(9): 759–780. DOI: 10.1080/08839514.2013.835230.

Günther, S. and M. Hagen. (2021). "Assessing Query Suggestions for Search Session Simulation". In: *Joint Proceedings of the Causality in Search and Recommendation (CSR) and Simulation of Information Retrieval Evaluation (Sim4IR) Workshops 2021*. 38–45.

Guo, F., C. Liu, A. Kannan, T. Minka, M. Taylor, Y.-M. Wang, and C. Faloutsos. (2009). "Click Chain Model in Web Search". In: *Proceedings of the 18th International Conference on World Wide Web. WWW '09*. 11–20. DOI: 10.1145/1526709.1526712.

Guo, H., R. Tang, Y. Ye, Z. Li, and X. He. (2017). "DeepFM: A Factorization-Machine Based Neural Network for CTR Prediction". In: *Proceedings of the 26th International Joint Conference on Artificial Intelligence. IJCAI'17*. 1725–1731.

Guo, Q., R. W. White, Y. Zhang, B. Anderson, and S. T. Dumais. (2011). "Why Searchers Switch: Understanding and Predicting Engine Switching Rationales". In: *Proceedings of the 34th international ACM SIGIR conference on Research and development in Information Retrieval*. 335–344. DOI: 10.1145/2009916.2009964.

Guo, X., H. Wu, Y. Cheng, S. Rennie, G. Tesauro, and R. Feris. (2018). "Dialog-based Interactive Image Retrieval". In: *Advances in Neural Information Processing Systems. NeurIPS '18*.

Gür, I., D. Hakkani-Tür, G. Tür, and P. Shah. (2018). "User Modeling for Task Oriented Dialogues". In: *2018 IEEE Spoken Language Technology Workshop. SLT '18*. 900–906. DOI: 10.1109/SLT.2018.8639652.

Harman, D. (1992). "Relevance Feedback Revisited". In: *Proceedings of the 15th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. SIGIR '92*. 1–10. DOI: 10.1145/133160.133167.

Harman, D. (1995). "Overview of the Second Text Retrieval Conference (TREC-2)". *Information Processing & Management*. 31(3): 271–289. DOI: 10.1016/0306-4573(94)00047-7.

Hauptmann, A., J. Magalhaes, R. G. Sousa, and J. P. Costeira. (2020). "MuCAI'20: 1st International Workshop on Multimodal Conversational AI". In: *Proceedings of the 28th ACM International Conference on Multimedia. MM '20*. 4767–4768. DOI: 10.1145/3394171.3421900.

Hausman, D. M. (2011). *Preference, Value, Choice, and Welfare*. Cambridge University Press. DOI: 10.1017/CBO9781139058537.

Hawking, D., B. Billerbeck, P. Thomas, and N. Craswell. (2020). *Simulating Information Retrieval Test Collections*. Morgan and Claypool.

Hayati, S. A., D. Kang, Q. Zhu, W. Shi, and Z. Yu. (2020). "INSPIRED: Toward Sociable Recommendation Dialog Systems". In: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing. EMNLP '20*. 8142–8152. DOI: 10.18653/v1/2020.emnlp-main.654.

Hazrati, N. and F. Ricci. (2022). "Recommender Systems Effect on the Evolution of Users' Choices Distribution". *Information Processing & Management*. 59(1). DOI: 10.1016/j.ipm.2021.102766.

Hazrati, N. and F. Ricci. (2024). "Choice Models and Recommender Systems Effects on users' Choices". *User Modeling and User-Adapted Interaction*. 34: 109–145. DOI: 10.1007/s11257-023-09366-x.

Hearst, M. A. (2009). *Search User Interfaces*. Cambridge University Press. DOI: 10.1017/CBO9781139644082.

Hersh, W., A. Turpin, S. Price, B. Chan, D. Kramer, L. Sacherek, and D. Olson. (2000). "Do Batch and User Evaluations Give the Same Results?" In: *Proceedings of the 23rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. SIGIR '00*. 17–24. DOI: 10.1145/345508.345539.

Hofmann, K., L. Li, and F. Radlinski. (2016). "Online Evaluation for Information Retrieval". *Foundations and Trends in Information Retrieval*. 10(1): 1–117. DOI: 10.1561/1500000051.

Hofmann, K., S. Whiteson, and M. de Rijke. (2011). "A Probabilistic Method for Inferring Preferences from Clicks". In: *Proceedings of the 20th ACM International Conference on Information and Knowledge Management. CIKM '11*. 249–258. DOI: 10.1145/2063576.2063618.

Hopfgartner, F., K. Balog, A. Lommatzsch, L. Kelly, B. Kille, A. Schuth, and M. Larson. (2019). "Continuous Evaluation of Large-Scale Information Access Systems: A Case for Living Labs". In: *Information Retrieval Evaluation in a Changing World - Lessons Learned from 20 Years of CLEF*. Ed. by N. Ferro and C. Peters. Vol. 41. *The Information Retrieval Series*. Springer. 511–543. DOI: 10.1007/978-3-030-22948-1_21.

Hu, B., Z. Lu, H. Li, and Q. Chen. (2014). "Convolutional Neural Network Architectures for Matching Natural Language Sentences". In: *Proceedings of the 27th International Conference on Neural Information Processing Systems - Volume 2. NIPS '14*. 2042–2050.

Hu, Z., Y. Feng, A. T. Luu, B. Hooi, and A. 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*. 3953–3957. DOI: 10.1145/3583780.3615220.

Huang, J., R. W. White, G. Buscher, and K. Wang. (2012). "Improving Searcher Models Using Mouse Cursor Activity". In: *Proceedings of the 35th International ACM SIGIR Conference on Research and Development in Information Retrieval. SIGIR '12*. 195–204. DOI: 10.1145/2348283.2348313.

Huang, J., H. Oosterhuis, M. de Rijke, and H. van Hoof. (2020). "Keeping Dataset Biases out of the Simulation: A Debiased Simulator for Reinforcement Learning Based Recommender Systems". In: *Fourteenth ACM Conference on Recommender Systems. RecSys '20*. 190–199. DOI: 10.1145/3383313.3412252.

Hussein, A., M. M. Gaber, E. Elyan, and C. Jayne. (2017). "Imitation Learning: A Survey of Learning Methods". *ACM Computing Surveys*. 50(2): 1–35. DOI: 10.1145/3054912.

Ie, E., C.-w. Hsu, M. Mladenov, V. Jain, S. Narvekar, J. Wang, R. Wu, and C. Boutilier. (2019). "RecSim: A Configurable Simulation Platform for Recommender Systems". arXiv: 1909.04847 [cs.LG].

Ingwersen, P. (1982). "Search Procedures in the Library—Analysed from the Cognitive Point of View". *Journal of Documentation*. 38(3): 165–191. DOI: 10.1108/eb026727.

Ingwersen, P. (1996). "Cognitive Perspectives of Information Retrieval Interaction: Elements of a Cognitive IR Theory". *Journal of Documentation*. 52(1): 3–50. DOI: 10.1108/eb026960.

Ingwersen, P. and K. Järvelin. (2005). *The Turn: Integration of Information Seeking and Retrieval in Context*. Vol. 18. *The Information Retrieval Series*. Springer. DOI: 10.1007/1-4020-3851-8.

Jagerman, R., K. Balog, and M. D. Rijke. (2018). "OpenSearch: Lessons Learned from an Online Evaluation Campaign". *Journal of Data and Information Quality*. 10(3): 13:1–13:15. DOI: 10.1145/3239575.

Jameson, A., B. Berendt, S. Gabrielli, F. Cena, C. Gena, F. Vernero, and K. Reinecke. (2014). "Choice Architecture for Human-Computer Interaction". *Foundations and Trends in Human-Computer Interaction*. 7(1–2): 1–235. DOI: 10.1561/1100000028.

Jameson, A., M. C. Willemsen, and A. Felfernig. (2022). "Individual and Group Decision Making and Recommender Systems". In: *Recommender Systems Handbook*. Ed. by F. Ricci, L. Rokach, and B. Shapira. 3rd ed. Springer US. 789–832. DOI: 10.1007/978-1-0716-2197-4_21.

Jameson, A., M. C. Willemsen, A. Felfernig, M. de Gemmis, P. Lops, G. Semeraro, and L. Chen. (2015). "Human Decision Making and Recommender Systems". In: *Recommender Systems Handbook*. Ed. by F. Ricci, L. Rokach, and B. Shapira. 2nd ed. Springer US. 611–648. DOI: 10.1007/978-1-4899-7637-6_18.

Jannach, D., L. Lerche, I. Kamehkhosh, and M. Jugovac. (2015). "What Recommenders Recommend: An Analysis of Recommendation Biases and Possible Countermeasures". *User Modeling and User-Adapted Interaction*. 25(5): 427–491. DOI: 10.1007/s11257-015-9165-3.

Jannach, D., A. Manzoor, W. Cai, and L. Chen. (2021). "A Survey on Conversational Recommender Systems". *ACM Computing Surveys*. 54(5): 1–36. DOI: 10.1145/3453154.

Jannach, D., S. Naveed, and M. Jugovac. (2017). "User Control in Recommender Systems: Overview and Interaction Challenges". In: *Proceedings of the 17th International Conference on Electronic Commerce and Web Technologies. EC-Web '16*. 21–33. DOI: 10.1007/978-3-319-53676-7_2.

Jansen, B. J., A. Spink, C. Blakely, and S. Koshman. (2007). "Defining a Session on Web Search Engines: Research Articles". *Journal of the American Society for Information Science and Technology*. 58(6): 862–871. DOI: doi.org/10.1002/asi.20564.

Jansen, B. J., A. Spink, and T. Saracevic. (2000). "Real Life, Real Users, and Real Needs: A Study and Analysis of User Queries on the Web". *Information Processing & Management*. 36(2): 207–227. DOI: 10.1016/S0306-4573(99)00056-4.

Järvelin, K. and J. Kekäläinen. (2002). "Cumulated Gain-Based Evaluation of IR Techniques". *ACM Transactions on Information Systems*. 20(4): 422–446. DOI: 10.1145/582415.582418.

Järvelin, K., S. L. Price, L. M. L. Delcambre, and M. L. Nielsen. (2008). "Discounted Cumulated Gain Based Evaluation of Multiple-Query IR Sessions". In: *Proceedings of the 30th European Conference on Advances in Information Retrieval. ECIR '08*. 4–15. DOI: 10.1007/978-3-540-78646-7_4.

Joachims, T., L. Granka, B. Pan, H. Hembrooke, F. Radlinski, and G. Gay. (2007). "Evaluating the Accuracy of Implicit Feedback from Clicks and Query Reformulations in Web Search". *ACM Transactions on Information Systems*. 25(2). DOI: 10.1145/1229179.1229181.

Jordan, C., C. Watters, and Q. Gao. (2006). "Using Controlled Query Generation to Evaluate Blind Relevance Feedback Algorithms". In: *Proceedings of the 6th ACM/IEEE-CS Joint Conference on Digital Libraries. JCDL '06*. 286–295. DOI: 10.1145/1141753.1141818.

Jurafsky, D. and J. H. Martin. (2023). *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition, 3nd Edition draft*. Prentice Hall, Pearson Education International.

Kanoulas, E., B. Carterette, P. D. Clough, and M. Sanderson. (2011). "Evaluating Multi-Query Sessions". In: *Proceedings of the 34th International ACM SIGIR Conference on Research and Development in Information Retrieval. SIGIR '11*. 1053–1062. DOI: 10.1145/2009916.2010056.

Karmaker, S. (K., P. Sondhi, and C. Zhai. (2020). "Empirical Analysis of Impact of Query-Specific Customization of NDCG: A Case-Study with Learning-to-Rank Methods". In: *Proceedings of the 29th ACM International Conference on Information & Knowledge Management. CIKM '20*. 3281–3284. DOI: 10.1145/3340531.3417454.

Käser, T. and G. Alexandron. (2023). "Simulated Learners in Educational Technology: A Systematic Literature Review and a Turing-like Test". *International Journal of Artificial Intelligence in Education*: 1–41. DOI: 10.1007/s40593-023-00337-2.

Kaya, M. and H. Ş. Bilge. (2019). "Deep Metric Learning: A Survey". *Symmetry*. 11(9): 1066. DOI: 10.3390/sym11091066.

Kelly, D. (2009). "Methods for Evaluating Interactive Information Retrieval Systems with Users". *Foundations and Trends in Information Retrieval*. 3(1–2): 1–224. DOI: 10.1561/1500000012.

Keskustalo, H., K. Järvelin, and A. Pirkola. (2006). "The Effects of Relevance Feedback Quality and Quantity in Interactive Relevance Feedback: A Simulation Based on User Modeling". In: *Proceedings of the 28th European Conference on Advances in Information Retrieval. ECIR '06*. 191–204. DOI: 10.1007/11735106_18.

Keskustalo, H., K. Järvelin, and A. Pirkola. (2008). "Evaluating the Effectiveness of Relevance Feedback Based on a User Simulation Model: Effects of a User Scenario on Cumulated Gain Value". *Information Retrieval*. 11(3): 209–228. DOI: 10.1007/s10791-007-9043-7.

Keskustalo, H., K. Järvelin, A. Pirkola, T. Sharma, and M. Lykke. (2009). "Test Collection-Based IR Evaluation Needs Extension toward Sessions — A Case of Extremely Short Queries". In: *Proceedings of the 5th Asia Information Retrieval Symposium on Information Retrieval Technology. AIRS '09*. 63–74. DOI: 10.1007/978-3-642-04769-5_6.

Khashabi, D., S. Min, T. Khot, A. Sabharwal, O. Tafjord, P. Clark, and H. Hajishirzi. (2020). "UNIFIEDQA: Crossing Format Boundaries with a Single QA System". In: *Findings of the Association for Computational Linguistics: EMNLP 2020. EMNLP '20*. 1896–1907. DOI: 10.18653/v1/2020.findings-emnlp.171.

Kiesel, J., M. Gohsen, N. Mirzakhmedova, M. Hagen, and B. Stein. (2024). "Simulating Follow-up Questions in Conversational Search". In: *Proceedings of the 46th European Conference on IR Research. ECIR '24*. 382–398. DOI: 10.1007/978-3-031-56060-6_25.

Kim, S., M. Chang, and S.-W. Lee. (2021). "NeuralWOZ: Learning to Collect Task-Oriented Dialogue via Model-Based Simulation". In: *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers). ACL '21*. 3704–3717. DOI: 10.18653/v1/2021.acl-long.287.

Kim, T. E. and A. Lipani. (2022). "A Multi-Task Based Neural Model to Simulate Users in Goal Oriented Dialogue Systems". In: *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval. SIGIR '22*. 2115–2119. DOI: 10.1145/3477495.3531814.

Kincaid, J. P., R. P. Fishburne, R. L. Rogers, and B. S. Chissom. (1975). "Derivation of New Readability Formulas (Automated Readability Index, Fog Count and Flesch Reading Ease Formula) for Navy Enlisted Personnel". *Tech. rep.* Institute for Simulation and Training.

Klöckner, K., N. Wirschum, and A. Jameson. (2004). "Depth- and Breadth-First Processing of Search Result Lists". In: *CHI '04 Extended Abstracts on Human Factors in Computing Systems. CHI EA '04*. 1539. DOI: 10.1145/985921.986115.

Koren, Y. and R. Bell. (2015). "Advances in Collaborative Filtering". In: *Recommender Systems Handbook*. Ed. by F. Ricci, L. Rokach, and B. Shapira. 2nd ed. Springer US. 77–118. DOI: 10.1007/978-1-4899-7637-6_3.

Kostric, I., K. Balog, T. A. Aresvik, N. Bernard, E. T. Dørheim, P. Hantula, S. Havn-Sørensen, R. Henriksen, H. Hosseini, E. Khlybova, W. Lajewska, S. E. Mosand, and N. Orujova. (2022). "DAGFiNN: A Conversational Conference Assistant". In: *RecSys '22*. 628–631. DOI: 10.1145/3523227.3551467.

Kraft, D. and T. Lee. (1979). "Stopping Rules and their Effect on Expected Search Length". *Information Processing & Management*. 15(1): 47–58. DOI: 10.1016/0306-4573(79)90007-4.

Krauth, K., S. Dean, A. Zhao, W. Guo, M. Curmei, B. Recht, and M. I. Jordan. (2020). "Do Offline Metrics Predict Online Performance in Recommender Systems?" arXiv: 2011.07931 [cs.LG].

Krebs, J. R. (1973). "Behavioral Aspects of Predation". In: *Perspectives in Ethology*. Ed. by P. P. G. Bateson and P. H. Klopfer. 73–111. DOI: 10.1007/978-1-4615-7569-6_3.

Krebs, J. R., J. C. Ryan, and E. L. Charnov. (1974). "Hunting by Expectation or Optimal Foraging? A Study of Patch use by Chickadees". *Animal Behaviour*. 22: 953–964. DOI: 10.1016/0003-3472(74)90018-9.

Kreyssig, F., I. Casanueva, P. Budzianowski, and M. Gašić. (2018). "Neural User Simulation for Corpus-based Policy Optimisation of Spoken Dialogue Systems". In: *Proceedings of the 19th Annual SIGdial Meeting on Discourse and Dialogue. SIGDIAL '18*. 60–69. DOI: 10.18653/v1/W18-5007.

Kuhlthau, C. C. (1991). "Inside the Search Process: Information Seeking from the User's Perspective." *Journal of the American Society for Information Science*. 42(5): 361–371. DOI: 10.1002/(SICI)1097-4571(199106)42:5<361::AID-ASI6>3.0.CO;2-\%23.

Kuhlthau, C. C. (1988). "Developing a Model of the Library Search Process: Cognitive and Affective Aspects". *RQ*. 28(2): 232–242.

Kunkel, J., B. Loepp, and J. Ziegler. (2017). "A 3D Item Space Visualization for Presenting and Manipulating User Preferences in Collaborative Filtering". In: *Proceedings of the 22nd International Conference on Intelligent User Interfaces. IUI '17*. 3–15. DOI: 10.1145/3025171.3025189.

Labhishetty, S. (2023). "Models and Evaluation of User Simulation in Information Retrieval". *PhD thesis*. University of Illinois at Urbana-Champaign.

Labhishetty, S. and C. Zhai. (2022a). "PRE: A Precision-Recall-Effort Optimization Framework for Query Simulation". In: *Proceedings of the 2022 ACM SIGIR International Conference on Theory of Information Retrieval. ICTIR '22*. 51–60. DOI: 10.1145/3539813.3545136.

Labhishetty, S. and C. Zhai. (2022b). "RATE: A Reliability-Aware Tester-Based Evaluation Framework of User Simulators". In: *Proceedings of the 44th European Conference on IR Research. ECIR '22*. 336–350. DOI: 10.1007/978-3-030-99736-6_23.

Labhishetty, S. and C. Zhai. (2021). "An Exploration of Tester-Based Evaluation of User Simulators for Comparing Interactive Retrieval Systems". In: *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval. SIGIR '21*. 1598–1602. DOI: 10.1145/3404835.3463091.

Labhishetty, S., C. Zhai, S. Ranganath, and P. Ranganathan. (2020). "A Cognitive User Model for E-Commerce Search". In: *Proceedings of the Data Science for Retail and E-Commerce Workshop*.

Lagun, D. and E. Agichtein. (2014). "Effects of Task and Domain on Searcher Attention". In: *Proceedings of the 37th International ACM SIGIR Conference on Research and Development in Information Retrieval. SIGIR '14*. 1087–1090. DOI: 10.1145/2600428.2609516.

Lee, S., Q. Zhu, R. Takanobu, Z. Zhang, Y. Zhang, X. Li, J. Li, B. Peng, X. Li, M. Huang, and J. Gao. (2019). "ConvLab: Multi-Domain End-to-End Dialog System Platform". In: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: System Demonstrations. ACL '19*. 64–69. DOI: 10.18653/v1/P19-3011.

Leuski, A. (2000). "Relevance and Reinforcement in Interactive Browsing". In: *Proceedings of the Ninth International Conference on Information and Knowledge Management. CIKM '00*. 119–126. DOI: 10.1145/354756.354809.

Levin, E., R. Pieraccini, and W. Eckert. (2000). "A Stochastic Model of Human-machine Interaction for Learning Dialog Strategies". *IEEE Transactions on Speech and Audio Processing*. 8(1): 11–23. DOI: 10.1109/89.817450.

Li, J., S. Huffman, and A. Tokuda. (2009). "Good Abandonment in Mobile and PC Internet Search". In: *Proceedings of the 32nd International ACM SIGIR Conference on Research and Development in Information Retrieval. SIGIR '09*. 43–50. DOI: 10.1145/1571941.1571951.

Li, J., W. Monroe, T. Shi, S. Jean, A. Ritter, and D. Jurafsky. (2017a). "Adversarial Learning for Neural Dialogue Generation". In: *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing. EMNLP '17*. 2157–2169. DOI: 10.18653/v1/D17-1230.

Li, R., S. Kahou, H. Schulz, V. Michalski, L. Charlin, and C. Pal. (2018). "Towards Deep Conversational Recommendations". In: *Proceedings of the 32nd International Conference on Neural Information Processing Systems. NIPS '18*. 9748–9758.

Li, X., Z. C. Lipton, B. Dhingra, L. Li, J. Gao, and Y.-N. Chen. (2017b). "A User Simulator for Task-Completion Dialogues". arXiv: 1612.05688 [cs.LG].

Li, Z., W. Chen, S. Li, H. Wang, J. Qian, and X. Yan. (2022). "Controllable Dialogue Simulation with In-context Learning". In: *Findings of the Association for Computational Linguistics: EMNLP 2022. EMNLP '22*. 4330–4347. DOI: 10.18653/v1/2022.findings-emnlp.318.

Liao, L., L. H. Long, Z. Zhang, M. Huang, and T.-S. Chua. (2021). "MMConv: An Environment for Multimodal Conversational Search across Multiple Domains". In: *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval. SIGIR '21*. 675–684. DOI: 10.1145/3404835.3462970.

Lin, H.-c., C. Geishauser, S. Feng, N. Lubis, C. van Niekerk, M. Heck, and M. Gasic. (2022). "GenTUS: Simulating User Behaviour and Language in Task-oriented Dialogues with Generative Transformers". In: *Proceedings of the 23rd Annual Meeting of the Special Interest Group on Discourse and Dialogue. SIGDIAL '22*. 270–282. DOI: 10.18653/v1/2022.sigdial-1.28.

Lin, H.-c., N. Lubis, S. Hu, C. van Niekerk, C. Geishauser, M. Heck, S. Feng, and M. Gasic. (2021). "Domain-independent User Simulation with Transformers for Task-oriented Dialogue Systems". In: *Proceedings of the 22nd Annual Meeting of the Special Interest Group on Discourse and Dialogue. SIGDIAL '21*. 445–456. DOI: 10.18653/v1/2021.sigdial-1.47.

Lipani, A., B. Carterette, and E. Yilmaz. (2021). "How Am I Doing?: Evaluating Conversational Search Systems Offline". *ACM Transactions on Information Systems*. 39(4). DOI: 10.1145/3451160.

Liu, Y., M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov. (2019). "RoBERTa: A Robustly Optimized BERT Pretraining Approach". arXiv: 1907.11692 [cs.CL].

Liu, Z., H. Wang, Z.-Y. Niu, H. Wu, and W. Che. (2021a). "DuRecDial 2.0: A Bilingual Parallel Corpus for Conversational Recommendation". In: *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing. EMNLP '21*. 4335–4347. DOI: 10.18653/v1/2021.emnlp-main.356.

Liu, Z., Y. Liu, K. Zhou, M. Zhang, and S. Ma. (2015). "Influence of Vertical Result in Web Search Examination". In: *Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval. SIGIR '15*. 193–202. DOI: 10.1145/2766462.2767714.

Liu, Z., K. Zhou, and M. L. Wilson. (2021b). "Meta-Evaluation of Conversational Search Evaluation Metrics". *ACM Transactions on Information Systems*. 39(4). DOI: 10.1145/3445029.

Louis, A., D. Roth, and F. Radlinski. (2020). ""I'd rather just go to bed": Understanding Indirect Answers". In: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing. EMNLP '20*. 7411–7425. DOI: 10.18653/v1/2020.emnlp-main.601.

Luger, E. and A. Sellen. (2016). "“Like Having a Really Bad PA”: The Gulf between User Expectation and Experience of Conversational Agents". In: *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems. CHI '16*. 5286–5297. DOI: 10.1145/2858036.2858288.

Lyu, S., A. Rana, S. Sanner, and M. R. Bouadjenek. (2021). "A Workflow Analysis of Context-Driven Conversational Recommendation". In: *Proceedings of the Web Conference 2021. WWW '21*. 866–877. DOI: 10.1145/3442381.3450123.

Machmouchi, W., A. H. Awadallah, I. Zitouni, and G. Buscher. (2017). "Beyond Success Rate: Utility as a Search Quality Metric for Online Experiments". In: *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*. 757–765. DOI: 10.1145/3132847.3132850.

Marchionini, G. (1995). *Information Seeking in Electronic Environments*. Cambridge University Press. DOI: 10.1017/CBO9780511626388.

Marchionini, G. (2006). "Exploratory Search: From Finding to Understanding". *Communications of the ACM*. 49(4): 41–46. DOI: 10.1145/1121949.1121979.

Markey, K. (2007). "Twenty-Five Years of End-User Searching, Part 1: Research Findings". *Journal of the American Society for Information Science and Technology*. 58(8): 1071–1081. DOI: 10.1002/asi.20462.

Maxwell, D. (2019). "Modelling Search and Stopping in Interactive Information Retrieval". *PhD thesis*. University of Glasgow.

Maxwell, D. and L. Azzopardi. (2016a). "Agents, Simulated Users and Humans: An Analysis of Performance and Behaviour". In: *Proceedings of the 25th ACM International on Conference on Information and Knowledge Management. CIKM '16*. 731–740. DOI: 10.1145/2983323.2983805.

Maxwell, D. and L. Azzopardi. (2016b). "Simulating Interactive Information Retrieval: SimIIR: A Framework for the Simulation of Interaction". In: *Proceedings of the 39th International ACM SIGIR Conference on Research and Development in Information Retrieval. SIGIR '16*. 1141–1144. DOI: 10.1145/2911451.2911469.

Maxwell, D. and L. Azzopardi. (2018). "Information Scent, Searching and Stopping: Modelling SERP Level Stopping Behaviour". In: *Proceedings of the 40th European Conference on IR Research. ECIR '18*. 210–222. DOI: 10.1007/978-3-319-76941-7_16.

Maxwell, D., L. Azzopardi, K. Järvelin, and H. Keskustalo. (2015a). "An Initial Investigation into Fixed and Adaptive Stopping Strategies". In: *Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval. SIGIR '15*. 903–906. DOI: 10.1145/2766462.2767802.

Maxwell, D., L. Azzopardi, K. Järvelin, and H. Keskustalo. (2015b). "Searching and Stopping: An Analysis of Stopping Rules and Strategies". In: *Proceedings of the 24th ACM International on Conference on Information and Knowledge Management. CIKM '15*. 313–322. DOI: 10.1145/2806416.2806476.

McGinty, L. and J. Reilly. (2011). "On the Evolution of Critiquing Recommenders". In: *Recommender Systems Handbook*. Ed. by F. Ricci, L. Rokach, B. Shapira, and P. B. Kantor. Springer US. 419–453. DOI: 10.1007/978-0-387-85820-3_13.

McInerney, J., E. Elahi, J. Basilico, Y. Raimond, and T. Jebara. (2021). "Accordion: A Trainable Simulator for Long-Term Interactive Systems". In: *Fifteenth ACM Conference on Recommender Systems. RecSys '21*. 102–113. DOI: 10.1145/3460231.3474259.

McLaughlin, G. H. (1969). "SMOG Grading-a New Readability Formula". *Journal of Reading*. 12(8): 639–646.

McTear, M. (2021). *Conversational AI: Dialogue Systems, Conversational Agents, and Chatbots*. Springer Nature. DOI: 10.1007/978-3-031-02176-3.

McTear, M. and M. Ashurkina. (2024). *Transforming Conversational AI: Exploring the Power of Large Language Models in Interactive Conversational Agents*. Apress Berkeley, CA. DOI: 10.1007/979-8-8688-0110-5.

McTear, M., Z. Callejas, and D. Griol. (2016). *The Conversational Interface: Talking to Smart Devices*. Springer Publishing Company, Incorporated.

Meho, L. I. and H. R. Tibbo. (2003). "Modeling the Information-Seeking Behavior of Social Scientists: Ellis's Study Revisited". *Journal of the American Society for Information Science and Technology*. 54(6): 570–587. DOI: 10.1002/asi.10244.

Meij, E., W. Weerkamp, and M. de Rijke. (2009). "A Query Model Based on Normalized Log-Likelihood". In: *Proceedings of the 18th ACM Conference on Information and Knowledge Management. CIKM '09*. 1903–1906. DOI: 10.1145/1645953.1646261.

Merinov, P., D. Massimo, and F. Ricci. (2023). "Behaviour-aware Tourist Profiles Data Generation". In: *Proceedings of the 13th Italian Information Retrieval Workshop. IIR '23*. 3–8.

Meyer, S., D. Elsweiler, B. Ludwig, M. Fernandez-Pichel, and D. E. Losada. (2022). "Do We Still Need Human Assessors? Prompt-Based GPT-3 User Simulation in Conversational AI". In: *Proceedings of the 4th Conference on Conversational User Interfaces. CUI '22*. 1–6. DOI: 10.1145/3543829.3544529.

Miller, A. H., W. Feng, A. Fisch, J. Lu, D. Batra, A. Bordes, D. Parikh, and J. Weston. (2018). "ParlAI: A Dialog Research Software Platform". arXiv: 1705.06476 [cs.CL].

Mladenov, M., C.-W. Hsu, V. Jain, E. Ie, C. Colby, N. Mayoraz, H. Pham, D. Tran, I. Vendrov, and C. Boutilier. (2021). "RecSim NG: Toward Principled Uncertainty Modeling for Recommender Ecosystems". arXiv: 2103.08057 [cs.LG].

Moerland, T. M., J. Broekens, A. Plaat, C. M. Jonker, *et al.* (2023). "Model-based reinforcement learning: A survey". *Foundations and Trends® in Machine Learning*. 16(1): 1–118.

Moffat, A., P. Bailey, F. Scholer, and P. Thomas. (2017). "Incorporating User Expectations and Behavior into the Measurement of Search Effectiveness". *ACM Transactions on Information Systems*. 35(3). DOI: 10.1145/3052768.

Moffat, A., J. Mackenzie, P. Thomas, and L. Azzopardi. (2022). "A Flexible Framework for Offline Effectiveness Metrics". In: *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval. SIGIR '22*. 578–587. DOI: 10.1145/3477495.3531924.

Moffat, A., P. Thomas, and F. Scholer. (2013). "Users versus Models: What Observation Tells Us about Effectiveness Metrics". In: *Proceedings of the 22nd ACM International Conference on Information and Knowledge Management. CIKM '13*. 659–668. DOI: 10.1145/2505515.2507665.

Moffat, A. and J. Zobel. (2008). "Rank-biased Precision for Measurement of Retrieval Effectiveness". *ACM Transactions on Information Systems*. 27(1): 1–27. DOI: 10.1145/1416950.1416952.

Mohapatra, B., G. Pandey, D. Contractor, and S. Joshi. (2021). "Simulated Chats for Building Dialog Systems: Learning to Generate Conversations from Instructions". In: *Findings of the Association for Computational Linguistics: EMNLP 2021*. 1190–1203. DOI: 10.18653/v1/2021.findings-emnlp.103.

Moniz, L., A. L. Buczak, L. Hung, S. Babin, M. Dorko, and J. Lombardo. (2009). "Construction and Validation of Synthetic Electronic Medical Records". *Online Journal of Public Health Informatics*. 1(1). DOI: 10.5210/ojphi.v1i1.2720.

Mostafa, J., S. Mukhopadhyay, and M. Palakal. (2003). "Simulation Studies of Different Dimensions of Users' Interests and their Impact on User Modeling and Information Filtering". *Information retrieval*. 6(2): 199–223. DOI: 10.1023/A:1023932221048.

Navalpakkam, V., L. Jentzsch, R. Sayres, S. Ravi, A. Ahmed, and A. Smola. (2013). "Measurement and Modeling of Eye-Mouse Behavior in the Presence of Nonlinear Page Layouts". In: *Proceedings of the 22nd International Conference on World Wide Web. WWW '13*. 953–964. DOI: 10.1145/2488388.2488471.

Nguyen, T. N., F. Ricci, A. Delic, and D. Bridge. (2019). "Conflict Resolution in Group Decision Making: Insights from a Simulation Study". *User Modeling and User-Adapted Interaction*. 29: 895–941. DOI: 10.1007/s11257-019-09240-9.

Nguyen, T. T., D. Kluver, T.-Y. Wang, P.-M. Hui, M. D. Ekstrand, M. C. Willemsen, and J. Riedl. (2013). "Rating Support Interfaces to Improve User Experience and Recommender Accuracy". In: *Proceedings of the 7th ACM Conference on Recommender Systems. RecSys '13*. 149–156. DOI: 10.1145/2507157.2507188.

Ni, J., T. Young, V. Pandelea, F. Xue, and E. Cambria. (2022). "Recent Advances in Deep Learning Based Dialogue Systems: A Systematic Survey". *Artificial Intelligence Review*. 56(4): 3055–3155. DOI: 10.1007/s10462-022-10248-8.

Nickles, K. R. (1995). "Judgment-based and Reasoning-based Stopping Rules in Decision Making Under Uncertainty". *PhD thesis*. University of Minnesota.

Novikova, J., O. Dušek, A. Cercas Curry, and V. Rieser. (2017). "Why We Need New Evaluation Metrics for NLG". In: *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing. EMNLP '17*. 2241–2252. DOI: 10.18653/v1/D17-1238.

O'Day, V. L. and R. Jeffries. (1993). "Orienteering in an Information Landscape: How Information Seekers Get from Here to There". In: *Proceedings of the INTERACT '93 and CHI '93 Conference on Human Factors in Computing Systems. CHI '93*. 438–445. DOI: 10.1145/169059.169365.

Olston, C. and E. H. Chi. (2003). "ScentTrails: Integrating Browsing and Searching on the Web". *ACM Transactions on Computer-Human Interaction*. 10(3): 177–197. DOI: 10.1145/937549.937550.

Oosterhuis, H. and M. de Rijke. (2018). "Ranking for Relevance and Display Preferences in Complex Presentation Layouts". In: *The 41st International ACM SIGIR Conference on Research and Development in Information Retrieval. SIGIR '18*. 845–854. DOI: 10.1145/3209978.3209992.

Ouyang, L., J. Wu, X. Jiang, D. Almeida, C. Wainwright, P. Mishkin, C. Zhang, S. Agarwal, K. Slama, A. Ray, J. Schulman, J. Hilton, F. Kelton, L. Miller, M. Simens, A. Askell, P. Welinder, P. F. Christiano, J. Leike, and R. Lowe. (2022). "Training Language Models to Follow Instructions with Human Feedback". In: *Advances in Neural Information Processing Systems. NeurIPS '22*. 27730–27744.

Over, P. (2001). "The TREC Interactive Track: An Annotated Bibliography". *Information Processing & Management*. 37(3): 369–381. DOI: 10.1016/S0306-4573(00)00053-4.

Owoicho, P., I. Sekulic, M. Alianejadi, J. Dalton, and F. Crestani. (2023). "Exploiting Simulated User Feedback for Conversational Search: Ranking, Rewriting, and Beyond". In: *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval. SIGIR '23*. 632–642. DOI: 10.1145/3539618.3591683.

Pääkkönen, T., K. Järvelin, J. Kekäläinen, H. Keskustalo, F. Baskaya, D. Maxwell, and L. Azzopardi. (2015). "Exploring Behavioral Dimensions in Session Effectiveness". In: *Proceedings of the 6th International Conference on Experimental IR Meets Multilinguality, Multimodality, and Interaction. CLEF '15*. 178–189. DOI: 10.1007/978-3-319-24027-5_15.

Pääkkönen, T., J. Kekäläinen, H. Keskustalo, L. Azzopardi, D. Maxwell, and K. Järvelin. (2017). "Validating Simulated Interaction for Retrieval Evaluation". *Information Retrieval*. 20(4): 338–362. DOI: 10.1007/s10791-017-9301-2.

Papangelis, A., M. Namazifar, C. Khatri, Y.-C. Wang, P. Molino, and G. Tur. (2020). "Plato Dialogue System: A Flexible Conversational AI Research Platform". arXiv: 2001.06463 [cs.HC].

Parapar, J. and F. Radlinski. (2021). "Towards Unified Metrics for Accuracy and Diversity for Recommender Systems". In: *Proceedings of the 15th ACM Conference on Recommender Systems. RecSys '21*. 75–84. DOI: 10.1145/3460231.3474234.

Park, J. S., J. O'Brien, C. J. Cai, M. R. Morris, P. Liang, and M. S. Bernstein. (2023). "Generative Agents: Interactive Simulacra of Human Behavior". In: *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology. UIST '23*. DOI: 10.1145/3586183.3606763.

Pejtersen, A. (1980). "Design of a Classification Scheme for Fiction Based on an Analysis of Actual User-librarian Communication, and Use of the Scheme for Control of Librarians' Search Strategies". *Theory and Application of Information Research*: 146–159.

Pietquin, O. (2004). "A Framework for Unsupervised Learning of Dialogue Strategies". *PhD thesis*. Faculté Polytechnique de Mons, Belgium.

Pietquin, O. and H. Hastie. (2013). "A Survey on Metrics for the Evaluation of User Simulations". *The Knowledge Engineering Review*. 28(1): 59–73. DOI: 10.1017/S0269888912000343.

Pirolli, P. (2007). *Information Foraging Theory: Adaptive Interaction with Information*. Oxford University Press. DOI: 10.1093/acprof:oso/9780195173321.001.0001.

Pirolli, P. and S. Card. (1999). "Information Foraging". *Psychological Review*. 106(4): 643–675. DOI: 10.1037/0033-295X.106.4.643.

Qin, T., T.-Y. Liu, J. Xu, and H. Li. (2010). "LETOR: A Benchmark Collection for Research on Learning to Rank for Information Retrieval". *Information Retrieval*. 13(4): 346–374. DOI: 10.1007/s10791-009-9123-y.

Qu, C., L. Yang, W. B. Croft, J. R. Trippas, Y. Zhang, and M. Qiu. (2018). "Analyzing and Characterizing User Intent in Information-Seeking Conversations". In: *The 41st International ACM SIGIR Conference on Research and Development in Information Retrieval. SIGIR '18*. 989–992. DOI: 10.1145/3209978.3210124.

Radford, A., J. Wu, R. Child, D. Luan, D. Amodei, and I. Sutskever. (2019). "Language Models are Unsupervised Multitask Learners".

Radlinski, F. and N. Craswell. (2017). "A Theoretical Framework for Conversational Search". In: *Proceedings of the 2017 Conference on Conference Human Information Interaction and Retrieval. CHIIR '17*. 117–126. DOI: 10.1145/3020165.3020183.

Raffel, C., N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, and P. J. Liu. (2020). "Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer". *Journal of Machine Learning Research*. 21(140): 1–67.

Rahdari, B. and P. Brusilovsky. (2022). "Simulation-Based Evaluation of Interactive Recommender Systems". In: *Proceedings of the 9th Joint Workshop on Interfaces and Human Decision Making for Recommender Systems. IntRS '22*. 122–136.

Rahdari, B., P. Brusilovsky, and B. Kveton. (2024). "Towards Simulation-Based Evaluation of Recommender Systems with Carousel Interfaces". *ACM Transactions on Recommender Systems*. 2(1). DOI: 10.1145/3643709.

Rahdari, B., B. Kveton, and P. Brusilovsky. (2022). "The Magic of Carousels: Single vs. Multi-List Recommender Systems". In: *Proceedings of the 33rd ACM Conference on Hypertext and Social Media. HT '22*. 166–174. DOI: 10.1145/3511095.3531278.

Rastogi, A., X. Zang, S. Sunkara, R. Gupta, and P. Khaitan. (2020). "Towards Scalable Multi-domain Conversational Agents: The Schema-Guided Dialogue Dataset". In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 34. *AAAI '20*. No. 05. 8689–8696. DOI: 10.1609/aaai.v34i05.6394.

Rohde, D., S. Bonner, T. Dunlop, F. Vasile, and A. Karatzoglou. (2018). "RecoGym: A Reinforcement Learning Environment for the problem of Product Recommendation in Online Advertising". arXiv: 1808.00720 [cs.IR].

Roy, N., D. Maxwell, and C. Hauff. (2022). "Users and Contemporary SERPs: A (Re-)Investigation". In: *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval. SIGIR '22*. 2765–2775. DOI: 10.1145/3477495.3531719.

Ruthven, I. (2008). "Interactive Information Retrieval". *Annual Review of Information Science and Technology*. 42(1): 43–91. DOI: 10.1002/aris.2008.1440420109.

Sacks, H., E. A. Schegloff, and G. D. Jefferson. (1974). "A Simplest Systematics for the Organization of Turn-Taking for Conversation". *Language*. 50(4): 696–735.

Sakai, T., D. W. Oard, and N. Kando. (2021). *Evaluating Information Retrieval and Access Tasks: NTCIR's Legacy of Research Impact*. Springer Nature.

Sakai, T. and R. Song. (2011). "Evaluating Diversified Search Results Using Per-Intent Graded Relevance". In: *Proceedings of the 34th International ACM SIGIR Conference on Research and Development in Information Retrieval. SIGIR '11*. 1043–1052. DOI: 10.1145/2009916.2010055.

Salle, A., S. Malmasi, O. Rokhlenko, and E. Agichtein. (2021). "Studying the Effectiveness of Conversational Search Refinement Through User Simulation". In: *Proceedings of the 43rd European Conference on IR Research. ECIR '21*. 587–602. DOI: 10.1007/978-3-030-72113-8_39.

Salton, G. (1970). “Evaluation problems in Interactive Information Retrieval”. *Information Storage and Retrieval*. 6(1): 29–44. DOI: 10.1016/0020-0271(70)90011-2.

Sanderson, M. (2010). “Test Collection Based Evaluation of Information Retrieval Systems”. *Foundations and Trends in Information Retrieval*. 4(4): 247–375. DOI: 10.1561/1500000009.

Sansone, C. and G. Sperlí. (2022). “Legal Information Retrieval Systems: State-of-the-art and Open Issues”. *Information Systems*. 106: 101967. DOI: 10.1016/j.is.2021.101967.

Savolainen, R. (2018). “Pioneering Models for Information Interaction in the Context of Information Seeking and Retrieval”. *Journal of Documentation*. 74(5): 966–986. DOI: 10.1108/JD-11-2017-0154.

Schatzmann, J., K. Georgila, and S. J. Young. (2005). “Quantitative Evaluation of User Simulation Techniques for Spoken Dialogue Systems”. In: *Proceedings of the 6th SIGdial Workshop on Discourse and Dialogue. SIGDIAL ’05*. 45–54.

Schatzmann, J., B. Thomson, K. Weilhammer, H. Ye, and S. Young. (2007). “Agenda-Based User Simulation for Bootstrapping a POMDP Dialogue System”. In: *Human Language Technologies 2007: The Conference of the North American Chapter of the Association for Computational Linguistics; Companion Volume, Short Papers. NAACL-HLT ’07*. 149–152.

Schatzmann, J., K. Weilhammer, M. Stuttle, and S. Young. (2006). “A Survey of Statistical User Simulation Techniques for Reinforcement-Learning of Dialogue Management Strategies”. *The Knowledge Engineering Review*. 21(2): 97–126. DOI: 10.1017/S0269888906000944.

Schatzmann, J. and S. Young. (2009). “The Hidden Agenda User Simulation Model”. *IEEE Transactions on Audio, Speech, and Language Processing*. 17(4): 733–747. DOI: 10.1109/TASL.2008.2012071.

Schedl, M., P. Knees, B. McFee, D. Bogdanov, and M. Kaminskas. (2015). “Music Recommender Systems”. In: *Recommender Systems Handbook*. Ed. by F. Ricci, L. Rokach, and B. Shapira. 2nd ed. Springer US. 453–492. DOI: 10.1007/978-1-4899-7637-6_13.

Sekulić, I., M. Aliannejadi, and F. Crestani. (2022). "Evaluating Mixed-Initiative Conversational Search Systems via User Simulation". In: *Proceedings of the Fifteenth ACM International Conference on Web Search and Data Mining. WSDM '22*. 888–896. DOI: 10.1145/3488560.3498440.

Sekulić, I., M. Aliannejadi, and F. Crestani. (2024). "Analysing Utterances in LLM-based User Simulation for Conversational Search". *ACM Transactions on Intelligent Systems and Technology*. DOI: 10.1145/3650041.

Serban, I. V., A. Sordoni, Y. Bengio, A. Courville, and J. Pineau. (2016). "Building End-to-End Dialogue Systems Using Generative Hierarchical Neural Network Models". In: *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence. AAAI '16*. 3776–3783. DOI: 10.1609/aaai.v30i1.9883.

Shao, Y., J. Mao, Y. Liu, M. Zhang, and S. Ma. (2022). "From Linear to Non-linear: Investigating the Effects of Right-rail Results on Complex SERPs". *Advances in Computational Intelligence*. 2(1): 14. DOI: 10.1007/s43674-021-00028-2.

Shi, B., M. G. Ozsoy, N. Hurley, B. Smyth, E. Z. Tragos, J. Geraci, and A. Lawlor. (2019a). "PyRecGym: A Reinforcement Learning Gym for Recommender Systems". In: *Proceedings of the 13th ACM Conference on Recommender Systems. RecSys '19*. 491–495. DOI: 10.1145/3298689.3346981.

Shi, W., K. Qian, X. Wang, and Z. Yu. (2019b). "How to Build User Simulators to Train RL-based Dialog Systems". In: *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing. EMNLP-IJCNLP '19*. 1990–2000. DOI: 10.18653/v1/D19-1206.

Singh, S., M. Kearns, D. Litman, and M. Walker. (1999). "Reinforcement Learning for Spoken Dialogue Systems". In: *Proceedings of the 12th International Conference on Neural Information Processing Systems. NIPS '99*. 956–962.

Singh, S., D. Litman, M. Kearns, and M. Walker. (2002). "Optimizing Dialogue Management with Reinforcement Learning: Experiments with the NJFun System". *Journal of Artificial Intelligence Research*. 16(1): 105–133. DOI: 10.1613/jair.859.

Smith, C. L. and P. B. Kantor. (2008). "User Adaptation: Good Results from Poor Systems". In: *Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. SIGIR '08*. 147–154. DOI: 10.1145/1390334.1390362.

Smucker, M. D. (2011). "An Analysis of User Strategies for Examining and Processing Ranked Lists of Documents". In: *Proceedings of the Fifth Workshop on Human-Computer Interaction and Information Retrieval. HCIR '11*.

Smucker, M. D. and C. L. A. Clarke. (2012a). "Modeling User Variance in Time-Biased Gain". In: *Proceedings of the Symposium on Human-Computer Interaction and Information Retrieval. HCIR '12*. DOI: 10.1145/2391224.2391227.

Smucker, M. D. and C. L. Clarke. (2012b). "Time-Based Calibration of Effectiveness Measures". In: *Proceedings of the 35th International ACM SIGIR Conference on Research and Development in Information Retrieval. SIGIR '12*. 95–104. DOI: 10.1145/2348283.2348300.

Smucker, M. D. and C. L. Clarke. (2016). "Modeling Optimal Switching Behavior". In: *Proceedings of the 2016 ACM on Conference on Human Information Interaction and Retrieval. CHIIR '16*. 317–320. DOI: 10.1145/2854946.2854981.

Smyth, B. (2007). "Case-Based Recommendation". In: *The Adaptive Web: Methods and Strategies of Web Personalization*. Ed. by P. Brusilovsky, A. Kobsa, and W. Nejdl. Springer Berlin Heidelberg. 342–376. DOI: 10.1007/978-3-540-72079-9_11.

Sondhi, P., M. Sharma, P. Kolari, and C. Zhai. (2018). "A Taxonomy of Queries for E-Commerce Search". In: *The 41st International ACM SIGIR Conference on Research and Development in Information Retrieval. SIGIR '18*. 1245–1248. DOI: 10.1145/3209978.3210152.

Spärck Jones, K. (1979). "Search Term Relevance Weighting given Little Relevance Information". *Journal of Documentation*. 35(1): 30–48. DOI: 10.1108/eb026672.

Stavinova, E., A. Grigorievskiy, A. Volodkevich, P. Chunaev, K. Bochenina, and D. Bugaychenko. (2022). "Synthetic Data-Based Simulators for Recommender Systems: A Survey". arXiv: 2206.11338 [cs.IR].

Steck, H. (2010). "Training and Testing of Recommender Systems on Data Missing not at Random". In: *Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. KDD '10*. 713–722. DOI: 10.1145/1835804.1835895.

Steck, H. (2013). "Evaluation of Recommendations: Rating-prediction and Ranking". In: *Proceedings of the 7th ACM Conference on Recommender Systems. RecSys '13*. 213–220. DOI: 10.1145/2507157.2507160.

Stephens, D. W. and J. R. Krebs. (1986). *Foraging Theory*. Princeton University Press.

Su, L. T. (1992). "Evaluation Measures for Interactive Information Retrieval". *Information Processing & Management*. 28(4): 503–516. DOI: 10.1016/0306-4573(92)90007-M.

Sun, W., S. Guo, S. Zhang, P. Ren, Z. Chen, M. de Rijke, and Z. Ren. (2023). "Metaphorical User Simulators for Evaluating Task-oriented Dialogue Systems". *ACM Transactions on Information Systems*. 42(1). DOI: 10.1145/3596510.

Sun, W., S. Zhang, K. Balog, Z. Ren, P. Ren, Z. Chen, and M. de Rijke. (2021). "Simulating User Satisfaction for the Evaluation of Task-Oriented Dialogue Systems". In: *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval. SIGIR '21*. 2499–2506. DOI: 10.1145/3404835.3463241.

Szlávik, Z., W. Kowalczyk, and M. C. Schut. (2011). "Diversity Measurement of Recommender Systems under Different User Choice Models". In: *Proceedings of the International AAAI Conference on Web and Social Media. ICWSM '11*. 369–376. DOI: 10.1609/icwsm.v5i1.14116.

Tague, J., M. Nelson, and H. Wu. (1980). "Problems in the Simulation of Bibliographic Retrieval Systems". In: *Proceedings of the 3rd Annual ACM Conference on Research and Development in Information Retrieval. SIGIR '80*. 236–255.

Taylor, A. R., C. Cool, N. J. Belkin, and W. J. Amadio. (2007). "Relationships between Categories of Relevance Criteria and Stage in Task Completion". *Information Processing & Management*. 43(4): 1071–1084. DOI: 10.1016/j.ipm.2006.09.008.

Terragni, S., M. Filipavicius, N. Khau, B. Guedes, A. Manso, and R. Mathis. (2023). "In-Context Learning User Simulators for Task-Oriented Dialog Systems". arXiv: 2306.00774 [cs.CL].

Thomas, P., M. Czerwinksi, D. Mcduff, and N. Craswell. (2021). "Theories of Conversation for Conversational IR". *ACM Transactions on Information Systems*. 39(4). DOI: 10.1145/3439869.

Thomas, P., A. Moffat, P. Bailey, and F. Scholer. (2014). "Modeling Decision Points in User Search Behavior". In: *Proceedings of the 5th Information Interaction in Context Symposium. IIiX '14*. 239–242. DOI: 10.1145/2637002.2637032.

Thomas, P., A. Moffat, P. Bailey, F. Scholer, and N. Craswell. (2018). "Better Effectiveness Metrics for SERPs, Cards, and Rankings". In: *Proceedings of the 23rd Australasian Document Computing Symposium. ADCS '18*. 1–8. DOI: 10.1145/3291992.3292002.

Thomas, P., F. Scholer, and A. Moffat. (2013). "What Users Do: The Eyes Have It". In: *Proceedings of the 9th Asian Information Retrieval Societies Conference. AIRS '13*. 416–427. DOI: 10.1007/978-3-642-45068-6_36.

Trippas, J. R., D. Spina, L. Cavedon, H. Joho, and M. Sanderson. (2018). "Informing the Design of Spoken Conversational Search: Perspective Paper". In: *Proceedings of the 2018 Conference on Human Information Interaction and Retrieval. CHIIR '18*. 32–41. DOI: 10.1145/3176349.3176387.

Trippas, J. R., D. Spina, P. Thomas, M. Sanderson, H. Joho, and L. Cavedon. (2020). "Towards a Model for Spoken Conversational Search". *Information Processing & Management*. 57(2). DOI: 10.1016/j.ipm.2019.102162.

Tseng, B.-H., Y. Dai, F. Kreyssig, and B. Byrne. (2021). "Transferable Dialogue Systems and User Simulators". In: *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers). ACL '21*. 152–166. DOI: 10.18653/v1/2021.acl-long.13.

Turpin, A. and F. Scholer. (2006). "User Performance versus Precision Measures for Simple Search Tasks". In: *Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. SIGIR '06*. 11–18. DOI: 10.1145/1148170.1148176.

Turpin, A., F. Scholer, K. Jarvelin, M. Wu, and J. S. Culpepper. (2009). "Including Summaries in System Evaluation". In: *Proceedings of the 32nd International ACM SIGIR Conference on Research and Development in Information Retrieval. SIGIR '09*. 508–515. DOI: 10.1145/1571941.1572029.

Turpin, A. H. and W. Hersh. (2001). "Why Batch and User Evaluations Do Not Give the Same Results". In: *Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. SIGIR '01*. 225–231. DOI: 10.1145/383952.383992.

Ultes, S., L. M. Rojas-Barahona, P.-H. Su, D. Vandyke, D. Kim, I. Casanueva, P. Budzianowski, N. Mrkšić, T.-H. Wen, M. Gašić, and S. Young. (2017). "PyDial: A Multi-domain Statistical Dialogue System Toolkit". In: *Proceedings of ACL 2017, System Demonstrations. ACL '17*. 73–78.

Vakkari, P. (2016). "Searching as Learning: A Systematization based on Literature". *Journal of Information Science*. 42(1): 7–18.

Vakulenko, S., K. Revoredo, C. Di Ciccio, and M. de Rijke. (2019). "QRFA: A Data-Driven Model of Information-Seeking Dialogues". In: *Proceedings of the 41st European Conference on IR Research. ECIR '19*. 541–557. DOI: 10.1007/978-3-030-15712-8_35.

Van Rijsbergen, C. J. (1979). *Information Retrieval*. 2nd ed. Butterworth-Heinemann.

Vaswani, A., N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin. (2017). "Attention is All You Need". In: *Proceedings of the 31st International Conference on Neural Information Processing Systems. NIPS '17*. 6000–6010.

Verberne, S., M. Sappelli, K. Järvelin, and W. Kraaij. (2015). "User Simulations for Interactive Search: Evaluating Personalized Query Suggestion". In: *Proceedings of the 37th European Conference on IR Research. ECIR '15*. 678–690. DOI: 10.1007/978-3-319-16354-3_75.

Verbert, K., D. Parra, and P. Brusilovsky. (2016). "Agents Vs. Users: Visual Recommendation of Research Talks with Multiple Dimension of Relevance". *ACM Transactions on Information Systems*. 6(2). DOI: 10.1145/2946794.

Victor, P., M. De Cock, and C. Cornelis. (2011). "Trust and Recommendations". In: *Recommender Systems Handbook*. Ed. by F. Ricci, L. Rokach, B. Shapira, and P. B. Kantor. Springer US. 645–675. DOI: 10.1007/978-0-387-85820-3_20.

Vlachou, M. and C. Macdonald. (2024). "What Else Would I Like? A User Simulator using Alternatives for Improved Evaluation of Fashion Conversational Recommendation Systems". arXiv: 2401.05783 [cs.IR].

Voorhees, E. M. (2000). "Variations in Relevance Judgments and the Measurement of Retrieval Effectiveness". *Information processing & management*. 36(5): 697–716. DOI: 10.1016/S0306-4573(00)00010-8.

Voorhees, E. M. and D. K. Harman. (2005). *TREC: Experiment and Evaluation in Information Retrieval*. The MIT Press.

Wadhwa, S. and H. Zamani. (2021). "Towards System-Initiative Conversational Information Seeking". In: *Proceedings of the Second International Conference on Design of Experimental Search and Information REtrieval Systems. DESIRES '21*. 102–116.

Walker, M. A., J. C. Fromer, and S. Narayanan. (1998). "Learning Optimal Dialogue Strategies: A Case Study of a Spoken Dialogue Agent for Email". In: *COLING 1998 Volume 2: The 17th International Conference on Computational Linguistics*.

Wan, D., Z. Zhang, Q. Zhu, L. Liao, and M. Huang. (2022). “A Unified Dialogue User Simulator for Few-shot Data Augmentation”. In: *Findings of the Association for Computational Linguistics: EMNLP 2022*. 3788–3799. DOI: 10.18653/v1/2022.findings-emnlp.277.

Wang, C., Y. Liu, M. Zhang, S. Ma, M. Zheng, J. Qian, and K. Zhang. (2013a). “Incorporating Vertical Results into Search Click Models”. In: *Proceedings of the 36th International ACM SIGIR Conference on Research and Development in Information Retrieval. SIGIR ’13*. 503–512. DOI: 10.1145/2484028.2484036.

Wang, H., C. Zhai, A. Dong, and Y. Chang. (2013b). “Content-Aware Click Modeling”. In: *Proceedings of the 22nd international conference on World Wide Web. WWW ’13*. 1365–1376. DOI: 10.1145/2488388.2488508.

Wang, L., C. Ma, X. Feng, Z. Zhang, H. Yang, J. Zhang, Z. Chen, J. Tang, X. Chen, Y. Lin, W. X. Zhao, Z. Wei, and J. Wen. (2024a). “A Survey on Large Language Model based Autonomous Agents”. *Frontiers of Computer Science*. 18. DOI: 10.1007/s11704-024-40231-1.

Wang, L., J. Zhang, H. Yang, Z. Chen, J. Tang, Z. Zhang, X. Chen, Y. Lin, R. Song, W. X. Zhao, J. Xu, Z. Dou, J. Wang, and J.-R. Wen. (2024b). “User Behavior Simulation with Large Language Model based Agents”. arXiv: 2306.02552 [cs.IR].

Wang, X., X. Tang, X. Zhao, J. Wang, and J.-R. Wen. (2023a). “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. DOI: 10.18653/v1/2023.emnlp-main.621.

Wang, X., K. Zhou, X. Tang, W. X. Zhao, F. Pan, Z. Cao, and J.-R. Wen. (2023b). “Improving Conversational Recommendation Systems via Counterfactual Data Simulation”. In: *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining. KDD ’23*. 2398–2408. DOI: 10.1145/3580305.3599387.

Wang, Y., D. Yin, L. Jie, P. Wang, M. Yamada, Y. Chang, and Q. Mei. (2016). "Beyond Ranking: Optimizing Whole-Page Presentation". In: *Proceedings of the Ninth ACM International Conference on Web Search and Data Mining. WSDM '16*. 103–112. DOI: 10.1145/2835776.2835824.

Wang, Z., Z. Xu, V. Srikumar, and Q. Ai. (2024c). "An In-depth Investigation of User Response Simulation for Conversational Search". In: *Proceedings of the ACM on Web Conference 2024. WWW '24*. 1407–1418. DOI: 10.1145/3589334.3645447.

Weld, H., X. Huang, S. Long, J. Poon, and S. C. Han. (2022). "A Survey of Joint Intent Detection and Slot Filling Models in Natural Language Understanding". *ACM Computing Surveys*. 55(8): 1–38. DOI: 10.1145/3547138.

Wen, T.-H., M. Gašić, N. Mrkšić, P.-H. Su, D. Vandyke, and S. Young. (2015). "Semantically Conditioned LSTM-based Natural Language Generation for Spoken Dialogue Systems". In: *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing. EMNLP '15*. 1711–1721. DOI: 10.18653/v1/D15-1199.

White, R. W. and S. T. Dumais. (2009). "Characterizing and Predicting Search Engine Switching Behavior". In: *Proceedings of the 18th ACM conference on Information and knowledge management. CIKM '09*. 87–96. DOI: 10.1145/1645953.1645967.

White, R. W. (2006). "Using Searcher Simulations to Redesign a Polyrepresentative Implicit Feedback Interface". *Information Processing & Management*. 42(5): 1185–1202. DOI: 10.1016/j.ipm.2006.02.005.

White, R. W. (2016). *Interactions with Search Systems*. Cambridge University Press. DOI: 10.1017/CBO9781139525305.

White, R. W. and J. Huang. (2010). "Assessing the Scenic Route: Measuring the Value of Search Trails in Web Logs". In: *Proceedings of the 33rd International ACM SIGIR Conference on Research and Development in Information Retrieval. SIGIR '10*. 587–594. DOI: 10.1145/1835449.1835548.

White, R. W., J. M. Jose, C. J. van Rijsbergen, and I. Ruthven. (2004). "A Simulated Study of Implicit Feedback Models". In: *Proceedings of the 26th European Conference on IR Research. ECIR '04*. 311–326. DOI: 10.1007/978-3-540-24752-4_23.

White, R. W., I. Ruthven, J. M. Jose, and C. J. V. Rijsbergen. (2005). "Evaluating Implicit Feedback Models Using Searcher Simulations". *ACM Transactions on Information Systems*. 23(3): 325–361. DOI: 10.1145/1080343.1080347.

Williams, J., A. Raux, and M. Henderson. (2016). "The Dialog State Tracking Challenge Series: A Review". *Dialogue Discourse*. 7(3): 4–33.

Wilson, T. D. (1984). "The Cognitive Approach to Information-seeking Behaviour and Information Use". *Social Science Information Studies*. 4(2): 197–204. DOI: 10.1016/0143-6236(84)90076-0.

Wilson, T. D. (1999). "Models in Information Behavior Research". *Journal of Documentation*. 55(3): 249–270. DOI: 10.1108/EUM0000000007145.

Woodruff, A., A. Faulring, R. Rosenholtz, J. Morrsion, and P. Pirolli. (2001). "Using Thumbnails to Search the Web". In: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems. CHI '01*. 198–205. DOI: 10.1145/365024.365098.

Wu, H., Y. Gao, X. Guo, Z. Al-Halah, S. Rennie, K. Grauman, and R. Feris. (2021a). "Fashion IQ: A New Dataset Towards Retrieving Images by Natural Language Feedback". In: *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition. CVPR '21*. 11302–11312. DOI: 10.1109/CVPR46437.2021.01115.

Wu, W.-C., D. Kelly, and A. Sud. (2014). "Using Information Scent and Need for Cognition to Understand Online Search Behavior". In: *Proceedings of the 37th International ACM SIGIR Conference on Research and Development in Information Retrieval. SIGIR '14*. 557–566. DOI: 10.1145/2600428.2609626.

Wu, Y., C. Macdonald, and I. Ounis. (2021b). "Partially Observable Reinforcement Learning for Dialog-based Interactive Recommendation". In: *Proceedings of the 15th ACM Conference on Recommender Systems. RecSys '21*. 241–251. DOI: 10.1145/3460231.3474256.

Wu, Y., C. Macdonald, and I. Ounis. (2022). "Multi-Modal Dialog State Tracking for Interactive Fashion Recommendation". In: *Proceedings of the 16th ACM Conference on Recommender Systems. RecSys '22*. 124–133. DOI: 10.1145/3523227.3546774.

Wu, Y., C. Macdonald, and I. Ounis. (2023). "Goal-Oriented Multi-Modal Interactive Recommendation with Verbal and Non-Verbal Relevance Feedback". In: *Proceedings of the 17th ACM Conference on Recommender Systems. RecSys '23*. 362–373. DOI: 10.1145/3604915.3608775.

Wu, Y., W. Wu, C. Xing, M. Zhou, and Z. Li. (2017). "Sequential Matching Network: A New Architecture for Multi-turn Response Selection in Retrieval-Based Chatbots". In: *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). ACL '17*. 496–505. DOI: 10.18653/v1/P17-1046.

Wu, Z., M. Sanderson, B. B. Cambazoglu, W. B. Croft, and F. Scholer. (2020). "Providing Direct Answers in Search Results: A Study of User Behavior". In: *Proceedings of the 29th ACM International Conference on Information and Knowledge Management. CIKM '20*. 1635–1644. DOI: 10.1145/3340531.3412017.

Xu, Y. C. and Z. Chen. (2006). "Relevance Judgment: What Do Information Users Consider beyond Topicality?" *Journal of the American Society for Information Science and Technology*. 57(7): 961–973. DOI: 10.1002/asi.20361.

Yang, Y. and P. Zhai. (2022). "Click-through Rate Prediction in Online Advertising: A Literature Review". *Information Processing & Management*. 59(2): 102853. DOI: 10.1016/j.ipm.2021.102853.

Yang, Y. and A. Lad. (2009). "Modeling Expected Utility of Multi-Session Information Distillation". In: *Proceedings of the 2nd International Conference on Theory of Information Retrieval: Advances in Information Retrieval Theory. ICTIR '09*. 164–175. DOI: 10.1007/978-3-642-04417-5_15.

Yang, Y., A. Lad, N. Lao, A. Harpale, B. Kisiel, and M. Rogati. (2007). "Utility-Based Information Distillation over Temporally Sequenced Documents". In: *Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. SIGIR '07*. 31–38. DOI: 10.1145/1277741.1277750.

Yao, F., C. Li, D. Nekipelov, H. Wang, and H. Xu. (2023). "How Bad is Top-K Recommendation under Competing Content Creators?" In: *Proceedings of the 40th International Conference on Machine Learning. ICML '23*. 39674–39701.

Yao, S., Y. Halpern, N. Thain, X. Wang, K. Lee, F. Prost, E. H. Chi, J. Chen, and A. Beutel. (2021). "Measuring Recommender System Effects with Simulated Users". arXiv: 2101.04526 [cs.LG].

Yates, J. F. (1990). *Judgment and Decision Making*. Prentice Hall Englewood Cliffs, N.J.

Yilmaz, E., M. Shokouhi, N. Craswell, and S. Robertson. (2010). "Expected Browsing Utility for Web Search Evaluation". In: *Proceedings of the 19th ACM International Conference on Information and Knowledge Management. CIKM '10*. 1561–1564. DOI: 10.1145/1871437.1871672.

Yoon, S.-e., Z. He, J. M. Echterhoff, and J. McAuley. (2024). "Evaluating Large Language Models as Generative User Simulators for Conversational Recommendation". In: *Annual Conference of the North American Chapter of the Association for Computational Linguistics*.

Young, S. (1999). "Probabilistic Methods in Spoken Dialogue Systems". *Philosophical Transactions of the Royal Society (Series A)*. 358(1769): 1389–1402. DOI: 10.1098/rsta.2000.0593.

Young, S., M. Gašić, S. Keizer, F. Mairesse, J. Schatzmann, B. Thomson, and K. Yu. (2010). "The Hidden Information State Model: A Practical Framework for POMDP-based Spoken Dialogue Management". *Computer Speech & Language*. 24(2): 150–174. DOI: 10.1016/j.csl.2009.04.001.

Zach, L. (2005). "When is "Enough" Enough? Modeling the Information-seeking and Stopping Behavior of Senior Arts Administrators". *Journal of the American Society for Information Science and Technology*. 56(1): 23–35. DOI: 10.1002/asi.20092.

Zamani, H. and W. B. Croft. (2018). "Joint Modeling and Optimization of Search and Recommendation". In: *Proceedings of the First Biennial Conference on Design of Experimental Search & Information Retrieval Systems. DESIRES '18*. 36–41.

Zamani, H., S. Dumais, N. Craswell, P. Bennett, and G. Lueck. (2020). "Generating Clarifying Questions for Information Retrieval". In: *Proceedings of The Web Conference 2020. WWW '20*. 418–428. DOI: 10.1145/3366423.3380126.

Zamani, H., J. R. Trippas, J. Dalton, and F. Radlinski. (2023). "Conversational Information Seeking". *Foundations and Trends in Information Retrieval*. 17(3-4): 244–456. DOI: 10.1561/1500000081.

Zerhoudi, S., S. Günther, K. Plassmeier, T. Borst, C. Seifert, M. Hagen, and M. 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. CIKM '22*. 4661–4666. DOI: 10.1145/3511808.3557711.

Zhai, C. X., W. W. Cohen, and J. Lafferty. (2003). "Beyond Independent Relevance: Methods and Evaluation Metrics for Subtopic Retrieval". In: *Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Informaion Retrieval. SIGIR '03*. 10–17. DOI: 10.1145/860435.860440.

Zhai, C. (2016). "Towards a Game-Theoretic Framework for Text Data Retrieval". *IEEE Data Eng. Bull.* 39(3): 51–62.

Zhai, C. and S. Massung. (2016). *Text Data Management and Analysis: A Practical Introduction to Information Retrieval and Text Mining*. Morgan & Claypool.

Zhang, A., L. Sheng, Y. Chen, H. Li, Y. Deng, X. Wang, and T.-S. Chua. (2023). "On Generative Agents in Recommendation". arXiv: 2310.10108 [cs.IR].

Zhang, J., G. Adomavicius, A. Gupta, and W. Ketter. (2020). "Consumption and Performance: Understanding Longitudinal Dynamics of Recommender Systems via an Agent-Based Simulation Framework". *Information Systems Research*. 31(1): 76–101. DOI: 10.1287/isre.2019.0876.

Zhang, J., J. Mao, Y. Liu, R. Zhang, M. Zhang, S. Ma, J. Xu, and Q. Tian. (2019a). "Context-Aware Ranking by Constructing a Virtual Environment for Reinforcement Learning". In: *Proceedings of the 28th ACM International Conference on Information and Knowledge Management. CIKM '19*. 1603–1612. DOI: 10.1145/3357384.3357945.

Zhang, R., T. Yu, Y. Shen, H. Jin, and C. Chen. (2019b). "Text-Based Interactive Recommendation via Constraint-Augmented Reinforcement Learning". In: *Advances in Neural Information Processing Systems. NeurIPS '19*.

Zhang, S. and K. Balog. (2020). "Evaluating Conversational Recommender Systems via User Simulation". In: *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. KDD '20*. 1512–1520. DOI: 10.1145/3394486.3403202.

Zhang, S., M.-C. Wang, and K. Balog. (2022a). "Analyzing and Simulating User Utterance Reformulation in Conversational Recommender Systems". In: *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval. SIGIR '22*. 133–143. DOI: 10.1145/3477495.3531936.

Zhang, W., X. Zhao, L. Zhao, D. Yin, and G. H. Yang. (2021). "DRL4IR: 2nd Workshop on Deep Reinforcement Learning for Information Retrieval". In: *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval. SIGIR '21*. 2681–2684. DOI: 10.1145/3404835.3462818.

Zhang, Y., X. Liu, and C. Zhai. (2017). "Information Retrieval Evaluation as Search Simulation: A General Formal Framework for IR Evaluation". In: *Proceedings of the ACM SIGIR International Conference on Theory of Information Retrieval. ICTIR '17*. 193–200. DOI: 10.1145/3121050.3121070.

Zhang, Y. and C. Zhai. (2015). "Information Retrieval as Card Playing: A Formal Model for Optimizing Interactive Retrieval Interface". In: *Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval. SIGIR '15*. 685–694. DOI: 10.1145/2766462.2767761.

Zhang, Y., J. Zhang, M. Lease, and J. Gwizdka. (2014). "Multidimensional Relevance Modeling via Psychometrics and Crowdsourcing". In: *Proceedings of the 37th International ACM SIGIR Conference on Research and Development in Information Retrieval. SIGIR '14*. 435–444. DOI: 10.1145/2600428.2609577.

Zhang, Y., X. Chen, Q. Ai, L. Yang, and W. B. Croft. (2018). "Towards Conversational Search and Recommendation: System Ask, User Respond". In: *Proceedings of the 27th ACM International Conference on Information and Knowledge Management. CIKM '18*. 177–186. DOI: 10.1145/3269206.3271776.

Zhang, Y., W. Chen, D. Wang, and Q. Yang. (2011). "User-Click Modeling for Understanding and Predicting Search-Behavior". In: *Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. KDD '11*. 1388–1396. DOI: 10.1145/2020408.2020613.

Zhang, Y., E. Chen, B. Jin, H. Wang, M. Hou, W. Huang, and R. Yu. (2022b). "Clustering Based Behavior Sampling with Long Sequential Data for CTR Prediction". In: *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval. SIGIR '22*. 2195–2200. DOI: 10.1145/3477495.3531829.

Zhang, Y., L. A. Park, and A. Moffat. (2010). "Click-Based Evidence for Decaying Weight Distributions in Search Effectiveness Metrics". *Information Retrieval*. 13(1): 46–69. DOI: 10.1007/s10791-009-9099-7.

Zhou, G., X. Zhu, C. Song, Y. Fan, H. Zhu, X. Ma, Y. Yan, J. Jin, H. Li, and K. Gai. (2018). "Deep Interest Network for Click-Through Rate Prediction". In: *Proceedings of the 24th ACM SIGKDD international conference on knowledge discovery and data mining. KDD '18*. 1059–1068. DOI: 10.1145/3219819.3219823.

Zhou, M., J. Zhang, and G. Adomavicius. (2021). "Longitudinal Impact of Preference Biases on Recommender Systems' Performance". *Kelley School of Business Research Paper*. (2021-10). DOI: 10.2139/ssrn.3799525.

Zhu, Q., C. Geishauser, H.-c. Lin, C. van Niekerk, B. Peng, Z. Zhang, S. Feng, M. Heck, N. Lubis, D. Wan, X. Zhu, J. Gao, M. Gasic, and M. Huang. (2023). "ConvLab-3: A Flexible Dialogue System Toolkit Based on a Unified Data Format". In: *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: System Demonstrations. EMNLP '23*. 106–123. DOI: 10.18653/v1/2023.emnlp-demo.9.

Zhu, Q., Z. Zhang, Y. Fang, X. Li, R. Takanobu, J. Li, B. Peng, J. Gao, X. Zhu, and M. Huang. (2020). "ConvLab-2: An Open-Source Toolkit for Building, Evaluating, and Diagnosing Dialogue Systems". In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations. ACL '20*. 142–149. DOI: 10.18653/v1/2020.acl-demos.19.

Zhu, Z. A., W. Chen, T. Minka, C. Zhu, and Z. Chen. (2010). "A Novel Click Model and Its Applications to Online Advertising". In: *Proceedings of the Third ACM International Conference on Web Search and Data Mining. WSDM '10*. 321–330. DOI: 10.1145/1718487.1718528.

Zipf, G. K. (1949). *Human Behaviour and the Principle of Least Effort*. Addison-Wesley.

Zuccon, G. (2016). "Understandability Biased Evaluation for Information Retrieval". In: *Proceedings of the 38th European Conference on IR Research. ECIR '16*. 280–292. DOI: 10.1007/978-3-319-30671-1_21.