# Simulation for Information Retrieval *(Sim4IR)* at SIGIR 2021: Workshop Report

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

David Maxwell
Delft University of Technology
The Netherlands
*d.m.maxwell@tudelft.nl*

Paul Thomas
Microsoft Canberra
Australia
*pathom@microsoft.com*

Shuo Zhang
Bloomberg London
England
*imsure318@gmail.com*

**Abstract**

Simulation is used as a low-cost and repeatable means of experimentation. As *Information Retrieval (IR)* researchers, we are no strangers to the idea of using simulation within our own field—such as the traditional means of IR system evaluation as manifested through the *Cranfield paradigm.* While simulation has been used in other areas of IR research (such as the study of user behaviours), we argue that the potential for using simulation has been recognised by relatively few IR researchers so far.

To this end, the ***Sim4IR*** workshop was held online on July 15th, 2021 in conjunction with ACM SIGIR 2021. Building on past efforts, the goal of the workshop was to create a forum for researchers and practitioners to promote methodology and development of more widespread use of simulation for IR evaluation. Around 80 participants took part over two sessions. A total of two keynotes, three original paper presentations, and eight *'encore talks'* were presented. The main conclusions from the resultant discussion were that simulation has the potential to offer solutions to the limitations of existing evaluation methodologies, but there is more research needed toward developing realistic user simulators; and the development and sharing of simulators, in the form of toolkits and online services, is critical for successful uptake.

## 1 Introduction

*Simulation* is defined as the imitation of the operation of some real-world phenomenon over time [Azzopardi et al., 2011]. Equipped with some underlying *model* [Fishwick, 1995] of the said phenomenon, simulation allows us to conduct carefully designed and controlled experiments, with the aim of providing precise answers to specific research questions [Azzopardi, 2011]. Using

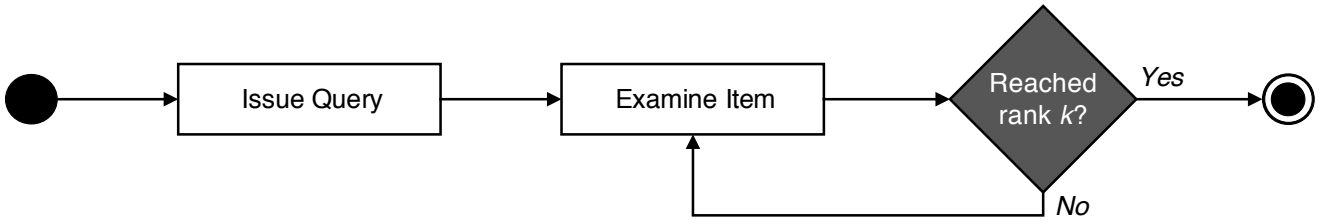


Figure 1: Example of a high-level, simplified searcher model under the *Cranfield paradigm*. Searchers, for a given topic, will issue a single query, and assess *all* documents to some rank, $k$. While advantageous for system-sided evaluation, this abstraction leaves much to be desired for evaulating user behaviours. Figure adapted from Figure 2.7 of Maxwell [2019].

simulation, high levels of experimental control are complemented with a number of other advantages. With this level of control, we can run *what-if* experiments [Kellner et al., 1999], where a variety of different *scenarios* can be explored to determine their effects. These scenarios can be run in such a way to ensure reproducible results, with all this being achieved at a low cost to researchers.

## 1.1 Background

As researchers in the *Information Retrieval (IR)* community, we are all attuned to the idea of using simulation in our research. The technique has a long history within the field, having first being used as early as the 1970s to evaluate early computerised retrieval systems [Cooper, 1973; Tague et al., 1980]. Perhaps the use of simulation within the field is best known through the *Cranfield paradigm* [Cleverdon et al., 1966], the *de facto* approach to IR evaluation. The paradigm spearheaded the notion of standardised test collections (amongst other concepts), and today still forms the basis of many evaluation forums such as the *Text REtrieval Conference (TREC)* [Harman, 1993]. The Cranfield paradigm can be argued to be a form of simulation, whereby underlying models constitute a series of implicit and explicit *assumptions* about the retrieval systems and their users. These assumptions are employed to reduce the complexities in comparing systems against one another. However, as illustrated in Figure 1, these simplifications ultimately lead to highly abstract (and patently unrealistic) models of the search process.

These issues were highlighted at a prior ACM SIGIR workshop. Azzopardi et al. [2011] reported on the *SimInt* workshop, run with the idea of spurring motivation for using simulation as a technique for evaluating *Interactive Information Retrieval (IIR)* systems. Although now a decade old, the subsequent report of the workshop is a highly useful resource for framing the benefits of employing simulation within IR and IIR contexts. Conclusions documented from the SimInt workshop were that *"...simulation offers great potential for the field of IR; and that simulations of user interaction can make explicit the user and the user interface while maintaining the advantages of the Cranfield paradigm"* [Azzopardi et al., 2011]. Since this workshop, we have seen a growing body of literature employing simulation in various aspects of IR and IIR research, which includes (but are not limited to) the following.

- **Evaluation of interactive tasks, such as search sessions** [Baskaya et al., 2012; Carterette et al., 2015; Luo et al., 2014, 2015];

- **Analysing search or browsing behaviours** [Maxwell and Azzopardi, 2016a, 2018; Carterette et al., 2015; Chuklin et al., 2015; Pääkkönen et al., 2015; Smucker, 2011];
- **Query formulation, suggestions, and querying behaviours** [Baskaya et al., 2013; Carterette et al., 2015; Cai and de Rijke, 2016; Verberne et al., 2015];
- **The influence of costs and time** [Azzopardi, 2011; Baskaya et al., 2013; Roy et al., 2021];
- **Filling values in a table** [Zhang and Balog, 2017];
- **Conversational search and recommendation** [Zhang and Balog, 2020; Salle et al., 2021; Lipani et al., 2021]; and
- **Generating synthetic test collections** [Hawking et al., 2020].

Indeed, simulation was prominently discussed in the SWIRL 2012 report [Allan et al., 2012], primarily as a means of studying the interactions between a user and retrieval system. Lacking from the most recent SWIRL 2018 report [Culpepper et al., 2018], we argue that there needs to be a renewed focus from the IR community on the merits of simulation, and what it can be used to achieve. This belief is reinforced with the emergence of research areas where simulation analyses would be highly suitable, such as **conversational information access scenarios**, such as conversational item recommendations [Gao et al., 2019; Huang et al., 2020; Zhang and Balog, 2020; Ie et al., 2019]. To evaluate this particular setup, human-in-the-loop evaluation would be regarded as both very time- and resource-intensive at scale. A further example is the case of test collections. As an example—as highlighted in the list above—work by Hawking et al. [2020] is trailblazing the idea of generating test collections—collections that cannot be shared with researchers due to privacy concerns.

We therefore argue that the time is right for researchers within the IR community to revisit the potential benefits that we can exploit through simulation. We are uniquely suited to drive research and development with this approach, given the rigorous focus on evaluation methodology that dates back to the inception of the field. The goal of the ***Sim4IR*** workshop was to create a forum for researchers and practitioners to present and discuss methods, tools, techniques, and experiences related to the use of simulation as a means to evaluate IR systems (and their users), and to develop a research agenda that drives methodological development—as well as unlocking the potential of simulation techniques.

## 1.2 The Workshop

Organised in conjunction with ACM SIGIR 2021, the ***Sim4IR*** workshop was run on July 15$^{\text{th}}$, 2021. The workshop's call for papers included regular papers, position papers, and demo papers—each of which were reviewed by at least three members of the program committee. Additionally, we invited the submission of a series of so-called *'encore talks'* to present relevant work that had recently been published in a leading IR/IIR conference or journal. In total, the committee accepted three regular papers and eight encore talks to be presented at the workshop. Outlines of the papers and talks can be found in Section 3. In addition to the regular paper presentations, the

***Sim4IR*** programme also included two keynote talks by David Hawking and ChengXiang Zhai. Outlines of the two keynotes are provided in Section 2.

Taking place fully online due to the ongoing COVID-19 pandemic, the ***Sim4IR*** workshop was organised over two sessions during the course of the day. The aim was cater for all time zones and allow as many individuals as possible to attend. To simplify organisation, the workshop was co-ordinated from the *CEST* timezone; we found that running a morning session (for CEST) could cater for Asia and Oceania; a subsequent early evening session catered for those wishing to attend from the Americas. Speakers were polled beforehand to provide their availability, with the schedule of events broadly being determined from these results.

Over both sessions, approximately 80 attendees were present. Each session began with one of the keynote presentations, followed by presentation of regular papers and encore talks. Short five- to ten-minute question and answer periods were provided after each talk to encourage discussion between attendees. Upon the completion of keynotes and paper presentations, each session concluded with a 30-40 minute breakout session. Attendees split up into one of three virtual breakout rooms, with summaries of each provided in Section 4.

The complete list of accepted contributions, presentation slides, and all other outcomes of the ***Sim4IR*** workshop are available online at `https://sim4ir.org`. The workshop proceedings have been published in *CEUR-WS* [Balog et al., 2021]. As we provide outlines of the talks and keynotes delivered in this report, we also provide direct URLs to access the abstracts, slides, and papers (where appropriate) for each entry.[1]

## 1.3 Committee Members

The ***Sim4IR*** was supported by eight researchers who volunteered their time to review the submissions. Committee members are listed below. Our thanks goes out to each of the members for their commitment to the workshop.

- **Leif Azzopardi** (University of Strathclyde, Scotland)
- **Nicola Ferro** (Università degli Studi di Padova, Italy)
- **Christophe van Gysel** (Apple, USA)
- **Claudia Hauff** (Technische Universiteit Delft, The Netherlands)
- **Djoerd Hiemstra** (Radboud Universiteit, The Netherlands)
- **Jaana Kekäläinen** (Tampereen yliopisto, Finland)
- **Heikki Keskustalo** (Tampereen yliopisto, Finland)
- **Mohamed Yahya** (AI Group, Bloomberg)

[1] Listed URLs are correct as of November 30$^{\text{th}}$, 2021.

# 2 Keynotes

***Sim4IR*** invited two keynote speakers. The keynotes headlined the two sessions of the workshop.

## 2.1 How Useful are Results from Simulated Test Collections?

*Abstract and slides available from https://sim4ir.org/speakers/#hawking*

Our first keynote of the ***Sim4IR*** workshop was delivered by **David Hawking**, reporting work with colleagues at Microsoft on simulating test collections [Hawking et al., 2020]. There are many reasons we might want to simulate documents and corpora, as well as searchers and their behaviour. Cloud providers are usually prevented from seeing the corpora they serve, as well as the associated query and interaction logs, but still need to test search algorithms on similar-enough data. We might want to simulate a larger corpus than we have, to test scalability. Or we might want to ship a large "test collection" by shipping random seeds and a generation process, rather than many large files.

Dave introduced SynthaCorpus,[2] an open-source system for emulating a base corpus and a query log, as well as for generating corpus-appropriate known-item test sets. For example, it could be used to emulate a private collection or to scale up an existing collection. He then turned to evaluation: How well do experimental results over a simulated collection predict the performance of a real system, in either efficiency or effectiveness? This evaluation run several open retrieval systems (Indri, Terrier, and ATIRE) over four real TREC datasets (AP, FR, Patents, and WT10g) and then over corpora synthesised by SynthaCorpus, using five methods aiming to match the TREC originals in salient ways. This included investigating the trade-off between privacy and predictive accuracy by including two forms of emulation by encryption, and investigating the influence of "noise" by including `/bin/cp` as an "emulation."

The emulation methods were judged by how closely each retrieval system on an emulated corpus matched that system on the real underlying corpus, both in efficiency (indexing time, indexing memory use, query processing time) and in effectiveness (MRR). For example, an emulation method was judged accurate if a retrieval system used almost the same memory to index an emulated corpus as the underlying real corpus. Dave argued that the findings showed that there are simulation methods which are capable of making predictions accurate enough for practical use while adequately preserving privacy. He also pointed to interactions between emulation method, corpus, and retrieval system.

## 2.2 User Simulation for Information Retrieval Evaluation: Opportunities and Challenges

*Abstract and slides available from https://sim4ir.org/speakers/#zhai*

Our second keynote was delivered by **ChengXiang Zhai**, who talked about simulation from an evaluation perspective. He began by highlighting the differences between two main evaluation goals:

[2] https://bitbucket.org/davidhawking/synthacorpus/

- Measuring *absolute performance* to assess the actual utility of a system, i.e. how useful a system is to a real user who is to perform a real task. This type of evaluation is needed to make decisions whether a system should be deployed in production. Examples include A/B tests [Kohavi et al., 2020] and small-scale user studies [Kelly, 2009]. These experiments, however, are expensive, non-reproducible, and non-reusable.

- Measuring *relative performance* to assess the relative strengths/weaknesses of different systems/methods (i.e. whether a proposed system/method is better than existing ones). The vast majority of published IR research follows the Cranfield paradigm, based on test collections and associated evaluation measures [Sanderson, 2010; Voorhees, 2019]. The limitations of this methodology include the inaccurate representation of users, limited aspects of utility, and the inability to evaluate interactive systems.

Having a relative measure, that is correlated with the absolute difference, is often sufficient, therefore allowing for a weaker requirement. On the other hand, experiments need to be reproducible. In order to make a fair comparison of multiple interactive IR systems using reproducible experiments, we must control the user. User simulation allows us to do exactly that.

Next in his talk, Cheng outlined a general simulation-based evaluation methodology, which consists of a collection of user simulators that are constructed to approximate real users, and a collection of task simulators that are constructed to approximate real tasks. Both user simulators and task simulators can be parameterised to enable modelling of variation in users and tasks. The evaluation of a system is conducted by having a simulated user perform a simulated task by using (interacting with) the system, and then computing various measures based on the entire interaction history.

The rest of the talk featured some of his recent work on simulation. In [Zhang et al., 2017], they propose a general formal framework for evaluating interactive IR systems based on search session simulation. Such a simulation-based evaluation framework is, in fact, a generalisation of the Cranfield paradigm, and existing evaluation measures can be derived as specific instantiations of the framework. Zhang et al. [2017] show that the proposed framework enables the evaluation of sophisticated interfaces using reproducible experiments and produces results that are consistent with real user experiments.

Labhishetty et al. [2020] present a user model for e-commerce search that explicitly models the user's cognitive state (information need and knowledge state) as well as all major user actions (query formulation, query reformulation, and clicks). This is an interpretable model, with parameters that meaningfully correlate with different user behaviours. They show that having an interpretable simulator is a valuable tool in an e-commerce setting to mine and identify interesting user behaviour patterns.

Labhishetty and Zhai [2021] introduce a tester-based approach to evaluate the reliability of user simulators. A Tester is based on a set of IR systems with an expected performance pattern about the order of performance. A Tester is then applied to a user simulator to check whether the simulator would generate the expected performance pattern. They show that this is an effective and feasible approach to evaluate the reliability of user simulators, albeit challenges remain, including the reliability of Testers themselves.

Cheng concluded his talk by identifying some future research directions. These included: *(i)* developing realistic and interpretable user simulators; *(ii)* evaluating user simulators; and *(iii)*

creating a sustainable ecosystem to "publish" user simulators so that they can be improved over time (either as a web service [Hopfgartner et al., 2018; Jagerman et al., 2018; Labhishetty and Zhai, 2021] or as a toolkit [Maxwell and Azzopardi, 2016c]).

# 3 Papers and Presentations

The ***Sim4IR*** workshop committee accepted a total of three regular papers and eight encore talks for presentation. Summaries of each can be found below.

## 3.1 Regular Papers

Regular papers were presented over the course of the two workshop sessions. We present outlines of each talk in the order they were presented; due to scheduling concerns, the first two presentations were delivered in the first session.

**Assessing Query Suggestions for Search Session Simulation**
***Sebastian Günther*** *and Matthias Hagen* [Günther and Hagen, 2021]

Our first regular paper presentation was delivered by Sebastian Günther. Sebastian began his presentation by discussing the querying process, and described a pilot study to assess the applicability of search engine query suggestions to simulate search sessions (i.e., sequences of topically related queries). This is opposed to simulating search behaviours, which has traditionally dealt with result list interactions. In automatic and manual assessments, Günther and Hagen [2021] evaluated to what extent a session detection approach considers the simulated query sequences as "authentic," and how humans perceive the quality of queries in the sense of coherence, realism, and how well the underlying topic is represented. As for actual suggestion-based simulations, Sebastian highlighted the different approaches to selecting the next query in a sequence (from always selecting the first suggestion, randomly sampling, or topic-informed selection) to the human TREC Session Track sessions, and a previously suggested simulation scheme. Results showed that while it is easy to create query logs that are authentic to both users and automated evaluation, keeping the sessions related to an underlying topic can be difficult when relying purely on given suggestions.

**ULTRE Framework: a Framework for Unbiased Learning to Rank Evaluation based on Simulation of User Behavior**
***Yurou Zhao****, Jiaxin Mao, Qingyao Ai* [Zhao et al., 2021]

Our second regular paper was presented by Yurou Zhao. In this paper, Zhao et al. [2021] studied how to evaluate and compare different *Unbiased Learning to Rank (ULTR)* approaches, which have not been systematically investigated and lack a shared task or benchmark. Yurou went on to show the proposed *Unbiased Learning to Rank Evaluation (ULTRE)* framework. The proposed framework utilises multiple click models to generate simulated click logs, and supports the evaluation of both the offline, counterfactual, and online, bandit-based ULTR models. Experimental results showed that the ULTRE framework is indeed effective in simulating click behaviours and

comparing different ULTR models. Yurou concluded the presentation by highlighting that their ULTRE model will be used as a pilot task in the upcoming NTCIR-16 evaluation effort.

**State of the Art of User Simulation Approaches for Conversational IR**
***Pierre Erbacher****, Laure Soulier, Ludovic Denoyer* [Erbacher et al., 2021]

Our final regular paper presentation was given by Pierre Erbacher. Pierre introduced us to *Conversational Information Retrieval (CIR)*, and highlighted that in order to optimise the interactions in such a system and enhance user experiences, they took sequential heterogeneous user-system interactions into account. *Reinforcement Learning (RL)* has emerged as a paradigm particularly suited to optimise sequential decision making in many domains, and have recently been applied to IR problems. However, training such systems by RL on users is not feasible. Pierre showed us that their work presented a potential alternative solution that trains IR systems on user simulations. These simulations model the behaviour of real-world users. Specifically, the work presented here reviewed the literature on user modelling and user simulation for information access, and discussed the different research perspectives for user simulations in the context of CIR.

## 3.2 Encore Talks

Our eight encore talks were split throughout both sessions, and covered a number of recent ways in which simulation techniques have been applied to IR-related problems.

**SIREN: A Simulation Framework for Understanding the Effects of Recommender Systems in Online News Environments**
*Dimitrios Bountouridis, Jaron Harambam, Mykola Makhortykh, Mónica Marrero, Nava Tintarev, and* ***Claudia Hauff*** [Bountouridis et al., 2019]

The first encore talk was presented by Claudia Hauff. While news recommender systems help consumers deal with information overload and increase their engagement, their use also raises an increasing number of societal concerns, such as "Matthew effects," "filter bubbles," and the overall lack of transparency. Claudia argued that focusing on transparency for content providers is an under-explored avenue. As such, the work Claudia presented included a simulation framework called *SIREN (SImulating Recommender Effects in online News environments)*. It allows content providers to: *(i)* select and parameterise different recommenders; and *(ii)* analyse and visualise their effects with respect to two diversity metrics. Taking the U.S. news media as a case study, this work presented an analysis of the recommender effects with respect to long-tail novelty and unexpectedness using SIREN. The analysis offers a number of interesting findings, such as the similar potential of certain algorithmically simple (item-based $k$-nearest neighbour) and sophisticated strategies (based on Bayesian personalised ranking) to increase diversity over time. Overall, Claudia argued that simulating the effects of recommender systems can help content providers to make more informed decisions when choosing algorithmic recommenders, and as such can help mitigate the aforementioned societal concerns.

### Context-Aware Ranking by Constructing a Virtual Environment for Reinforcement Learning

***Junqi Zhang**, Jiaxin Mao, Yiqun Liu, Ruizhe Zhang, Min Zhang, Shaoping Ma, Jun Xu, Qi Tian* [Zhang et al., 2019]

The second encore talk was presented by Junqi Zhang. In this work, Junqi explained that they had focused on result rankings for Web search technologies. They proposed a better ranking strategy should be a context-aware process, and optimise result ranking globally. Specifically, the proposed framework aims to improve context-aware listwise ranking performance by optimising online evaluation metrics. The ranking problem is formalised as a *Markov Decision Process (MDP)*, and solved with the reinforcement learning paradigm. To avoid the great cost to online systems during the training of the ranking model, a virtual environment was constructed with millions of historical click logs to simulate the behaviour of real-world users. Extensive experiments on both simulated and real datasets show that: *(i)* constructing a virtual environment can effectively leverage the large-scale click logs and capture some important properties of real users; and *(ii)* the proposed framework can improve search ranking performance by a large margin.

### Towards User-Oriented Privacy for Recommender System Data: A Personalization-based Approach to Gender Obfuscation for User Profiles

***Manel Slokom*** [Slokom, 2018]

Our third encore talk was presented by Manel Slokom. In her presentation, Manel proposed a new privacy solution for the data used to train a recommender system, i.e., *the user-item matrix.* This solution, called *Personalised Blurring (PerBlur)*, is a simple yet effective approach to adding and removing items from user profiles to generate an obfuscated user-item matrix. PerBlur is formulated within a user-oriented paradigm of recommender system data privacy that aims at making privacy solution understandable, unobtrusive, and useful for the user. When obfuscated data is used for training, a recommender system can reach performance comparable to what is attained when it is trained on the original, unobfuscated data. At the same time, a classifier can no longer reliably use the obfuscated data to predict the gender of users, indicating that implicit gender information has been removed. In addition to introducing PerBlur, this work make several key contributions. First, it proposed an evaluation protocol that creates a fair environment to compare between different obfuscation conditions. Second, it carried out experiments that show that gender obfuscation impacts the fairness and diversity of recommender system results. The experiments showed that PerBlur maintains fairness by not causing a gender-specific drop in recommender system performance. It also demonstrated the ability of PerBlur, through its greedy removal, to recommend a smaller proportion of gender-stereotypical items, i.e., items that are highly specific to a particular gender.

### Evaluating Conversational Recommender Systems via User Simulation

***Shuo Zhang** and Krisztian Balog* [Zhang and Balog, 2020]

The fourth encore talk was presented by Shuo Zhang. Shuo argued that human evaluation is used for end-to-end system evaluation, which is both very time and resource intensive at scale, and thus becomes a bottleneck of progress. As an alternative, this work proposed automated evaluation

employing simulating users. The proposed user simulator aims to generate responses that a real human would give by considering both individual preferences and the general flow of interaction with the system. Shuo evaluated the simulation approach on an item recommendation task by comparing three existing conversational recommender systems. They showed that preference modelling and task-specific interaction models both contribute to more realistic simulations, and can help achieve high correlation between automatic evaluation measures and manual human assessments.

**How Am I Doing?: Evaluating Conversational Search Systems Offline**
*Aldo Lipani,* ***Ben Carterette****, and Emine Yilmaz* [Lipani et al., 2021]

The fifth encore talk was presented by Ben Carterette. Ben argued that conversational search shares some features with traditional search, but differs in some important respects: conversational search systems are less likely to return ranked lists of results (a *Search Engine Results Page (SERP)*), more likely to involve iterated interactions, and more likely to feature longer, well-formed user queries in the form of natural language questions. Because of these differences, traditional methods for search evaluation (such as the Cranfield paradigm) do not translate easily to conversational search. To fill these gaps, Ben highlighted the proposed framework for offline evaluation of conversational search, which includes a methodology for creating test collections with relevance judgements, an evaluation measure based on a user interaction model, and an approach to collecting user interaction data to train the model. The framework is based on the idea of "subtopics," often used to model novelty and diversity in search and recommendation, and the user model is similar to the geometric browsing model introduced by *Rank Biased Precision (RBP)* and used in ERR.

**Studying the Effectiveness of Conversational Search Refinement Through User Simulation**
***Alexandre Salle****, Shervin Malmasi, Oleg Rokhlenko, and Eugene Agichtein* [Salle et al., 2021]

Alexandre Salle presented our sixth encore talk. This work focused on refining a user's search intent by asking a series of clarification questions, aiming to improve the relevance of search results. To support robust training/evaluation of such systems, Alexandre discussed how a simulation framework called *CoSearcher* was proposed. This framework includes a parameterised user simulator controlling key behavioural factors like cooperativeness and patience. Based on experiments with a range of user behaviours, semantic policies, and dynamic facet generation, Alexandre reported the results that quantify the effects of user behaviours and identify critical conditions required for conversational search refinement to be effective.

**Keeping Dataset Biases out of the Simulation: A Debiased Simulator for Reinforcement Learning based Recommender Systems**
***Jin Huang****, Harrie Oosterhuis, Maarten de Rijke, and Herke van Hoof* [Huang et al., 2020]

The seventh encore talk was presented by Jin Huang. Jin argued that previous simulation works ignored the interaction biases present in logged user data, and consequently, these biases affect the resulting simulation. To address this issue, they introduced a debiasing step in the simulation

pipeline, which corrects the biases present in the logged data before it is used to simulate user behaviour. To evaluate the effects of bias on RL4Rec simulations, they proposed a novel evaluation approach for simulators that considers the performance of policies optimised with the simulator. The results revealed that the biases from logged data negatively impact the resulting policies unless corrected with our debiasing method.

**Modelling Search and Stopping in Interactive Information Retrieval**
***David Maxwell*** [Maxwell, 2019]

The eight and final encore talk was presented by David Maxwell. David presented a high-level overview of his PhD thesis, which examined stopping behaviours. While stopping behaviours have been examined in the past by researchers, individuals usually would report that they stop examining content when what they have found feels "good enough." This isn't good enough for us as researchers—and David's work considered operationalising a number of stopping heuristics as proposed in the literature, and using the simulation of interaction to examine how each of the different implemented stopping strategies performed: in terms of overall performance (what-if) and comparisons to real-world searcher behaviours on average. From an ad-hoc search context, simple strategies appeared to work best (such as the *frustration strategy*, considering one's tolerance to non-relevant content). The work motivates further research on stopping behaviours, potentially leading to improved interfaces for future searchers.

# 4 Breakout Discussions

Upon completion of all the keynotes and presentations, ***Sim4IR*** attendees then split up into three breakout groups. Each of the breakout groups were assigned a specific theme for considering simulation in the context of IR. Themes included consideration for the **requirements for simulation**; considering the necessity of **humans-in-the-loop**; and **evaluation using simulation**. Below, we provide a brief summary of the discussions that entailed from attendees for each theme.

**Requirements for Simulation** It remains an open question as to how realistic (i.e. human-like) simulators *can be*, or indeed *should be*. It is important to note that simulators do not need to be perfect mirrors of human behaviour, but instead simply need to be *"good enough."* By this, we mean that output from simulations should correlate well with human assessments on a given task with respect to some evaluation metric. The main requirement is reproducibility. If the simulator is non-deterministic, the random seed numbers may need to be provided along with the reported experimental results to account for stochastic behaviours. This has already been considered in works such as those by Maxwell and Azzopardi [2018], for example: a seed value was used to instantiate stochastic components of the simulation framework used, ensuring reproducible results.

**Humans-in-the-Loop** One of the main motivations behind using simulation is that employing human judges in the evaluation process is expensive, time-consuming, and does not scale. Can humans then be removed from the loop altogether? Thoughts from attendees was that the answer is *no*. Simulators need to be validated periodically to show that they are mimicking realistic user

behaviour. This can only be done by comparing them against humans. Validation in practice could mean, for example, evaluating systems using a small number of users and showing that the evaluation measures obtained with simulated users are within a certain error tolerance from human assessments.

**Evaluation using Simulation** Offline (test collection-based) and online evaluation have been used in IR for a long time [Sanderson, 2010; Hofmann et al., 2016]. Such methodologies are well understood. For example, it is clear what procedures need to be followed and what measures need to be reported in order to have a scientifically sound outcome evaluation—such as selecting test queries, pooling, handling inter-annotator disagreement, performing significance testing, etc. However, this is not the case for simulation. It appears that there may be some scepticism among members of the IR community regarding the validity of simulation-based evaluation, while others may simply be unsure or in disagreement about specific elements/components of the methodology. Either way, there would be a great need for community benchmarking efforts. For example, simulation-based evaluation could be used at TREC or CLEF. This idea may involve the creation of a separate track dedicated to simulation (for example, similar to the Crowdsourcing track at TREC [Smucker et al., 2012]). Alternatively, specific tracks may consider including simulation-based evaluation. This would have the advantage of allowing a direct comparison between simulated users and human assessors. Community-wide benchmarks could also facilitate the shared development of simulators: for instance, each participant could contribute their own user simulators to a pool. Then, each participating system would be evaluated against a sample of simulated users drawn from this shared pool. As a practical first step, we could start looking into what would be needed in order to turn existing test collections into simulators. The recent work by Lipani et al. [2021] represents an effort in this direction. Consideration will also need to be given towards tooling. Simulation software can be complex. With various research groups introducing their own simulators for various IR tasks (such as *SimIIR* [Maxwell and Azzopardi, 2016b] for the simulation of interaction), can we co-ordinate development to produce a standardised, accepted framework/tool for all members of the community to use for given tasks?

# 5 Summary

The ***Sim4IR*** workshop was very successful in bringing researchers together that are interested in the use and development of simulation techniques for evaluation. The invited keynotes, papers accepted at the workshop, as well as previously published papers presented at encore talks provided a high-quality mixture of topics on simulation. There was broad agreement between participants that simulation has potential, and it should be added to the toolbox of IR researchers—not to replace existing evaluation methodologies, but to complement them. Clearly, there is more research needed toward developing realistic user simulators, understanding the limitations of simulation techniques, and arriving at best experimental practices. The development and sharing of simulators, in the form of toolkits and online services, is critical for successful uptake.

It is good to be reminded of the fact that many other communities build on the evaluation methodology that originates from IR. Simulation is one of those areas where the IR community has an opportunity to lead the way.

## References


James Allan, Bruce Croft, Alistair Moffat, and Mark Sanderson. Frontiers, challenges, and opportunities for Information Retrieval: Report from SWIRL 2012, the second strategic workshop on Information Retrieval in Lorne. In *ACM SIGIR Forum*, volume 46, pages 2–32, 2012.

Leif Azzopardi. The economics in interactive information retrieval. In *Proceedings of the 34th ACM SIGIR*, pages 15–24, 2011.

Leif Azzopardi, Kalervo Järvelin, Jaap Kamps, and Mark D Smucker. Report on the SIGIR 2010 workshop on the simulation of interaction. In *ACM SIGIR Forum*, volume 44, pages 35–47, 2011.

Krisztian Balog, Xianjie Chen, Xu Chen David Maxwell, Paul Thomas, Shuo Zhang, Yi Zhang, and Yongfeng Zhang, editors. *CSR-Sim4IR 2021: Joint Proceedings of the Causality in Search and Recommendation (CSR) and Simulation of Information Retrieval Evaluation (Sim4IR) Workshops*, volume 2911 of *CEUR Workshop Proceedings*, 2021. CEUR-WS.org. URL http://ceur-ws.org/Vol-2911.

Feza Baskaya, Heikki Keskustalo, and Kalervo Järvelin. Time drives interaction: Simulating sessions in diverse searching environments. In *Proceedings of the 35th ACM SIGIR*, page 105–114, 2012.

Feza Baskaya, Heikki Keskustalo, and Kalervo Järvelin. Modeling behavioral factors in interactive information retrieval. In *Proceedings of the 22nd ACM CIKM*, pages 2297–2302, 2013.

Dimitrios Bountouridis, Jaron Harambam, Mykola Makhortykh, Mónica Marrero, Nava Tintarev, and Claudia Hauff. Siren: A simulation framework for understanding the effects of recommender systems in online news environments. In *Proceedings of FAT**, page 150–159, 2019. URL https://doi.org/10.1145/3287560.3287583.

Fei Cai and Maarten de Rijke. A survey of query auto completion in information retrieval. *Foundations and Trends in Information Retrieval*, 10(4):273–363, 2016.

Ben Carterette, Ashraf Bah, and Mustafa Zengin. Dynamic test collections for retrieval evaluation. In *Proceedings of the 1st ACM ICTIR*, pages 91–100, 2015.

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

Cyril W. Cleverdon, Jack Mills, and Michael Keen. *Factors Determining the Performance of Indexing Systems*, volume 1:2 of *Cranfield Research Projects*. 1966.

Michael D Cooper. A simulation model of an information retrieval system. *Information Storage and Retrieval*, 9(1):13–32, 1973.

J Shane Culpepper, Fernando Diaz, and Mark D Smucker. Report from the Third Strategic Workshop on Information Retrieval (SWIRL). 52(1):34–90, 2018.

Pierre Erbacher, Laure Soulier, and Ludovic Denoyer. State of the art of user simulation approaches for conversational information retrieval. In Balog et al. [2021], pages 32–37. URL http://ceur-ws.org/Vol-2911/paper5.pdf.

Paul A. Fishwick. Computer simulation: The art and science of digital world construction. Technical report, University of Florida, 1995.

Jianfeng Gao, Michel Galley, and Lihong Li. Neural Approaches to Conversational AI. *Foundations and Trends in Information Retrieval*, 13(2-3):127–298, 2019.

Sebastian Günther and Matthias Hagen. Assessing query suggestions for search session simulation. In Balog et al. [2021], pages 38–45. URL http://ceur-ws.org/Vol-2911/paper6.pdf.

Donna Harman. Overview of the first trec conference. In *Proceedings of the 16$^{th}$ ACM SIGIR*, pages 36–47, 1993.

David Hawking, Bodo Billerbeck, Paul Thomas, and Nick Craswell. *Simulating Information Retrieval Test Collections*. Morgan and Claypool, 2020.

Katja Hofmann, Lihong Li, and Filip Radlinski. Online evaluation for information retrieval. *Foundations and Trends in Information Retrieval*, 10(1):1–117, 2016.

Frank Hopfgartner, Allan Hanbury, Henning Müller, Ivan Eggel, Krisztian Balog, Torben Brodt, Gordon V. Cormack, Jimmy Lin, Jayashree Kalpathy-Cramer, Noriko Kando, Makoto P. Kato, Anastasia Krithara, Tim Gollub, Martin Potthast, Evelyne Viegas, and Simon Mercer. Evaluation-as-a-service for the computational sciences: Overview and outlook. *Journal of Data and Information Quality*, 10(4):15:1–15:32, October 2018.

Jin Huang, Harrie Oosterhuis, Maarten de Rijke, and Herke van Hoof. Keeping dataset biases out of the simulation: A debiased simulator for reinforcement learning based recommender systems. In *Proceedings of the 14$^{th}$ ACM RecSys*, page 190–199, 2020. URL https://doi.org/10.1145/3383313.3412252.

Eugene Ie, Chih-Wei Hsu, Martin Mladenov, Vihan Jain, Sanmit Narvekar, Jing Wang, Rui Wu, and Craig Boutilier. Recsim: A configurable simulation platform for recommender systems. *Computing Research Repository*, abs/1909.04847, 2019.

Rolf Jagerman, Krisztian Balog, and Maarten De Rijke. Opensearch: Lessons learned from an online evaluation campaign. *Journal of Data and Information Quality*, 10(3):13:1–13:15, September 2018.

Marc I Kellner, Raymond J Madachy, and David M Raffo. Software process simulation modeling: Why? What? How? *Journal of Systems and Software*, 46(2-3):91–105, 1999.

Diane Kelly. Methods for evaluating interactive information retrieval systems with users. *Foundations and Trends in Information Retrieval*, 3:1–224, 1 2009.

R. Kohavi, D. Tang, and Y. Xu. *Trustworthy Online Controlled Experiments: A Practical Guide to A/B Testing*. Cambridge University Press, 2020.

Sahiti Labhishetty and Chengxiang Zhai. An exploration of tester-based evaluation of user simulators for comparing interactive retrieval systems. In *Proceedings of the 44th ACM SIGIR*, pages 1598–1602, 2021.

Sahiti Labhishetty, Chengxiang Zhai, Suhas Ranganath, and Pradeep Ranganathan. A cognitive user model for e-commerce search. In *Proceedings of the Data Science for Retail and E-Commerce Workshop*, 2020.

Aldo Lipani, Ben Carterette, and Emine Yilmaz. How am I Doing?: Evaluating conversational search systems offline. *ACM Transactions on Intelligent Systems and Technology*, 2021. URL https://doi.org/10.1145/3451160.

Jiyun Luo, Sicong Zhang, and Hui Yang. Win-win search: Dual-agent stochastic game in session search. In *Proceedings of the 37th ACM SIGIR*, pages 587–596, 2014.

Jiyun Luo, Sicong Zhang, Xuchu Dong, and Hui Yang. Designing states, actions, and rewards for using pomdp in session search. In *Advances in Information Retrieval*, pages 526–537, 2015.

David Maxwell. *Modelling Search and Stopping in Interactive Information Retrieval*. PhD thesis, University of Glasgow, Scotland, April 2019. URL https://theses.gla.ac.uk/41132/.

David Maxwell and Leif Azzopardi. Agents, simulated users and humans: An analysis of performance and behaviour. In *Proceedings of the 25th ACM CIKM*, pages 731–740, 2016a.

David Maxwell and Leif Azzopardi. Simulating interactive information retrieval: SimIIR: A framework for the simulation of interaction. In *Proceedings of the 39th ACM SIGIR*, pages 1141–1144, 2016b.

David Maxwell and Leif Azzopardi. Simulating interactive information retrieval: SimIIR: A framework for the simulation of interaction. In *Proceedings of the 39th ACM SIGIR*, pages 1141–1144, 2016c.

David Maxwell and Leif Azzopardi. Information scent, searching and stopping: Modelling SERP level stopping behaviour. In *Advances in Information Retrieval*, pages 210–222, 2018.

Teemu Pääkkönen, Kalervo Järvelin, Jaana Kekäläinen, Heikki Keskustalo, Feza Baskaya, David Maxwell, and Leif Azzopardi. Exploring behavioral dimensions in session effectiveness. In *Proceedings of the 6th CLEF*, pages 178–189, 2015.

Nirmal Roy, Arthur Câmara, David Maxwell, and Claudia Hauff. Incorporating widget positioning in interaction models of search behaviour. In *Proceedings of the 7th ACM ICTIR*, pages 53–62, 2021.

Alexandre Salle, Shervin Malmasi, Oleg Rokhlenko, and Eugene Agichtein. Studying the effectiveness of conversational search refinement through user simulation. In *Advances in Information Retrieval*, pages 587–602, 2021. URL https://doi.org/10.1007/978-3-030-72113-8_39.

Mark Sanderson. Test collection based evaluation of information retrieval systems. *Foundations and Trends in Information Retrieval*, 4(4):247–375, 2010.

Manel Slokom. Comparing recommender systems using synthetic data. In *Proceedings of the 12th ACM RecSys*, page 548–552, 2018. URL https://doi.org/10.1145/3240323.3240325.

Mark D. Smucker. An analysis of user strategies for examining and processing ranked lists of documents. In *Proceedings of the 5th HCIR*, 2011.

Mark D. Smucker, Gabriella Kazai, and Matthew Lease. Overview of the TREC 2012 crowdsourcing track. In *Proceedings of the 21st TREC*, 2012.

Jean Tague, Michael Nelson, and Harry Wu. Problems in the simulation of bibliographic retrieval systems. In *Proceedings of the 3rd ACM SIGIR*, pages 236–255, 1980.

Suzan Verberne, Maya Sappelli, Kalervo Järvelin, and Wessel Kraaij. User simulations for interactive search: Evaluating personalized query suggestion. In *Advances in Information Retrieval*, volume 9022, pages 678–690. 2015.

Ellen M. Voorhees. The evolution of Cranfield. In Nicola Ferro and Carol Peters, editors, *Information Retrieval Evaluation in a Changing World - Lessons Learned from 20 Years of CLEF*, volume 41 of *The Information Retrieval Series*, pages 45–69. Springer, 2019.

Junqi Zhang, Jiaxin Mao, Yiqun Liu, Ruizhe Zhang, Min Zhang, Shaoping Ma, Jun Xu, and Qi Tian. Context-aware ranking by constructing a virtual environment for reinforcement learning. In *Proceedings of the 28th ACM CIKM*, page 1603–1612, 2019. URL https://doi.org/10.1145/3357384.3357945.

Shuo Zhang and Krisztian Balog. EntiTables: Smart assistance for entity-focused tables. In *Proceedings of the 40th ACM SIGIR*, pages 255–264, 2017.

Shuo Zhang and Krisztian Balog. Evaluating conversational recommender systems via user simulation. In *Proceedings of the 26th ACM SIGKDD*, KDD '20, pages 1512–1520, 2020. URL https://doi.org/10.1145/3394486.3403202.

Yinan Zhang, Xueqing Liu, and ChengXiang Zhai. Information retrieval evaluation as search simulation: A general formal framework for ir evaluation. In *Proceedings of ICTIR'17*, pages 193–200, 2017.

Yurou Zhao, Jiaxin Mao, and Qingyao Ai. ULTRE framework: a framework for unbiased learning to rank evaluation based on simulation of user behavior. In Balog et al. [2021], pages 46–51. URL http://ceur-ws.org/Vol-2911/paper7.pdf.