The SimIIR 2.0 Framework: User Types, Markov Model-Based Interaction Simulation, and Advanced Query Generation

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

Sebastian Günther*
Martin-Luther-Universität
Halle-Wittenberg
Germany
sebastian.guenther@informatik.uni-halle.de

Kim Plassmeier
Timo Borst
ZBW Leibniz Information Centre for Economics
Germany
k.plassmeier@zbw.eu,t.borst@zbw.eu

Christin Seifert
University of Duisburg-Essen
Germany
christin.seifert@uni-due.de

Matthias Hagen
Martin-Luther-Universität
Halle-Wittenberg
Germany
matthias.hagen@informatik.uni-halle.de

Michael Granitzer
University of Passau
Germany
michael.granitzer@uni-passau.de

ABSTRACT

Simulated user–retrieval system interactions enable studies with controlled user behavior. To this end, the SimIIR framework offers static, rule-based methods. We present an extended SimIIR 2.0 version with new components for dynamic user type-specific Markov model-based interactions and more realistic query generation. A flexible modularization ensures that the SimIIR 2.0 framework can serve as a platform to implement, combine, and run the growing number of proposed search behavior and query simulation ideas.

CCS CONCEPTS

• **Information systems → Users and interactive retrieval**; • **Computing methodologies → Modeling and simulation**.

KEYWORDS

Simulation, Search Behavior, User Modeling, Software Framework

ACM Reference Format:
Saber Zerhoudi, Sebastian Günther, Kim Plassmeier, Timo Borst, Christin Seifert, Matthias Hagen, and Michael Granitzer. 2022. The SimIIR 2.0 Framework: User Types, Markov Model-Based Interaction Simulation, and Advanced Query Generation. In *Proceedings of the 31st ACM International Conference on Information and Knowledge Management (CIKM '22), October 17–21, 2022, Atlanta, GA, USA.* ACM, New York, NY, USA, 6 pages. https://doi.org/10.1145/3511808.3557711

1 INTRODUCTION

Behavior analyses are key to understand how searchers interact with a retrieval system and to assess whether changes to the interface or the retrieval model help to improve the user experience. Still, traditional academic retrieval evaluation often follows the Cranfield paradigm [9] with static test collections (documents, queries, relevance judgments) to ensure a controlled and reusable setup. But the Cranfield paradigm does not really cover dynamic interactions (e.g., query reformulation [7]) or the evaluation of user interface variants (e.g., with or without facets and filters). Such more realistic evaluations today are mostly conducted via large-scale A/B tests [28] or via smaller controlled user studies [11, 24]. However, user studies are costly and hard to reproduce, while A/B testing requires a large enough user base to draw meaningful conclusions.

For evaluation scenarios with a smaller number of users (e.g., in digital libraries), simulation offers an alternative beyond the Cranfield paradigm or controlled user studies. In a way, simulation also allows to "A/B-test" different back end configurations or interface variants by monitoring interactions of parameterized user types. Clearly, results from such artificial A/B-tests strongly depend on the realism and representativeness of the simulated behavior.

The open-source SimIIR framework [31] supports repeatable simulated retrieval experiments with static interaction modules for user behavior within the Complex Searcher Model. In this paper, we present an extended SimIIR 2.0 framework with dynamic and user type-specific simulation components. We include improved query formulation approaches and Markov modeling for global and search-type specific behavior. From the simulated interactions, various metrics can be computed that indicate how well a system assists users in completing their tasks. When comparing simulations from the existing framework to our extended version, one can observe that the new dynamic components support more diverse developments of users' information needs during sessions.

Our extended SimIIR 2.0 framework is updated to the latest Python version, comes with additional dynamic user and query modules, and—like the original SimIIR framework—is available as an open-source resource with a permissive license that allows others to easily contribute further components, modules, or adjustments.[1] SimIIR 2.0 can thus serve as a modern platform to implement, configure, combine, run, and compare the growing number of user models and query simulation ideas from the literature.

*Both authors contributed equally to the paper.

[1] GitHub: https://github.com/padre-lab-eu/simiir-2

CIKM '22, October 17–21, 2022, Atlanta, GA, USA

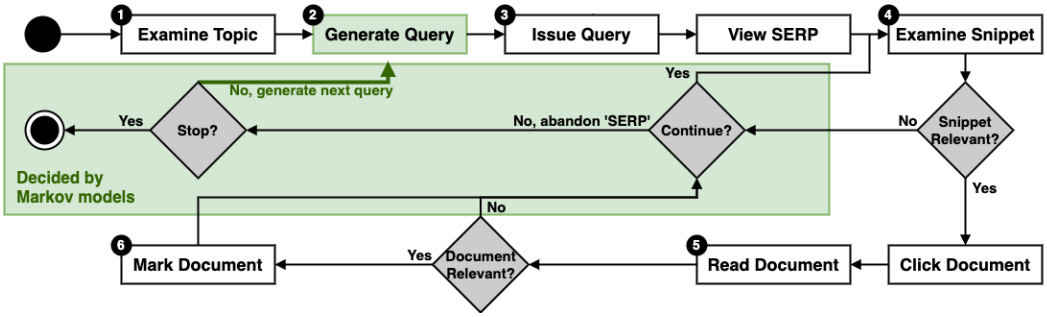

Figure 1: Flowchart of the Complex Searcher Model (CSM) that is the basis for interaction simulation in the SimIIR framework [31]. The key components are shown as boxes with numbered steps, while a simulated user's decision points are indicated in gray (SERP: search engine result page). Components that we improve in our extension are shown in green.

2 RELATED WORK

Evaluation has always been a core IR topic; Harman [19], Kelly [24], and Sanderson [36] nicely cover the history. Today, the Cranfield paradigm from the 1960s [9] is still often used even though this usually means to evaluate systems with static queries and no user interactions. Some studies thus employed simulation but mostly for single aspects like click behavior [8], query (re-)formulation [2, 6, 23, 42], relevance feedback [20, 25], or stopping [32, 41].

Later, the importance of simulating the search process as a whole has been emphasized [31, 48] even though Cole [10] had collected several challenges when developing realistic simulations against "real" retrieval systems [10]. Cole's challenges are based on Boring's five step operationalist approach [3] and essentially state that simulations need to be aligned with real user behavior.

The SimIIR framework [31] provides tools to simulate user–system interactions as a whole (queries, clicks, stopping, etc.) for different configurations of simulation components, experimental conditions, and retrieval systems. However, the simulation components originally implemented in SimIIR produce rather static behavior sequences. We thus extend the SimIIR framework by including more dynamic components and allowing the simulation modules to influence each other (e.g., to take the interaction history into account for the next simulation steps) .

To align simulations with real user behavior, SimIIR 2.0 contains components to train Markov models on real log data to simulate global or user type-specific behavior. Markov models are based on a well-established theory, are rather simple and compact, and have been used for simulations in various areas: first-order or higher-order Markov models [37], partially observable Markov decision processes (POMDPs) [40], and hidden Markov models (HMMs) [12]. For now, we rely on first-order Markov models in SimIIR 2.0 due to their simplicity but other variants can later be included.

3 THE EXISTING SIMIIR FRAMEWORK

SimIIR [31] is a Python-based framework for simulating search sessions following the Complex Searcher Model (CSM). The CSM has components for the decision points and activities in search sessions (cf. Figure 1; from formulating a query on a topic over examining some documents to stopping the search).

To run a SimIIR simulation, the following four main elements must be configured. (1) *Topics* represent the simulated users' information needs and consist of a title and a description. In SimIIR, the standard topics come from TREC tracks (e.g., the TREC 2005 Robust track). (2) A *retrieval system* that returns a ranked list of documents with snippets for a query. In SimIIR, Whoosh is used as the standard retrieval system.[2] (3) An *output controller* that generates output files for a simulation run compatible with evaluation programs like trec_eval.[3] Finally, a simulation requires (4) a *series of simulated users*, each possibly with differently configured but still rather static characteristics for the decision points and activities in the CSM. During the simulation, the users attempt to complete a session on a given topic while interacting with the retrieval system.

4 SIMIIR 2.0 EXTENSION

After describing our conceptual CSM extensions, we give details on the newly added query generation and Markov model components.

4.1 Extended Complex Searcher Model

We add two novel elements to the CSM to improve the realism of the simulated sessions: advanced query generation and user type-specific Markov model-based stopping (green blocks in Figure 1).

Query generation. In the original CSM [31], a pool of queries is generated once at the start of a simulated session. From this static pool, a query is selected whenever the simulated user decides to submit a new query. However, in real sessions, the seen results will often influence subsequent queries (e.g., a user may acquire new vocabulary from a read document) [18, 21, 38, 43]. In the extended CSM, we thus enable the query generation to access the session history and to dynamically generate new query candidates based on this information. When a 'dynamic' simulated user wants to submit a new query, it is selected from an updated pool of candidates.

User types. The original CSM does not include a possibility to group different simulated users as kind of a user type with possibly specific search behavior. For instance, 'exploratory searchers' will explore a search result list more exhaustively than 'lookup

[2] https://pypi.python.org/pypi/Whoosh
[3] http://trec.nist.gov/trec_eval

searchers' who will only investigate the first few results and then rephrase their queries rather quickly [1, 45]. We thus include user types in the extended CSM in the sense that the components of the CSM can be initialized with user type-specific characteristics to support the simulation of user type-specific sessions.

Markov models. In our extended CSM, the stopping decisions on the SERP and session level are made by user type-specific Markov models instead of the original stochastic heuristics with stopping threshold variables. To this end, we categorize users into different types and simulate their search process using specific Markov models. At the stopping decisions, these models also predict a user's next likely step by taking the session history into account.

Besides stopping, we also employ Markov model-based decisions for query generation to let later queries in a session depend on the content of previously viewed search results. User type-specific Markov models predict the next likely query 'change' direction (e.g., generalization or specialization) based on how a particular user group reformulates their queries and the predicted direction is then used to select an appropriate query from the (updated) pool.

4.2 Realization and Implementation

Query generation. We add several types of query generation approaches. The first type uses an actual search engine's API to obtain query suggestions—a technique earlier demonstrated to yield realistic sessions [16]. Our second type extends the original SimIIR query generation approaches—that determine query terms from the static topic information—by additionally giving them access to the session history in form of examined snippets and documents as a resource for new query terms. The third type of approaches implements Markov model-based query change prediction (e.g., did a generalization or specialization happen before) as this was demonstrated to reliably simulate specific querying behavior [44]. The model then guides the selection of a next query from the (possibly dynamic) query pool based on a user's previous changes.

User types. In order to simulate user type-specific behavior, we categorize users into different groups based on their search and stopping behavior. So far, we use the previously introduced contextual search types [1, 46] of 'exploratory searchers' who tend to fully explore a search result list and extensively use potentially available result filters, and 'lookup searchers' who only investigate the first few results and quickly rephrase their queries. But also other user types like 'fast and liberal' vs. 'slow and picky' users [27, 39] or 'build up' vs. 'boil down' behavior [35] could be the basis.

Markov models. For using a Markov model in a simulation, one simply indicates files with the respective states and transition probabilities. SimIIR 2.0 offers the possibility to derive the probabilities from some existing search logs but they can also be set manually.

Extended configuration. Listing 1 shows an example of how the configuration attributes for SimIIR 2.0 have been extended. The file extends the SimIIR `trec_user`[4] with a new section `behaviorModel` for the user type-specific behavior. The `user_type` attribute indicates to use the `exploratory` Markov model with the states and transition probabilities given in the attributes `states` and `transition_matrix`.

[4]https://github.com/leifos/simiir/blob/master/example_sims/users/trec_user.xml

Listing 1: Configuration file `markov_google_trec_user.xml` with the new options `behaviorModeler` and `GoogleSuggestGenerator`.

```
<userConfiguration id="markovgoogletrecuser">
  <behaviorModeler class="Markov">
    <attribute name="user_type" value="exploratory"/>
    <attribute name="states" value="<..>/states.data"/>
    <attribute name="transition_matrix"
               value="<..>/matrix.data"/>
  </behaviorModeler>
  <queryGenerator class="GoogleSuggestGenerator">
    <attribute name="stopword_file" value="<..>/stopwords.txt"/>
    <attribute name="max_depth" type="integer" value="5" />
  </queryGenerator>
  <relevanceAssessor>
    <snippetAssessor class="TrecAssessor">
    </snippetAssessor>
    <documentAssessor class="TrecAssessor">
    </documentAssessor>
  </relevanceAssessor>
  <stoppingDecisionMaker class="FixedDepthDecisionMaker">
    <attribute name="depth" value="10" />
  </stoppingDecisionMaker>
  <costCalculator class="FixedCostCalculator">
    <attribute name="time_limit" value="600" />
    <attribute name="query_cost" value="10" />
    <attribute name="document_cost" value="20" />
    <attribute name="snippet_cost" value="3" />
    <attribute name="serp_results_cost" value="5" />
    <attribute name="mark_document_cost" value="3" />
  </costCalculator>
  <searchContext class="SearchContext">
    <attribute name="relevance_revision" value="1"/>
  </searchContext>
  <serpImpression class="SimpleSerpImpression">
    <attribute name="qrel_file" value="<..>/trec2005.qrels.all"/>
  </serpImpression>
</userConfiguration>
```

The value of `user_type` can also be set to `None` to resemble original SimIIR configurations without the behavior section.

5 SIMIIR 2.0 IN ACTION

The transition probabilities of our new Markov model-based components can be instantiated manually or be derived from search logs. For our experiments, we use the Sowiport User Search Session dataset (SUSS) [33] that includes 558,008 sessions with about 8 million interactions (179,796 of them queries) collected from April 2014 to April 2015 from users of the Sowiport digital library search system. Within the sessions, 58 different actions were logged while users interacted with the system (e.g., formulating a query or clicking on a document). Following Zerhoudi et al. [45], we split the SUSS data into exploratory and lookup subsets to train respective individual models. Of course, also any other search log or splitting strategy could be used to train the Markov models.

The simulation process in the original SimIIR framework is triggered by a single XML file.[5] It defines the output options, topics (i.e., titles and descriptions available to the query generation strategies), the search interface, and the configuration files of the simulated users. Each individual simulated user can mimic an individual participant of a user study with a specific static query generation strategy, document/snippet relevance assessment method, and stopping criterion. Examples are fixed depth users (stopping at a certain

[5]https://github.com/padre-lab-eu/simiir-2/blob/main/example_sims/trec_test_simulation.xml

Figure 2: Excerpt of simulated sessions for the topic `extinction wildlife` generated by (left) the standard SimIIR user, (middle) an exploratory SimIIR 2.0 user, and (right) a lookup SimIIR 2.0 user. A session includes the actions of the simulated user (e.g., QUERY, SERP, SNIPPET), the session's time limit (600 seconds), the cumulated elapsed time (e.g., 10, 15, 18 seconds), and an action's metadata (e.g., the query string in green or the relevance assessment for some snippet or document ID).

```
QUERY   600  10 extinction wildlife
SERP    600  15 EXAMINE_SERP
SNIPPET 600  18 SNIPPET_NOT_RELEVANT APW19...0801
SNIPPET 600  21 SNIPPET_NOT_RELEVANT APW19...1129
SNIPPET 600  24 SNIPPET_NOT_RELEVANT APW19...0405
SNIPPET 600  27 SNIPPET_RELEVANT APW19...1290
DOC     600  47 EXAMINING_DOCUMENT APW19...1290
SNIPPET 600  50 SNIPPET_NOT_RELEVANT APW19...0561
QUERY   600  60 extinction species wildlife
SERP    600  65 EXAMINE_SERP
SNIPPET 600  68 SNIPPET_RELEVANT APW19...0801
DOC     600  88 EXAMINING_DOCUMENT APW19...0801
MARK    600  91 CONSIDERED_RELEVANT APW19...0801
SNIPPET 600  94 SNIPPET_NOT_RELEVANT APW19...1129
SNIPPET 600  97 SNIPPET_RELEVANT APW19...0561
DOC     600 117 EXAMINING_DOCUMENT APW19...0561
MARK    600 120 CONSIDERED_RELEVANT APW19...0561
SNIPPET 600 123 SNIPPET_NOT_RELEVANT APW19...1668
SNIPPET 600 126 SNIPPET_NOT_RELEVANT APW19...0166
QUERY   600 136 extinction prevent wildlife
SERP    600 141 EXAMINE_SERP
SNIPPET 600 144 SNIPPET_NOT_RELEVANT APW19...0801
SNIPPET 600 147 SNIPPET_RELEVANT APW19...1129
DOC     600 167 EXAMINING_DOCUMENT APW19...1129
MARK    600 170 CONSIDERED_RELEVANT APW19...1129
SNIPPET 600 173 SNIPPET_NOT_RELEVANT APW19...0405
SNIPPET 600 176 SNIPPET_NOT_RELEVANT APW19...1290
SNIPPET 600 179 SNIPPET_NOT_RELEVANT APW19...0561
...
```

```
QUERY   600  10 extinction wildlife
SERP    600  15 EXAMINE_SERP
SNIPPET 600  18 SNIPPET_NOT_RELEVANT APW19...0861
SNIPPET 600  21 SNIPPET_NOT_RELEVANT APW19...1129
SNIPPET 600  24 SNIPPET_NOT_RELEVANT APW19...0405
SNIPPET 600  27 SNIPPET_RELEVANT APW19...1290
DOC     600  47 EXAMINING_DOCUMENT APW19...1290
SNIPPET 600  50 SNIPPET_NOT_RELEVANT APW19...0561
SNIPPET 600  53 SNIPPET_NOT_RELEVANT APW19...1434
SNIPPET 600  56 SNIPPET_RELEVANT APW19...0030
DOC     600  76 EXAMINING_DOCUMENT APW19...0030
QUERY   600  86 wildlife extinction in the
                philippines
SERP    600  91 EXAMINE_SERP
SNIPPET 600  94 SNIPPET_RELEVANT APW19...0801
DOC     600 114 EXAMINING_DOCUMENT APW19...0801
MARK    600 117 CONSIDERED_RELEVANT APW19...0801
SNIPPET 600 120 SNIPPET_NOT_RELEVANT APW19...1129
SNIPPET 600 123 SNIPPET_RELEVANT APW19...0561
DOC     600 143 EXAMINING_DOCUMENT APW19...0561
MARK    600 146 CONSIDERED_RELEVANT APW19...0561
SNIPPET 600 149 SNIPPET_NOT_RELEVANT APW19...1668
SNIPPET 600 152 SNIPPET_NOT_RELEVANT APW19...0166
SNIPPET 600 155 SNIPPET_NOT_RELEVANT APW19...0986
SNIPPET 600 158 SNIPPET_RELEVANT APW19...0738
DOC     600 178 EXAMINING_DOCUMENT APW19...0738
MARK    600 181 CONSIDERED_RELEVANT APW19...0738
SNIPPET 600 184 SNIPPET_NOT_RELEVANT APW19...0566
...
```

```
QUERY   600  10 extinction wildlife
SERP    600  15 EXAMINE_SERP
SNIPPET 600  18 SNIPPET_NOT_RELEVANT APW19...0801
SNIPPET 600  21 SNIPPET_NOT_RELEVANT APW19...1129
QUERY   600  26 extinction species wildlife
SERP    600  31 EXAMINE_SERP
SNIPPET 600  34 SNIPPET_NOT_RELEVANT APW19...0801
SNIPPET 600  37 SNIPPET_NOT_RELEVANT APW19...1129
SNIPPET 600  40 SNIPPET_NOT_RELEVANT APW19...0561
SNIPPET 600  43 SNIPPET_RELEVANT APW19...1668
DOC     600  63 EXAMINING_DOCUMENT APW19...1668
SNIPPET 600  66 SNIPPET_NOT_RELEVANT APW19...0166
QUERY   600  76 extinction animals wildlife
SERP    600  81 EXAMINE_SERP
SNIPPET 600  84 SNIPPET_NOT_RELEVANT APW19...0801
SNIPPET 600  87 SNIPPET_RELEVANT APW19...1129
DOC     600 107 EXAMINING_DOCUMENT APW19...1129
MARK    600 110 CONSIDERED_RELEVANT APW19...1129
QUERY   600 120 extinction prevent wildlife
SERP    600 125 EXAMINE_SERP
QUERY   600 130 extinction spotted wildlife
SERP    600 135 EXAMINE_SERP
SNIPPET 600 138 SNIPPET_RELEVANT APW19...0801
DOC     600 158 EXAMINING_DOCUMENT APW19...0801
MARK    600 161 CONSIDERED_RELEVANT APW19...0801
SNIPPET 600 164 SNIPPET_NOT_RELEVANT APW19...1129
SNIPPET 600 167 SNIPPET_RELEVANT APW19...0405
DOC     600 187 EXAMINING_DOCUMENT APW19...0405
...
```

threshold), TREC users (following the relevance judgments), and IFT users (maximizing their gain).

The simulation process in SimIIR 2.0 allows for more complex experimental settings. Simulated users are defined by an elaborated search behavior like the user type-specific Markov models for exploratory and lookup users. These models can determine the stopping behavior instead of the original threshold-based strategies.[6] In the new SimIIR 2.0 setup, Markov model-based decisions can also be combined with the stopping strategies of the original SimIIR framework. For instance, while predicting the next actions of a simulated exploratory user using the respective Markov model, the search result examination can be stopped when the gained knowledge drops below a user's average gain rate.

Figure 2 (left) shows an excerpt of a session generated by a basic simulation configuration of the original SimIIR framework.[5] Given the topic `extinction wildlife` and its description, the simulated fixed-depth smart user starts by submitting the topic title as the first query, examines some snippets and a document before submitting a second query, inspecting further snippets, etc.

The simulated session in Figure 2 (middle) is generated by a new exploratory user,[7] while the session in Figure 2 (right) comes from a new lookup user.[8] Both use the Google suggest API to select a next query from the up to ten suggestions. Just like in the example, we observed that simulated exploratory users tend to more exhaustively explore the search result list and reformulate the query as they learn more about the topic while lookup users only investigate the first few results and quickly rephrase their queries.

6 CONCLUSION

We have presented SimIIR 2.0: an extended and updated version of the SimIIR search behavior simulation framework. Since the rather static components of the original framework do not take session history into account, we add this ability to the components for query formulation and stopping decisions—also including Markov modeling abilities to reflect different dynamic user types.

In future work, we plan to enable more influence between the different components of the extended Complex Searcher Model for more realistic simulated sessions. We also experiment with other non-Markov models for interaction simulation [15] and plan to include respective components in upcoming SimIIR 2.0 versions. Furthermore, so far, only single sessions are simulated but also cross-session search [29] or search missions [17, 22], as well as cross-device search [14] could be interesting simulation targets.

With the SimIIR 2.0 framework open-sourced under a permissive license, others can also easily contribute further simulation components. SimIIR 2.0 thus can become a platform for accessible and reproducible retrieval simulation. Ideally, it can directly support or otherwise quickly include components for new simulation ideas in "classic" search box-based but also in conversational scenarios; for instance, the various simulation-based studies published recently at ECIR 2022 [4, 5, 30, 34] or at SIGIR 2022 [13, 26, 47].

ACKNOWLEDGMENTS

This work has been partially supported by the DFG (German Research Foundation) through the project 408022022 "SINIR – Simulating INteractive Information Retrieval".

[6] https://github.com/padre-lab-eu/simiir-2/blob/main/example_sims/users/exploratory_user.xml
[7] https://github.com/padre-lab-eu/simiir-2/blob/main/example_sims/trec_exploratory_simulation.xml
[8] https://github.com/padre-lab-eu/simiir-2/blob/main/example_sims/trec_lookup_simulation.xml

REFERENCES

[1] Kumaripaba Athukorala, Dorota Glowacka, Giulio Jacucci, Antti Oulasvirta, and Jilles Vreeken. 2016. Is exploratory search different? A comparison of information search behavior for exploratory and lookup tasks. *Journal of the Association for Information Science and Technology* 67, 11 (2016), 2635–2651.

[2] Leif Azzopardi, Maarten de Rijke, and Krisztian 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 2007, Amsterdam, The Netherlands, July 23–27, 2007.* ACM, 455–462.

[3] Edwin G. Boring. 1946. Mind and mechanism. *The American Journal of Psychology* 59, 2 (1946), 173–192.

[4] Timo Breuer, Norbert Fuhr, and Philipp Schaer. 2022. Validating simulations of user query variants. In *Advances in Information Retrieval - 44th European Conference on IR Research, ECIR 2022, Stavanger, Norway, April 10–14, 2022, Proceedings, Part I (Lecture Notes in Computer Science, Vol. 13185).* Springer, 80–94.

[5] Arthur Câmara, David Maxwell, and Claudia Hauff. 2022. Searching, learning, and subtopic ordering: A simulation-based analysis. In *Advances in Information Retrieval - 44th European Conference on IR Research, ECIR 2022, Stavanger, Norway, April 10-14, 2022, Proceedings, Part I (Lecture Notes in Computer Science, Vol. 13185).* Springer, 142–156.

[6] Ben Carterette, Ashraf Bah, and Mustafa Zengin. 2015. Dynamic test collections for retrieval evaluation. In *Proceedings of the 2015 International Conference on The Theory of Information Retrieval, ICTIR 2015, Northampton, Massachusetts, USA, September 27–30, 2015.* ACM, 91–100.

[7] Jia Chen, Jiaxin Mao, Yiqun Liu, Fan Zhang, Min Zhang, and Shaoping Ma. 2021. Towards a better understanding of query reformulation behavior in web search. In *The Web Conference 2021, WWW 2021, Virtual Event, Ljubljana, Slovenia, April 19–23, 2021.* ACM / IW3C2, 743–755.

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

[9] Cyril W. Cleverdon, Jack Mills, and Michael E. Keen. 1966. *Factors determining the performance of indexing systems; Volume 1: Design.* Technical Report. College of Aeronautics, Cranfield.

[10] Michael J. Cole. 2010. Simulation of the IIR user: Beyond the automagic. In *Proceedings of the SIGIR 2010 Workshop on the Simulation of Interaction: Automated Evaluation of Interactive IR (SimInt 2010).* 1–2.

[11] Susan T. Dumais, Edward Cutrell, Jonathan J. Cadiz, Gavin Jancke, Raman Sarin, and Daniel 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 Information Retrieval, SIGIR 2003, July 28 – August 1, 2003, Toronto, Canada.* ACM, 72–79.

[12] Sean R. Eddy. 1995. Multiple alignment using hidden Markov models. In *Proceedings of the Third International Conference on Intelligent Systems for Molecular Biology, Cambridge, United Kingdom, July 16–19, 1995.* AAAI, 114–120.

[13] Pierre Erbacher, Ludovic Denoyer, and Laure Soulier. 2022. Interactive query clarification and refinement via user simulation. In *The 45th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2022, Madrid, Spain, July 11–15, 2022.* ACM, 2420–2425.

[14] Sebastian Gomes, Miriam Boon, and Orland Hoeber. 2022. A study of cross-session cross-device search within an academic digital library. In *The 45th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2022, Madrid, Spain, July 11–15, 2022.* ACM, 384–394.

[15] Sebastian Günther, Paul Göttert, and Matthias Hagen. 2022. Exploring LSTMs for simulating search sessions in digital libraries. In *Linking Theory and Practice of Digital Libraries - 26th International Conference on Theory and Practice of Digital Libraries, TPDL 2022, Padua, Italy, September 20–23, 2022, Proceedings, (to appear).* 5 pages.

[16] Sebastian Günther and Matthias Hagen. 2021. Assessing query suggestions for search session simulation. In *Causality in Search and Recommendation (CSR) and Simulation of Information Retrieval Evaluation (Sim4IR) workshops at SIGIR 2021 (CEUR Workshop Proceedings, Vol. 2911).* CEUR-WS.org, 8 pages.

[17] Matthias Hagen, Jakob Gomoll, Anna Beyer, and Benno Stein. 2013. From search session detection to search mission detection. In *Open research Areas in Information Retrieval, OAIR 2013, Lisbon, Portugal, May 15-17, 2013.* ACM, 85–92.

[18] Matthias Hagen, Martin Potthast, Michael Völske, Jakob Gomoll, and Benno Stein. 2016. How writers search: Analyzing the search and writing logs of nonfictional essays. In *Proceedings of the 2016 ACM Conference on Human Information Interaction and Retrieval, CHIIR 2016, Carrboro, North Carolina, USA, March 13–17, 2016.* ACM, 193–202.

[19] Donna Harman. 2011. *Information Retrieval Evaluation.* Morgan & Claypool Publishers.

[20] Kalervo Järvelin. 2009. Interactive relevance feedback with graded relevance and sentence extraction: Simulated user experiments. In *Proceedings of the 18th ACM Conference on Information and Knowledge Management, CIKM 2009, Hong Kong, China, November 2–6, 2009.* ACM, 2053–2056.

[21] Jiepu Jiang and Chaoqun Ni. 2016. What affects word changes in query reformulation during a task-based search session?. In *Proceedings of the 2016 ACM Conference on Human Information Interaction and Retrieval, CHIIR 2016, Carrboro, North Carolina, USA, March 13–17, 2016.* ACM, 111–120.

[22] Rosie Jones and Kristina Lisa Klinkner. 2008. Beyond the session timeout: Automatic hierarchical segmentation of search topics in query logs. In *Proceedings of the 17th ACM Conference on Information and Knowledge Management, CIKM 2008, Napa Valley, California, USA, October 26–30, 2008.* ACM, 699–708.

[23] Chris Jordan, Carolyn R. Watters, and Qigang Gao. 2006. Using controlled query generation to evaluate blind relevance feedback algorithms. In *ACM/IEEE Joint Conference on Digital Libraries, JCDL 2006, Chapel Hill, NC, USA, June 11–15, 2006, Proceedings.* ACM, 286–295.

[24] Diane Kelly. 2009. Methods for evaluating interactive information retrieval systems with users. *Foundations and Trends in Information Retrieval* 3, 1-2 (2009), 1–224.

[25] Heikki Keskustalo, Kalervo Järvelin, and Ari 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 (2008), 209–228.

[26] To Eun Kim and Aldo Lipani. 2022. A multi-task based neural model to simulate users in goal oriented dialogue systems. In *The 45th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2022, Madrid, Spain, July 11–15, 2022.* ACM, 2115–2119.

[27] Julia Kiseleva, Hoang Thanh Lam, Mykola Pechenizkiy, and Toon Calders. 2013. Predicting current user intent with contextual Markov models. In *13th IEEE International Conference on Data Mining Workshops, ICDM Workshops, TX, USA, December 7–10, 2013.* IEEE Computer Society, 391–398.

[28] Ron Kohavi, Diane Tang, and Ya Xu. 2020. *Trustworthy Online Controlled Experiments: A Practical Guide to A/B Testing.* Cambridge University Press.

[29] Alexander Kotov, Paul N. Bennett, Ryen W. White, Susan T. Dumais, and Jaime Teevan. 2011. Modeling and analysis of cross-session search tasks. In *Proceedings of the 34th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2011, Beijing, China, July 25–29, 2011.* ACM, 5–14.

[30] Sahiti Labhishetty and ChengXiang Zhai. 2022. RATE: A reliability-aware tester-based evaluation framework of user simulators. In *Advances in Information Retrieval - 44th European Conference on IR Research, ECIR 2022, Stavanger, Norway, April 10-14, 2022, Proceedings, Part I (Lecture Notes in Computer Science, Vol. 13185).* Springer, 336–350.

[31] David Maxwell and Leif Azzopardi. 2016. 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 2016, Pisa, Italy, July 17–21, 2016.* ACM, 1141–1144.

[32] David Maxwell, Leif Azzopardi, Kalervo Järvelin, and Heikki Keskustalo. 2015. 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 2015, Santiago, Chile, August 9–13, 2015.* ACM, 903–906.

[33] Philipp Mayr. 2016. Sowiport User Search Sessions Data Set (SUSS). GESIS - Leibniz-Institute for the Social Sciences. Data File Version 1.0.0. https://doi.org/10.7802/1380

[34] Gustavo Penha, Arthur Câmara, and Claudia Hauff. 2022. Evaluating the robustness of retrieval pipelines with query variation generators. In *Advances in Information Retrieval - 44th European Conference on IR Research, ECIR 2022, Stavanger, Norway, April 10–14, 2022, Proceedings, Part I (Lecture Notes in Computer Science, Vol. 13185).* Springer, 397–412.

[35] Martin Potthast, Matthias Hagen, Michael Völske, and Benno Stein. 2013. Crowdsourcing Interaction Logs to Understand Text Reuse from the Web. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics, ACL 2013, 4–9 August 2013, Sofia, Bulgaria, Volume 1: Long Papers.* The Association for Computer Linguistics, 1212–1221.

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

[37] Ahmad Shamshad, M. A. Bawadi, W. M. A. Wan Hussin, Taksiah A. Majid, and S. A. M. Sanusi. 2005. First and second order Markov chain models for synthetic generation of wind speed time series. *Energy* 30, 5 (2005), 693–708.

[38] Marc Sloan, Hui Yang, and Jun Wang. 2015. A term-based methodology for query reformulation understanding. *Information Retrieval Journal* 18, 2 (2015), 145–165.

[39] Mark D. Smucker. 2011. An analysis of user strategies for examining and processing ranked lists of documents. In *Human-Computer Information Retrieval Symposium, HCIR 2011, Mountain View, CA, USA, October 20, 2011.* 4 pages.

[40] Matthijs T. J. Spaan. 2012. Partially observable Markov decision processes. In *Reinforcement Learning.* Adaptation, Learning, and Optimization, Vol. 12. Springer, 387–414. https://doi.org/10.1007/978-3-642-27645-3_12

[41] Paul Thomas, Alistair Moffat, Peter Bailey, and Falk Scholer. 2014. Modeling decision points in user search behavior. In *Fifth Information Interaction in Context Symposium, IIiX '14, Regensburg, Germany, August 26–29, 2014.* ACM, 239–242.

[42] Suzan Verberne, Maya Sappelli, Kalervo Järvelin, and Wessel Kraaij. 2015. User simulations for interactive search: Evaluating personalized query suggestion. In *Advances in Information Retrieval - 37th European Conference on IR Research,*

ECIR 2015, Vienna, Austria, March 29 – April 2, 2015, Proceedings (Lecture Notes in Computer Science, Vol. 9022). Springer, 678–690.

[43] Hui Yang, Dongyi Guan, and Sicong Zhang. 2015. The query change model: Modeling session search as a Markov decision process. *ACM Transactions on Information Systems* 33, 4 (2015), 20:1–20:33.

[44] Saber Zerhoudi, Michael Granitzer, Jörg Schlötterer, and Christin Seifert. 2021. Query change as a contextual Markov model for simulating user search behaviour. In *Forum for Information Retrieval Evaluation, FIRE 2021, Virtual Event, India, December 13–17, 2021*. ACM, 43–51.

[45] Saber Zerhoudi, Michael Granitzer, Christin Seifert, and Joerg Schloetterer. 2022. Evaluating simulated user interaction and search behaviour. In *Advances in Information Retrieval - 44th European Conference on IR Research, ECIR 2022, Stavanger, Norway, April 10–14, 2022, Proceedings, Part II (Lecture Notes in Computer Science, Vol. 13186)*. Springer, 240–247.

[46] Saber Zerhoudi, Michael Granitzer, Christin Seifert, and Jörg Schlötterer. 2022. Simulating user interaction and search behaviour in digital libraries. In *Proceedings of the 18th Italian Research Conference on Digital Libraries, Padua, Italy, February 24–25, 2022 (hybrid event) (CEUR Workshop Proceedings, Vol. 3160)*. CEUR-WS.org, 15 pages.

[47] Shuo Zhang, Mu-Chun Wang, and Krisztian Balog. 2022. Analyzing and simulating user utterance reformulation in conversational recommender systems. In *The 45th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2022, Madrid, Spain, July 11–15, 2022*. ACM, 133–143.

[48] Yinan Zhang, Xueqing Liu, and ChengXiang 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 2017, Amsterdam, The Netherlands, October 1–4, 2017*. ACM, 193–200.