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

Saber Zerhoudi[1,⋆] Sebastian Günther[2,⋆] Kim Plassmeier[3]
Timo Borst[3] Christin Seifert[4] Matthias Hagen[2] Michael Granitzer[1]

⋆contributed equally [1]University of Passau [2]Martin-Luther-Universität Halle-Wittenberg
[3]ZBW Leibniz Information Centre for Economics [4]University of Duisburg-Essen

ABSTRACT

Simulating user–retrieval system interactions enables evaluation studies with controlled user behavior variations. To this end, the prominent SimIIR framework offers static, rule-based user models. We present an extended SimIIR 2.0 version with new components for dynamic user type-specific Markov model-based interaction simulation and more realistic query generation. A flexible modularization concept ensures that the SimIIR 2.0 framework can serve as a platform to implement, combine, run, and compare the growing number of proposed user models 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, Michael Granitzer. 2022. Resource: The SimIIR 2.0 Framework: User Types, Markov Model-Based Interaction Simulation, and Advanced Query Generation. In *Proceedings of ACM International Conference on Information and Knowledge Management (CIKM '22).* ACM, New York, NY, USA, 5 pages. https://doi.org/XXXXXXX.XXXXXXX

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 [7] 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., searchers reformulating queries [5]) 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 [19] or via smaller controlled user studies [9, 16]. 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 [20] 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 the simulated users in completing their tasks. In an experimental comparison of the existing framework and our extended version, we show that the new dynamic components help to capture the evolution of a user's information need during a session.

Our extended SimIIR 2.0 framework is updated to the latest Python version and—like the original SimIIR framework—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.

2 RELATED WORK

Evaluation has always been a core theme in IR; Kelly [16], Sanderson [23], and Harman [12] nicely cover the history. Today, the Cranfield paradigm from the 1960s [7] is still often used even though this usually means to evaluate systems against static ad-hoc queries without user interactions. Some evaluation studies employed simulation but usually focused on simulating single aspects like click behavior [6], query (re-)formulation [2, 4, 15, 29], relevance feedback [13, 17], or stopping behavior [21, 28].

Recently, two studies have emphasized the importance of simulating the search process as a whole [20, 34] even though Cole [8] had collected several challenges when developing realistic simulations against "real" retrieval systems [8]. Cole's challenges are based

CIKM '22, October 17–22, 2022, Atlanta, Georgia, USA

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM.
https://doi.org/XXXXXXX.XXXXXXX

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

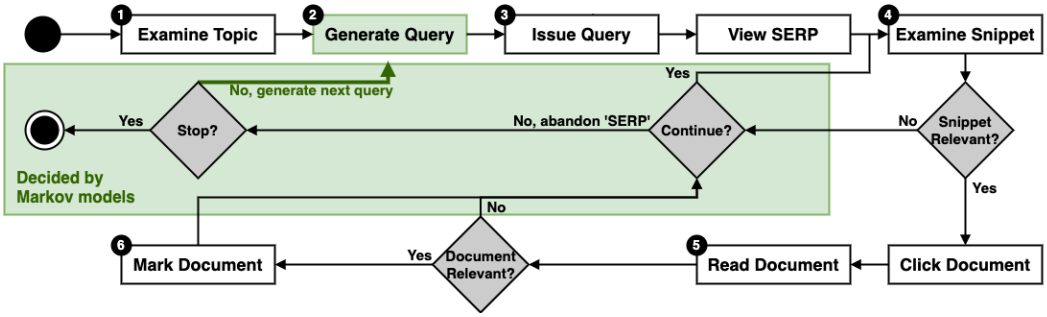

Figure 1: Flowchart of the Complex Searcher Model (CSM) that is the basis for interaction simulation in the SimIIR framework [20]. The key components are shown as white boxes with numbered steps and a simulated user's decision points are indicated in gray (SERP: search engine result page). Components that we improve in our extension are shown in green.

on Boring's five step operationalist approach [3] and essentially state that simulations need to be aligned with real user behavior.

The SimIIR framework [20] 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. In this paper, we thus extend the SimIIR framework by including more dynamic components and allowing more interaction between the simulation modules (e.g., to take the interaction history into account for the next simulation steps) .

In order to align the simulation with real user behavior, SimIIR 2.0 contains components that can fit Markov models on real log data to simulate global or search type-specific behavior. Markov models are based on a well-established theory and are rather simple and compact. Multiple approaches have applied Markov models to search behavior: first-order [24] or kth-order Markov models, partially observable Markov decision processes (POMDPs) [27], and hidden Markov models (HMMs) [10]. We apply first-order Markov models due to their relative simplicity.

3 THE EXISTING SIMIIR FRAMEWORK

SimIIR [20] 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.

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

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 [20], 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). 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" users will explore a search result list more exhaustively than "lookup" users who will only investigate the first few results and then rephrase their queries rather quickly [1, 33]. 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 result page level and on the session level are made by Markov model-based user type-specific models instead of the original stochastic heuristics with stopping threshold variables. To this end, we categorize the simulated users into different types and model their search process using specific Markov models. At the stopping decisions, these models are used to 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. Several studies have shown that later queries in a session often depend on the content of previously viewed search results [14, 25, 30]. We thus use user type-specific Markov approaches to model how a particular user group changes their queries from query q_{i-1} to q_i (e.g., generalization or specialization) and to predict the next likely query "change" direction.

4.2 Realization and Implementation

Query generation. We add several types of query generation approaches. The first type queries an actual search engine's API to obtain query suggestions—a technique earlier demonstrated to yield realistic sessions [11]. 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 [31]. 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 reliably simulate user type-specific behavior, we categorize users into different groups based on their search and stopping behavior. We use the previously introduced contextual search types [1, 32] and categorize users into (1) 'exploratory searchers' who tend to fully explore a search result list and extensively use potentially available result filters, and (2) 'lookup searchers' who only investigate the first few results and quickly rephrase their queries. Also other user types like 'fast and liberal' vs. 'slow and picky' users [18, 26] can easily be configured.

Extended Configuration. The configuration for SimIIR 2.0 has been extended to accommodate additional attributes. Configuration 1 showcases an example of an advanced simulated user configuration file (i.e., advanced TREC user behavior). The file's structure is similar to the original,[4] with the inclusion of a new query generation approach (i.e., GoogleSuggestGenerator), and the addition of a new section (i.e., `algorithm`) that describes the user type-specific behavior which have been used to predict the user's next action. The `algorithm` section contains three different attributes: `model_type` to define the used method, `transition_matrix` and `states` to provide the transition probability matrix with a list of states. The value of the `model_type` attribute can be left as `None` is we wish to simulate user search session using the original configuration.

Configuration 1: advanced_configuration.xml

```
<userConfiguration id="advancedtrecuser">
  <algorithm class="MarkovChain">
    <attribute name="model_type" value="exploratory"/>
    <attribute name="transition_matrix"
               value="<..>/dataset_matrix.data"/>
    <attribute name="states" value="<..>/dataset_states.data"/>
  </algorithm>
  <queryGenerator class="GoogleSuggestGenerator">
    <attribute name="stopword_file" value="<..>/stopwords.txt"/>
    <attribute name="max_depth" type="integer" value="5" />
  </queryGenerator>
  <textClassifiers>
    <snippetClassifier class="TrecTextClassifier">
    </snippetClassifier>
    <documentClassifier class="TrecTextClassifier">
    </documentClassifier>
  </textClassifiers>
  <stoppingDecisionMaker class="FixedDepthDecisionMaker">
    <attribute name="depth" value="10" />
  </stoppingDecisionMaker>
  <logger class="FixedCostLogger">
    <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" />
  </logger>
  <searchContext class="SearchContext">
    <attribute name="relevance_revision" value="1"/>
  </searchContext>
  <serpImpression class="SimpleSERPImpression">
    <attribute name="qrel_file" value="<..>/trec2005.qrels.all"/>
  </serpImpression>
</userConfiguration>
```

5 EXPERIMENTAL SCENARIOS

Our new Markov model-based query selection and stopping decision components require data in form of search engine logs to train the models. Such logs also enable the identification of general user types whose behavior a simulation could mimic. In our below experiments, we use the Sowiport User Search Session dataset (SUSS).[5] [22] It includes 558,008 sessions with about 8 million interactions (179,796 of these are queries) and was collected from April 2014 to April 2015 from users of the Sowiport digital library search system. Within the sessions, 58 different actions describe the users' activities while interacting with the interface of the search engine (e.g., formulating a query or clicking on a document). Alternatively, also any other search session dataset could be used to analyze the user types and train the Markov models for exploratory-vs-lookup behavior and the query changes (cf. Section 4.2). Following Zerhoudi et al. [33], we divide the SUSS dataset into exploratory and lookup subsets to train respective Markov models.

The simulation process in the original SimIIR framework is triggered by a single XML file.[6] 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 threshold), TREC users (following the relevance judgments), and IFT users (maximizing their gain).

The simulation process in the SimIIR 2.0 also 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

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

[5] Publicly available at http://dx.doi.org/10.7802/1380

[6] https://github.com/padre-lab-eu/extended-simiir/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/document ID).

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

```
QUERY   600  10 extinction wildlife
SERP    600  15 EXAMINE_SERP
SNIPPET 600  18 SNIPPET_NOT_RELEVANT b'APW19...0801'
SNIPPET 600  21 SNIPPET_NOT_RELEVANT b'APW19...1129'
SNIPPET 600  24 SNIPPET_NOT_RELEVANT b'APW19...0405'
SNIPPET 600  27 SNIPPET_RELEVANT b'APW19...1290'
DOC     600  47 EXAMINING_DOCUMENT APW19...1290
SNIPPET 600  50 SNIPPET_NOT_RELEVANT b'APW19...0561'
SNIPPET 600  53 SNIPPET_NOT_RELEVANT b'APW19...1434'
SNIPPET 600  56 SNIPPET_RELEVANT b'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 b'APW19...0801'
DOC     600 114 EXAMINING_DOCUMENT APW19...0801
MARK    600 117 CONSIDERED_RELEVANT APW19...0801
SNIPPET 600 120 SNIPPET_NOT_RELEVANT b'APW19...1129'
SNIPPET 600 123 SNIPPET_RELEVANT b'APW19...0561'
DOC     600 143 EXAMINING_DOCUMENT APW19...0561
MARK    600 146 CONSIDERED_RELEVANT APW19...0561
SNIPPET 600 149 SNIPPET_NOT_RELEVANT b'APW19...1668'
SNIPPET 600 152 SNIPPET_NOT_RELEVANT b'APW19...0166'
SNIPPET 600 155 SNIPPET_NOT_RELEVANT b'APW19...0986'
SNIPPET 600 158 SNIPPET_RELEVANT b'APW19...0738'
DOC     600 178 EXAMINING_DOCUMENT APW19...0738
MARK    600 181 CONSIDERED_RELEVANT APW19...0738
SNIPPET 600 184 SNIPPET_NOT_RELEVANT b'APW19...0566'
SNIPPET 600 187 SNIPPET_NOT_RELEVANT b'APW19...0552'
...
```

```
QUERY   600  10 extinction wildlife
SERP    600  15 EXAMINE_SERP
SNIPPET 600  18 SNIPPET_NOT_RELEVANT b'APW19...0801'
SNIPPET 600  21 SNIPPET_NOT_RELEVANT b'APW19...1129'
QUERY   600  26 extinction species wildlife
SERP    600  31 EXAMINE_SERP
SNIPPET 600  34 SNIPPET_NOT_RELEVANT b'APW19...0801'
SNIPPET 600  37 SNIPPET_NOT_RELEVANT b'APW19...1129'
SNIPPET 600  40 SNIPPET_NOT_RELEVANT b'APW19...0561'
SNIPPET 600  43 SNIPPET_RELEVANT b'APW19...1668'
DOC     600  63 EXAMINING_DOCUMENT APW19...1668
SNIPPET 600  66 SNIPPET_NOT_RELEVANT b'APW19...0166'
QUERY   600  76 extinction animals wildlife
SERP    600  81 EXAMINE_SERP
SNIPPET 600  84 SNIPPET_NOT_RELEVANT b'APW19...0801'
SNIPPET 600  87 SNIPPET_RELEVANT b'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 b'APW19...0801'
DOC     600 158 EXAMINING_DOCUMENT APW19...0801
MARK    600 161 CONSIDERED_RELEVANT APW19...0801
SNIPPET 600 164 SNIPPET_NOT_RELEVANT b'APW19...1129'
SNIPPET 600 167 SNIPPET_RELEVANT b'APW19...0405'
DOC     600 187 EXAMINING_DOCUMENT APW19...0405
MARK    600 190 CONSIDERED_RELEVANT APW19...0405
...
```

the stopping behavior instead of the original threshold-based strategies.[7] 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.[6] Given the topic `extinction wildlife` and its description, the simulated user starts their session by submitting the topic title as the first query, examining some snippets and a document before submitting the second query `extinction species wildlife`, inspecting further snippets, etc. until the stopping criterion is met.

In Figure 2 (middle), the simulated session is generated by an exploratory user,[8] while the session in Figure 2 (right) comes from a lookup user.[9] Both simulations use the Google Suggest API to generate the next queries from the up to 10 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.

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

6 CONCLUSION

In this work, 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 cannot take session history into account, we add this ability to the components for stopping decisions and query formulation. We also extend the stopping decisions by including Markov modeling abilities that can reflect different dynamic user types.

As future work, we will enable even more connections and interactions between the different components of the extended Complex Searcher Model to improve the overall realism of the simulated sessions. With the SimIIR 2.0 framework open-sourced under a permissive license, also others can easily contribute further simulation components so that SimIIR 2.0 can become a platform for accessible and reproducible retrieval simulation.

ACKNOWLEDGMENTS

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

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. *J. Assoc. Inf. Sci. Technol.* 67, 11 (2016), 2635–2651. https://doi.org/10.1002/asi.23617

[2] Leif Azzopardi, Maarten de Rijke, and Krisztian Balog. 2007. Building simulated queries for known-item topics: an analysis using six european languages. In *SIGIR 2007: Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Amsterdam, The Netherlands, July 23-27, 2007*, Wessel Kraaij, Arjen P. de Vries, Charles L. A. Clarke, Norbert Fuhr, and Noriko Kando (Eds.). ACM, 455–462. https://doi.org/10.1145/1277741.1277820

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

[4] 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*, James Allan, W. Bruce Croft, Arjen P. de Vries, and Chengxiang Zhai (Eds.). ACM, 91–100. https://doi.org/10.1145/2808194.2809470

[5] 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 *WWW '21: The Web Conference 2021, Virtual Event / Ljubljana, Slovenia, April 19-23, 2021*, Jure Leskovec, Marko Grobelnik, Marc Najork, Jie Tang, and Leila Zia (Eds.). ACM / IW3C2, 743–755. https://doi.org/10.1145/3442381.3450127

[6] Aleksandr Chuklin, Ilya Markov, and Maarten de Rijke. 2015. *Click Models for Web Search*. Morgan & Claypool Publishers. https://doi.org/10.2200/S00654ED1V01Y201507ICR043

[7] Cyril W Cleverdon, Jack Mills, and Michael E Keen. 1966. Factors determining the performance of indexing systems,(Volume 1: Design). *Cranfield: College of Aeronautics* 28 (1966).

[8] Michael J Cole. 2010. Simulation of the IIR user: Beyond the automagic. *Simulation of Interaction* (2010), 1.

[9] 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 *SIGIR 2003: Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, July 28 - August 1, 2003, Toronto, Canada*, Charles L. A. Clarke, Gordon V. Cormack, Jamie Callan, David Hawking, and Alan F. Smeaton (Eds.). ACM, 72–79. https://doi.org/10.1145/860435.860451

[10] 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*, Christopher J. Rawlings, Dominic A. Clark, Russ B. Altman, Lawrence Hunter, Thomas Lengauer, and Shoshana J. Wodak (Eds.). AAAI, 114–120. http://www.aaai.org/Library/ISMB/1995/ismb95-014.php

[11] Sebastian Günther and Matthias Hagen. 2021. Assessing Query Suggestions for Search Session Simulation. *Balog et al.[2021]* (2021), 38–45.

[12] Donna Harman. 2011. *Information Retrieval Evaluation*. Morgan & Claypool Publishers. https://doi.org/10.2200/S00368ED1V01Y201105ICR019

[13] 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*, David Wai-Lok Cheung, Il-Yeol Song, Wesley W. Chu, Xiaohua Hu, and Jimmy Lin (Eds.). ACM, 2053–2056. https://doi.org/10.1145/1645953.1646299

[14] 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*, Diane Kelly, Robert Capra, Nicholas J. Belkin, Jaime Teevan, and Pertti Vakkari (Eds.). ACM, 111–120. https://doi.org/10.1145/2854946.2854978

[15] 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*, Gary Marchionini, Michael L. Nelson, and Catherine C. Marshall (Eds.). ACM, 286–295. https://doi.org/10.1145/1141753.1141818

[16] Diane Kelly. 2009. Methods for Evaluating Interactive Information Retrieval Systems with Users. *Found. Trends Inf. Retr.* 3, 1-2 (2009), 1–224. https://doi.org/10.1561/1500000012

[17] 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. *Inf. Retr.* 11, 3 (2008), 209–228. https://doi.org/10.1007/s10791-007-9043-7

[18] 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*, Wei Ding, Takashi Washio, Hui Xiong, George Karypis, Bhavani M. Thuraisingham, Diane J. Cook, and Xindong Wu (Eds.). IEEE Computer Society, 391–398. https://doi.org/10.1109/ICDMW.2013.143

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

[20] 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*, Raffaele Perego, Fabrizio Sebastiani, Javed A. Aslam, Ian Ruthven, and Justin Zobel (Eds.). ACM, 1141–1144. https://doi.org/10.1145/2911451.2911469

[21] 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, Santiago, Chile, August 9-13, 2015*, Ricardo Baeza-Yates, Mounia Lalmas, Alistair Moffat, and Berthier A. Ribeiro-Neto (Eds.). ACM, 903–906. https://doi.org/10.1145/2766462.2767802

[22] Philipp Mayr. 2016. Sowiport user search sessions data set (SUSS) (Version: 1.0.0). https://doi.org/10.7802/1380

[23] Mark Sanderson. 2010. Test Collection Based Evaluation of Information Retrieval Systems. *Found. Trends Inf. Retr.* 4, 4 (2010), 247–375. https://doi.org/10.1561/1500000009

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

[25] Marc Sloan, Hui Yang, and Jun Wang. 2015. A term-based methodology for query reformulation understanding. *Inf. Retr. J.* 18, 2 (2015), 145–165. https://doi.org/10.1007/s10791-015-9251-5

[26] Mark D Smucker. 2011. An analysis of user strategies for examining and processing ranked lists of documents. *Proc. of 5th HCIR* (2011).

[27] Matthijs T. J. Spaan. 2012. Partially Observable Markov Decision Processes. In *Reinforcement Learning*, Marco A. Wiering and Martijn van Otterlo (Eds.). Adaptation, Learning, and Optimization, Vol. 12. Springer, 387–414. https://doi.org/10.1007/978-3-642-27645-3_12

[28] Paul Thomas, Alistair Moffat, Peter Bailey, and Falk Scholer. 2014. Modeling decision points in user search behavior. In *Fifth Information Interaction in Context Symposium, IliX '14, Regensburg, Germany, August 26-29, 2014*, David Elsweiler, Bernd Ludwig, Leif Azzopardi, and Max L. Wilson (Eds.). ACM, 239–242. https://doi.org/10.1145/2637002.2637032

[29] 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)*, Allan Hanbury, Gabriella Kazai, Andreas Rauber, and Norbert Fuhr (Eds.). 678–690. https://doi.org/10.1007/978-3-319-16354-3_75

[30] Hui Yang, Dongyi Guan, and Sicong Zhang. 2015. The Query Change Model: Modeling Session Search as a Markov Decision Process. *ACM Trans. Inf. Syst.* 33, 4 (2015), 20:1–20:33. https://doi.org/10.1145/2747874

[31] Saber Zerhoudi, Michael Granitzer, Jörg Schlötterer, and Christin Seifert. 2021. Query Change as a Contextual Markov Model for Simulating User Search Behaviour. In *FIRE 2021: Forum for Information Retrieval Evaluation, Virtual Event, India, December 13 - 17, 2021*, Debasis Ganguly, Surupendu Gangopadhyay, Mandar Mitra, and Prasenjit Majumder (Eds.). ACM, 43–51. https://doi.org/10.1145/3503162.3503165

[32] Saber Zerhoudi, Michael Granitzer, Jörg Schlötterer, and Christin Seifert. 2022. Simulating user interaction and search behaviour in digital libraries. In *Proceedings of the 18th Italian Research Conference on Digital Libraries*.

[33] 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)*, Matthias Hagen, Suzan Verberne, Craig Macdonald, Christin Seifert, Krisztian Balog, Kjetil Nørvåg, and Vinay Setty (Eds.). Springer, 240–247. https://doi.org/10.1007/978-3-030-99739-7_28

[34] 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*, Jaap Kamps, Evangelos Kanoulas, Maarten de Rijke, Hui Fang, and Emine Yilmaz (Eds.). ACM, 193–200. https://doi.org/10.1145/3121050.3121070