Report on the SIGIR 2013 Workshop on Modeling User Behavior for Information Retrieval Evaluation (MUBE 2013)

Charles L. A. Clarke[1] Luanne Freund[2] Mark D. Smucker[3] Emine Yilmaz[4]

[1] School of Computer Science, University of Waterloo, Canada
[2] School of Library, Archival and Information Studies, University of British Columbia, Canada
[3] Department of Management Sciences, University of Waterloo, Canada
[4] Department of Computer Science, University College London, UK

Abstract

The SIGIR 2013 Workshop on Modeling User Behavior of Information Retrieval Evaluation brought together researchers interested in improving Cranfield-style evaluation of information retrieval through the modeling of user behavior. The workshop included two invited talks, ten short paper presentations, and breakout groups. Workshop participants brainstormed research questions of interest and formed breakout groups to explore these questions in greater depth. In addition to summarizing the invited talks and presentations, this report details the results of the breakout groups, which provide a set of research directions for the improvement of information retrieval evaluation.

1 Introduction

A goal of Cranfield-style evaluation of information retrieval systems is to produce an accurate estimate of retrieval quality without the expense of laboratory-based user studies or the risk of deploying poor quality retrieval algorithms in a live retrieval system. There is a growing body of research on ways to improve Cranfield-style evaluation through the modeling of user behavior.

While there are many approaches towards the modeling of user behavior for information retrieval evaluation, much of the work related to the workshop can be understood in terms of two main approaches. One approach has been to create effectiveness measures that better model user behavior with a simplified user interface that consists of a ranked list of results and may include summaries that when clicked take the user to the full result. Another approach has been to simulate user behavior over a varied set of user interfaces to investigate the potential performance improvement offered by alternative ways of interacting with the user or to use simulation to investigate theories of user behavior. Selected examples of work from both of these approaches are listed in Table 1, and many of the works could be classified as being of either approach.

In addition to these two main approaches, there is other related work that utilizes the building of user models for evaluation purposes [2, 10, 24, 36]. Likewise, much of the work

Year	Selected Work - Effectiveness Measures	Selected Work - Simulation
1992		Aalbersberg [1]
1997		Dunlop [21]
2001		Chi, Pirolli, Chen, and Pitkow [17]
2002	Järvelin and Kekäläinen [25]	
2003		Ruthven [38]
2004	de Vries, Kazai, and Lalmas [20]	White, Jose, van Rijsbergen, and Ruthven [47]
2006		Smucker and Allan [42] O'Brien, Keane, and Smyth [34]
2007		Pirolli [35] Lin [30]
2008	Moffat and Zobel [33] Robertson [37] Sakai and Robertson [40]	Lin and Smucker [31] Lin and Wilbur [32]
2009	Chapelle, Metlzer, Zhang, and Grinspan [16] Turpin, Scholer, Järvelin, Wu, and Culpepper [46] Yang and Lad [48]	Keskustalo, Järvelin, Sharma, and Nielsen [28] Azzopardi [5]
2010	Zhang, Park, and Moffat [52] Dupret and Piwowarski [23] Yilmaz, Shokouhi, Craswell, and Robertson [49]	Arvola, Keklinen, and Junkkari [4]
2011	Carterette [13] Carterette, Kanoulas, and Yilmaz [14] Dupret [22]	Azzopardi [6] Kanoulas, Carterette, Clough, and Sanderson [27]
2012	Smucker and Clarke [43, 44] Carterette, Kanoulas, and Yilmaz [15]	Baskaya, Keskustalo, and Järvelin [11]
2013	Chuklin, Serdyukov, and de Rijke [18] Sakai and Dou [39]	Azzopardi, Kelly, and Brennan [8]

Table 1: There has been increasing amounts of work that has involved the modeling of user behavior for information retrieval evaluation. This table represents only selected examples of work. One approach has been to incorporate user models into effectiveness measures and another has been to simulate user interaction with retrieval systems.

on relevance feedback can also be understood in terms of the simulation of users, which was discussed by Donna Harman in her keynote at the SIGIR 2010 workshop on the simulation of interaction [7].

This workshop succeeded in bringing together researchers pursuing these two main approaches as well as many other interested researchers.

The morning of the workshop consisted of two invited talks and 6 short paper presentations in addition to a brainstorming session that generated topics of interest for the later breakout groups. Following lunch were the 4 remaining short paper presentations, breakout sessions for a final set of topics, and then presentations from each of the breakout groups.

2 Invited Talks

The workshop opened with an invited talk by Ben Carterette, who focused on the importance of incorporating variability in user behavior into automatic (batch-style) retrieval evaluation. He showed that information retrieval evaluation is currently highly reliant on averages: that

typically an engine is evaluated by (1) testing for a significant difference in *average effectiveness* that is computed by using judgments that may be *averaged over assessors*, and/or (2) taking one or more parameters *estimated as averages* from user data. Dr. Carterette argued that using averages at various stages of the evaluation process this way presents a false certainty. He showed that in reality there is so much potential variability at each point that even long-held conventional wisdom about effectiveness-enhancing techniques must be questioned. Dr. Carterette then described three different ways in which variability of user behavior can be incorporated into automatic (batch-style) retrieval evaluation: (1) using user logs to incorporate variability in user behavior, (2) using many preferences-based assessments to incorporate variability about relevance, and (3) using different classes of tasks to incorporate variability about topics. Dr. Carterette concluded his talk by describing how the TREC 2013 Session Track includes a pilot study simulating users performing tasks with search engines and how the methods described throughout his talk will be used in this pilot study.

The second invited talk of the day by Leif Azzopardi raised a number of issues regarding the simulation of users and argued that users are *assimilated* into our evaluations through the models we create. The talk started with some definitions about what measures, models and simulations are, and how they relate. Given these definitions, Dr. Azzopardi argued that simulation has been a central component in most evaluations. He then reviewed some of the different kinds of simulations used in evaluation of retrieval systems, as well as the advantages, pitfalls and challenges of performing these simulations. Finally, he focused on the goal of using these simulations and argued that we need to look towards developing more explanatory models of user behavior so that we can obtain a better understanding of users and their interaction with information systems. Dr. Azzopardi concluded his talk by raising a few open questions regarding the simulation of users in retrieval evaluation, such as (1) Is it sensible to include models of the user within the measures that we create?, (2) How do we delineate between the user and the system?, and (3) Should we try to decouple the measures and models? The talk by Dr. Azzopardi was followed by a discussion on the open questions raised during the talk.

3 Workshop Papers

Ten short papers were selected for presentation at the workshop, covering a wide range of topics and approaches related to user behavior modeling, simulation and evaluation. Papers were presented in brief talks in three sessions over the course of the day. Below we summarize each paper, highlighting key themes and findings. Papers are in alphabetical order by first author's last name.

Ageev, Lagun, and Agichtein [3], *Towards Task-Based Snippet Evaluation: Preliminary Results and Challenge*. The authors propose a novel approach to snippet evaluation that takes into account the extent to which a searcher's information need, or task, is satisfied. This approach addresses the increased use of rich snippets in Web IR. The proposed Search Success Ratio (SSR) metrics were employed in a crowdsourced experiment in the form of a game that compared outcomes for various snippet generation algorithms. While the results of the study are inconclusive, the authors suggest that task based evaluation of snippets is a promising future research direction.

Azzopardi, Wilkie, and Russell-Rose [9], *Towards Measures and Models of Findability*. This paper outlines work in progress focusing on the development of models of searching and browsing within websites that will be used to define measures of document findability. Future work will take into account measures of retrievability and navigability as well as the user's information need and task, and will be validated through user studies.

Bruza, Zuccon, and Sitbon [12], *Modelling the information seeking user by the decisions they make*. Relevance assessment is one of the central user behaviors in information retrieval. This paper proposes a novel approach to modeling human relevance assessments that is based on quantum probabilities and takes into account the possible influence or incompatibility of one document assessment on another. An experimental design is proposed to test for effects of assessing multiple documents together or in isolation. The overall goal of this work is to model human cognition using quantum theory.

Jiang and He [26], *Simulating User Selections of Query Suggestions*. The authors propose a preliminary solution for the problem of simulating user selections of query suggestions for use in the evaluation of query suggestion algorithms. The proposed method takes into account the utility of the suggested query, the likelihood that the user will select the best query, and the user's persistence and ability as a judge. Future work will focus on testing and refining the proposed model.

Kharitonov, Macdonald, Serdyukov, and Ounis [29], *Incorporating Efficiency in Evaluation*. The central claim of this work is that evaluation metrics should take into account both effectiveness and efficiency. The specific focus is the need to account for users' responses to system delays. The authors present preliminary experiments using data from the Yandex Browser and Toolbar, which show that searcher abandonment increases approximately linearly with increases in system delay times. The final section of the paper suggests how this efficiency effect could be incorporated into the ERR measure [16] or used to modify the time-biased gain measure [43], and proposes an approach to evaluating the metric.

Scholer, Thomas, and Moffat [41], *Observing Users to Validate Models*. This paper raises the question of the accuracy of the implicit models of user behavior that are expressed in existing IR metrics, such as DCG and Prec@k. A laboratory experiment using gaze transitions was conducted to test one of the most prevalent assumptions of user behavior: linear, top-to-bottom reading with both high quality and degraded search results. While linear reading is confirmed as the strongest trend, the analysis reveals considerable variability in reading patterns that are masked in the general model.

Tran and Fuhr [45], *Markov Modeling for User Interaction in Retrieval*. Working from within the framework of the Interactive Probability Ranking Principle, the authors derive Markov models from user interaction data gathered through logging and eye-tracking. They propose that these Markov models, which identify transition probabilities and user effort, can be used in several ways: (1) for evaluation, as a means of estimating the time needed to find relevant documents, (2) for simulation of user interaction by varying the model parameters and observing the effects, and (3) for guiding the user to avoid sub-optimal behaviors.

Younus, Qureshi, O'Riordan, and Pasi [50], *Personalization for Difficult Queries*. This paper investigates whether certain types of queries benefit more from personalization. The authors present a naturalistic study with real users and relevance judgments using a side by side evaluation technique comparing non-personalized results with personalized results based on the searchers' twitter profiles. Results suggest that the personalized system performs well for difficult queries. The authors note that user profiles and preferences have been overlooked in evaluation measures to date, despite the important role they play in shaping

search behavior.

Zengin and Carterette [51], *User Judgements of Document Similarity*. This paper follows in the vein of assessing the validity of commonly used metrics through comparison with actual user behavior. The authors report on a crowdsourcing study that investigated the level of agreement between cosine similarity scores and human assessments of similarity between two documents. The study finds very low levels of correlation between user ratings and cosine similarity, leading the authors to suggest that cosine similarity captures only some aspect of human judgments.

Zhou, Sakai, Lalmas, Dou, and Jose [53], *Evaluating Heterogeneous Information Access*. This position paper makes the claim that evaluating search is more challenging in heterogeneous information environments. Three challenges are identified for evaluation in this context: (1) users may engage in non-linear traversal browsing when encountering aggregated or clustered results, rather than following a linear top-to-bottom pattern; (2) users may be carrying out complex search tasks requiring more interaction and shifting between different types of results; and (3) the coherence of search results may be reduced when they are blended from different sources. The authors suggest that user studies need to be conducted to build up user models and associated evaluation frameworks for these types of search environments.

4 Breakouts

The workshop included a breakout session, allowing small groups to focus closely on a number of specific topics. Just before lunch, we did some brainstorming to identify questions of interest, which were then grouped into six categories. Groups formed around five of these six categories, with one group member volunteering to lead the discussion as "champion" and one group member volunteering to record the discussion as "scribe". Each group spent an hour in the late afternoon discussing their chosen topics.

After the breakouts were over, each group reported back with a few slides and a short talk outlining their discussion. The overview below is based on those slides. The discussions generated a great deal of "raw material" for future research, which we have attempted to preserve in as much detail as possible. We are responsible for all errors in the interpretation of the slides. We also attempted to record the names of all participating groups members, and we apologize if we missed you.

Breakout Topic 1: Validation of user models for simulation

Group members: Mikhail Ageev, Aleksandr Chuklin, Kathy Brennan, Gleb Gusev, Miguel Martinez, Iadh Ounis, Andrei Rikitianskii, Tetsuya Sakai (scribe), and Aleksei Tikhonov.

Questions identified during brainstorming:

- How do we validate evaluation measures?
- How do we establish limits or bounds on the difference between simulated users and real users?
- How do we quantify the accuracy of user models?
- How do we best use people to make better simulations?

- How do we model a system's "usefulness" to its users?
- How do we isolate system components from the whole?

The group refined this initial set of questions into three key questions: 1) What kind of information (besides relevance) do we have to put into the test collections? 2) Can we provide guidance on the number of real users needed to validate simulation results? 3) Can we ensure that component performances add up to end-to-end experience?

Existing test collections typically comprise a topic set, a document collection, and relevance assessments. The first question considers the additional information required to enable a more realistic IR evaluation. For example, if the collection provides document length and snippet length statistics, new metrics such as TBG and U-measure can be computed. Dwell time, eye-tracking statistics, etc., may be useful for even more fine-grained evaluation of user satisfaction.

With respect to the second question, the group noted that simulation is simulation; there will always be a gap between the underlying model and reality. Iterating simulation may mean multiplying the errors that arise due to this gap, and these errors need to be quantified. For this purpose, a small-to-medium scale user study is inevitable. How many users should be involved in the validation experiment? How often should the user-simulation comparison study be conducted? Can the IR community provide a guideline?

With respect to the third question, the group noted that effective components do not necessarily add up to an optimal end-to-end IR system. For example, a highly effective query suggestion system is useless if it is never used by the users when presented as a component of a search engine result page. How do we efficiently and effectively conduct component-based and end-to-end IR system evaluation? Can both be done seamlessly within a single evaluation framework?

Breakout topic 2: Re-usable test collections

Group members: Matthew Ekstrand-Abueg (scribe), Frank Hopfgartner, Mark Smucker (champion), and Emine Yilmaz.

Questions identified during brainstorming:

- How do we construct re-usable test collections for user-model based metrics?
- How do we reduce variance in results?
- Does evaluation of individual system components improve re-usability?

The group noted that a test collection could be used not just to evaluate IR systems, but also to evaluate user models. The group viewed re-usability on a spectrum of difficulty with the representation of individual user characteristics at one end, and a full representation of of every aspect of an interactive retrieval session at the other. The re-usability of document relevance judgments and summary judgments fell between these extremes. Along with queries and relevance judgments, a re-usable collection might include click probabilities, result-dependent reformulations, scanning times, and decision times. In addition, the collection might model characteristics of individual users, which might be based on cognitive tests or include personal data.

Breakout topic 3: Modeling user intent

Group members: Peter Bruza (champion), Luanne Freund (scribe), Yi Chen, Maya Subramanian, Daniel Tunkelang, and Colin Wilkie.

Questions identified during brainstorming:

- How do we determine intent based on queries?
- Can we use games-with-a-purpose to understand and/or generate query intents?
- How do we best use people to make better simulations?
- Can we create models of users that are cognitively situated?

The group identified problems with the use of queries as a proxy for intent, particularly with limits to their realism. The group asked: Is intent measurable? Is intent changeable? Can we identify a change in intent? Can we ensure that the users we are studying are motivated by a genuine or natural intents? Games might reasonably take the place of (or augment) naturalistic studies to simulate work task scenarios. Unfortunately, these approaches cannot be used for all problems. In particular, games require measurable outcomes.

Many intents may be in play at the same time. Moreover, user goals and system goals may not be fully aligned. Tacit and implicit intents are problematic, and it is difficult to define formal models for them. Navigational queries are also problematic because the intent is hidden. Attempts to mine query intent must consider the complexity of intent. For a given common set of queries there are several patterns of behavior. Within each of these patterns, users select different document sets. For example, the query "excel" may mean that the user may want to buy excel, want to use excel, need support for excel, etc.

The group concluded that in an ideal world we would have clearly defined user intents. We would have the control and cleanliness of imposed intents. We would have the authenticity of natural intents. We would understand the dimensions of intent. Games might get us part-way there.

Breakout topic 4: Interactive measures

Group members: Matthew Crane, Djoerd Hiemstra, Gabriella Kazai, Haiming Liu, and Adam Ke Zhou.

Questions identified during brainstorming:

- Should interactive measures consider the user interface?
- How does the user interface affect users?
- Should we model user learning?
- How do we handle interface differences?

A major concern of the group was modeling and measuring the effect of the user interface, which immediately raised a enormous number of additional questions. Models must consider novices vs. experts, cognition, attention spam, context, and risk tolerance. Once we model these factors, how do we validate and instantiate the models? Is one user more important than another? How much do we need to know about users? Where do we get user signals

from? Can we use social media? Do we need very detailed user behavior models or do success type metrics already estimate most of these?

Models must consider modes of interaction, such as exploratory, entertainment, long term task, and answer finding. Do we need a fixed taxonomy where each class needs to be evaluated? Would each of these need a dedicated test collection? We may need to consider properties and constraints of different devices. What is the best and most scalable way to evaluate (e.g., feedback, A/B testing, side-by-side, crowdsourcing)?

Even if we know what to measure, how do we measure? Do we want unified models and measures, or component-based models that separately model features such as snippets? Intent is important. Spending more time is good if need is entertainment, but not if the user is looking for an answer. What is the benefit to the user? To assist the user? To satisfy the user's intent? To keep them engaged on the site longer? Do we measure user satisfaction or success in adapting to different user types? What about measures of user engagement? Can we measure user's learning through the search process, particularly for informational needs? Does it matter how quickly they learn, or how much, or how they benefit from what was learned and what their recall is later?

We also must consider the attractiveness of result presentations. What is the trade-off between happiness and accuracy? Happy users are not necessarily getting the right answers.

What are the properties of rigorous evaluation? More samples, different designs? What is the ultimate goal of a user interface? What signals can we use to derive its value to the user? How do we aggregate these signals? Does machine learning help? How can we get more implicit signals of user behavior? What if interaction patterns are the same for two systems but we know their quality is different?

Breakout topic 5: Whole session evaluation

Group members: Leif Azzopardi (scribe), Pablo Castells, Charles Clarke, Henry Feild, and Craig Macdonald.

Questions identified during brainstorming:

- What are the limits to modeling/simulating whole sessions?
- Can we construct joint economic models for users and search engines?
- Can we use social benefit as the measure of gain?
- Is it meaningful to model "usefulness"?

As it did for other groups, the discussion raised many additional questions. The group generally agreed that we must start with an intent model to meaningfully simulate a whole session, but how do we model intents and estimate parameters for the simulation? How do we calculate gain, efficiency and effectiveness across entire sessions? How do we compute user costs and gains? How about system costs and gains? How do we combine these costs and gains?

Breakout topic 6: User variability

This last category was dropped from the breakout sessions. While there was interest in this topic, it was not the first choice of sufficient participants to make a viable group.

Questions identified during brainstorming:

- How do we reduce variance in results?
- How do we sample users to minimize variance?
- How do we sample queries to minimize variance?
- Do we educate users to make them better searchers?

5 Workshop Summary

In proposing the workshop, we identified two main approaches to the modeling of user behavior for information retrieval evaluation. One approach focuses on the incorporation of user models into effectiveness measures. Another approach focuses on the simulation of user interaction with varied interfaces to retrieval systems. Roughly representing the two approaches, the invited talks by Ben Carterette and Leif Azzopardi both showed the overlap of the approaches and provided significant insight from two active researchers in this area. The short papers represent current research in this area and give examples of the latest approaches being taken to improve evaluation via the modeling of user behavior. The workshop attendees identified the following key research areas / questions for the future of Cranfield-style evaluation:

- Validation: How should we validate user models and effectiveness measures?
- Re-Usable Test Collections: In what ways can we create re-usable test collections that support evaluation methods based on modeled user behavior?
- User Intents: What can we do to better model the many user intents represented today by search topics?
- Interaction and Interfaces: How can we best measure the effect of user interfaces on retrieval quality?
- Whole Sessions: How do we estimate effectiveness for whole sessions?
- User Variance: What interpretation should we make of the variability in performance caused by more realistic modeling of users?

In summary, the workshop succeeded in bringing together researchers following varied approaches to the modeling of user behavior for information retrieval evaluation and set forth several research areas for future work.

6 Acknowledgments

We thank the ACM and SIGIR for their support and assistance with making this workshop possible. In particular, we thank the workshop chairs Arjen de Vries and Vincent Wade and the Local Arrangements Chair, Séamus Lawless. We thank our excellent program committee: Eugene Agichtein, James Allan, Javed Aslam, Leif Azzopardi, Nicholas Belkin, Pia Borlund, Ben Carterette, Arjen de Vries, Norbert Fuhr, Donna Harman, Hideo Joho, Joemon Jose, Jaap Kamps, Evangelos Kanoulas, Mounia Lalmas, Alistair Moffat, Virgil Pavlu, Stephen Robertson, Ian Ruthven, Falk Scholer, Pavel Serdyukov, and Ellen Voorhees. We

are grateful to our invited speakers, Ben Carterette and Leif Azzopardi, and to all of the authors for providing thought-provoking presentations. We thank the many attendees for their participation and especially thank those who took an active role with the breakouts for charting future directions for research. Further details, including the workshop proceedings, are available at `http://www.mansci.uwaterloo.ca/~msmucker/mube2013/`.

This work was supported in part by the Natural Sciences and Engineering Research Council of Canada (NSERC), in part by GRAND NCE, and in part by the University of Waterloo.

References

[1] I. J. Aalbersberg. Incremental relevance feedback. In *SIGIR*, pages 11–22, 1992.

[2] M. Ageev, Q. Guo, D. Lagun, and E. Agichtein. Find it if you can: a game for modeling different types of web search success using interaction data. In *SIGIR*, pages 345–354, 2011.

[3] M. Ageev, D. Lagun, and E. Agichtein. Towards task-based snippet evaluation:preliminary results and challenges. In Clarke et al. [19], pages 1–2.

[4] P. Arvola, J. Keklinen, and M. Junkkari. Expected reading effort in focused retrieval evaluation. *Information Retrieval*, 13:460–484, 2010.

[5] L. Azzopardi. Usage based effectiveness measures: monitoring application performance in information retrieval. In *CIKM*, pages 631–640, 2009.

[6] L. Azzopardi. The economics in interactive information retrieval. In *SIGIR*, pages 15–24, 2011.

[7] L. Azzopardi, K. Järvelin, J. Kamps, and M. D. Smucker. Report on the SIGIR 2010 workshop on the simulation of interaction. *SIGIR Forum*, 44:35–47, January 2011.

[8] L. Azzopardi, D. Kelly, and K. Brennan. How query cost affects search behavior. In *SIGIR*, pages 23–32, 2013.

[9] L. Azzopardi, C. Wilkie, and T. Russell-Rose. Towards measures and models of findability. In Clarke et al. [19], pages 3–4.

[10] R. Baeza-Yates, C. Hurtado, M. Mendoza, and G. Dupret. Modeling user search behavior. In *Proceedings of the Third Latin American Web Conference*, pages 242–251. IEEE, 2005.

[11] F. Baskaya, H. Keskustalo, and K. Järvelin. Time drives interaction: simulating sessions in diverse searching environments. In *SIGIR*, pages 105–114, 2012.

[12] P. Bruza, G. Zuccon, and L. Sitbon. Modelling the information seeking user by the decisions they make. In Clarke et al. [19], pages 5–6.

[13] B. Carterette. System effectiveness, user models, and user utility: A conceptual framework for investigation. In *SIGIR*, pages 903–912, 2011.

[14] B. Carterette, E. Kanoulas, and E. Yilmaz. Simulating simple user behavior for system effectiveness evaluation. In *CIKM*, pages 611–620, 2011.

[15] B. Carterette, E. Kanoulas, and E. Yilmaz. Incorporating variability in user behavior into systems based evaluation. In *CIKM*, pages 135–144, 2012.

[16] O. Chapelle, D. Metlzer, Y. Zhang, and P. Grinspan. Expected reciprocal rank for graded relevance. In *CIKM*, pages 621–630, 2009.

[17] E. H. Chi, P. Pirolli, K. Chen, and J. Pitkow. Using information scent to model user information needs and actions and the web. In *SIGCHI*, pages 490–497, 2001.

[18] A. Chuklin, P. Serdyukov, and M. de Rijke. Click model-based information retrieval metrics. In *SIGIR*, pages 493–502, 2013.

[19] C. L. A. Clarke, L. Freund, M. D. Smucker, and E. Yilmaz, editors. *Proceedings of the SIGIR 2013 Workshop on Modeling User Behavior for Information Retrieval Evaluation (MUBE 2013)*, 2013.

[20] A. P. de Vries, G. Kazai, and M. Lalmas. Tolerance to irrelevance: A user-effort oriented evaluation of retrieval systems without predefined retrieval unit. In *RIAO*, pages 463–473, 2004.

[21] M. D. Dunlop. Time, relevance and interaction modelling for information retrieval. In *SIGIR*, pages 206–213, 1997.

[22] G. Dupret. Discounted cumulative gain and user decision models. In *SPIRE*, pages 2–13, 2011.

[23] G. Dupret and B. Piwowarski. A user behavior model for average precision and its generalization to graded judgments. In *SIGIR*, pages 531–538, 2010.

[24] A. Hassan, R. Jones, and K. L. Klinkner. Beyond DCG: user behavior as a predictor of a successful search. In *WSDM*, pages 221–230, 2010.

[25] K. Järvelin and J. Kekäläinen. Cumulated gain-based evaluation of IR techniques. *TOIS*, 20(4):422–446, 2002.

[26] J. Jiang and D. He. Simulating user selections of query suggestions. In Clarke et al. [19], pages 7–8.

[27] E. Kanoulas, B. Carterette, P. D. Clough, and M. Sanderson. Evaluating multi-query sessions. In *SIGIR*, pages 1053–1062, 2011.

[28] H. Keskustalo, K. Järvelin, T. Sharma, and M. L. Nielsen. Test collection-based IR evaluation needs extension toward sessions: A case of extremely short queries. In *AIRS*, pages 63–74, 2009.

[29] E. Kharitonov, C. Macdonald, P. Serdyukov, and I. Ounis. Incorporating efficiency in evaluation. In Clarke et al. [19], pages 9–10.

[30] J. Lin. User simulations for evaluating answers to question series. *IPM*, 43(3):717–729, 2007.

[31] J. Lin and M. D. Smucker. How do users find things with PubMed? Towards automatic utility evaluation with user simulations. In *SIGIR*, pages 19–26, 2008.

[32] J. Lin and W. J. Wilbur. Modeling actions of PubMed users with n-gram language models. *Information Retrieval*, 2008.

[33] A. Moffat and J. Zobel. Rank-biased precision for measurement of retrieval effectiveness. *TOIS*, 27(1):1–27, 2008.

[34] M. O'Brien, M. T. Keane, and B. Smyth. Predictive modeling of first-click behavior in web-search. In *WWW*, pages 1031–1032, 2006.

[35] P. Pirolli. *Information Foraging Theory*. Oxford University Press, 2007.

[36] K. Punera and S. Merugu. The anatomy of a click: modeling user behavior on web information systems. In *CIKM*, pages 989–998, 2010.

[37] S. Robertson. A new interpretation of average precision. In *SIGIR*, pages 689–690, 2008.

[38] I. Ruthven. Re-examining the potential effectiveness of interactive query expansion. In *SIGIR*, pages 213–220, 2003.

[39] T. Sakai and Z. Dou. Summaries, ranked retrieval and sessions: a unified framework for information access evaluation. In *SIGIR*, pages 473–482, 2013.

[40] T. Sakai and S. Robertson. Modelling a user population for designing information retrieval metrics. In *The Second International Workshop on Evaluating Information Access (EVIA)*, 2008.

[41] F. Scholer, P. Thomas, and A. Moffat. Observing users to validate models. In Clarke et al. [19], pages 11–12.

[42] M. D. Smucker and J. Allan. Find-similar: Similarity browsing as a search tool. In *SIGIR*, pages 461–468, 2006.

[43] M. D. Smucker and C. L. A. Clarke. Time-based calibration of effectiveness measures. In *SIGIR*, 2012.

[44] M. D. Smucker and C. L. A. Clarke. Modeling user variance in time-biased gain. In *HCIR*, pages 3:1–3:10, 2012.

[45] V. T. Tran and N. Fuhr. Markov modeling for user interaction in retrieval. In Clarke et al. [19], pages 13–14.

[46] A. Turpin, F. Scholer, K. Järvelin, M. Wu, and J. S. Culpepper. Including summaries in system evaluation. In *SIGIR*, pages 508–515, 2009.

[47] R. W. White, J. M. Jose, C. J. van Rijsbergen, and I. Ruthven. A simulated study of implicit feedback models. In *ECIR*, 2004.

[48] Y. Yang and A. Lad. Modeling expected utility of multi-session information distillation. In L. Azzopardi, G. Kazai, S. Robertson, S. Rger, M. Shokouhi, D. Song, and E. Yilmaz, editors, *Advances in Information Retrieval Theory*, volume 5766 of *LNCS*, pages 164–175. Springer Berlin / Heidelberg, 2009.

[49] E. Yilmaz, M. Shokouhi, N. Craswell, and S. Robertson. Expected browsing utility for web search evaluation. In *CIKM*, pages 1561–1564, 2010.

[50] A. Younus, M. A. Qureshi, C. O'Riordan, and G. Pasi. Personalization for difficult queries. In Clarke et al. [19], pages 15–16.

[51] M. Zengin and B. Carterette. User judgements of document similarity. In Clarke et al. [19], pages 17–18.

[52] Y. Zhang, L. A. Park, and A. Moffat. Click-based evidence for decaying weight distributions in search effectiveness metrics. *Information Retrieval*, 13:46–69, Feb 2010.

[53] K. Zhou, T. Sakai, M. Lalmas, Z. Dou, and J. M. Jose. Evaluating heterogeneous information access (position paper). In Clarke et al. [19], pages 19–20.