Report on the SIGIR 2010 Workshop on the Simulation of Interaction

Leif Azzopardi[1] **Kalervo Järvelin**[2] **Jaap Kamps**[3] **Mark D. Smucker**[4]

[1] University of Glasgow, Scotland
[2] University of Tampere, Finland
[3] University of Amsterdam, The Netherlands
[4] University of Waterloo, Canada

Abstract

All search in the real-world is inherently interactive. Information retrieval (IR) has a firm tradition of using simulation to evaluate IR systems as embodied by the Cranfield paradigm. However, to a large extent, such system evaluations ignore user interaction. Simulations provide a way to go beyond this limitation. With an increasing number of researchers using simulation to evaluate interactive IR systems, it is now timely to discuss, develop and advance this powerful methodology within the field of IR. During the SimInt 2010 workshop around 40 participants discussed and presented their views on the simulation of interaction. The main conclusion and general consensus was that simulation offers great potential for the field of IR; and that simulations of user interaction can make explicit the user and the user interface while maintaining the advantages of the Cranfield paradigm.

1 Introduction

The use of simulation to evaluate retrieval systems has a long history in IR, especially before the availability of large-scale test collections developed in the 1990s [6, 20]. In recent years, simulation has mainly been used to overcome limitations of traditional test collections, in particular to evaluate adaptive or interactive IR [e.g., 22]. The different types of experiments that can be performed to examine Interactive IR may be classified into four classes [14, p.210]:

1. observing users in real situations (real users; no simulation);
2. observing users performing simulated tasks;
3. performing simulations in the lab without users (simulation of interaction; no users); and
4. traditional lab research (no users and no simulation).

Interactive IR may be studied experimentally with real searchers performing real or simulated work tasks (class 1 and 2 respectively). However, these types of experiments require a lot effort to setup in order to conduct reliable studies, and they tend to be tedious and

costly to run. If several rounds of experimentation are needed, new sets of test subjects are required due to fatigue and learning effects. While studies with real users performing real or simulated IR tasks provide rich data, they are often very limited, covering only a few test cases. On the other extreme, laboratory studies (class 4) can be seen as limited-perspective abstractions of user search. While, they are reusable facilitating comparison and re-use, they fail to model many aspects of users and interaction properly or adequately.

Therefore, it makes sense to explore various IIR problems through user simulation (class 3). Laboratory-based simulation of interaction provides a rapid means of exploring the potential and limits of real interaction, at a low cost, without wasting the valuable time of real searchers. For example, in the case of relevance feedback (RFB), one may find out what kind of user RFB effort or RFB method is most effective (or if it is even effective) before deploying it. Of course whether these results translate to real world situations depends on the realism of the simulation. Hence, the assumptions underlying the simulation need to be explicitly modelled, and based on or tested against user experiments. This will in itself contribute to a better understanding of Interactive IR. For example, one could (1) efficiently test the resulting findings in a "real" user study, with greater depth and focus, (2) obtain clearer insights into what to test specifically, and (3) determine whether new methods are potentially worthwhile, and worth testing on real users. However, while the advantages and benefits of the simulation of interaction are many, there are also many challenges and potential pitfalls that need to be addressed.

2 Workshop

2.1 Workshop Objectives

The main objective of this workshop was to kick off discussions on the simulation of interaction. With the increase in researchers adopting a simulation based methodology for IR evaluation, the workshop aimed to provide a forum where:

1. pressing issues about the implications of simulation could be discussed,
2. a common ground between researchers could be established, and
3. new developments and techniques could be shared.

In addition, we aimed to promote the development of resources for the simulation of interaction.

2.2 Many Open Questions

Before the workshop we gathered a broad range of questions on the simulation of interaction, with the aim to bring together both proponents of simulation and those critical of this approach. The topics discussed included:

- **What is Simulation of Interaction?**
  - Definitions of simulation of interaction
  - Methods and methodologies for simulation of interaction
  - The ideal simulation (akin to the ideal test collection)
- **Where can we apply it?**

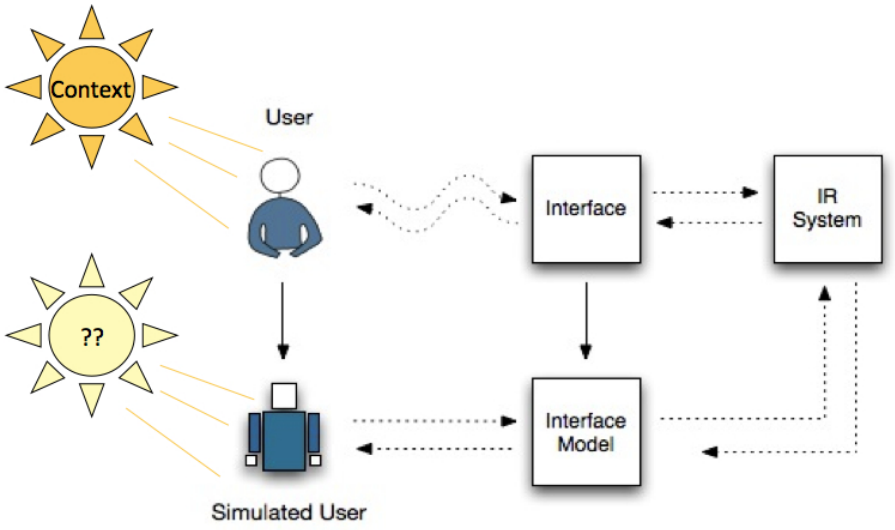

Figure 1: The simulation of Interaction: one approach could be through the explicit definition of a simulated user, their actions and behaviors along with a model of the interface.

  - Types of experiments that could be performed
  - Control and limitations of the different approaches
  - Why and why not, cases for and against such evaluation
- **What and how to model?**
  - Simulating queries, judgments, clicks, etc.
  - User modeling and estimating models of interaction
  - Requirements for future development
- **What are the limits of simulation?**
  - Where is the user or her task in a simulation?
  - What is interaction in a simulation?
  - Realistic and unrealistic assumptions, barriers to success
- **Where are we, where do we want to get to, and how do we get there?**
  - Research agenda with a road map of future challenges

2.3 Overview the Workshop

To set the scene for the workshop and provide some background we invited two notable researchers who have performed numerous simulations to provide their personal reflections on the subject. The first keynote of the workshop was given by Donna Harman, and the second was by Ryen White. Donna started the day with her presentation entitled, "User Study → What-IF Experiment → User Study". While, Ryen followed, with his talk on "Simulations, Logs and their Synergies".

After a break, Leif Azzopardi provided an overview of simulation within information retrieval, presenting the case for simulation and the possibilities that it afforded for interactive

IR research (see Figure 1). He also stressed the limitations involved, and the necessity to collaborate with those doing traditional observational studies of real users.

Before lunch, the authors of workshop posters were invited to boast about their work during a 'minute madness' session. Then over lunch the 17 posters were discussed with participants (see Section 4 for an overview of the different posters).

After lunch, the workshop broke out into four groups, where we discussed the following topics:

(a) Making Simulations Work, led by Leif Azzopardi,

(b) Generating and Modeling Queries and Interaction, led by Charlie Clark,

(c) Creating Simulations with Search Logs, led by Jaap Kamps, and,

(d) Simulated Browsing and User Interfaces, led by Mark Smucker.

Rounding up the day, Daniel Tunkelang chaired the final session where the leader from each breakout group reported their findings.

3 Keynotes Addresses

We invited the keynote speakers to report on their experience with simulations, to help frame the main research questions, and to set the stage for the rest of the day.

3.1 User Simulations as What-IF Studies

Donna Harman (NIST) started the day with her presentation entitled, "User Study → What-IF Experiment → User Study." She discussed some of the early work performed at NIST, before TREC started, on relevance feedback. This work used the existing small scale test collections to simulate the effectiveness of relevance feedback across a wide range of parameters [8]. E.g., term weighting versus term expansion, number of expansion terms and the term selection method, and the effectiveness of multiple iterations of relevance feedback. The simulations yielded relatively clear conclusions (called "user test recommendations"). The original plan was to then validate these in a user experiment – however, this was never conducted. The importance of such a validation experiment was demonstrated later when in other user experiments it was found that users were reluctant to provide relevance feedback to the system, i.e. the users acted differently to the simulations, even through relevance feedback could improve system performance.

The general message from the talk was to view user simulations as "What-IF" experiments, which would enable a wide range of configurations to be tested, without subjecting users to all the different possibilities. The process is as follows: first, start with some questions based on user studies or observations. Investigate those questions using user simulations based on, for example, an existing test collection. Then, validate the simulation results by conducting a focused user study. This last step is important, and should be done to complete the cycle; either performed in conjunction with the simulation or as a subsequent follow up study. If the relevance feedback findings had been subjected to user testing, they would have failed despite the potential! That said, user experiments have to be very carefully design and carried out in order to not reject "progress" falsely.

3.2 Building Shareable Simulations from Log Data

Ryen White (Microsoft Research) talked about "Simulations, Logs and their Synergies." Ryen reflected upon his own work both in academia, using simulations of interaction to study interactive IR, and in industry, mainly on characterizing search behaviors via log analysis. The work on simulations of interaction, for example to evaluate implicit feedback models [22] or to re-design search interfaces, clearly demonstrated the benefits of studying interaction in a highly controlled experimental setup.

The log-based studies allow for large-scale analysis with greater diversity in vocabularies, sites, tasks than lab-based user studies. The log analysis focuses on characterizing web search behavior, studying especially user variability, such as different interaction patterns in search sessions and browse trials, and trying to link this behavior to for example search tactics (navigators versus explorers) or domain expertise. The resulting characterizations can be used as building blocks for searcher simulation models. However, search logs keep IR researchers in a double-bind. On the one hand, rich log data provides an invaluable resource for studying searcher interactions. On the other hand, search engines respect user privacy and are unable to share these logs to the wider research community. Simulation may offer a way out: searcher simulations based on these logs could be shared, and simulation models of sufficient realism can be immensely valuable for research.

4 Workshop Papers

The workshop saw a range of submissions which broadly fell into four main themes: challenges and directions; seed data and queries for simulations, using log data for simulations, the simulation of browsing and interfaces. A brief summary of each paper is provided below.

4.1 Making Simulations Work

A selection of the accepted papers discussed aspects of the overall challenges and directions of simulations: Cole [5], Simulation of the IIR user: beyond the automagic. This paper outlines a number of issues when trying to develop realistic simulations. The case for the arguments presented is based upon Boring's five step operationalist approach for creating an objective model of the mind's functions. A key point made in this position paper is that for simulations to be realistic, they need to be seeded with real data.

Kato et al. [12], Bridging evaluations: inspiration from dialog system research. This paper considers the evaluation of exploratory information access environments, and proposes the use of a semi-automated approach to perform such evaluations. The approach is based upon the PARADISE framework used in dialogue systems in conjunction with a simulated user. The paper makes an important point (which is also raised by [5]), that to obtain an accurate and realistic simulation, data based on real interactions should be used to seed the simulations.

Mulwa et al. [16], A proposal for the evaluation of simulated interactive information retrieval in customer support. This paper proposes a methodology for evaluating adaptive information retrieval systems that is underpinned by the simulation of interaction and personalization. An overview of some of the different metrics taken from IR and Adaptive Hypermedia are also presented, followed by some of the challenges associated with employing such a simulation.

Tunkelang [21], Modeling communication between users and information retrieval systems methodologies. This paper suggests considering the interaction between users and systems as a form of communication. The main suggestion is that a standard test collection can be used in conjunction with different types of simulated users to investigate various interactions/tasks.

4.2 Generating and Modeling Queries and Interaction

Several papers presented methods and models for seeding simulations and generating queries/judgements:

Geva and Chappell [7], Focused relevance feedback evaluation. This paper suggests how focused feedback can be incorporated within the IR process, in the context of INEX where relevant passages of text can be used to simulate such feedback. This paper also argues for the development of a reusable evaluation platform.

Jethani and Smucker [10], Modeling the time to judge document relevance. This paper attempts to model the time it takes for a user to judge the relevance of document with a linear function where the total judgement time is proportional to the length of the document plus an overhead. While, this simple model explains 26%-45% of the variance given real user judgement times, it provides the first steps for modeling this aspect of the interaction.

Kanoulas et al. [11], Session track at TREC 2010. This paper presents a manual method for the construction of query reformulations (as performed at the Session Track at TREC 2010). The aim of the session track was to examine the difference between the original query and the reformulated query, when the system either considered the original query when responding to the reformulated query, or not. Three types of query reformulations are created as part of the query set: specifications, parrallel reformulations and generalizations.

Keskustalo et al. [15], Graph-based query session exploration based on facet analysis. This paper proposes that concept graphs and expression graphs be created for test topics. This allows systematic formulation of queries and sessions as traversals of the graphs: each query is one vertex of an expression graph of a topic; query reformulations manifest as edges in the graph; and sessions as paths in the graph.

Zhang et al. [23], A probabilistic automaton for the dynamic relevance judgement of users. This paper aims to extend standard TREC based test collections by generating simulated user judgements of relevance given the user's interactions with previously examined documents, i.e. seeing one relevant document may influence the relevance of a subsequent document. The proposed method uses transition probabilities from one judgement to the next.

4.3 Creating Simulation using Search Logs

Three papers discussed the use of log data for simulations:

Clough et al. [4], Creating re-useable log files for interactive CLIR. A methodology for generating log files is described and used as part of image CLEF. These logs are of multilingual searches for known-items conducted using the Flickling system (which searchers Flickr.com) by 435 participants, consisting of 6,182 searches. The logs are available for re-use at: `http://nlp.uned.es/iCLEF/`

Huurnink et al. [9], Simulating searches from transaction logs. This paper discusses the potential of transaction log data for developing and validating simulation-based IR experiments. There are two scenarios: evaluation testbeds consisting of artificial queries and clicks;

and simulations of session behavior in terms of sequences of queries and clicks.

Nanas et al. [17], A methodology for simulated experiments in interactive search. This paper puts forward a methodology for evaluating query recommendations / term suggestions within the web search context based on using query log data.

4.4 Browsing and User Interfaces

Another selection of accepted papers focused on the simulation of browsing behavior and user interfaces:

Alonso and Pedersen [1], Recovering temporal context for relevance assessments. In this paper, the authors propose a method to simulate the interaction of a user and system over time for a temporal query (i.e. the same query submitted over and over again during a given period, for example, "Germany World Cup"). This is so they can collect relevance judgements given the results presented as they change over time. This is not exactly what we had in mind, in terms of the simulation of interaction, but it does present an interesting way method for obtaining judgements for temporal queries.

Arvola and Kekäläinen [2], Simulating user interaction in result document browsing. This paper proposes a simulation approach to within-document browsing. The result of a simulation depends on whether or not the user found relevant information in the document, whether this was enough, and how much effort (s)he expended in exploring the document.

Keskustalo and Järvelin [13], Query and browsing-based interaction simulation in test collections. This paper focuses on sessions based on very short queries where the (simulated) searchers are assumed to probe the collection with short queries and limited browsing until the search goal is reached or the searcher gives up. Keywords for the queries may have a searcher warrant (collected from real users) or literary warrant (collected from relevant documents). Session strategies may have searcher warrant, i.e. based on real user behavior.

Preminger [18], Evaluating a visualization approach by user simulation. This paper presents the evaluation methodology of a 3-dimensional visualizations of the document space.

Stober and Nuernberger [19], Automatic evaluation of user adaptive interfaces for information organization and exploration. This paper outlines a method for evaluating different visualizations/structuring of documents (like clustering methods) by simulating different kinds of structuring behavior of users. That is, simulating how users would go about organizing a collection of documents to reflect their personal interests.

4.5 Annotated Bibliography

There have been a number of works in that past that have employed or developed simulation techniques within IR. Generally, research has focused on five main areas, the simulation of Test Collections, Implicit & Relevance Feedback, Browsing and Foraging, Querying and Interfaces. A list of references can be found on the SimInt website at `http://www.dcs.gla.ac.uk/access/simint/` and SimInt Google group at `http://groups.google.com/group/simint`.

5 Breakout Sessions

5.1 Making Simulations Work

In this breakout group the aim was to discuss higher level issues regarding the use of simulation in interactive IR. One of the main issues discussed was how simulations are defined, and should be defined, in order to build a common ground among researchers. For example, in other fields that employ simulations there are known core parameters (for instance, when modeling a chemical reaction or the aerodynamics of a plane, the variables and laws that govern these are well defined and often fully known a priori). To simulate interaction there was a consensus that we need to explicitly define the simulation parameters, variables and functions within an experiment. The main components within a simulation (at a high level) are shown in Figure 1 , and these include a model of the user (or simulated user) and a model of the interface. More precisely, the simulated user encapsulates the user and their behavior i.e. the actions and responses to events, how they assess documents for relevance, their criteria for success or failure for the given task (or an evaluation model), etc. While the interface model provides an abstraction of the interface itself, and exposes the main functionality of the system to the simulated user, who can then act/react accordingly. Working towards providing a common framework for the simulation of interaction would greatly improve and accelerate the use of this methodology within IR.

Another important issue that was discussed was how simulations are instantiated. It was felt that it was very important that we draw upon the literature (in particularly, and especially) from information seeking, to help seed and motivate simulations. Once a set of parameters and variables have been constructed to describe users and their tasks, then it is important to seed simulations with realistic values based on previous findings. And similarly with the interactions of users and systems. One major thing to point out though is that not all simulations need to be grounded or based on reality – it depends on the research questions and aims of the experiments. For instance, to run what-if experiments may require hypothesizing about different courses of interaction and different user models. Or in "why" experiments that aim to explain certain interactions or behaviors, then strategies which users do not perform may need to be evaluated. However, if the goal of the simulation is to provide supporting evidence for how users behave in real life then seeding the models with data from user studies will enable better predictions to be made. Another benefit of working towards a common framework for simulations is that it provides a way for user focused researchers to contribute and seed simulations. While lab-based researchers can move towards working on interactive IR capitalizing on the wealth of knowledge provided by user studies. Formalizing and creating realistic user models and interaction strategies is a definite way to crystallize our understanding of interaction within IR.

To summarize the main points from the breakout group, we felt that simulations had an important role to play in the evaluation of IR. For the road ahead, defining a common framework for developing simulations would be very worthwhile. This would provide a way to connect disparate areas of research, and provide the basis for more simulations to be conducted in a principled, coordinated and strategic fashion. However, it may also lead to a lot of simulations which are not well considered. So before we get carried away running millions of simulations, it is important for researchers to understand the implications of their simulations and understand the results of such simulations, and use them appropriately. Thus, for simulations to be a powerful evaluation tool, care needs to be taken when designing simula-

tions, i.e. the main components need to explicitly described, and when running simulations they need to be appropriately seeded for the given experiment. The four main stages of experimentation are: (1) initial observations, (2) hypothesizes, (3) simulations, and then (4) confirmation/validation, where appropriate. The sentiment behind this was summed up by one participant, who stated that a good theory is one that is testable, and at some point it should be tested.

5.2 Generating and Modeling Queries and Interaction

The second breakout group, chaired by Charlie Clarke, discussed issues regarding Querying and Interaction. There was substantial discussion on what to simulate when performing an experiment. Some promising candidates included a range of different possibilities, such as (1) modeling short queries with lots of interaction, (2) modeling how interaction time varies for given tasks, and depending on the context, (3) modeling different searching strategies that users employ, or (4) modeling specifics like how people (re)formulate queries. However, the simplest interaction we could model would be the clicks associated with a query and a result list. Each of the participants of had different facets in mind, and it was clear that there were many factors that needed to be taken into account when building models of queries and interaction (even for this small subset of possibilities). These included the following:

The scale and granularity of the models: do we model how an individual user makes a single decision in her information seeking episode, or just some overall aspect of interaction behavior like an average click probability.

The impact of presentation order: do we take into account relations between different results (such as similar results, domains, snippets).

The user population: Do we model user characteristics and hence differentiate between different user groups or user stereotypes?

The system response: Depending on the (simulated) searcher's actions, the system reacts by e.g. showing different results, modeling this also requires modeling (parts of) the system, and their interaction. So in terms of Figure 1, we may model our user and her context, the system, the user interface, and any combination of these.

Other factors: Do we factor in non-informational aspects, like e.g. how long it takes to complete the task?

Another question discussed during the session was, how to build simulations? One approach would be to base the simulations on handcrafted rules learned from available interaction data. In fact, we may be able to generate simulated interactions as captured in query logs. This could lead to a kind of "Turing test" that determines whether we would be able to discriminate between real user behavior and simulated user behavior in logs. Depending on the interaction steps of interest, simulation models may capture either a single "user in a box," or try to model the behavior of a "population of users in a box."

With regards to building simulations, it was noted that doing so requires proper software engineering to ensure the separation of concerns in the development and testing of simulations. That is, independent creation of simulation objects (i.e. users, interfaces, interactions, evaluation, etc), independent validation of the models, independent use of the models, and allowing for both operational (does the output of the model resemble real behavior close

enough) and conceptual validation (do the elements of the model resonate closely with reality).

If simulations are available, they can at least be instrumental to weed out some failures before wasting valuable user time and costs. Following the *Katja principle*: before you go to real users you need a pretty good system. However, by the same token proper validation of the simulated results remains valuable: there always has to be "a man behind the curtain."

5.3 Creating Simulations with Search Logs

This breakout group focused on the role that search logs can play for seeding simulations of interaction. There is an abundance of interaction data available in search logs, both on the Web and in Digital Libraries or other domain-specific collections. Can we use these to derive, or learn specific "user models"? What features of interaction should such a model contain? What aspect of interaction should it predict? And how do we know if it is any good?

The key idea was to setup a task that evaluates the "user models" directly, on their ability to predict aspects of interaction behavior (say, clicks or search termination) or user-based performance indicators (say, user satisfaction). This could result in a TREC-like task where participants do not submit the results of a system (a retrieval model), but of their user simulation (a user model). Such models could be trained and evaluated on observational interaction data in logs, with possible further annotations.

There was substantial discussion on what to model. There is unlikely to be a one-size-fits-all solution and so this will require us to address wither individual or group differences. There are many aspects of interest to model, such as multi-tasking behavior, topic drift, domain and task differences, and temporal dynamics. There are also generic differences in cognitive abilities, such as age, education level, search experience, patience and mood.

There was also substantial discussion on what to predict. Clear candidates are observable interaction features, such as clicks and skips/click reversal, the switch between searching and browsing, or search abandonment. Generally speaking, such features are great for averages, but more difficult to interpret for individual sessions. The logs themselves provide limited information, and should be complemented with information about the context (task, user, etc.) from other sources. Also editorial judgments from opt-in self reported questionnaires, or from paid annotators, are useful. There are also substantial differences in the completeness of search, HTTP, and proxy logs.

The evaluation and validation is also complex. For the observable features, intrinsic evaluation by the ability to predict an observable feature is possible. It is not clear to what extent this generalizes from one particular log to other settings. Should we use many logs? We may be able to build profiles that predict which results will transfer or not.

Finally, it was discussed that simulations are "models" that model some parts or aspects of interaction but purposefully ignore others. Being non-perfect is their strength. It allows us to focus on particular aspects without having to open up the Pandora's box of every possible element of context that may or may not impact the behaviors. This allows us to incrementally build models that are an increasingly close approximation of user interaction. External validation will be needed to make sure our interpretation of the simulation results is correct.

As for concrete plans for the next year, there was strong support for Ryen's suggestion during his keynote. Since obtaining and sharing log data is highly sensitive, academics may team up with industrial partners who can evaluate build models against a slice of real data.

Table 1: The browsing and interfaces breakout group identified three of many possible types of browsing and several dimensions of user satisfaction that one might want to measure.

	Types of Browsing		
Dimension of Satisfaction	**Exploratory**	**Directed search**	**Drifting search**
Coverage of topic	X		
Novelty	X	X	X
Diversity	X	X	X
Entertainment	X		
Surprise	X		X
Time	X	X	X
Topical relevance		X	
Lack of frustration		X	X
Total knowledge gain			X
Decisions satisfied			X

In this way, log data need not be distributed, only the evaluation results can be made public. There is still an open question on what data to use for training purposes, but perhaps even simulated data may be sufficient to capture a particular aspect of interest.

5.4 Simulated Browsing and User Interfaces

In this breakout group, the discussion was initially focused on three aspects of the problem of simulation of browsing and interfaces: how to do the modeling, what to measure, and how to validate simulations. The group focused its discussions around browsing.

In looking at the problem of browsing, the group came to the realization that in many cases it is non-trivial to determine what should be measured. As part of a user study, one can imagine using a questionnaire to collect information on the satisfaction with one exploratory browsing system vs. another, but these sorts of measures cannot be used with a simulation. Thus in the case of browsing, the group decided that before proceeding with simulations of browsing, it would be important to develop measures of browsing success that can be had from the observation of behavior and therefore also obtained as part of a simulation.

The group then discussed the nature of measures for browsing. An insight coming out of this discussion was that there are many dimensions to user satisfaction and that only some of these dimensions are likely to be applicable to different forms of browsing. Table 1 summarizes the measures and types of browsing discussed.

6 Summary

Given the limitations of the traditional IR evaluation, simulation of interaction is the next frontier for the evaluation of IR/IIR. It is important to be able to understand the interaction between users and IR applications/systems and to precisely study their behavior. The simulation of interaction can answer such questions, which are extremely difficult or near impossible to study otherwise. It changes the focus in IR experimentation — away from systems to interaction and usage. It will also help to create testable theories of user interaction.

Simulation provides a powerful tool to: (1) hypothesize about the outcome of different interactions, user models, and interfaces i.e. the "What IF" experiments, (2) test theories about IIR, (3) examine and explore the evaluation of the user and the interface, not just the ranking of documents, and (4) perform experiments in a controlled environment where interaction can be reproduced and replicated. It allows a system designer to explore far more design alternatives than is possible in user experiments and to identify the combinations of system components and parameters that are likely to work best. Furthermore, simulation of interaction invites lab-based IR to consider more explicitly the user and their interaction with the user interface and IR application/system. This will help build bridges between their work and more user-oriented research.

Aside from the benefits, there are also many challenges in the simulation of interaction. How do we know if a simulation is any good? The setup and validation of simulations will be important in order to decide this. When designing simulation experiments to answer particular research questions appropriate justification will be required. For instance, the components within the simulation should be seeded from previous observations and past user experiments, while the validation of the conclusions of simulations will be needed to be able to transfer the findings back to real life. In all, due care will be needed to design and conduct scientifically valid and meaningful simulations.

In summary, the simulation of interaction provides a powerful experimental methodology for researching interactive IR, if done properly, it will prove to be invaluable in the years to come.

Acknowledgements We would like to thank ACM and SIGIR for hosting this workshop, in particular Gianni Amati, Omar Alonso, and Stéphane Marchand-Maillet for their outstanding support in the organization. We would also like to thank the program committee: Nicholas Belkin, Pia Borlund, Ben Carterette, Paul Clough, Georges Dupret, Donna Harman, Claudia Hauff, Evangelos Kanoulas, Jaana Kekäläinen, Diane Kelly, Heikki Keskustalo, Birger Larsen, Jeremy Pickens, Benjamin Piwowarski, Ian Ruthven, Mark Sanderson, Falk Scholer, Andrew Turpin, Ryen White, Max Wilson, and the four program chairs. Final thanks are due to the paper authors, the invited speakers Ryen White and Donna Harman, the session chairs Charlie Clarke and Daniel Tunkelang, and the participants for a great and lively workshop. Details about the workshop are online at `http://www.dcs.gla.ac.uk/access/simint/`. The contributed papers, a literature overview, and ongoing discussion on SimInt is available on-line at `http://groups.google.com/group/simint`.

References

[1] O. Alonso and J. Pedersen. Recovering temporal context for relevance assessments. In Azzopardi et al. [3], pages 25–26.

[2] P. Arvola and J. Kekäläinen. Simulating user interaction in result document browsing. In Azzopardi et al. [3], pages 27–28.

[3] L. Azzopardi, K. Järvelin, J. Kamps, and M. D. Smucker, editors. *Proceedings of the SIGIR 2010 Workshop on the Simulation of Interaction: Automated Evaluation of Interactive IR (SimInt 2010)*, 2010. ACM Press.

[4] P. Clough, J. Gonzalo, and J. Karlgren. Creating re-useable log files for interactive CLIR. In Azzopardi et al. [3], pages 19–20.

[5] M. J. Cole. Simulation of the IIR user: Beyond the automagic. In Azzopardi et al. [3], pages 1–2.

[6] M. D. Cooper. A simulation model of an information retrieval system. *Information Storage and Retrieval*, 9:13–32, 1973.

[7] S. Geva and T. Chappell. Focused relevance feedback evaluation. In Azzopardi et al. [3], pages 9–10.

[8] D. Harman. Relevance feedback revisited. In *SIGIR 1992*, pages 1–10, 1992.

[9] B. Huurnink, K. Hofmann, and M. de Rijke. Simulating searches from transaction logs. In Azzopardi et al. [3], pages 21–22.

[10] C. Jethani and M. D. Smucker. Modeling the time to judge document relevance. In Azzopardi et al. [3], pages 11–12.

[11] E. Kanoulas, P. Clough, B. Carterette, and M. Sanderson. Session track at TREC 2010. In Azzopardi et al. [3], pages 13–14.

[12] T. Kato, M. Matsushita, and N. Kando. Bridging evaluations: Inspiration from dialog system research. In Azzopardi et al. [3], pages 3–4.

[13] H. Keskustalo and K. Järvelin. Query and browsing-based interaction simulation in test collections. In Azzopardi et al. [3], pages 29–30.

[14] H. Keskustalo, K. Järvelin, and A. Pirkola. Effectiveness of relevance feedback based on a user simulation model: Effects of a user scenario on cumulated gain value. *Journal of Information Retrieval*, 11:209–228, 2008.

[15] H. Keskustalo, K. Järvelin, and A. Pirkola. Graph-based query session exploration based on facet analysis. In Azzopardi et al. [3], pages 15–16.

[16] C. Mulwa, W. Li, S. Lawless, and G. Jones. A proposal for the evaluation of adaptive information retrieval systems using simulated interaction. In Azzopardi et al. [3], pages 5–6.

[17] N. Nanas, U. Kruschwitz, M.-D. Albakour, M. Fasli, D. Song, Y. Kim, U. C. Beresi, and A. D. Roeck. A methodology for simulated experiments in interactive search. In Azzopardi et al. [3], pages 23–24.

[18] M. Preminger. Evaluating a visualization approach by user simulation. In Azzopardi et al. [3], pages 31–32.

[19] S. Stober and A. Nuernberger. Automatic evaluation of user adaptive interfaces for information organization and exploration. In Azzopardi et al. [3], pages 33–34.

[20] J. Tague, M. Nelson, and H. Wu. Problems in the simulation of bibliographic retrieval systems. In *SIGIR 1980*, pages 236–255, 1980.

[21] D. Tunkelang. Using QPP to simulate query refinement. In Azzopardi et al. [3], pages 7–8.

[22] R. W. White, I. Ruthven, J. M. Jose, and C. J. Van Rijsbergen. Evaluating implicit feedback models using searcher simulations. *ACM Transactions on Information Systems*, 23:325–361, 2005.

[23] P. Zhang, U. C. Beresi, D. Song, and Y. Hou. A probabilistic automaton for the dynamic relevance judgement of users. In Azzopardi et al. [3], pages 17–18.