# Simulating Searches from Transaction Logs

Bouke Huurnink
bhuurnink@uva.nl

Katja Hofmann
k.hofmann@uva.nl

Maarten de Rijke
derijke@uva.nl

ISLA, University of Amsterdam
Science Park 107, Amsterdam, The Netherlands

## ABSTRACT

Computer simulations have become key to modeling human behavior in many disciplines. They can be used to explore, and deepen our understanding of, new algorithms and interfaces, especially when real-world data is too costly to obtain or unavailable due to privacy or competitiveness reasons.

In information retrieval, simulators can be used to generate inputs from simulated users—including queries, clicks, reformulations, and judgments—which can then be used to develop a deeper understanding of user behavior and to evaluate (interactive) retrieval systems.

The trust that we put in simulators depends on their validity, which, in turn, depends on the data sources used to inform them. In this paper we present our views on future directions and challenges for simulation of queries and clicks from transaction logs.

**Categories and Subject Descriptors:** H.3 [**Information Storage and Retrieval**]: H.3.3 Information Search and Retrieval

**General Terms:** Experimentation

**Keywords:** Simulation, Transaction Logs, Evaluation

## 1. INTRODUCTION

Simulation offers a source of experimental data when real-world information is either unobtainable or too expensive to acquire. This makes simulation a valuable solution for evaluating information retrieval (IR) theories and systems, especially for interactive settings.

One approach to creating simulators for IR evaluation is to build simulation models that incorporate manually created queries and relevance judgments, as in the Cranfield tradition. A problem here is that it is not clear in how far such explicit judgments reflect how users would interact with a real IR system.

In this paper we discuss search engine transaction logs as a source of data for building simulators for IR evaluation. Transaction logs typically record, among other things, sequences of queries and result clicks issued by a user while using a search engine [6]. The data is collected unobtrusively, thereby capturing the actions of users "in the wild." In addition, large quantities of searches and clicks can be gathered. This makes them a rich source of information.

For privacy and competitiveness reasons transaction logs are rarely made publicly available. However, simulators can be developed to generate artificial transaction log data not linked to any real users. Such a simulator can then be released to outside parties, for example in order to test our theories about searcher behavior or to evaluate the performance of (interactive) retrieval systems. In the process of creating such simulators, theories about search engine users could be tested by incorporating them in the simulation models.

So far, relatively little research has been conducted into developing simulators based on the searches in transaction logs. Below, we will discuss future directions for such research. We start by giving a brief overview of the state-of-the-art in simulation for retrieval evaluation in Section 2. Next, we outline possible application areas for transaction log-based simulators, and discuss some of the challenges that need to be addressed, in Section 3. We summarize our views in Section 4.

## 2. SIMULATION FOR IR

We now sketch some recent developments in simulation for IR purposes. In particular, we focus on the specific aspects of user interaction that have been simulated, and on how simulators have been validated.

A common form of data used to inform simulators in IR is manually created sets of queries and explicit relevance judgments. There are multiple scenarios where such data has been exploited. For example, an approach for evaluating novel interfaces for presenting search results is to create a simulated user who clicks on relevant documents that appear on the screen [4, 10]. Another example is in simulating queries for retrieval evaluation, for example by creating new queries or sequences of queries from an existing query and set of relevant documents [1, 8]. Here all simulated queries generated from the same truth data are associated with the same set of truth judgments.

When simulators are not informed by manually created queries and document judgments, key challenges are to (1) create queries, and (2) identify relevant documents for a given query. One solution is to use document labels to identify groups of related "relevant" documents: these documents are then considered relevant to the queries generated from their combined text [7] (the document labels themselves are not used in the query generation process). Another solution is to address retrieval tasks where only one document is considered relevant to a query, as in for example known-item search. Here Azzopardi et al. [2] used document collections in multiple European languages to simulate pairs of queries and relevant documents, and compared them to

*SIGIR Workshop on the Simulation of Interaction,* July 23, 2010, Geneva.
.

sets of manually created known-item queries.

For the simulation approaches mentioned above, it is unclear to what extent they reflect real-world queries and result interactions. This can be addressed through the use of transaction logs, as we discuss in the next section.

## 3. LOG-BASED SIMULATORS

Following on from our brief description of simulation in IR in general, we turn to directions for research in the development of transaction log-based simulators for IR.

The first direction for research is the realistic simulation of queries and clicks. We ourselves investigated this task using transaction logs from an archive search engine [5]. We validated each simulator by ranking different retrieval systems on its output data, and comparing this to a "gold-standard" retrieval system ranking obtained by evaluating the systems on actual log data. We found that incorporating information about users in the simulation model improved the simulator output. Another approach to simulation of queries and clicks was taken by Dang and Croft [3], who worked in a web setting. Here, anchor texts were available, which they used as simulated queries. The purpose here was to evaluate the effect of different query reformulation techniques. The authors compared retrieval performance on the simulated queries to retrieval performance on queries and clicks taken from an actual search log, and found that the simulated queries showed retrieval performance similar to real queries.

A second direction for research is the simulation of *sessions*—sequences of queries and clicks. Retrieval evaluation with explicitly judged queries and documents generally considers each query in isolation. Transaction logs, however, offer us a wealth of information about query modification behavior, and there is broad interest in the IR community about using such information for retrieval system evaluation. What is needed here is to develop and incorporate into the simulator insights about possible "moves" in a session, based on different assumptions about user intent.

A key problem when designing a simulator is ensuring that it is valid for the purpose for which that simulator is developed. Sargent [9] identifies three types of validity in the simulation model: *conceptual model validity*, the validity of the underlying assumptions, theories, and representations of the model; *model verification*, the correctness of the programming and implementation of a conceptual model; and *operational validity*, the accuracy of the output data created by a simulation model. Transaction logs, due to the large number of interactions that they record, offer a wealth of data for quantitatively determining operational validity. The measure of validity will vary according to the purpose of the simulator. For example, when generating simulated queries and clicks for comparing retrieval systems, a valid simulator is one that produces output data that scores different retrieval systems in the same way as data derived from actual transaction logs. Here a rank correlation coefficient such as Kendall's $\tau$ can be used to compare the rankings of retrieval systems on real and simulated output [5].

An open question in developing transaction log-based simulators is whether those simulators are transferable to new domains. The data contained in a transaction log represents the actions of a specific set of users on a specific search engine. A simulator that captures general aspects of user behavior could successfully be applied to new collections.

## 4. SUMMARY

In this position paper we have discussed the potential of transaction log data for developing and validating simulators for IR experiments. In particular we have discussed the creation of simulators for two scenarios: generating evaluation testbeds consisting of artificial queries and clicks; and creating simulations of session behavior in terms of sequences of queries and clicks. We discussed some of the challenges in creating such simulators, including the validation of simulation output. It is our view that transaction logs pose a rich source of information for simulator development and validation. By producing simulators that accurately reproduce the queries and clicks contained in transaction logs, we will not only be able to generate data for different retrieval tasks, but we will also obtain a better understanding of the behavior of users "in the wild."

**Acknowledgements** This research was supported by the European Union's ICT Policy Support Programme as part of the Competitiveness and Innovation Framework Programme, CIP ICT-PSP under grant agreement nr 250430, by the DuOMAn project carried out within the STEVIN programme which is funded by the Dutch and Flemish Governments under project nr STE-09-12, by the Center for Creation, Content and Technology (CCCT), by the Dutch Ministry of Economic Affairs and Amsterdam Topstad under the Krant van Morgen project, and by the Netherlands Organisation for Scientific Research (NWO) under project nrs 612.066.-512, 612.061.814, 612.061.815, 640.004.802.

## REFERENCES

[1] L. Azzopardi. Query side evaluation: an empirical analysis of effectiveness and effort. In *SIGIR '09*, pages 556–563. ACM, 2009.

[2] L. Azzopardi, M. de Rijke, and K. Balog. Building simulated queries for known-item topics: an analysis using six European languages. In *SIGIR '07*, pages 455–462. ACM, 2007.

[3] V. Dang and B. W. Croft. Query reformulation using anchor text. In *WSDM '10*, pages 41–50. ACM, 2010.

[4] O. de Rooij and M. Worring. Browsing video along multiple threads. *IEEE Trans. Multimedia*, 12(2):121–130, 2010.

[5] B. Huurnink, K. Hofmann, M. de Rijke, and M. Bron. Validating query simulators: An experiment using commercial searches and purchases. In *CLEF '10*, 2010.

[6] B. J. Jansen and U. Pooch. A review of web searching studies and a framework for future research. *JASIS&T*, 52(3):235–246, 2001.

[7] C. Jordan, C. Watters, and Q. Gao. Using controlled query generation to evaluate blind relevance feedback algorithms. In *DL '06*, pages 286–295. ACM, 2006.

[8] H. Keskustalo, K. Järvelin, A. Pirkola, T. Sharma, and M. Lykke. Test collection-based IR evaluation needs extension toward sessions—a case of extremely short queries. In *AIRS '09*, pages 63–74. Springer-Verlag, 2009.

[9] R. Sargent. Verification and validation of simulation models. In *WSC '05*, pages 130–143. Winter Simulation Conference, 2005.

[10] R. White, I. Ruthven, J. Jose, and C. Rijsbergen. Evaluating implicit feedback models using searcher simulations. *ACM TOIS*, 23(3):361, 2005.