Contextual Simulations for Information Retrieval Evaluation

Ryen W. White
Human-Computer Interaction Laboratory
Institute for Advanced Computer Studies
University of Maryland
College Park, MD 20742, USA
ryen@umd.edu

ABSTRACT

Non-interactive evaluations of Information Retrieval (IR) systems do not model many of the contextual factors that influence real users' information seeking. As such, they may give overly-simplified grounds for IR system comparison. This paper advocates the use of rich contextual simulations (i.e., simulations of user behavior and the factors that influence it) to extend and enhance the non-interactive evaluation of IR systems and the iterative development of interfaces that use their algorithms.

1. INTRODUCTION

The evaluation of IR systems is usually performed without human subjects where queries are derived for a given topic set and system performance is assessed using metrics such as *precision* and *recall*. Whilst these metrics allow system performance to be objectively measured, the model upon which non-interactive evaluations is based simulates only minimal user interaction such as the submission of search queries or the provision of additional relevance information through Relevance Feedback (RF) [1]. In search systems users can interact ways other than represented by this model (e.g., browsing result pages and requesting additional support from the system) and system performance can be affected by contextual factors such as users, tasks, and environmental and situational variables. A more robust representation of user interaction and the contextual constraints that affect this interaction is therefore desirable for non-interactive IR evaluation.

In this paper we suggest that evaluation work for the development of IR systems and interfaces could be aided by the use of simulations of user interaction behavior that incorporate a model of the search context. Rather than modeling minimal interaction and ignoring other factors, contextual simulations model all ways in which users could interact, retrieval strategies they could employ and external factors that could influence their interaction decisions. Simulation-based methods have already been used to predict user interaction behavior or evaluate Web site usability, to test components that improve search queries or to detect shifts in the interests of computer users, e.g., [2]. Although simulations cannot directly test interfaces from a user's perspective, nor can they fully capture the cognitive processes that motivate search decisions, they can test algorithms that underlie interfaces in circumstances that influence interface design.

2. OUR APPROACH

Our research in this area has mainly focused on simulating user interaction strategies to select algorithms for IR systems [4]. In this section we briefly describe interfaces we have created and simulations of user behavior developed to emulate interaction with these interfaces in pre-determined retrieval scenarios.

2.1 Interfaces

We developed search interfaces to encourage users to interact closely with their search results and improve the quality of their searching [3]. These interfaces differ from traditional RF interfaces in at least two ways: (i) they do not use a traditional "ranked-list" style of results display, instead showing multiple representations of each document and interactive paths between these documents, (ii) all information users view is used as evidence of their interests and new queries are generated automatically by RF algorithms. It was therefore necessary to develop an evaluation methodology that could model more complex user interaction and feedback behaviors. We created simulations that modeled interaction with our interfaces, presented this interaction to candidate RF algorithms and selected the best algorithm for use in the final versions of our interfaces.

2.2 Simulations

Our simulation models one part of the search context: *user interaction* such as document selections and browsing behaviours. The approach assumes the role of a user interacting with retrieved information (e.g., clicking on hyperlinks and viewing parts of documents) and emulating their behavior at the search interface (i.e., in their style of interaction and in the type of information they interact with). Interfaces are represented in the simulation by all possible interaction permutations. As the simulated user interacts, new query words are chosen by RF algorithms based on the information viewed and we evaluate algorithm performance based on the quality of the search queries the algorithms create and how quickly they "learn" what information is relevant.

In our simulation users are modeled with at least the following strategies: (i) assuming they only viewed relevant or non-relevant information, (ii) assuming they viewed *all* relevant or *all* non-relevant information, (iii) exhibiting differing degrees of "wandering" behaviour, i.e., trying to view relevant information but also viewing different amounts of non-relevant information. Although not represented in our simulations, the strategy users adopt can be affected by factors such as search experience, topic familiarity and demands placed upon them by the task setter or environment in which they are searching.

We selected the best performing RF algorithm and deployed it in some experimental interfaces. Experimentation with human subjects using the interfaces (described in [3]) yielded data about user interaction patterns, recorded in interaction log files. These data were then used to improve the realism of the simulations. This represents one way in which the quality of such simulations can be iteratively improved.

3. FUTURE DIRECTIONS

Since simulations allow many aspects of the search context to be modeled and parameterised (e.g., users, systems, tasks, search strategies, interaction) they provide experimenters with a way of evaluating IR systems more completely in non-interactive settings than the simple contextual model presently employed in TREC[1]. Our work in this area is only beginning and in this section we identify four future directions for our research.

3.1 Verify Validity of Approach

Since simulations are being used as the basis for evaluation it is important to evaluate their correctness. To do this we plan to conduct user studies with interactive versions of the simulation and leverage user feedback on the correctness of a representative sample of simulation decisions. We will improve the simulation where weaknesses are identified and are currently working closely with users to develop visualizations of simulation decisions.

3.2 Develop a Generic Evaluation Framework

A framework is required to allow simulations of this kind to be developed in a robust, generic and extensible way, suitable for the evaluation of other systems and interfaces. This framework will model different search contexts and use interaction data gathered during experimentation with human subjects to enhance realism of the simulations. It will be flexible and allow additional components to be built and added to suit the IR system and interface under investigation. The main components in such a framework will be those to model the user's characteristics, behaviors and search strategies, model the search task and model the search interface. The contribution of these components in the simulation can be varied as part of an experimental design.

An advantage of a generic evaluation framework is the potential for comparability between designers employing the framework to evaluate their respective systems. All participants can use the same set of simulations and compare the performance of their system against those of others for a given set of simulation parameters (*specific* performance) or across a variety of parameter settings (*general* performance). The use of a framework of this nature could therefore facilitate collaboration and competition between IR system designers. As part of a "plug-and-evaluate" methodology system designers only need to create an extension to the framework specific to the interactions afforded by the interface to their IR system. We have begun developing this framework and are currently investigating the capture and representation of contextual factors.

3.3 Create a Collection of Simulated Users

IR systems may be created for a specific group of users (as with those that search restricted domains such as library reference catalogues) or for general use by all users regardless of their level of searching experience (as with the World Wide Web). It may therefore be useful to control the characteristics of the simulated users and only use a certain type in the evaluation of IR systems. To allow an IR system to be analysed with a particular user group a suite of simulated users would be developed, each with their own search behavior constructed from real user logs or from stereotypes of users. A possible characterisation could be *novices*, *occasional users* and *experts*, and factors varied will include the style of interaction and the nature of search queries.

As demonstrated in Section 2.2, this collection of users can leverage data from human experimentation to supplement their user stereotypes. Human subjects will interact with our systems, and interaction captured and stored in log files. The contents of these files could then be used to modify simulation parameters and improve the realism of simulated user interaction. A collection of this nature could be part of the generic framework already discussed in this section, with those designers using the framework contributing log data from their subjects' interaction.

3.4 Prototype Interface Components

There is a large amount of possible interaction with search interfaces; however users only cover a small set. Experimentation with human subjects therefore does not allow for a full analysis of which interactions improve or degrade system performance. Simulations can be used either after a prototype interface is built, or before the interface is built, to test its performance with every possible set of user interactions prior to development.

If there was an association between certain interactions and poor system performance (e.g., the selection of poor feedback terms), it may be appropriate to redesign the interface to ensure these interactions were no longer possible. The creation of an approach to evaluate the performance of components on an interface also ensures that interface designers think about how users *could* interact with their system (i.e., all possible interactions); this is useful by-product of our methodology. We are currently employing this technique in the refinement of our interfaces.

3.5 Summary

We have described four future directions to further our research and develop contextual simulations that could influence design decisions and encourage collaboration between designers. Although simulations facilitate completeness in evaluation they should are not a replacement for human experimentation; it is still of the utmost importance to gather qualitative feedback on interfaces and search processes from users. Simulations are best used as a complementary technique designed to make human experimentation more worthwhile by being used before a user evaluation and limiting the number of system variations subjects must use. Our work proposes an extension of the standard non-interactive model of evaluation where simulations can be used to model complex interactions and assist developers in making IR system design choices.

REFERENCES

[1] Buckley, C., Salton, G., and Allan, J. (1994). The effect of adding relevance information in a relevance feedback environment. *Proceedings of the 17th Annual ACM SIGIR Conference*, 292-300.

[2] Lam, W., Mukhopadhyay, S., Mostafa, J., and Palakal, M. (1996). Detection of shifts in user interests for personalised information filtering. *Proceedings of the 18th Annual ACM SIGIR Conference*, 317-325.

[3] White, R.W., Ruthven, I. and Jose, J.M. (2005). A study of factors affecting the utility of implicit relevance feedback. *Proceedings of the 28th Annual ACM SIGIR Conference*, in press.

[4] White, R.W., Ruthven, I., Jose, J.M. and Van Rijsbergen, C.J. (2005). Evaluating implicit feedback models using searcher simulations. *ACM Transactions on Information Systems*, in press.

[1] http://trec.nist.gov