A Proposal for the Evaluation of Adaptive Information Retrieval Systems using Simulated Interaction

Catherine Mulwa[1], Wei Liu[2], Séamus Lawless[1], Gareth J. F. Jones[2]

Centre for Next Generation Localisation
School of Computer Science and Statistics, Trinity College Dublin, Dublin, Ireland[1]
School of Computing, Dublin City University, Dublin, Ireland[2]

mulwac@sccs.tcd.ie, wli@computing.dcu.ie, seamus.lawless@sccs.tcd.ie, gjones@computing.dcu.ie

ABSTRACT
The Centre for Next Generation Localisation (CNGL) is involved in building interactive adaptive systems which combine Information Retrieval (IR), Adaptive Hypermedia (AH) and adaptive web techniques and technologies. The complex functionality of these systems coupled with the variety of potential users means that the experiments necessary to evaluate such systems are difficult to plan, implement and execute. This evaluation requires both component-level scientific evaluation and user-based evaluation. Automated replication of experiments and simulation of user interaction would be hugely beneficial in the evaluation of adaptive information retrieval systems (AIRS). This paper proposes a methodology for the evaluation of AIRS which leverages simulated interaction. The hybrid approach detailed combines: (i) user-centred methods for simulating interaction and personalisation; (ii) evaluation metrics that combine Human Computer Interaction (HCI), AH and IR techniques; and (iii) the use of qualitative and quantitative evaluations. The benefits and limitations of evaluations based on user simulations are also discussed.

Categories and Subject Descriptors

H.3 [**Information Storage and Retrieval**]: Information Search and Retrieval; H.5 [**Information Interfaces and Presentation**]: Multimedia Information Systems; H.5 [**Information Interfaces and Presentation**]: Hypertext/Hypermedia;

General Terms
Experimentation, Measurement, Performance

Keywords
Relevance Feedback, Simulation,

1. INTRODUCTION
The Centre for Next Generation Localisation (CNGL) is developing novel technologies which address the key challenges in localisation. Localisation refers to the process of adapting digital content to culture, locale and linguistic environments at high quality and speed. The technologies being developed combine techniques from natural language processing, information retrieval and Adaptive Hypermedia. The complex functionality offered by these systems and the variety of users who interact with them, mean that evaluation can be extremely difficult to plan, implement and execute. Both component-level scientific evaluation and extensive user-based evaluation are required to comprehensively assess the performance of an application. It is critically important that such experiments are thoroughly planned and conducted to ensure the quality of application produced. The potential number of experiments needed to gain a full understanding of the systems being developed means that carrying out these repeated investigations using real interactive user studies is impractical. As a result, simulated interaction is vital to enable these experiments to be replicated and recursively executed in a controlled manner.

2. EVALUATION USING SIMULATED INTERACTION
IR evaluation experiments can be divided into four classes: i) observing users in real situations, ii) observing users performing simulated tasks, iii) performing simulations in the laboratory without users and iv) traditional laboratory research (no users and no interaction simulation) [1]. When simulating user interaction and replicating experiments it is essential that performance is measured using the most suitable evaluation metrics. The following sections detail metrics which can be using in the evaluation of AIRS, particularly experiments which use simulated interaction.

2.1 IR Evaluation Metrics
IR is classically evaluated in terms of precision and recall, which tell us about the accuracy and scope of the retrieval of relevant documents. These metrics are, of course, very valuable in measuring the effectiveness of real world search tasks. They are also used to evaluate retrieval effectiveness with test collections in laboratory IR experimental settings. However, the standard assumption, in laboratory IR experiments, that the relevance of individual documents is constant for multiple search interactions limits the suitability of such test collections for the evaluation of simulated interactive search.

An experimental framework is needed which can capture simulated explicit or implicit feedback from a user and exploit this for relevance feedback and subsequent experiments. This framework could also potentially modify the identified set of relevant documents to reflect: (i) relevant information found in previous iterations of the experiment; and (ii) the development of the user's information need. For example, in some situations documents may become relevant as the search progresses and the user's knowledge of a subject grows having seen previous relevant documents. This concept of a user interacting with an IR system and providing feedback which modifies the systems response has similarities with the AH systems from which we next consider relevant evaluation principles.

2.2 AH Evaluation Metrics
Numerous measures of the performance of adaptivity in adaptive systems have been proposed [2]. These metrics aim to address both component-level scientific evaluation and user-based evaluation of the adaptivity offered by the system.

Personalised Metrics: Personalisation in IR can be achieved using a range of contextual information such as information about the user, the task being conducted and the device being used. Contextual information is increasingly being used to facilitate personalisation in IR. The personalised identification, retrieval and presentation of resources can provide the user with a tailored information seeking experience [2]. Personalisation metrics aim to express the effort necessary to exploit a system [3] e.g. **MpAC:** Minimum personalisation Adaptive Cost which indicates the percentage of entities which are personalised in an AIRS system. This metric considers only the minimum number of entities necessary to make a system adaptive.

Interaction Metrics: These metrics aim to provide information on the quality of the AIRS system's functionality. This is achieved by evaluating the variation in the interaction between administrators or users and the adaptive and non-adaptive versions of a system [4]. Examples include: i) **AiAI:** Administrator Interaction Adaptivity Index. This metric compares the actions performed by administrator to manage the system before and after the addition of adaptivity; ii) **UiAI**: User interaction Adaptivity Index. This metric compares the actions performed by a user to access the functionality of a system both before and after the addition of adaptivity. Whenever an action differs, an additional action is needed or an action is missing, this index increases by one. Interaction metrics assist in the comparative evaluation of AIRS systems from an adaptive perspective.

Performance metrics*:* Many metrics can be used to measure performance e.g., knowledge gain (AEHS), amount of requested materials, duration of interaction, number of navigation steps, task success, usability (e.g., effectiveness, efficiency and user satisfaction). Such metrics concern aspects of the system related to response time, improvement of response quality in the presence of adaptivity and the influence of performance factors on the adaptive strategies.

2.3 Simulation of Interaction Techniques

Simulation techniques enable multiple changes of system configuration, running of extensive experiments and analysing results. The simulation assumes the role of a searcher, browsing the results of an initial retrieval [5]. The information content of the top-ranked documents in the first retrieved document set constitutes the information space that the searcher must explore. All the interaction in this simulation is with this set and it is assumed that searchers will only view relevant information. The authors are interested in the use of this technique to determine how to evaluate the change in retrieval effectiveness when an AIRS system adapts to a query in a standard way, and also to incorporate user and domain models and investigate how to exploit these.

2.4 Simulation-Based Evaluation Challenges

The main challenges in the use of simulation methods include: i) determining what data must be gathered in order to replicate experiments; ii) deciding how to gather this data; iii) identifying how to replicate the variety of user behaviours and personalisation offered by the system; iv) the simulation of relevance, for instance simulating the characteristics of relevant documents successfully over a search session; v) validating the simulation's query evaluation times against the actual implementation; vi) selecting what method to use to collect implicit feedback; and vii) deciding how to filter the collected implicit feedback.

3. PROPOSED METHODOLOGY

It is essential that the correct methods are used when evaluating AIRS systems [6]. In order to sufficiently evaluate both the adaptive functionality and the retrieval performance of these systems a hybrid approach is proposed which combines IR, AH and Simulation-based evaluation methods. The techniques and metrics required are: i) **simulation-based techniques** where simulation assumes the role of a searcher, browsing the results of an initial retrieval; ii) **user-centred methods** for simulating interaction and personalisation; and iii) evaluation metrics borrowed from **AH** and **IR**. During a search the information state and need of the user changes and this must be modelled in each simulation so that the information viewed so far by the user can be used to influence the generation of a subsequent query. An objective of AIRS is to minimise the amount of information that must be viewed in order to gain a certain amount of knowledge. Thus the user must be shown relevant information in correct order. This is related to both IR and AH, where personalised responses are created for a domain-specific information need. Thus, for an information need, it is necessary to assess not only the relevance of documents to a topic, but also the order in which these should be presented. The number of documents which must be viewed over a search session to satisfy the information need can be further measured. At each point, search effectiveness can be measured with respect to the current information state of the simulated user. One of the main objectives of this work is to explore the potential of using user and domain models to reduce the user search effort. The potential benefits of the proposed methodology include: retrieval accuracy, completeness of system functionality, cost saving, user satisfaction, adaptivity, time, satisfied customer goal, user ratings, quality, appropriateness, accessibility, assistance, richness, availability, completeness, self-evidence, usability, user-retention, consistency, functionality, performance, predictability, portability, reliability and reuse.

4. CONCLUSION AND FUTURE WORK

Simulation-driven evaluation is not new, but the effects of personalisation on creating reproducible, large scale experiments can be addressed by incorporating AH and IR techniques and evaluation metrics. Further work is required in order to test the proposed methodology using systems being developed by CNGL.

5. ACKNOWLEDGMENTS

This research is based upon works supported by Science Foundation Ireland (Grant Number: 07/CE/I1142) as part of the Centre for Next Generation Localisation (www.cngl.ie).

6. REFERENCES

[1] H. Keskustalo et al. "Evaluating the effectiveness of relevance feedback based on a user simulation model: effects of a user scenario on cumulated gain value". Information Retrieval, vol.11, pp.209-228, 2008.

[2] Séamus Lawless, et al., "A Proposal for the Evaluation of Adaptive Personalised Information Retrieval," presented at the CIRSE 2010 Workshop on Contextual Information Access, Seeking and Retrieval Evaluation, Milton Keynes, UK, 2010.

[3] L. Masciadri and C. Raibulet, "Frameworks for the Development of Adaptive Systems: Evaluation of Their Adaptability Feature Through Software Metrics,"4th International conference SEA 2009, pp. 309-312.

[4] C. Raibulet and L. Masciadri, "Evaluation of Dynamic Adaptivity Through Metrics: an Achievable Target?," 4th International conference SEA 2009.

[5] R. White, et al., "A simulated study of implicit feedback models," Advances in Information Retrieval, pp. 311-326, 2004.

[6] P. Brusilovsky, et al., "Layered evaluation of adaptive learning systems," International Journal of Continuing Engineering Education and Life Long Learning, vol. 14, pp. 402-421, 2004.