Building Realistic Simulations for Interactive Information Retrieval

David Maxwell
School of Computing Science
University of Glasgow
Glasgow, Scotland
d.maxwell.1@research.gla.ac.uk [*]

ABSTRACT

Simulation has been used within the field of *Information Retrieval (IR)* for many years to evaluate retrieval models and other aspects of the wider IR process. In recent years, there has been a renewed interest towards using simulation for *Interactive Information Retrieval (IIR)*, an area which focuses on the study of human interactions with IR systems. A variety of different interaction models (e.g. click models) associated with behavioural aspects of searchers have over time been developed and evaluated using simulation in order for us to better understand the complex processes involved. Despite these advances, such models are still relatively naïve, and further work is required to make simulations of searchers more realistic. To this end, this project seeks to build more realistic simulations, using a more *Complex Searcher Model (CSM)*. Within the CSM, each component and decision point can be varied and customised as required. The CSM can then be instantiated using components that are grounded from empirical evidence based upon actual real-world searcher behaviour and interaction data.

Categories and Subject Descriptors: H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval:Search Process H.3.4 [**Information Storage and Retrieval**]:Systems and Software:Performance Evaluation

Keywords Simulation; User Modeling; Search Strategies; Search Behavior; Querying Strategies; Stopping Strategies; Evaluation

1. INTRODUCTION AND MOTIVATION

The *Information Retrieval (IR)* community has centred much of its recent research upon the so-called *Cranfield Paradigm*. The paradigm revolves around the idea of a test collection and associated relevance judgements for documents within said collection. This approach provides a standardised way in which one can evaluate their retrieval system against a given baseline. While the general concepts of the paradigm have remained in place since the 1960s, components have over the years evolved as the associated data and tasks have become ever more complex in nature [8]. Examples of use today include the *NIST*-sponsored *Text REtrieval Conference (TREC)* and other evaluation forums.

The Cranfield Paradigm today still largely remains the *de facto* means of IR evaluation. Despite this however, the approach possesses a simplistic means of examining the actions with which a real-world searcher undertakes. As such, several different scientific approaches have been developed to better understand the complex sequence of interactions taking place, and are readily used in the study of *Interactive Information Retrieval (IIR)* which deals specifically with the interactions between humans and search engines. As outlined by Keskustalo et al. [9], the approaches can be split into four distinct categories:

1. obtaining data from searchers in real-world situations (e.g. log data from a commercial search engine);
2. observing searchers perform simulated search tasks (e.g. a user study involving a search engine);
3. performing simulations in a lab environment (e.g. simulations of interaction, without real-world searchers); and
4. traditional lab-based research, sans real-world searchers.

Experimentation with real-world searchers undertaking either real or simulated search tasks is the preferred way to study IIR (approaches *1* and *2*). However, such experiments require a significant level of effort to organise and setup. They are also laborious to run, and are usually very costly - both for the researcher and subject involved [3]. Approach *4* can be argued as a simplistic form of *'TREC-style'* simulation, assuming a single query with a fixed number of documents examined in a linear fashion. This approach however is not interactive, and can be considered naïve.

This therefore leaves the simulation of real-world searchers, incorporating *interactive* components such as relevance feedback and other interaction components (approach *3*) as a means of exploring IIR. As the main focus of this project, simulation provides a rapid means of exploring the potential limits of real-world searcher interactions at a low cost. Current means of simulation are limited because they assume searchers act in a fixed way, or act stochastically by examining content with fixed probabilities. Research has shown that in reality, searchers tend to adapt their interactions based upon the quality of the presented ranked list [14].

[*] The author is currently a third year PhD student, supervised by Dr Leif Azzopardi (Leif.Azzopardi@glasgow.ac.uk) and Professor Roderick Murray-Smith (Roderick.Murray-Smith@glasgow.ac.uk).

CHIIR'16, March 13 - 17, 2016, Carrboro, NC, USA
Copyright is held by the owner/author(s).
ACM 978-1-4503-3751-9/16/03
http://dx.doi.org/10.1145/2854946.2854950

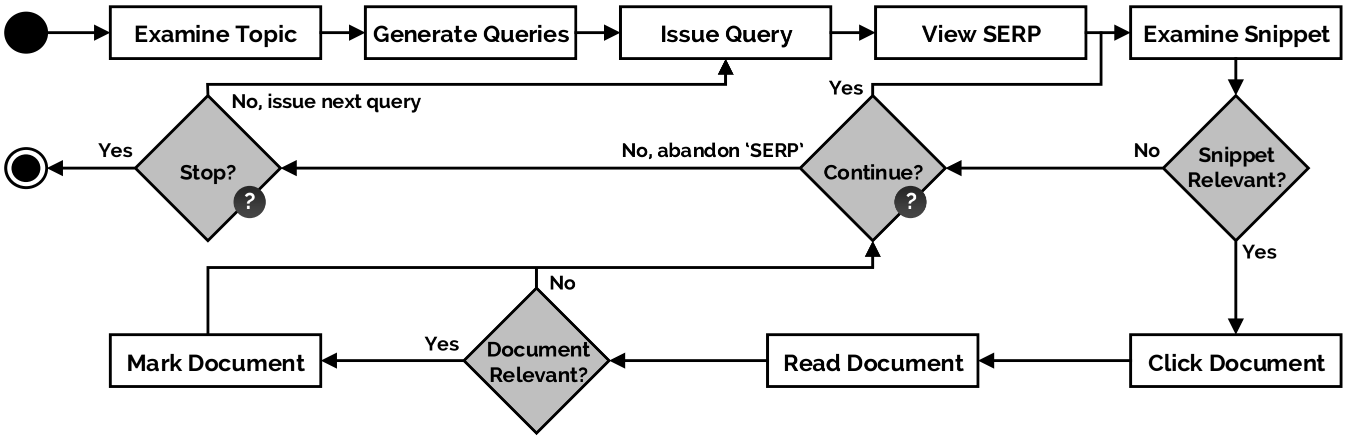

Figure 1: A flowchart of the current iteration of the CSM, complete with decision points (shown in grey) and tasks (shown in white) that a searcher undertakes within the wider search/IIR process. The model is adapted from Baskaya et al. [4] and Thomas et al. [16], with the figure itself adapted from Maxwell et al. [13].

Simulations of course would not be effective if the underlying models and assumptions do not adequately represent the actions that real-world searchers would be inclined to undertake [3]. While the IIR community has recently made advances in user modelling (e.g. [1, 2, 4, 16]), a significant amount of work is still required to make simulations of searchers more credible. To this end, this project aims to build more realistic simulations, starting with a *Complex Searcher Model (CSM)*, illustrated as a flowchart in Figure 1. Within the CSM, a complex search process is modelled, where each component (e.g. query generation) and decision point (e.g. deciding when to stop) can be varied and customised. We also instantiate the CSM with components grounded from empirical evidence based upon actual real-world searcher behaviour and interaction data, providing realism to the simulations.

2. RESEARCH QUESTIONS

The initial focus of this project will involve addressing the first three research questions, which revolve around improving querying and stopping components. The latter part of this project will then ensure the realism, flexibility and validity of the developed components as a whole.

RQ1 How can we create more realistic stopping components that reflect different types of searcher stopping behaviour?

Addressing this research question will require research to **RQ1.1** examine what factors specifically influence the stopping behaviours of searchers, and **RQ1.2** which operationalised stopping strategies best reflect the actual stopping behaviours of real-world searchers.

We also address the issue of query generation.

RQ2 How do we create more realistic querying components that reflect upon different types of searcher querying behaviour?

Specifically, we wish to examine **RQ2.1** what factors influence querying behaviours; **RQ2.2** how we can rank and select queries for issuing to the underlying search engine; and **RQ2.3** how we can revise the list of possible queries that can be issued as the simulated searcher 'learns' from examining snippets and documents. This final point is of particular importance to building a realistic query generation component for the CSM, but will however require a modification of the CSM to allow for the updating of potential query terms. These findings, in combination with the findings related to stopping behaviour, leads to the following overarching question.

RQ3 What is the interplay between querying and stopping strategies, and other components of the CSM?

Of course, the CSM presented in Figure 1 contains a variety of different components that can be explored. While the time available for this project does not permit a thorough exploration of all, there are still some areas in which potential improvements in realism can be made.

RQ4 What other components of the CSM can be improved or refined to make the underlying model more realistic and flexible?

For example, based upon the data available at a given point, how can we **RQ4.1** assess the quality of a provided *Search Engine Results Page (SERP)* from first impressions? **RQ4.2** What features do searchers look for that give them confidence the SERP may contain useful information related to their information need?

Our final research question can neatly encapsulate all of the efforts towards the previous four questions, examining the realism of the simulations run with the CSM.

RQ5 With grounded CSM components, will the actions performed by simulated searcher be indistinguishable from those performed by actual searchers?

This question revolves around the concept of the so-called *Belkin test*. When provided with the interaction logs of both a real-world searcher and simulated searcher under similar conditions (e.g. time constraints), would one be able to differentiate between the two?

3. METHODOLOGY AND PROGRESS

This section details the work that has been undertaken thus far towards this project, highlighting the key accomplishments and findings towards addressing the overarching research questions provided in Section 2. The main approach taken for this project entails: *(i)* the collection of real-world searcher interaction log data through a controlled user study [11]; *(ii)* the design and implementation of an IIR simulator (Section 3.1); and *(iii)* the use of the simulator and searcher interaction log data to implement and evaluate new components for the CSM (Section 3.2).

3.1 SimIIR and the CSM

Underpinning the remainder of this project, we have developed an open-source searcher simulator framework called *SimIIR*, available at `http://git.io/vZ5mH`. SimIIR operationalises the CSM, as detailed in Figure 1, and is based upon two salient interaction models by Baskaya et al. [4] and Thomas et al. [16]. Experimental setups for publications utilising SimIIR are also available at `http://git.io/vOBLz`.

3.2 Modelling Stopping Behaviours

We then began an investigation into a series of different stopping strategies, a means to describe the point at which searchers decide to stop examining a provided SERP. The stopping strategies are encoded within the CSM as the two decision points, highlighted with question marks (*?*) as seen in Figure 1. Many studies examining the stopping behaviours of searchers have found searchers use their intuition when deciding when to stop, with the belief that what they find is 'good enough' [13]. Despite this, several researchers proposed a series of stopping rules and heuristics [5, 6, 7, 10].

We used some of these defined rules and heuristics and operationalised them, examining the effectiveness of each with a simulated analysis [12]. We used a fixed-depth baseline stopping strategy (e.g. $P@k$) and two implementations of the *frustration point/disgust* stopping rules [6, 10], one considering the total number of snippets judged non-relevant, with the other considering the number of snippets judged non-relevant *contiguously.* The simulations were conducted over the AQUAINT and WT2g collections using topics associated with each collection, with each stopping strategy trialled over a wide range of thresholds (e.g. stop after x snippets are judged non-relevant). Our findings from the study showed that our implementations of the frustration point/disgust rules resulted in higher levels of gain per second across a number of different querying strategies. This work was then taken further with the implementation of three additional, more sophisticated stopping rules [13]. We also compared simulated behaviours against that of our real-world searcher log data [11], finding that the same, adaptive stopping strategies yielded the closest approximations.

4. FUTURE WORK

As much of the work related to **RQ1** is complete, some aspects do still remain. For example, Smucker [15] identified different categories of searcher depending on their interaction characteristics (e.g. searchers who are fast at examining snippets, and likely to consider them relevant). By clustering searchers from our user study [11], we could determine what stopping strategies and associated threshold values best approximates searcher behaviours by cluster/group.

Secondly, with regards to **RQ2**, work is required to determine how to construct better queries based upon content examined by a simulated searcher as they progress through a search task. A detailed analysis on the factors that influence querying behaviour, coupled with a closer examination between the links of stopping and querying behaviours, should provide us with a means to address **RQ3**.

With so many different components of the CSM that can be addressed, we will then explore the phase of *initial SERP examination*, addressing **RQ4**. Understanding behaviours exhibited by searchers at this phase would allow us to significantly improve the realism of the CSM, as skipping poor SERPs would save the simulated searcher time and effort.

With these implemented components and observed improvements in real-world searcher approximations, we should then be able to provide an answer to **RQ5**, or at least specify which aspects still need improvement. A final important question that can be raised is whether the findings from these simulated studies will generalise across different searchers. As such, it may be beneficial to conduct a further user study for comparison against our simulations and examine if the results generalise across studies.

Acknowledgments: The author wishes to thank his primary supervisor, Dr Leif Azzopardi, for all the help, guidance, patience and friendship shown during his studies at *UoG. Thank you so much, Leif!* The author also thanks the two reviewers for their feedback, and Professor Kalervo Järvelin and Dr Heikki Keskustalo at the University of Tampere, Finland for their co-operation and participation in a STSM, funded by the MUMIA Cost Action (reference `ECOST-STSM-IC1002-080914-049840`). The author is also financially supported by the UK Government through the EPSRC, grant number `1367507`.

References

[1] L. Azzopardi. The economics in interactive information retrieval. In *Proc. 34th ACM SIGIR*, pages 15–24, 2011.

[2] L. Azzopardi. Modelling interaction with economic models of search. In *Proc. 37th ACM SIGIR*, pages 3–12, 2014.

[3] L. Azzopardi, K. Järvelin, J. Kamps, and M.D. Smucker. Report on the sigir 2010 workshop on the simulation of interaction. *SIGIR Forum*, 44(2):35–47, 2011.

[4] F. Baskaya, H. Keskustalo, and K. Järvelin. Modeling behavioral factors in interactive information retrieval. In *Proc. 22nd ACM CIKM*, pages 2297–2302, 2013.

[5] G.J. Browne, M.G. Pitts, and J.C. Wetherbe. Stopping rule use during web-based search. In *Proc. HICSS-38*, page 271b, 2005.

[6] W.S. Cooper. On selecting a measure of retrieval effectiveness part ii. implementation of the philosophy. *J. of the American Soc. for Info. Sci.*, 24(6):413–424, 1973.

[7] W.S. Cooper. The paradoxical role of unexamined documents in the evaluation of retrieval effectiveness. *Info. Processing and Management*, 12(6):367 – 375, 1976.

[8] D. Harman. Is the cranfield paradigm outdated? In *Proceedings of SIGIR 2010*, pages 1–1, 2010.

[9] H. Keskustalo, K. Järvelin, and A. Pirkola. Evaluating the effectiveness of relevance feedback based on a user simulation model: Effects of a user scenario on cumulated gain value. *Information Retrieval*, 11(3):209–228, 2008.

[10] D.H. Kraft and T. Lee. Stopping rules and their effect on expected search length. *IPM*, 15(1):47 – 58, 1979.

[11] D. Maxwell and L. Azzopardi. Stuck in traffic: How temporal delays affect search behaviour. In *Proc. 5th IIiX*, pages 155–164, 2014.

[12] D. Maxwell, L. Azzopardi, K. Järvelin, and H. Keskustalo. An initial investigation into fixed and adaptive stopping strategies. In *Proc. 38th ACM SIGIR*, pages 903–906, 2015.

[13] D. Maxwell, L. Azzopardi, K. Järvelin, and H. Keskustalo. Searching and stopping: An analysis of stopping rules and strategies. In *Proc. 24th ACM CIKM*, pages 313–322, 2015.

[14] A. Moffat, P. Thomas, and F. Scholer. Users versus models: What observation tells us about effectiveness metrics. In *Proc. 22nd ACM CIKM*, pages 659–668, 2013.

[15] M.D. Smucker. An analysis of user strategies for examining and processing ranked lists of documents. In *Proc. of 5th HCIR*, 2011.

[16] P. Thomas, A. Moffat, P. Bailey, and F. Scholer. Modeling decision points in user search behavior. In *Proc. 5th IIiX*, pages 239–242, 2014.