Ways we can improve Simulated Personal Search Evaluation

David Elsweiler
Department of Computer Science (8 AI),
University of Erlangen, Germany
david@elsweiler.co.uk

David Losada
Department of Computer Science, University of
Santiago de Compostela, Spain
david.losada@usc.es

1. INTRODUCTION

Analysing how people perform personal search and evaluating the performance of a Personal Search algorithm in a controlled and repeatable way represent an important, but extremely difficult problem for researchers. In Personal Search Evaluation everyone has a unique collection of personal documents, which makes it difficult to compare the performance of one user against another. A second problem is that much of the information within individual collections is private so devising tasks for these collections is also a challenge. Even after overcoming these problems, there is still the issue of repeatability. An individual's relationship with his information changes constantly and the way he interacts is context-dependent. This means that any user study performed is almost impossible to re-perform under the same conditions.

A few methods have been proposed to address these issues. For example, Elsweiler and Ruthven [4] suggested a method of task creation for user-based laboratory re-finding experiments. Chernov and colleagues [2] proposed that researchers volunteer their own personal data to create a shared test collection for research purposes. Kim and Croft [5] use pseudo-desktop collections that have similar properties to personal collections to avoid privacy issues and utilise a simulated querying approach [1] to facilitate automated experiments for known-item tasks.

We believe that this third approach represents the best opportunity to run controlled and repeatable experiments to test retrieval models for Personal Search. That being said, this method, as has been applied to date, suffers from a number of limitations. It is oversimplified and is, consequently, unlikely to replicate user behaviour realistically. In this position statement we outline our views on the weaknesses of the approach and propose ways to improve the process.

2. OVERVIEW OF STATE OF THE ART

The pseudo-collections available in the community include three collections generated from TREC Enterprise track dataset [5], where prominent individuals were identified from the W3C mailing list. Documents were established for these people by taking the emails sent or received by these individuals on the mailing list. These mails were complemented by querying a web search engine with the name, organization and specialization of each target individual to obtain web pages and documents related to that person. A further collection was described in [6], where documents of various types were collected from many public sources in a particular Computer Science department. This collection contains emails from the department mailing list, news articles and blog postings on technology, calendar items of department announcements, web pages and office documents crawled from the department and research group web sites.

Strategies for building simulated queries have been proposed for known-item web page search [1] and for desktop search [5]. Essentially, they are based on randomly selecting a document (known-item) from the collection and algorithmically selecting query terms from the target document. This leads to the automatic generation of simulated queries and relevance judgments.

In the following sections we outline our thoughts on how the various aspects of this process may be improved. More specifically we offer suggestions to improve the query simulation process, the item selection process, and the collections used. We also discuss how we may evaluate the quality of the simulation.

3. IMPROVING QUERY SIMULATION

We posit that the query simulation process used in previous work may not reflect real life. The approaches used to date either randomly select terms from the documents to create queries of an allocated length or they draw terms independently based on how discriminative the terms are (using tf_idf-like weights). We believe this approach is overly simplistic and does not reflect the way queries would be generated in real life. This process does not take into account, for example, that:

- people may be more or less likely to choose query terms from different fields of a document (e.g., the subject or sender field of an email)
- spelling mistakes may be present
- queries may consist of phrases rather than just independent terms
- re-finding queries regularly contain named entities [3]

- queries may contain words not actually present in the document
- queries may be context- or situation-dependent. For example, the characteristics of the user or situation surrounding the task may influence the kind of queries submitted

We argue that to make the simulation process as accurate to real-life behaviour as possible the above aspects need to be accounted for. Our suggestion would be to seed the simulation with real query characteristics extracted from controlled or naturalistic user studies. For example, from a user study evaluating the use of a desktop search tool, e.g. [3], we can learn about how long queries tend to be, the document fields against which they are submitted to, the presence of spelling mistakes, etc. Further, a controlled laboratory-based evaluation, such as performed by [4], would allow researchers to control user and contextual variables to establish query profiles for different situations. This would offer the potential to test the hypothesis that query characteristics change in different scenarios and different algorithms may be offer better support in differing situations as a result.

4. IMPROVING ITEM SELECTION

In current implementations of the query simulation process items in the collection are chosen at random to create known-items. However, previous work has shown that only a small number of personal documents tend to be re-found [7] and that various document properties, such as whether or not it has been re-found before and the time that has lapsed since last access will influence whether or not it will be later re-found.

Further, current approaches treat documents independently, i.e, they do not consider the fact that they may be related and this may influence the likelihood that they will be re-found. If these kinds of properties could be built into the simulation process, we hypothesize that a much more realistic framework for evaluation could be achieved.

We propose to perform longitudinal, naturalistic investigations to establish predictors that documents will be re-used, i.e. document properties that make them more likely to be re-found. This could be achieved by using statistical modelling techniques, such as logistical regression.

5. IMPROVING PSEUDO COLLECTIONS

Due to the inherent difficulties in establishing an appropriate collection for this kind of work, with existing pseudo collections the main criteria has been on establishing any collection that looks like a personal collection, i.e. it is semi-structured and contains information largely created by or associated with one person. While this is a good starting place, we have to investigate whether this is really enough.

The first issue to address is collection size. The existing collections are very small. Second, it is important to ensure that pseudo collections cover a similar breadth of topics as real email collections. Third, the distributions of meta data e.g. senders in email collections should be comparable in real and artificially created collections.

6. EVALUATING THE SIMULATION

Evaluating how the methods suggested above affect the ecological validity of the process is again difficult.

In the literature, query simulations are often evaluated against manual queries (e.g. [1], in the context of known-item web search). Usually, given a pseudo collection, we do not have manual queries and, therefore, this limits the way in which we can assess the quality of the simulated queries. The few attempts done to evaluate the simulations in a pseudo collection environment were based on rather artificial ways to produce hand-written queries from the pseudo collection [5]. Therefore, we strongly argue that a proper method to evaluate simulated queries for pseudo collections is still to be found. Achieving this challenging objective would be a significant advance in this field.

7. CONCLUSIONS

In our view, the pseudo desktop collection approach with simulated queries is the best option to achieve a realistic, controlled and repeatable test environment for Personal Search. In this paper, we have enumerated a number of paths on which simulated evaluation for Personal Search needs to make progress.

8. ACKNOWLEDGMENTS

This work was partially funded by the Alexander von Humboldt Foundation. Further, the second author acknowledges the financial support given by "Ministerio de Ciencia e Innovación" through project grant TIN2010-18552-C03-03

9. REFERENCES

[1] L. Azzopardi, M. de Rijke, and K. Balog, *Building simulated queries for known-item topics: an analysis using six european languages*, Proc. ACM SIGIR conference on Research and development in information retrieval, 2007, pp. 455–462.

[2] S. Chernov, P. Serdyukov, P. Chirita, G. Demartini, and W. Nejdl, *Building a desktop search test-bed*, ECIR'07: Proceedings of the 29th European conference on IR research (Berlin, Heidelberg), Springer-Verlag, 2007, pp. 686–690.

[3] S. Dumais, E. Cutrell, J. Cadiz, G. Jancke, R. Sarin, and D.C. Robbins, *Stuff i've seen: a system for personal information retrieval and re-use*, Proc. ACM SIGIR '03:, 2003, pp. 72–79.

[4] D. Elsweiler and I. Ruthven, *Towards task-based personal information management evaluations*, Proc. ACM SIGIR conference on Research and development in information retrieval, 2007, pp. 23–30.

[5] J. Kim and W. B. Croft, *Retrieval experiments using pseudo-desktop collections*, CIKM '09: Proceeding of the 18th ACM conference on Information and knowledge management (New York, NY, USA), ACM, 2009, pp. 1297–1306.

[6] ———, *Ranking using multiple document types in desktop search*, Proceeding of the 33rd international ACM SIGIR conference on Research and development in information retrieval (New York, NY, USA), SIGIR '10, ACM, 2010, pp. 50–57.

[7] S. K. Tyler and J. Teevan, *Large scale query log analysis of re-finding*, Proc. WSDM '10, 2010.