Workshop on Evaluating Personal Search

Liadh Kelly
Dublin City University
Glasnevin, Dublin
Ireland
lkelly@computing.dcu.ie

Jinyoung Kim
Univ. of Massachusetts
Amherst
USA
jykim@cs.umass.edu

David Elsweiler
University of Erlangen
91058 Erlangen
Germany
david@elsweiler.co.uk

Abstract

The first ECIR workshop on Evaluating Personal Search was held on 18th April 2011 in Dublin, Ireland. The workshop consisted of 6 oral paper presentations and several discussion sessions. This report presents an overview of the scope and contents of the workshop and outlines the major outcomes.

1 Introduction

Personal Search (PS) refers to the process of searching within one's personal space of information. This space includes content that resides on an individual's personal computer (e.g., documents, emails, visited Web pages, and multimedia files), but extends to content on other personal devices, such as music players and mobile phones, personal information stored in the cloud e.g. Google documents, and personal information about an individual stored by some third party e.g. phone book. Such information has been referred to informally as "Stuff an individual has seen" [1] and "Stuff an individual should see" [2]. Technological developments and cultural changes mean that the quantities of these types of information are growing at an incredible rate and it is becoming increasingly difficult to manage and find the correct information when it is needed. Consequently, this has become an important research problem and there has been growing interest in the topic from within several research communities, including information retrieval, human computer interaction, library and information science and cognitive psychology.

Despite research interest and progress being made in terms of understanding user behaviour with respect to personal information [3] and the development of numerous tools to assist users manage and re-access their information [1, 4], it is well understood within the community that progression has been limited by a lack of evaluation methods [4, 5, 6, 7, 8, 9, 10]. This problem was highlighted at the recent SIGIR 2010 Desktop Search workshop [2]. An outcome of which was the need for community-exerted effort in reaching standardization in PS evaluation. Currently, there are no established or standardized baselines or evaluation metrics, and no commonly available test collections. Privacy concerns, the challenges of working with personal collections [5], and the individual differences in behaviour between users [11] all must be addressed in order to facilitate repeatable and comparable evaluation and to advance research in this domain.

While over-coming these problems is a big challenge, there have been some notable efforts in the past from which to build on:

- Elsweiler and Ruthven [5] suggested a method for creating semi-repeatable task-based user-study evaluations.
- Kelly and Teevan [8] outlined a number of ways of evaluating personal information behaviour.
- Chernov et al. [12] devised a means of creating a test collection for desktop search.
- Kim and Croft [13] proposed pseudo-desktop collections as a means of performing cheap and repeatable experiments to test algorithms etc.

However, none of this work has found a solution capable of delivering repeatable and comparable results that would become the standard method to evaluate personal search. Improved solutions for personal search evaluation that have lower cost, are more repeatable, and more realistic are required.

The workshop acted as a first step towards this objective, by focusing on standardized evaluation for the textual elements within personal desktop collections and known item keyword queries for these elements, with the goal of fostering collaborations and moving towards a strategy for repeatable evaluation in this space and forming plans to bring this strategy to fruition post workshop.

An international programme committee and 6 reviewed papers submitted to the workshop, and short papers and position papers were accepted for the workshop proceedings. These papers reflect many of the issues associated with moving towards standardization in this space and highlight many possible directions for such standardization. The proceedings are available at: http://www.cdvp.dcu.ie/iCLIPS/EPS2011/EPS2011.pdf

What follows is our personal interpretation of the workshop activities, including the presented papers and the various discussion sessions. We conclude by summarising the main points raised during the workshop, the main achievements during the day and the open points for future investigation.

2 Presented Papers

The first session consisted of presentation of the 6 accepted workshop papers interleaved with discussion.

The first paper "Pseudo-Desktop Collections & PIM" was presented by Daniel Gonçalves. The presented work overviewed the personal nature of individuals' desktop collections, the important role of context of interaction with desktop items, people's memories in the searching process and the need to identify users' intent. They put forward that individuals' actual collections, annotated with rich personal context of experience or meaning of data to the individual should be used in evaluation. Issues highlighted included the difficulty in anonymizing these personal collections (in order to allow them be shared with the community for evaluation purposes).

Gareth Jones presented "A Strategy for Evaluating Search of 'Real' Personal Information Archives" co-authored with Yi Chen. This work suggested a 'living lab' approach to evaluation in the domain. In this proposal infrastructure would be shared among the research community. Specifically, tools to index subjects' collections, search components and interface components would be provided to participating institutions, along with test search topics. This approach would allow individual institutions recruit their own participants, add their retrieval approach to be evaluated (for example) to the provided infrastructure and conduct experiments in a cross comparable way.

The third paper presentation was "Towards 'Cranfield' Test Collections for Personal Data Search Evaluation", presented by Liadh Kelly, and co-authored by Gareth Jones. This work proposed a means to create pseudo desktop test sets, user queries and target result sets which exhibit the characteristics of 'real' users' collections by conducting a detailed statistical analysis of the makeup of real user collections through user survey, questionnaire, diary study and predominantly through mining of individuals' collections to gather statistics on the makeup of the content of individuals' desktops, and through the use of a plugin to individuals' desktop search application (e.g. Google desktop) to gather statistics on the makeup of individuals user queries and target result items.

Jinyoung Kim then provided an overview of the contents of David Elsweiler and David Losada's paper on "Ways we can improve simulated personal search evaluation" and Claudia Hauff and Geert-Jan Houben's paper on "Simulating Memory Recall in Personal Search" in their absence. David Elsweiler and co-author's paper presented several ways of improving simulated collections based on the statistics from real users' collections. Claudia Hauff and co-author's paper presented a study where they showed how simulated collections can be used to test hypotheses on human memory's impact on search performance.

Jinyoung Kim gave the final presentation on his and W. Bruce Croft's "Three-stage model for evaluating personal search", where he presented a model of combining simulated collection and user study based on the progress of a research project. He argued that simulated evaluation can provide an early validation of research hypotheses, which can be validated further by more realistic user studies later.

3 Initial Discussion

The presented papers provided an excellent platform for discussion. An initial round of debate served to raise important themes in the work presented, led to further standardized evaluation thoughts and discussion on how this influences the field.

The importance of 'real' users and their collections in evaluation was discussed, along with the fact that users exhibit individual differences. The challenge of understanding the individual context associated with items in personal collections, and the important role this plays in individuals' querying and retrieval needs was also discussed. It was put forward that this leads to the need for real users and their collections in the evaluation process. Hence making moving towards standardization in evaluation a very difficult challenge. The important role of context and understanding users' interactions with their collections was a point that kept reappearing throughout the workshop, which is an indication of the importance of this topic and the challenges it presents. A related issue of personalized interaction was also discussed in this regard.

The possibility of anonymizing users' collections and then sharing them amongst the community was raised. However, it was conceded that it would be difficult to gather willing participants to share their collections for such an initiative, and that the anonymization process itself would be challenging and perhaps the degree of actual anonymization afforded dubious. Taking statistics however from anonymized data (as opposed to sharing such anonymized data with the larger research community) to gain greater understanding of the nature of personal collections may be possible. How much sense could be made from such anonymized collections for statistical analysis purposes was questioned.

The relationship between simulated personal collections and user studies was discussed, and whether simulated collections could be used in preliminary study before user studies are conducted. It was also raised that simulation cannot exist in isolation. User studies are required to create good simulated collections which exhibit the characteristics of 'real' users' collections and which will be useful in evaluating techniques intended for these 'real' user collections. Such things as the nature of collections, tasks performed by individuals on their collections and users' behavioral patterns need to be understood in greater detail.

The issue of users' being individuals and the need to model a variety of users was discussed. To address this, user studies need to include a variety of participants, or as large a cross section of the populous as possible. Simulation techniques then would need to vary the parameters to account for the differences observed across individuals in the populous sample.

It was noted that librarians already have access to many forms of personal collections. The possibility of researchers using these personal collections, perhaps to gain a greater understanding of the nature of personal collections was raised. The possibility of such librarians acting as mediators was also discussed, where they either house personal collections and provide the facility for researchers to conduct evaluations on these collections, or act as a mediator between personal collection owners and researchers for evaluation purposes was proposed.

4 Moving Towards a Standardized Evaluation Approach Discussion Session

Following the interactive paper presentation and discussion session, the second discussion session focused on moving towards a realistic approach for personal search evaluation. Following discussion, possible evaluation approaches consisted of:

1. Gain greater understanding of people's memories and of the questions (i.e. what we need to understand about users' collections and querying behaviours)
2. Improve simulation techniques
3. Perform an instrumentation study with shared infrastructure (code and users)
4. Create a real dataset that's sharable (by anonymization and annotation)
5. Create artificial users which exhibit the characteristics of real users
6. Use a mediator as a point of contact for researchers to gain access to subjects for evaluation purposes.

Point (3) relates to the 'living lab' notion. This approach was disregarded as a first step towards a standardization approach, due to the time involved in creating shared infrastructure, the need for a consortium of people for such an approach and the expected low buy in.

Point (5) was a new option raised. Consisting of the notion of creating artificial bots/agents which act like real users in personal collections, and move beyond the desktop to social networks. Such an approach would require detailed understanding of the behaviors of 'real' users, especially with respect to the collection and the access of documents.

Point (6) relates to the previously mentioned possibility of librarians acting as mediators between researchers and the owners of personal collections.

It was decided that a first move towards standardized evaluation in the personal search space should consist of points (1), (2) and (4). The first step (point (1)) consists of deciding what we need to understand about users' collections and their memories associated with these collections, and both designing and conducting studies to answer these questions.

The second step (point (4)) consists of creating real users collections to gain understanding of the nature of such collections. Challenges with this step include how participants would be gained or motivated to partake in the studies (this presents a huge challenge), the challenge of anonymizing personal collections, and identifying and extracting user intent (simulating topic change and drift were also proposed, however it was questioned whether it would be possible to simulate this). Another source of information proposed for this step was the possibility of obtaining data or statistics from corporations. Also involved in this step would be the requirement to conduct interviews, surveys, etc to gain an understanding of individuals' queries and items they retrieve.

The third, and final step (point (2)) consists of simulating 'real' users collections. Due to time constraints the actual simulation process was not discussed at the workshop, but left for post workshop discussion. The issue of how we would validate the pseudo collections and query generation was also raised. This three-step process would form a cyclical process.

5 Wrap-up Session

During a brief wrap-up session the need to progress towards repeatable standardized evaluation approaches for personal search techniques was reiterated, as evaluation is one of the key areas hindering research in this domain. Moving towards 'Cranfield' style evaluation approaches was confirmed as the most viable option, with first steps of deciding what a first pass at standardized evaluation in this space should seek to evaluate and designing a framework or umbrella for progress. The need for gaining greater understanding of the nature of individuals' desktop collections, their experiences with and memories of their collections, the types of items they re-access from their collections, and their personal querying styles and habits was highlighted as a subsequent step.

Towards progressing the goal of standardization in personal search evaluation, a consortium of people to engage in post workshop brainstorming and further development of the proposed 'Cranfield' style evaluation approach was formed at the workshop.

6 Concluding Remarks

Following an exciting, interactive workshop the take home message was that the research community needs to start making steps towards standardized, cross comparable evaluation approaches for personal search evaluation.

7 Acknowledgements

We would like to thank ECIR for hosting the workshop. Thanks also go to the program committee (Leif Azzopardi – University of Glasgow, UK; Daragh Byrne – Dublin City University, Ireland; Robert Capra – University of North Carolina, USA; Yi Chen – Dublin City University, Ireland; Sergey Chernov – L3S, Germany; Bruce Croft – Univ. of Massachusetts, Amherst, USA; Ronald Fernandez – University of Santiago de Compostela, Spain; Cathal Gurrin – Dublin City University, Ireland; Karl Gyllstrom – Katholieke Universiteit Leuven, Belgum; Donna Harman – NIST, USA; David Hawking – Funnelback, Australia; Sara Javanmardi – Bing, USA; Gareth Jones – Dublin City University, Ireland; Noriko Kando – National Institute of Informatics, Japan; Diane Kelly –

University of North Carolina, USA; David Losada – University of Santiago de Compostela, Spain; Ian Ruthven – University of Strathclyde, Glasgow, UK; Alan Smeaton – Dublin City University, Ireland; Jaime Teevan – Microsoft Research, Redmond, USA; Paul Thomas – CSIRO ICT Centre, Australia), paper authors and workshop attendees, without whom the workshop would not have been the success it was.

8 References

1. Dumais, S., Cutrell, E., Cadiz, J., Jancke, G., Sarin, R. and Robbins, D.C. (2003) Stuff I've seen: a system for personal information retrieval and re-use. *In SIGIR '03: The 26th annual international ACM SIGIR conference on Research and development in information retrieval*, pp. 72–79, New York, NY, USA. Toronto, Canada, ACM Press.
2. Elsweiler, D., Jones, G., Kelly, L. and Teevan, J. (2010) Report on SIGIR 2010 workshop on Desktop Search, SIGIR Forum, December 2010.
3. Jones, W. and Teevan, J., ed. (2007), Personal Information Management, Seattle: University of Washington Press.
4. Cutrell. E., Robbins, D.C, Dumais, S.T., and Sarin, R. (2006) Fast, Flexible Filtering with Phlat - Personal Search and Organization Made Easy. *In CHI 2006: Conference companion on Human factors in computing systems*, pp. 261–270, Montreal, Quebec, Canada, ACM Press.
5. Elsweiler, D. and Ruthven, I. (2007) Towards task-based personal information management evaluations. *In SIGIR '07: The 26th annual international ACM SIGIR conference on Research and development in information retrieval*, pp. 23–30, New York, NY, USA, 2007. ACM.
6. Boardman, R. (2004), Improving Tool Support for Personal Information Management. PhD thesis, Imperial College London.
7. Capra, R. G. and Perez-Quinones, M.A. (2006) Factors and Evaluation of Refinding Behaviors. *In SIGIR 2006 Workshop on Personal Information Management*, August 10-11, Seattle, Washington.
8. Kelly, D. and Teevan, J. (2007) Understanding what works: Evaluating personal information management tools. *In Personal Information Management*, Jones, W. and Teevan, J., ed., Seattle: University of Washington Press., pp. 190-204.
9. Kelly, L., Chen, Y., Fuller, M. and Jones, G. (2008) A Study of Remembered Context for Information Access from Personal Digital Archives. *In IIiX '08*, pp. 44-50, October, London, UK.
10. Jones, G., Gurrin C., Kelly, L, Byrne, D. and Chen, Y. (2008) Information Access Tasks and Evaluation from Personal Lifelogs. *In EVIA '08 – 2nd International Workshop on Evaluating Information Access*, 16 December.
11. Gwizdka, J. and Chignell, M.(2007) Individual differences. *In Personal Information Management*, Jones, W. and Teevan, J., ed., Seattle: University of Washington Press, pp. 206-220.
12. Chernov, S., Serdyukov, P., Chirita, P.-A., Demartini, G. and Nejdl, W. (2007) Building a desktop search test-bed. *In ECIR '07*, pp. 686–690.
13. Kim, J. and Croft W. B. (2009) Retrieval experiments using pseudo-desktop collections. *In CIKM '09*, pp. 1297–1306. ACM.
14. Kim, J. and Croft W. B. (2010) Ranking using multiple document types in desktop search. *In SIGIR '10*, New York, NY, USA, 2010. ACM.