A Strategy for Evaluating Search of "Real" Personal Information Archives

Gareth J. F. Jones
Centre for Digital Video Processing
School of Computing
Dublin City University, Dublin 9, Ireland
gjones@computing.dcu.ie

Yi Chen
Centre for Digital Video Processing
School of Computing
Dublin City University, Dublin 9, Ireland
ychen@computing.dcu.ie

ABSTRACT

Personal information archives (PIAs) can include materials from many sources, e.g. desktop and laptop computers, mobile phones, etc. Evaluation of personal search over these collections is problematic for reasons relating to the personal and private nature of the data and associated information needs and measuring system response effectiveness. Conventional information retrieval (IR) evaluation involving use of Cranfield type test collections to establish retrieval effectiveness and laboratory testing of interactive search behaviour have to be re-thought in this situation. One key issue is that personal data and information needs are very different to search of more public third party datasets used in most existing evaluations. Related to this, understanding the issues of how users interact with a search system for their personal data is important in developing search in this area on a well grounded basis. In this proposal we suggest an alternative IR evaluation strategy which preserves privacy of user data and enables evaluation of both the accuracy of search and exploration of interactive search behaviour. The general strategy is that instead of a common search dataset being distributed to participants, we suggest distributing standard expandable personal data collection, indexing and search tools to non-intrusively collect data from participants conducting search tasks over their own data collections on their own machines, and then performing local evaluation of individual results before central agregation.

1. INTRODUCTION

Personal information archives (PIAs) can include materials from many sources, e.g. content on personal computers and smartphones. The value of such archives can only be realised if they can be searched effectively. Development of suitable search technologies requires that their effectiveness be evaluated. Evaluation in information retrieval (IR) systems ideally includes the measurement of retrieval accuracy and users' satisfaction with the IR system. The former is particularly important for evaluation of IR algorithms, and is generally tested without actual user involvement, while the latter requires input from users. A standard IR evaluation collection includes a document collection, a test set of information needs expressed as search topics, and a set of judgments indicating the relevance of documents to each test topic. Use of this data in an interactive experimental setting enables standardized exploration of interactive IR [5]. However, for personal search, such a dataset is much more difficult to generate due to the heterogeneous nature of personal information space, practical challenges of collecting the data and, significantly, privacy concerns relating to the personal nature of this data. This latter issue creates problems for all aspects of evaluation for search of PIAs.

Current work on evaluation of PIA search is exploring the development of simulated personal Cranfield type search test collections [4]. However, this type of dataset only enables a limited range of research experimentation for PIA search [2]. For example, it cannot be used to explore how a user will query their own PIA or how they will interact with a particular search application. From the search perspective, the key difference between PIA search and standard search environments, is that only the owner of the PIA will be aware of the contents, and thus only they will be able to establish information needs which can be answered by the collection and to determine the relevance of returned content. In order to satisfy this requirement, real users are needed to perform test search tasks, preferably on their own personal data. This requires not only that a user participates in evaluation experiments, but also that they enable the archiving of their personal data and for it to be processed for use in a search system. In this paper we propose a strategy to support evaluation of PIA search based on real user data.

2. LIVING LABORATORY EVALUATION FRAMEWORK FOR PIAS

Our proposed PIA search evaluation methodology is similar to the idea of the *living laboratory* discussed in [3]. This is suggested in the context of evaluating information-seeking support systems which aim to assist users in carrying out open-ended search related tasks. The basic idea of the living laboratory is that rather than individual research groups independently developing experimental search infrastructures and gathering their own groups of test searchers, that an experimental environment is developed which facilitates sharing of resources. This might contain software for data collection, search and evaluation protocols, but also subjects who are available to participate in evaluation tests.

Within a living laboratory for PIA search evaluation researchers wishing to evaluate their technologies would participate in a collaborative evaluation effort. Common indexing and search components would be made available to individuals who agreed to take part in the evaluation exercise. This would then be used to gather PIAs locally and conduct search experiments as outlined in the following sections.

ECIR 2011 Workshop on Evaluating Personal Search April 18 2011, Dublin, Ireland

This proposal builds on the approach in existing work such as the *Stuff I've Seen* study described in [1]. In this investigation a desktop search system with rich user interface functionality was sent to about 230 participants to use as their daily desktop search tool. This tool was used to explore a series of questions on interactive desktop search, and collected a considerable amount of data for further analysis.

2.1 Data Collection and Indexing

Evaluation of personal search requires a document collection and a search system. In terms of developing a personal collection there are two options: to index the data currently on the participant's computer, or to collect data incrementally over a period of time. In either case it will be necessary to install one or more applications on the computer to index the data for search applications.

To do this, open source IR toolkit projects may provide suitable backend technologies for indexing and retrieving for PIA search. However most of these systems are too constrained to traditional IR tasks, e.g. only handling one file form at a time. A PIA search application must be able to index very heterogeneous data sources, and thus existing toolkits may need some extension to support this . In practical terms data may be collected via plugins to existing data management clients before being made available to the indexing application.

2.2 Search System

In order to explore the question of search effectiveness and user search behaviour, we suggest the development of standard search systems which could then be distributed to participants for installation on their own computer. The search client would then search the PIA data index created on their computer. Use of a component based framework for the search system would enable different interface elements and retrieval algorithms to be used in alternative instantiations of the system, which would then enable comparison of their usefulness in search.

2.3 Experimental Tasks

Search topics within a standard IR test collection are typically defined in terms of specific topics known to be covered by the documents in the collection. Since the specific details of an individual's personal collection will not be known and will vary between collections, broader search tasks would need to be defined for our porposed experimental scenario, e.g. referring to meetings with unnamed friends, relatives or colleagues. Even with these more general task statements, the searcher may sometimes find that they are unable to recall any relevant content in their PIA to search for. How to develop suitable task descriptions would obviously have to be clearly defined, and useful lessons in doing this may be gathered from work in designing less specific exploratory search tasks for evaluating information-seeking support systems [3].

2.4 Evaluation

A key part of an IR test collection is the relevance information indicating which items in the collection are relevant to the searcher's information need. Retrieval effectiveness is typically measured using metrics such as precision, recall and various averages where there are multiple relevant items, and average rank and mean reciprocal rank where there is a single relevant known-item. In order to gather relevance data for PIA search, the searcher could be asked to assess the relevance of items retrieved in response to their search in response to each task. If this were undertake at the end of searching for each task, it should not interfere with their search behaviour.

Assessing all items in a collection for relevance is impractical. However, assessing only the items retrieved at high rank using one retrieval method may not give a reasonable indication of the effectiveness with which available relevant documents are being retrieved. To address these issues, *pooling* of results from runs using multiple retrieval methods is often used to construct better approximate relevance sets for standard IR test collections. For the PIA search case, the users topic statement could be applied to multiple retrieval algorithms; only one of these being used by the searcher. The responses of multiple retrieval runs could be pooled and shown to the searcher for assessment.

Interactive search effectiveness can be explored using various measures such as numbers of actions required to locate a required item type, time taken to complete a task, amount of relevant content found, and also potentially subjective feedback of the user's search experiences. Details of user action and responses to questions could be used to explore the cognitive processes undertaken in forming queries and performing a search.

3. CONCLUSIONS

In this paper we have proposed a strategy for PIA search evaluation using a living laboratory approach. The scenario is based on users maintaining their own PIA on their own computer, and using standardized tools to index and search their collection. All relevance assessment and evaluation is also carried out on their computer with only the computed evaluation metrics being returned for aggregation thus preserving privacy of experimental subjects' personal data.

4. ACKNOWLEDGEMENTS

This work was supported by grant CMS023 under the Science Foundation Ireland Research Frontiers Programme 2006. The authors are grateful to the reviewers for their very useful feedback.

5. REFERENCES

[1] S. T. Dumais, E. Cutrell, J. J. Cadiz, G. Jancke, R. Sarin, and D. C. Robbins. Stuff i've seen: a system for personal information retrieval and re-use. In *Proceedings of the ACM SIGIR 2003*, pages 72–79, Toronto, Canada, 2003.

[2] P. Gomes, S. Gama, and D. Gonçalves. Designing a personal information visualization tool. In *Proceedings of the 6th Nordic Conference on Human-Computer Interaction*, pages 663–666, Reykjavik, Iceland, 2010.

[3] D. Kelly, S. Dumais, and J. O. Pedersen. Evaluation challenges and directions for information-seeking support systems. *IEEE Computer*, 42(3):60–66, 2009.

[4] J. Kim and W. B. Croft. Retrieval experiments using pseudo-desktop collections. In *Proceedings of CIKM 2009*, pages 1297–1306, Hong Kong, China, 2009.

[5] P. Over. The TREC interactive track: an annotated bibliography. *Information Processing and Management*, 37(3):369–381, 2001.