

# Personal information retrieval

Paul Thomas  
Department of Computer Science  
Australian National University  
paul.thomas@anu.edu.au

## ABSTRACT

A personal information retrieval tool, offering a unified search interface to all data sources available to a user and operating with some degree of knowledge of the user's preferences and context, seems desirable in light of the vast amount of information available in electronic form. Such a tool however raises challenging research questions in areas include source discovery and selection, result merging and presentation, making use of context, and evaluating the tool's success or failure.

New algorithms for server selection, which need little knowledge of server holdings, have proved useful in a "natural" testbed based on web tasks and data. A new technique has also been developed and tested for evaluating IR systems, or components, in a context-laden, dynamic environment with private information sources and queries. Future plans include using these new techniques to compare methods for result merging and presentation, and for exploiting context.

## Categories and Subject Descriptors

H.3.3 [Information Storage and Retrieval]: Information Search and Retrieval

## General Terms

Design, Experimentation

## 1. INTRODUCTION

Each of us has access to a vast amount of information in electronic form: personal files, source code, calendars, public and private Web sites, enterprise databases, email, and so forth (Figure 1). A personal information retrieval (IR) tool, operating with some degree of knowledge of its owner and providing a unified search interface to all accessible sources, seems highly desirable but raises challenging questions in a number of areas.



Figure 1: An example of the range of information sources available to an individual. A personal information retrieval system aims to provide a unified search interface to all of them.

### 1.1 Why a single tool?

At present, each information source typically offers its own search tool or tools, each with its own interface and each with different capabilities and restrictions. Given that a needed piece of information may come from any of these sources, a user needs to: (1) decide where it's likely to be found; (2) start or switch to the appropriate tool; (3) translate their information need into some appropriate syntax or sequence of actions; and (4) scan the resulting set for the information they need. If the need isn't satisfied, the user will have to either reformulate the query (and repeat steps 3-4) or, worse, start a different tool to look elsewhere (and repeat all of steps 1-4).

A single tool operating over all information sources would simplify this by eliminating step 1 and the prospect of guessing which source is needed; by providing a single interface, it would also save users having to learn several tools. Further, by searching all sources at once a single tool mitigates the risk of missing information by choosing the wrong source, or of accepting information while a more correct or up-to-date

version can be found elsewhere.

## 1.2 How is it different?

A number of free and commercial “desktop search” programs provide a single search interface to many of the typical information sources on a PC. There are however several differences between these existing tools and a personal IR tool as described above.

A personal IR tool may consider rich metadata as well as content. While current tools index simple metadata such as authorship and file name, we may wish to consider attributes such as source reliability, document genre, or relationship to a user’s role in the workplace.

Desktop search programs only consider local files, and in some cases the public Web. To these, a useful personal IR tool would have to add restricted or subscription Web sites, corporate databases, and others. Desktop search programs do not presently distinguish (*e.g.*) local websites from others, or do any form of source selection.

In order to work with this variety of sources, personal IR tools will need to handle access restrictions on data.

A personal IR tool can make use of knowledge about the user to inform the search and/or the presentation of results: for example, the user’s physical location, past preferences, languages, role, or friends and colleagues may be useful. At present, desktop search programs are not making use of these contextual clues.

## 2. OPEN PROBLEMS

There are several outstanding problems in building a personal IR tool, including:

- Discovering and indexing very different data sources;
- Selecting and searching these sources (§2.1);
- Combining results (§2.2);
- Designing a user interface;
- Making use of personal or task context (§2.3); and
- Measuring success or failure (§2.4).

For this project, I am assuming that information sources can be identified (probably by the user), and that each source provides a reliable search service; the personal IR tool is then a metasearcher. Several interesting problems remain, and these are discussed below.

### 2.1 Source selection

In the general case, a personal IR tool may need a process of source selection to determine which sources may be useful for each information need. Irrelevant sources, if included, may “contaminate” the result set with inappropriate or misleading documents; including irrelevant sources also impacts on response time, network traffic, compute load, and potentially on cost.

#### 2.1.1 Related work

Many server selection algorithms have been proposed and evaluated using testbeds derived by partitioning TREC Ad Hoc corpora — generally into one or two hundred collections, and with an eye to producing collections of approximately equal size (*e.g.* [5, 9]). French, Powell, *et al.* have evaluated several algorithms in this manner (*e.g.* [19]).

Relatively little work has examined these algorithms in environments more representative of real servers. Some work has used testbeds based on web crawls, such as that of Rasofo *et al.* [20] and Singhal and Kaszkiel [23].

Results here may (or may not) hold for personal IR, where there will be great variety in server characteristics.

#### 2.1.2 Experiments in Web search

The approximately equal size distribution in much previous work is not typical of the Web, which suggested further experiments. In work with Hawking [12] I compared several standard and two novel algorithms for server selection on a testbed derived from the .GOV corpus and TREC Web Track queries. This testbed features a natural partition of documents between servers, with a resulting document distribution like that in the Web, and queries and scoring functions which attempt to better capture realistic Web tasks.

We were able to demonstrate good results with standard techniques but improved results with newer techniques including our contributions. It also proved possible to get good results in the server selection task without too much knowledge of a server’s exact holdings: for example, ReDDE uses a small sample of documents and HARP and AWSUM the anchor text of already-crawled pages. This may be useful in an environment where building comprehensive models of a server’s contents is expensive, such as subscription services.

Although aimed primarily at Web environments, this work may provide some insight into server selection for personal IR inasmuch as the testbed more accurately represents a wide range of servers and the techniques developed make use of less information about server holdings.

#### 2.1.3 Other investigations

A current line of enquiry is trying to determine whether source selection is worth doing in personal IR. I suggest that it will be unnecessary if, for an individual user:

1. Result ranking algorithms can push poor-quality results far enough down the list that they do not intrude or mislead; and
2. Querying each information source is relatively cheap, in time and network traffic as well as money; or alternatively selecting a source is relatively expensive.

If these conditions above are not met, source selection will be warranted.

I am presently planning a survey of professional users of online search including business analysts and librarians; if possible I would like to extend this to other fields such as journalists, researchers, lawyers, and medical professionals. Recording what information sources individuals use and when, their information seeking habits, and costs and success rates associated with different sources, should provide insight into whether source selection is worthwhile as well as informing other areas of personal IR research.

## 2.2 Merging and presentation

Having selected servers to query, translated the query as appropriate, and retrieved results from each server, a tool must still integrate the results and present them to the user.

Many presentation options have been demonstrated in the literature, including a flat list ordered according to *e.g.* estimated utility; an arrangement by source, in a tree or in

different parts of the interface; an arrangement by thread, for example for results from email or news archives; clustering techniques; or timelines (*e.g.* [2, 27]). At present, the high cost of user experiments prevents easy comparison. The best integration and presentation technique may of course vary from user to user, source to source, or even query to query.

I note also that most work to date considering result merging has assumed that results are of the same basic type, which will not be the case in personal IR.

Little work has been carried out on this problem within this project thus far, although it is a challenging area for future research. Other aspects of user interface design will of course prove important, but I consider these outside the scope of this project.

## 2.3 Making use of context

The IR retrieval process can be improved by taking into account aspects of the user's context, such as workplace role, task type [8], or location [1]. As a personal IR tool acts as an agent for just one user, there is ample scope to gather and exploit this data.

Capturing full context information is decidedly non-trivial: besides a lack of agreement as to what context is (*e.g.* [14, 16]), collecting a large amount of data about a user will likely prove prohibitively difficult and expensive. It seems likely however that there are types of context, or information about context, which are easy to capture and can be used to significantly increase personal IR performance. I propose in this project to concentrate on these. As concrete examples, I may consider only immediate context such as which documents are currently open, after the manner of the Remembrance Agent [21] or Watson [4]. An alternative may be to consider easily encoded information which changes slowly or not at all, such as the user's location or organisational role.

Capturing and using context has not been considered in this project to date, although I expect it to be an important part of later work.

## 2.4 Evaluation

Assuming we can build a personal IR tool, how should we measure its effectiveness, or compare it with an alternative? Conventional evaluation methodologies seem ill-suited to this task. In recent work I have developed an alternative evaluation technique appropriate to personal IR, in which results from systems under review are presented side-by-side to a user in response to real information needs.

### 2.4.1 Related work

There are several established schools of evaluation in IR, which are candidates for evaluating personal IR tools: test collections, logfile analysis, human experimentation in the lab, and naturalistic observation.

**Test collections** The style of experiment introduced by Cleverdon [6] and notably taken up by the TREC conferences [25] relies on three elements: a well-defined corpus of documents, a large set of information needs which may be satisfied by documents in the corpus, and "complete" lists of relevant documents corresponding to each information need. There are strong advantages to this approach; unfortunately applying test collection methodology to evaluating personal information retrieval raises particular problems.

First, the nature of personal information corpora is significantly different to existing test collections. Personal information corpora almost always contain private or proprietary data, which may not be viewable by experimenters. Personal search will typically cover tens of billions of documents as most searches will include the Web. Some corpora will be rapidly evolving, and may change even from use to use.

Also, personal information needs are likely to be diverse and unarticulated. It is likely that future users of personal IR systems will use them for a range of purposes including question answering, known-item retrieval, service finding, and other. Further, in many cases users may be unable or unwilling to articulate their information need, at least at early stages of the search process.

Finally, judgements seem likely to be set-based and contextual. After-the-fact assessments of relevance to a written statement of need are very different to the way a person would judge the results of a search conducted in the course of their usual activities. A quick scan of part of a results set is often enough to judge its utility for the task at hand. Unlike relevance assessors, searchers very seldom read all the documents retrieved for them by a search engine.

**Logfile analysis** An alternative is to consider user selection of a document as an indication of expected utility, and use clickthrough logging to evaluate one or more systems. This does not entail any extra burden on users, and can capture judgements for a variety of information needs; however quality and trust biases need to be considered [15].

Further weaknesses in clickthrough data are relevant to our setting. First, although clickthrough data is known to correlate with utility in the Web [15], it has not been established that this is the case for other types of data. Second, many queries have no associated clicks. Finally, we note that search logs are generally maintained by individual search engines, which makes direct comparison difficult.

**Human experiments** A further method of evaluation involves observing test users in a laboratory setting, conducting searches in response to a simulated information need. The TREC Interactive track [13] has attempted to isolate the effect of different IR systems with a sophisticated design which controls for differences in searcher and topic, and for presentation order.

Other techniques include post-search or post-experiment questionnaires [26], and manual judgements of results or result sets (*e.g.* [22]). Borund and Ingwersen (*e.g.* [3]) also suggest "simulated work task situations" and individual interpretation of utility.

Although these techniques allow individual interpretation of utility, they still rely on artificial information needs and may be confounded by inter-subject and order effects. Further, there is generally a need for a large number of test users and a good deal of time for each.

**Naturalistic observation** Relatively few studies have placed an experimenter in the field to observe subjects in the course of their day-to-day information seeking. Beaulieu observed library users as they used catalogue services and continued to browse the shelves [10]. Nordli [18] and Hansen and Järvelin [11] have carried out similar studies.

An alternative approach which avoids the costs of experimenter time and minimises the chance of altering subjects' search behaviour is to use instrumented search software, which records aspects of interactions for later analysis. Kelly and Belkin [17] used monitoring software on specially-

configured laptops, and Dumais *et al.* [7] used pre- and post-search questionnaires and recordings of interface actions.

### 2.4.2 A proposed model

I propose an alternative model for personal IR evaluation, which combines aspects of naturalistic observations and search log analysis and allows comparisons between systems in a user's full context.

Using a metasearch technique, I provide a front-end to many different search services and can generate logs of user interactions. The front-end presents two or more panels, each with a different IR system (or results from the same system presented differently). By logging interactions with each panel, it is possible to infer which system or presentation is preferred. The front-end can also ask for explicit judgments of preference.

Using a live search system has several advantages. The corpora being searched are those available to our users, and since the experiments need never divulge details of any document we can use private or otherwise restricted corpora. Information needs are those users genuinely encounter day to day, and judgements can be made which account for context and the type of information required. Since we record user satisfaction with a result set, not relevance of each document, we can allow judgments of entire result sets.

The interface has two further advantages. Unlike an interleaved results list, it can offer two different result presentations: for example, it can compare two clustering algorithms. At the expense of greater intrusion, the interface also allows us to prompt for extra information. I currently do this in two cases: periodically after a result is selected, to ask whether it was useful, and when no result is selected for a query. Figure 2 has examples.

There are limitations of this approach. Like other forms of embedded observation, there is an experimenter effect. Subjects are inevitably aware that they are participating in an experiment and that their actions are being logged for study. Furthermore, even if the metasearcher delivers the same results as their standard service, it presents them in a different way in less screen area.

Second, experiments of this nature are not repeatable in particular ways: for example, without access to the corpus we cannot re-run queries with a different IR system.

Third, while it may be easy to show from a set of pairwise comparisons that system A is categorically better than system B, it is much harder to know by how much. It is also difficult to make multi-way comparisons.

### 2.4.3 Experiments

I have implemented two versions of the search system described above. One, an application installed on the user's computer, provides metasearch over any number of corpora; this version also implements the pop-ups described above and illustrated in Figure 2. A second implementation provides a Web interface to two or more Web search engines.

With Hawking [24], I have carried out three experiments to verify that the approach described above can provide a useful comparison between two search systems.

**First experiment** The first experiment addressed two questions: given two result sets with a difference in quality, do users' judgements reflect this difference? If so, can the two-panel design tell which is better?

Users were given the Web-based software described above.

One panel, chosen at random, displayed Google's first ten results; this was assumed to be a high-quality set. The other displayed Google's 21st–30th results. Users were given queries from popular searches and after each were prompted to indicate which set of results were "better", if either.

We observed a significant preference for the higher-quality set of results, but no significant difference in the preference for result sets in the left- or right-hand panels. Several attributes of clickthrough data were good predictors of final preference.

**Second experiment** A second experiment considered the same questions and used the same technique, but instead of assigned tasks participants were encouraged to use our software in place of their regular Web search engine. This gives a good indication of how well a two-panel evaluation method would work for personal IR systems.

The data proved similar to that from the first experiment: there was a significant preference for the high-quality result set, but no significant difference between the judgements in favour of the left- or right-hand panels. Clickthrough attributes were very good predictors of final judgement.

**Third experiment** These observations suggested a third experiment, still using the same system but with the voting buttons removed; users were told to use our system as they would any other Web search service. The question we then ask is: assuming users prefer the high-quality set, does clickthrough data still predict this preference if we remove the explicit voting step?

The clickthrough data again were good predictors of the high-quality result set. These results strongly suggest that using clickthrough data alone, in place of explicit judgements, in our method could provide a robust comparison of two IR systems. This would significantly reduce the burden on both test users and experimenters.

## 3. CONCLUSIONS AND QUESTIONS FOR DISCUSSION

A personal IR tool seems desirable for several reasons, and although it prompts many questions these questions are tractable and this project is attempting to answer some: particularly around source selection, result merging and presentation, making use of context, and evaluating systems or system components.

I am particularly interested in discussing techniques for evaluation in a context-laden, interactive environment; methods for capturing and making use of information about users, information needs, and background context; and user experiments and observation.

## 4. REFERENCES

- [1] M. S. Aktas, M. A. Nacar, and F. Menczer. Personalising PageRank based on domain profiles. In *Proc. KDD*, 2004.
- [2] J. A. Aslam and M. Montague. Models for metasearch. In *Proc. ACM SIGIR*, 2001.
- [3] P. Borlund. The IIR evaluation model: A framework for evaluation of interactive information retrieval systems. *Information Research*, 8(3), 2003.
- [4] J. Budzik, K. J. Hammond, and L. Birnbaum. Information access in context. *Knowledge-Based Systems*, 14(1–2):37–53, 2001.

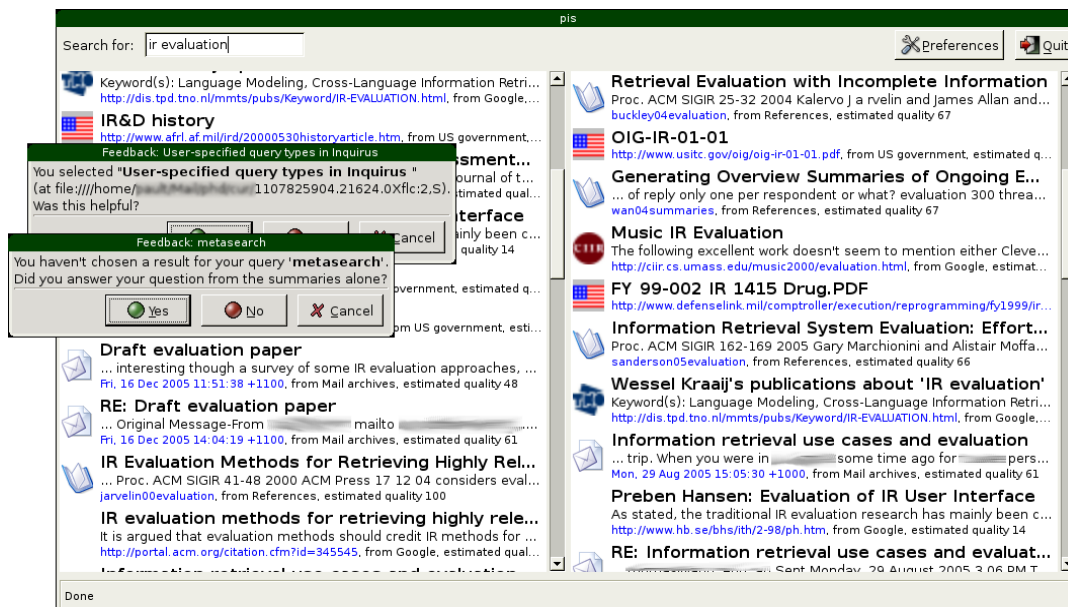


Figure 2: Sample two-panel interface, with optional feedback

- [5] J. P. Callan, Z. Lu, and W. B. Croft. Searching distributed collections with inference networks. In *Proc. ACM SIGIR*, 1995.
- [6] C. Cleverdon. The Cranfield tests on index language devices. In K. S. Jones and P. Willett, editors, *Readings in Information Retrieval*. Morgan Kaufmann, 1997.
- [7] S. 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 *Proc. ACM SIGIR*, 2003.
- [8] L. Freund, E. G. Toms, and C. L. A. Clarke. Modeling task-genre relationships for IR in the workplace. In *Proc. ACM SIGIR*, 2005.
- [9] L. Gravano, H. García-Molena, and A. Tomasic. GLOSS: Text-source discovery over the internet. *ACM TODS*, 24(2), 1999.
- [10] M. Hancock-Beaulieu. Evaluating the impact of an online library catalogue on subject searching behaviour at the catalogue and at the shelves. *Journal of Documentation*, 46:318–338, 1990.
- [11] P. Hansen and K. Järvelin. The information seeking and retrieval process at the Swedish Patent and Registration Office. In *Proc. ACM SIGIR Workshop on Patent Retrieval*, 2000.
- [12] D. Hawking and P. Thomas. Server selection methods in hybrid portal search. In *Proc. ACM SIGIR*, 2005.
- [13] W. Hersh and P. Over. TREC-9 interactive track report. In *Proc. TREC*, 2001.
- [14] P. Ingwersen. Selected variables for IR interaction in context. In *Proc. IRIx Workshop, ACM SIGIR*, 2005.
- [15] T. Joachims, L. Granka, B. Pan, H. Hembrooke, and G. Gay. Accurately interpreting clickthrough data as implicit feedback. In *Proc. ACM SIGIR*, 2005.
- [16] G. J. F. Jones and P. J. Brown. The role of context in information retrieval. In *Proc. IRIx Workshop, ACM SIGIR*, 2004.
- [17] D. Kelly and N. J. Belkin. Display time as implicit feedback: Understanding task effects. In *Proc. ACM SIGIR*, 2004.
- [18] R. Nordli. “User revelation” — a comparison of initial queries and ensuing question development in online searching and in human reference interactions. In *Proc. ACM SIGIR*, 1999.
- [19] A. L. Powell and J. C. French. Comparing the performance of collection selection algorithms. *ACM TOIS*, 21(4), 2003.
- [20] Y. Rasolofo, F. Abbaci, and J. Savoy. Approaches to collection selection and results merging for distributed information retrieval. In *Proc. CIKM*, 2001.
- [21] B. J. Rhodes. The wearable remembrance agent: A system for augmented memory. *Personal Technologies Journal*, 1, 1997.
- [22] X. Shen, B. Tan, and C. Zhai. Implicit user modeling for personalized search. In *Proc. CIKM*, 2005.
- [23] A. Singhal and M. Kaszkiel. A case study in web search using TREC algorithms. In *Proc. WWW10*, 2001.
- [24] P. Thomas and D. Hawking. Evaluation by comparing result sets in context. Manuscript submitted for publication, 2005.
- [25] E. M. Voorhees and D. K. Harman, editors. *TREC: Experiment and Evaluation in Information Retrieval*. MIT Press, 2005.
- [26] R. W. White, I. Ruthven, and J. M. Jose. A study of factors affecting the utility of implicit relevance feedback. In *Proc. ACM SIGIR*, 2005.
- [27] M. Wu, G. Muresan, A. McLean, M.-C. M. Tang, R. Wilkinson, Y. Li, H.-J. Lee, and N. J. Belkin. Human versus machine in the topic distillation task. In *Proc. ACM SIGIR*, 2004.

## **APPENDIX**

### **A. STUDENT'S STATEMENT**

The SIGIR Doctoral Consortium represents an excellent opportunity for me to discuss my recent and proposed work with both established IR researchers and other students. Since my future work is likely to involve user observation and experiments, it would be especially valuable to seek comment and feedback from experts in the field before committing to particular experimental designs or methods. In attempting to work with diverse data sources and take account of user and task context, this project touches on many active areas of research and SIGIR is a rare chance to meet and talk with researchers across the field.

To the extent that my recent work, especially in evaluation, would be useful to others this also represents a good opportunity to demonstrate this contribution.

— Paul Thomas

### **B. SUPERVISOR'S STATEMENT**

Paul is about eighteen months into his PhD program. He has written a thesis proposal and has made good progress, with one co-authored paper accepted at last year's SIGIR and another submitted this year.

The future direction of Paul's research in personal IR is heavily dependent upon finding an appropriate evaluation methodology reflecting the highly individual, highly contextual nature of personal IR, and capable of being executed within resources available. Paul has a proposal along these lines which it would be very valuable to discuss with external experts and with fellow grad students. He would also benefit from the opportunity to discuss with consortium participants means for extracting and exploiting relevant search context.

— David Hawking