Experiences evaluating personal metasearch

Paul Thomas
CSIRO ICT Centre
Canberra, Australia
paul.thomas@csiro.au

David Hawking
Funnelback Pty Ltd
Canberra, Australia
david.hawking@acm.org

ABSTRACT

Many current evaluation techniques for information retrieval, such as test collections and simulations, are difficult to apply in situations where queries and preferred results are context-dependent. This is particularly true in personal metasearch applications, which provide a person with unified search access to all their usual online sources. A recently-proposed technique, based on presenting two or more search results sets in a single comparison interface, offers an alternative.

We have embedded this technique in a working personal metasearch tool which we have distributed to volunteers. Initial experiments with server selection suggest that the technique is accepted by users, can operate over diverse and unarticulated contexts, and that the data it provides can provide a useful comparison to that from test collections. Further experimentation with the technique is continuing.

Themes: Case studies, field experiments, simulations, etc.

1. INTRODUCTION

Most computer users have access to a vast amount of information in electronic form: personal files, calendars, public and private web sites, corporate databases, email, and so forth. 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.

Our research addresses the problems in building a single search tool—a personal metasearcher—capable of selecting the sources most likely to provide good answers to a query, querying those sources and merging the results. Note that, unlike desktop search[1] and Stuff I've Seen [2], a personal metasearcher is not restricted to information local to the person's machine and does not build its own comprehensive index of the data to be searched. We have constructed a prototype personal information search (PIS) tool to support our experiments.

An important challenge arises in evaluating and comparing personal metasearch tools, which deal with tasks and data which may be confidential. Since these cannot be shared with experimenters, established evaluation techniques such as test collections and user studies are not likely to be appropriate.

Further, naturalistic observation of users in the workplace does not seem a feasible evaluation approach. As well as causing problems of confidentiality, the presence of an observer may affect outcomes; and we have found that intervals between successive search tasks are generally long, leading to much wasted observer time.

1.1 Embedded comparison

We have suggested embedded comparison [5] as an alternative evaluation technique when collections are dynamic, context is hard to capture, or documents are private.

The method uses an instrumented search tool which presents two sets of results, in independent panels, for each query. Each set of results is drawn from a different system. By recording user interactions with each panel, and assuming a click on a result is an indication of the quality of the containing set, user preferences for one set or the other can be inferred.[2] Figure 1 illustrates the technique, using our pilot software and showing two sets of results for a single query, each drawn from a separate retrieval algorithm.

Our early experiments indicated that a very simple analysis of click patterns reliably indicated user preference, and that embedded comparisons were a promising way to evaluate search in its full context. Although the technique does not share some of the desirable features of testbed evaluations, it does allow users to search personal data with naturally-occurring information needs and records preferences in light of a user's full context—even when that context cannot be captured or possibly even articulated. By contrast, user studies in the laboratory almost inevitably involve a simulated context.

These experiments suggested that embedded comparisons should provide a straightforward method for evaluating personal metasearch, without needing to capture or describe context.

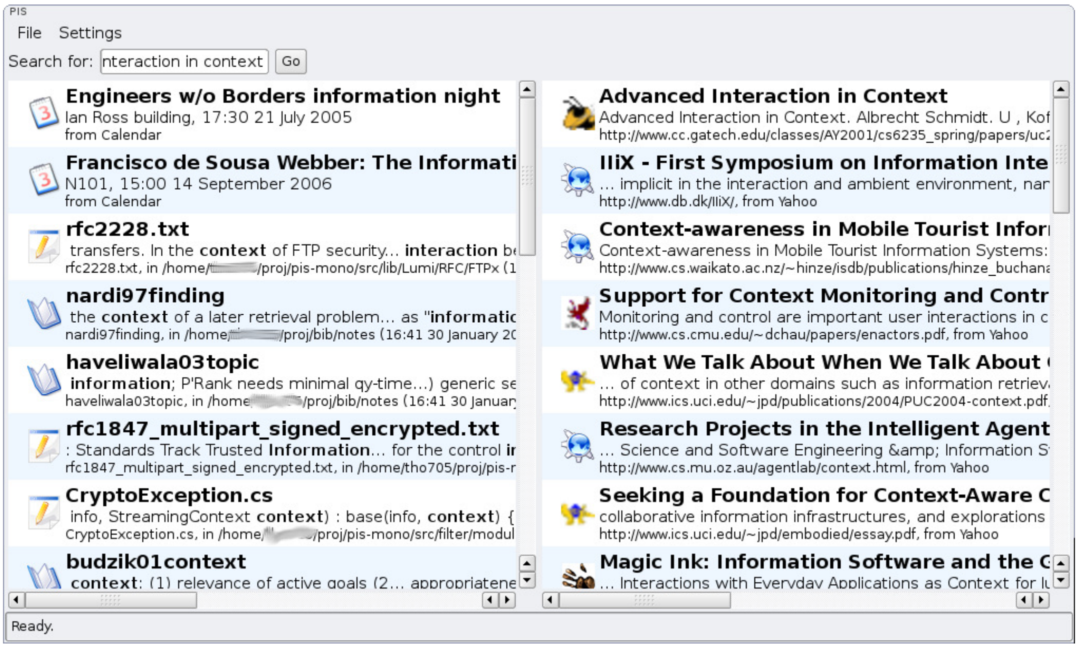
1.2 Server selection

Our personal metasearch tool, in common with similar tools, includes a process of *server selection* to choose the most promising servers with which to search. For example, a query apparently about current events might be directed to news websites, while a query about a particular research topic might be directed to bibliographic servers such as Citeseer.

A great many techniques have been proposed for server selection. To evaluate these, our earlier experiments [4] used a fixed set of "servers" which did not correspond to any real user's documents; and artificial queries which did not correspond to any real information need. This testbed approach, since it was inexpensive, allowed

[1] For example `desktop.google.com`.

[2] Alternatively, users can be asked to express explicit preferences.



PIS

Search for: nteraction in context

Engineers w/o Borders information night
Ian Ross building, 17:30 21 July 2005
from Calendar

Francisco de Sousa Webber: The Informati
N101, 15:00 14 September 2006
from Calendar

rfc2228.txt
transfers. In the **context** of FTP security... **interaction** b
rfc2228.txt, in /home/[redacted]/proj/pis-mono/src/lib/Lumi/RFC/FTPx (1

nardi97finding
the **context** of a later retrieval problem... as "**informatic**
nardi97finding, in /home/[redacted]/proj/bib/notes (16:41 30 January 20

haveliwala03topic
information; P'Rank needs minimal qy-time...) generic se
haveliwala03topic, in /home[redacted]/proj/bib/notes (16:41 30 Januar

rfc1847_multipart_signed_encrypted.txt
: Standards Track Trusted **Information**... for the control **ir**
rfc1847_multipart_signed_encrypted.txt, in /home/tho705/proj/pis-r

CryptoException.cs
info, StreamingContext **context**) : base(info, **context**) {
CryptoException.cs, in /home/[redacted]/proj/pis-mono/src/filter/modul

budzik01context
context: (1) relevance of active goals (2... appropriatene

Advanced Interaction in Context
Advanced Interaction in Context. Albrecht Schmidt. U , Kof
http://www.cc.gatech.edu/classes/AY2001/cs6235_spring/papers/uc2

IIiX - First Symposium on Information Inte
... implicit in the interaction and ambient environment, nar
http://www.db.dk/IIiX/, from Yahoo

Context-awareness in Mobile Tourist Infor
Context-awareness in Mobile Tourist Information Systems:
http://www.cs.waikato.ac.nz/~hinze/isdb/publications/hinze_buchana

Support for Context Monitoring and Contr
Monitoring and control are important user interactions in c
http://www.cs.cmu.edu/~dchau/papers/enactors.pdf, from Yahoo

What We Talk About When We Talk About
... of context in other domains such as information retriev
http://www.ics.uci.edu/~jpd/publications/2004/PUC2004-context.pdf

Research Projects in the Intelligent Agent
... Science and Software Engineering &amp; Information S
http://www.cs.mu.oz.au/agentlab/context.html, from Yahoo

Seeking a Foundation for Context-Aware C
collaborative information infrastructures, and explorations
http://www.ics.uci.edu/~jpd/embodied/essay.pdf, from Yahoo

Magic Ink: Information Software and the C
... Interactions with Everyday Applications as Context for I

Ready.

Figure 1: Sample two-panel interface, as implemented in PIS

us to evaluate twelve alternative algorithms on the standard $\mathcal{R}_n$ measure [3], an analogue of recall. Of the twelve, one—Kullback-Leibler divergence [7]—seemed promising and another—vGlOSS [3]—did not. Kullback-Leibler divergence performed greatly and significantly better than vGlOSS, across a range of conditions.

2. EXPERIMENT

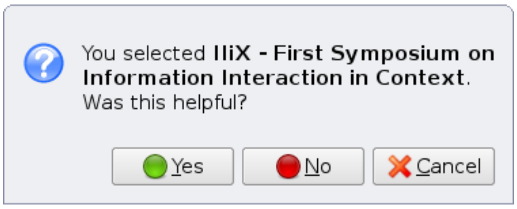
We implemented the two server selection algorithms above within PIS and used embedded comparison to answer the following questions:

1. Does the embedded comparisons technique work in a less controlled environment than that of the first experiments, despite the complications of a working personal metasearch tool?
2. If so, do users' preferences for server selection bear out testbed findings, which were based on simulated information needs and took no account of context?

Ten volunteer participants were recruited from a group who had expressed interest in PIS. These users were selected to demonstrate a range of possible search scenarios, as well as for their willingness to use prototype software; although a convenience sample, they are likely to resemble early adopters of any personal metasearch tool. Participants were all experienced computer users, but worked in a variety of fields and were presumed to have different information needs.

Each participant was given the two-panel version of PIS illustrated in Figure 1. Both panels used the same presentation, but servers were selected in two ways: one panel, chosen at random, scored servers using Kullback-Leibler divergence; and the other used vGlOSS.

Although participants were not asked what servers were used, and PIS did not record this information, informal feedback indicated that a variety were used including email search, of different types; local file search, also of different types; Wikipedia and other reference Web sites; intranets; wikis; the public Web; LDAP directories; and local databases.



You selected **IIiX - First Symposium on Information Interaction in Context**.
Was this helpful?

Figure 2: Extra feedback from the two-panel interface

PIS logged a small set of data for each user query: the fact that a query had been issued, the number of servers in use, the ordering of the panels for this query, and the panel and rank of any result opened. It was also configured to ask for additional feedback in the two situations described in earlier work [5]: first, if no document was selected, and second, in about 50% of cases, after a document was selected to confirm that it was in fact useful (Figure 2). To minimise work for participants, they were not explicitly asked which panel was preferred. Relying on implicit indicators limited the amount of data available, but demonstrates a plausible "bare-bones" implementation of embedded comparisons.

The experiment ran for 64 days, although not all participants were active for the full period.

2.1 Results

In the present work we consider preferences for selection algorithms briefly, but are more concerned with experiences in context-laden evaluation. Fuller details for the former can be found elsewhere [4].

Results recorded by PIS were automatically sent to a central server for analysis. If additional feedback indicated that a selected result was not helpful, which happened in about $\frac{1}{3}$ of cases, any associated clickthrough data was removed and was not considered as evidence of a preference for either selection algorithm. Further, since PIS always selected at least two servers regardless of the algorithm in use, clickthrough data was removed for any query issued when only one or two servers were configured.

Three participants withdrew from the study without submitting results, one because of a particular bug in PIS and two without giving a reason. In total, 273 queries were recorded from seven users; of these, 98 queries had usable clickthrough data with one to four clicks each.

Kullback-Leibler divergence was significantly preferred by two of the six participants in our experiment, and vGlOSS by none (one-sided binomial sign test, $\alpha = 0.05$). This is broadly consistent with the testbed result, but the effect is not as strong as might have been expected. This may be explained by differences between the two styles of evaluation: for example, set-based judgements in this experiment substitute for document-based judgements in earlier experiments, and judgments in this experiment take into account relevant aspects of each user's context.

It is also possible that both selection techniques are good enough, or poor enough, in a real setting that relative differences are less important; or that "information needs" in previous experiments are not like the queries used here. Since this experiment includes real information needs, real users, and real data, we must conclude that previous experiments were able to detect the direction of the difference between the two selection methods but overestimated the strength of this difference.

Large differences in traditional metrics such as $\mathcal{R}_n$ may be needed before users observe a difference in quality, a conclusion consistent with earlier work [1, 6]. In the particular case of $\mathcal{R}_n$, high scores depend upon selecting servers with the highest numbers of relevant documents. Users may however be satisfied with some small subset of these documents, in which case $\mathcal{R}_n$ is not an accurate measure of effectiveness.

2.2 Observations on using the method

The take-up rate (effectively seven out of nine) likely indicates that personal metasearch, and the PIS prototype in particular, does seem useful to most people; and that the tool, while only a research prototype, was sufficiently robust for day-to-day use.

This experiment included only a small number of participants as each was selected for using a range of collections, including collections of different sizes and data types. It would have been possible, for example, to install PIS across student machines or across an entire workgroup and thus aquire more users, although this would be at the expense of including a much more homogenous set of users and collections. Future work may consider this possibility.

Scaling the embedded comparisons technique to a larger group of users seems feasible: some effort was required helping participants install the software, and further effort was required to debug the software and issue new versions as bugs were exposed, but to support the participants in the present experiment little effort was needed overall. As more participants were added, with collections and usage patterns similar to existing users, experimenter effort would likely be relatively small.

3. OTHER APPLICATIONS

The PIS prototype can easily be extended to investigate and evaluate other personal metasearch components, such as result merging, query translation and presentation.

As well as personal search, we have been investigating the use of embedded comparisons in other situations where context is hard to capture. Experiments have investigated bibliographic search and the effect of search engine branding, and are considering the impact of spam results in Web search.

4. CONCLUSIONS

Results from this experiment were broadly in line with those of testbed experiments, which did not attempt to capture a user's context; however the participants in this experiment expressed a much weaker preference than would be expected from earlier results. This suggests that the test collection approach, while useful, may be somewhat inaccurate in predicting user preferences. The inaccuracy may be due to differences in judgements (result sets against documents); user context; both methods being good enough, or poor enough, that quality differences are small; a mismatch between the collections and queries used earlier and those used by test participants; or some combination of the above. This in turn suggests that evaluations based on test collections may be somewhat limited, and in particular that large differences on typical test collection metrics may be needed before users appreciate a difference in quality.

The embedded comparisons technique, while requiring significantly more experimenter effort than testbed techniques, appears feasible even with prototype search tools operating over a variety of user contexts and with minimal intervention. Although this experiment used a small number of participants, indications are that the technique could scale to larger uses without unreasonable effort on the part of experimenters.

5. ACKNOWLEDGEMENTS

We thank the participants in this study for their time and their willingness to help.

6. REFERENCES

[1] J. Allan, B. Carterette, and J. Lewis. When will information retrieval be "good enough"? In *Proc. ACM SIGIR*, 2005.

[2] 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.

[3] L. Gravano and H. García-Molina. Generalizing GlOSS to vector-space databases and broker hierarchies. In *Proc. VLDB*, 1995.

[4] P. Thomas. *Server Characterisation and Selection for Personal Metasearch*. PhD thesis, Australian National University, 2008.

[5] P. Thomas and D. Hawking. Evaluation by comparing result sets in context. In *Proc. CIKM*, 2006.

[6] A. Turpin and F. Scholer. User performance versus precision measures for simple search tasks. In *Proc. ACM SIGIR*, 2006.

[7] J. Xu and W. B. Croft. Cluster-based language models for distributed retrieval. In *Proc. ACM SIGIR*, 1999.